The McGraw·Hill Companies

MARKETING RESEARCH

TEXT AND CASES

THIRD EDITION

ABOUT THE AUTHOR

Dr. Rajendra Nargundkar is Dean, Continuing Education, IFIM Business School, Bangalore. Earlier, he was a Professor in the Marketing Area, at IIM Kozhikode and IIM Lucknow.

A postgraduate in Marketing from IIM Bangalore and a Ph.D. from Clemson University, USA, Dr. Nargundkar has been actively involved in teaching, training, consulting and research for over two decades in India and abroad. He has taught management and marketing courses at Clemson and Lander universities in USA, and Kirloskar Institute of Advanced Management Studies, Xavier Institute of Management and PES Institute of Management in India. He has published over 50 papers in several international conferences and journals including *Academy of Management Journal,* one of the top academic journals in USA. He has also written ***Services Marketing***, a popular text among MBA students, published by Tata McGraw-Hill.

Before entering academics, Dr. Nargundkar was with Marketing and Business Associates (MBA), one of India's leading Marketing Research agencies (now Gallup India Pvt. Ltd.). He can be reached at rnargundkar@gmail.com.

The McGraw·Hill Companies

MARKETING RESEARCH

TEXT AND CASES

THIRD EDITION

Rajendra Nargundkar
Dean, Continuing Education,
IFIM Business School, Bangalore

Tata McGraw-Hill Publishing Company Limited
NEW DELHI

McGraw-Hill Offices

New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

Tata McGraw-Hill

Published by Tata McGraw-Hill Publishing Company Limited,
7 West Patel Nagar, New Delhi 110 008.

Marketing Research: Text and Cases, 3/e

Copyright © 2008 by Tata McGraw-Hill Publishing Company Limited
No part of this publication may be reproduced or distributed in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise or stored in a database or retrieval system without the prior written permission of the publishers. The program listings (if any) may be entered, stored and executed in a computer system, but they may not be reproduced for publication.

This edition can be exported from India only by the publishers,
Tata McGraw-Hill Publishing Company Limited.

ISBN (13 digits): 978-0-07-022087-4
ISBN (10 digits): 0-07-022087-5

Managing Director: *Ajay Shukla*
General Manager—Publishing: B&E/HSSL and School: *V Biju Kumar*
Editorial Manager—B&E: *Tapas K Maji*
Junior Editorial Executive: *Hemant K Jha*
Junior Executive—Editorial Services: *Anubha Srivastava*
Senior Production Manager: *Manohar Lal*
General Manager—Marketing (Higher Education and School): *Michael J Cruz*
Asst. Product Manager: *Vijay S Jagannathan*
Controller—Production: *Rajender P Ghansela*
Asst. General Manager—Production: *B L Dogra*

Information contained in this work has been obtained by Tata McGraw-Hill, from sources believed to be reliable. However, neither Tata McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither Tata McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that Tata McGraw-Hill and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.

Typeset at Bharati Composers, A-1/402, Sector-VI, Rohini, Delhi 110 085, and printed at Avon Printers, 271 FIE, Patparganj, Delhi 110 092

Cover Design: Astral Pre Media Pvt. Ltd., Noida

Cover Printer: Rashtriya Printers

RQLYCRZXRQXQB

The McGraw·Hill Companies

To
my parents,
my wife Anuradha,
and my daughters Prarthana and Pooja

PREFACE TO THE THIRD EDITION

I have been receiving a lot of positive feedback on the earlier editions of this book. But I also felt that updating readers, both students and research scholars, on a couple of newer techniques like Structural Equation Modeling (SEM) and Logistic Regression was needed.

I acknowledge the help from two people in adding the new material. Dr. Satish Nargundkar of Georgia State University, USA, has contributed the chapter on Logistic Regression. This technique can be used instead of Discriminant Analysis for similar applications. SEM has gained a lot of popularity for modeling simultaneous equations similar to regression, with all the variables being latent variables (measured by a few manifest or observed variables each). In other words, a combination of Factor Analysis and many simultaneously run regressions. Packages like LISREL, AMOS from SPSS, or the SEPATH procedure in Statistica can perform the required modeling for SEM as well as Confirmatory Factory Analysis. I would like to thank Dr. V.S.R. Vijayakumar, currently with Everonn Systems, for help in developing the material on SEM.

I have also added a section called "Doing More with Data" on transforming data and using part files instead of complete data files, for those who are new to these SPSS procedures. Also, a brief primer on Reliability and Validity has been added, mainly to help doctoral students. Some SPSS commands have been updated, especially for use in doing ANOVA, due to changes in the newer versions of the software. My thanks are due to SPSS Inc. for their offer of the CD containing the software package.

I do hope the readers will appreciate the changes. I look forward to their feedback.

PREFACE TO THE FIRST EDITION

It gives me great pleasure to have reached this milestone and to be able to present a new textbook on marketing research. Authors have different reasons for writing books. In my case, the reason was purely a need I felt for a book that demystified marketing research and brought it within the easy grasp of students of the subject. Having worked in one of India's leading Marketing Research (MR) agencies started by a group of IIMA alumni, I was quite taken aback when I started teaching the subject in the early nineties. The same books (almost all foreign) that we used as textbooks in our MBA classes in the mid-eighties when I studied at IIM Bangalore, were still in use. Nothing wrong with that, I thought, if the books served the purpose. But as I went on, I discovered that many of the existing books made the subject of marketing research appear as complex as Rocket Science to a student. To my mind, it is not. Many books also appeared to introduce unnecessary complexity by trying to include too much detail on too many topics, leading to confusion among students who attempted to read the text. At the end of it all, we had students who could not grasp the essence of marketing research methods and analytical techniques. One other discovery, as time went by, was that many of the standard books still clung on to manual techniques of statistical analysis that had become obsolete with the advent of the computer. A third discovery was that, that few Indian books were available after almost four decades of management education in this country.

My attempt, through this book, is to address all the issues I have mentioned here. This book aims to provide a crisp, clear, easy-to-understand view of the methods, processes and statistical analysis techniques used in marketing research, free from excessive unrelated material that comes in the way of a student's understanding. In my view, a serious student of the subject can read, understand and assimilate most of the material presented in this book in one trimester or semester, as the case may be, even without too much help from a teacher. A student who goes through the material presented here, along with a couple of projects that take him into the field, can design, conduct and analyse almost any commonly occurring marketing research problem. It would be desirable that the student also uses a computer-based statistical package for all the data processing and statistical analysis. It must be mentioned that this book is not a text on research methodology but on marketing research. There is a major distinction between the two in my opinion, though there are many common aspects too. Research methodology lays a lot more emphasis on Reliability testing and Validation methods, and deals with the needs of practical as well as academic research that tends to be much more rigorous. Marketing research, while it recognizes the basic need for sound methodology, is more pragmatic and business practice-oriented than pure research.

A Blend of Practice and Theory

The emphasis in this book is on what can be done in practice, without sacrificing essential theory. In other words, it is aimed at making a practicing marketing researcher out of a student than merely an academic one. To borrow from hypothesis testing terminology, it is better to have done a marketing research study that is 95 percent correct but timely, rather than do a study that is 99 percent correct but takes too long and costs a lot more. To the student who is interested in more detail, many other books on research methodology are available as references. But my vision in writing this book was not to create an encyclopaedia that contains the answer to all possible questions. Rather, it is to provide 90 or 95 percent of what a student needs to know in an easy-to-understand format.

Relevant, Contextual and Realistic Cases

The teacher who uses this book as a text will find his job that much easier. The reason is that this book presents some illustrative case studies (over 30 in number) with either real or simulated realistic data at appropriate places to clarify and illustrate the analytical techniques covered. The uniqueness of the book lies in these illustrated case studies from contemporary India of around year 2000 and later. Most examples cited in the text are also from my own and my colleagues' experiences of real marketing research done in India. Even published research studies cited are of recent Indian origin. Therefore, the context is what the student would experience in the field.

Assumptions Made

An important assumption made in writing this book is that the student is aware of the material taught in a basic course on business statistics, which would generally cover normal distributions, hypothesis testing and non-parametric tests. Only as an illustration, a couple of commonly used 't' tests are covered in Chapter 7 (Planning the Data Analysis). It is possible that some students are also taught correlation and regression as a part of the basic statistics course, but I have tried to give that chapter a more marketing research-oriented touch. The same holds good for Analysis of Variance, where only marketing applications have been taken up.

Computer-based Data Processing and Analysis

One thing I have scrupulously adhered to in the book is that no manual statistical analysis or data processing has been mentioned or used anywhere. This is based on my strong belief that these methods are outdated, and the student should only use contemporary computer-based analytical methods. However, the student can use manual data processing methods (like tabulation) for small surveys that do not involve much statistical analysis, if computers are not available. I would still recommend that an effort should be made by the teacher to make computer assistance available if it is possible.

One outcome of my emphasis on computer usage is that I have completely excluded all statistical tables such as the 't' statistic table, the 'F' statistic table, the chi-squared table, the normal distribution 'z' values, and so on. This may seem odd to a student or teacher used to seeing them in all marketing research books, but I would only say that as you proceed through the various chapters, you would not

feel the need for any tables. I have taught the subject for about twelve years now without using any tables, except the standard normal tables in the chapter on sampling to determine the values of 'z' at 90 and 95 percent confidence levels. Since these values are well-known, the student who knows them does not need any tables for using this entire book. All hypotheses are tested by the computer and 'p' values are generated by the computer package, which makes it unnecessary to do any manual testing of hypotheses. This is also a unique feature of my book, and I hope it will be appreciated for its ease of usage when compared with manual tests of hypotheses.

Road-map of the Book

The book is organised into three parts. There is a reason for this. Part One deals with basic marketing research methodology, which is taught as a separate course in some institutions that I have seen. Part Two covers analytical techniques of many kinds, starting with chi-squared tests for cross-tabulations and closing after addressing many of the commonly used multivariate statistical techniques. Part Three consists of two Appendices—one on 'Industrial Marketing Research' and another on 'Careers in Marketing Research'. Part Three is optional and can be combined with Part One alone, or used as additional readings after Parts One and Two have been covered.

My recommendation to the teacher who wants to cover only Marketing Research basics is to use Chapters 1 to 8, because they cover the entire fundamentals of the marketing research process, secondary and primary data collection techniques, research design, sampling techniques, sample size calculations in theory and practice, the essentials of questionnaire design, organising field work, data preparation and planning for analysis, and basics of report writing. Simple and cross-tabulation is also covered in Chapter 8. If multivariate analysis is not required to be covered, this can be a logical place to stop. The three integrated case studies presented at the end of Part One can be used to illustrate how the final shape of a report may look like, after a study is complete. I would strongly recommend that students be asked to do a small field project as the course progresses, so that they grasp the fine art of doing marketing research and can experience for themselves the difficulties involved in translating a research brief into a useful, actionable marketing research report.

For those who teach a course on multivariate analysis, Chapters 9 to 15 would be the recommended ones. Each of these chapters contains, at the end, three case studies done with simulated, fictitious data that illustrates the multivariate technique covered in that chapter. The input data (realistic, but fictitious) as well as the output from a statistical package like SPSS or STATISTICA is presented along with interpretation done by a group of students from my Advanced Marketing Research course at Kirloskar Institute of Advanced Management Studies, where I taught until recently. The chapter itself contains one fully illustrated example and, coupled with three at the end of the chapter, there are a total of four complete illustrations of each multivariate technique. The student who has access to a computer and a statistical package like SPSS, should try these problems out by using the given data and try to get output similar to that given in the illustrated cases. The output may differ slightly in some cases, depending on the package used, but should be substantially the same.

The student who goes through Chapters 9 to 15 will gain considerable confidence in using the multivariate statistical techniques covered. In addition, I usually ask students, in groups of three or four, to formulate a problem (with data) appropriate for the use of each multivariate technique taught and interpret the output relative to the problem formulated, after running the data through a computer.

This prepares the student in all the important areas of use of these techniques—problem formulation, data collection (or simulation), computer usage, and interpretation of the output. Even if the student has no access to a package such as SPSS, the illustrations (numbering four for each technique covered in Part Two) should be sufficient for a comfortable level of understanding. Unless the student takes up a career in marketing research, he would rarely do the computer applications himself. Even in the M R companies, the MBA may not do the computer-based input and output but may get it done from others. So it would still give the student an adequate appreciation of the technique if he can formulate a problem (know when to use which technique), understand the scale and format of the input data, and interpret the output. This is the emphasis in all the chapters of Part Two of the book. In my opinion, the critical factor in the use of multivariate techniques is to know when to use them and when not to use them.

The teacher who intends to do an integrated course on marketing research including both basics of methodology and advanced multivariate analysis, could go through all the chapters in sequence, from Chapters 1 to 15.

Learning Objectives and Summary

All chapters begin with introductory bullet points that spell out the learning objectives for the chapter. Every chapter closes with a summary of the main points covered in the chapter.

Website/E-mail

The book has its website: `http://www.tatamcgrawhill.com/digital_solutions/nargundkar` for both students and teachers, where students can refer to updates and additional material that will be regularly added, or post his questions for the author to reply to. Alternatively, students can e-mail me their query related to any portion of the book, to which I will surely reply.

My e-mail ID is: `rnargundkar@iiml.ac.in`.

Teachers who adopt this book as a text will have access to teaching aids like Powerpoint Slides for each chapter, and my suggestions and learnings from my teaching the course. Teachers can post their views and suggestions as also e-mail me at the same address mentioned above, with feedback or queries.

Acknowledgements

This book has been in the making for a couple of years now and many people have contributed to its completion. First and foremost, my thanks to the publisher Tata McGraw-Hill for the confidence reposed in me, and the constant encouragement of Mr. V. Biju Kumar and Mr. Tapas K. Maji, who coordinated the project. My thanks to Mr. Raza Khan for his cooperation through the production stage of the book. I also appreciate the efforts of the production department, especially Ms. Medha Arora, for readying the proofs on time.

A unique feature of the book is the many illustrative case studies that feature in it. These have come from the projects done by almost the entire class of 2000–2002, who took the Advanced Marketing Research and the Basic Marketing Research courses with me at the Kirloskar Institute of Advanced

Management Studies, Harihar, where I taught for several years. I am thankful to the students who have contributed their work in the form of these cases and illustrations. Their names feature as authors of the cases at appropriate places in the book. In addition to the project reports, questionnaires from their projects were contributed by Jitender Singh (Jimmy), Shilpa George and Sharmistha Kundu, which feature at the end of Chapters 2, 3, and 4. My thanks to all of them.

Much of the typing and secretarial assistance came from Rajesh Dixit, M.Veena, Gururaj Khasnis and Raghavendra Dixit at Kirloskar Institute. I am grateful to all of them for the long hours they put in to convert my dream into reality.

All the authors of previous books on marketing research have influenced me to a degree, and I am especially fond of the books by Churchill, Luck and Rubin, and Green, Tull and Albaum. I would like to place on record my gratitude to my marketing professors Prof. P.N. Thirunarayana and Prof. J.D. Singh, who taught me marketing fundamentals at IIM Bangalore, which crystallised into a better understanding of marketing research. Prof. Subba Rao and Prof. M.R.Rao (currently, director, IIM Bangalore) taught me statistics courses at IIMB, which laid the foundation for my interest in the subject of marketing research later on. Prof. M.G. Korgaonker and Prof. S.N. Chary, my directors at Kirloskar Institute, were very supportive of my book-writing venture, and I am grateful to them for their support.

I am also grateful to the Kirloskar group executives and members of the Kirloskar family for supporting various case-writing initiatives and research projects during my stay at the institute. I have also had the opportunity to work on a consultancy project for the Shimoga Milk Union in Harihar and Davangere, and I have featured a questionnaire used in a related study as an exercise at the end of Chapter 4.

I learnt the basics of several statistical analysis techniques using computerised packages while pursuing my doctoral work at Clemson University, South Carolina, USA, and I am grateful to my professors and colleagues there, in particular Dr. Charles McNichols, Dr. Steve Cantrell, Dr. Alok Srivastava and Dr. Satish Nargundkar, for improving my understanding of statistical techniques.

My colleagues at Marketing and Business Associates (MBA), the marketing research company where I worked, have all contributed to my understanding of the ground realities of M R in India, and I am grateful to all of them, including Mr. C.K. Sharma, Mr. Shyam Sundar, Mr. Dhananjay Karopady and Mr. Nikhil Prabhu.

I am thankful to my colleague and friend Mr. N.S. Muthukumaran of ORG-MARG for his insightful comments on the manuscript of this book.

I am grateful to my family and all the well-wishers including friends from my batch at IIMB, and colleagues at Kirloskar Institute, who kept my interest in the project alive when I felt like giving it up.

Last but not the least, my thanks to IIM Lucknow, my current employer, for giving me support for completing my book. It has been a long but satisfying journey.

RAJENDRA NARGUNDKAR

The McGraw·Hill Companies

CONTENTS

Preface to the Third Edition *vii*
Preface to the First Edition *ix*

PART 1: FUNDAMENTALS OF MARKETING RESEARCH

1. Introduction, Evolution, and Emerging Issues **3**
Role of Marketing Research in a Marketing Plan *3*
Marketing Intelligence versus Marketing Research *4*
Who Does the Marketing Research? *6*
Typical Applications of Marketing Research *6*
Concept Research *7*
Product Research *8*
Pricing Research *8*
Distribution Research *9*
Advertising Research *10*
When To Do Marketing Research? *11*
Limitations of Marketing Research *11*
Differences in Methodology *12*
Complementary Inputs for Decision-making *12*
Secondary and Primary Research *12*
Ethical Considerations in Marketing Research *13*
Consumer's Right to Privacy *14*
Emerging Issues *14*
Marketing Research in the Internet Era *14*
Online Research *15*
Data Warehousing and Data Mining *16*

Summary 17
Assignment Questions 18

2. The Marketing Research Process—An Overview **19**
Information Need *19*
Defining the Research Objective *21*
Research Designs: Exploratory, Descriptive, and Causal *21*
Exploratory Research *22*
Descriptive Research *22*
Causal Research Designs *23*
Designing the Research Methodology *24*
Survey *24*
Observation *25*
Experimentation *25*
Qualitative Techniques *26*
Specialised Techniques *26*
Plan for Sampling, Field Work, and Analysis *27*
Sampling Plan *27*
Field Work Plan *28*
Briefing *28*
Debriefing *29*
Analysis Plan and Expected Outcome *29*
Expected Outcome *30*
Budget and Cost Estimation *30*
Presentation, Report, and Marketing Action *31*
Case Study 1 32
Summary 33
Assignment Questions 34

3. Research Methods and Design—Additional Inputs **36**
Sources of Secondary Data *36*
Disadvantages of Secondary Data *37*
Exploratory and Conclusive Research *38*
Major Qualitative Research Techniques *39*
Depth Interview *39*
Focus Group *40*
Projective Techniques *40*

Validity of Research *41*
Experiments *41*
Test Marketing *41*
Case Study: Consumer Perception of High-end IT Education *43*
Summary *45*
Assignment Questions *47*

4. Questionnaire Design: A Customer-centric Approach 48

Designing Questionnaires for Market Research 48
Language 48
Difficulty Level 49
Fatigue 49
Cooperation with Researcher *49*
Social Desirability Bias *49*
Ease of Recording *50*
Coding *50*
Purpose of a Questionnaire *50*
Sequencing of Questions *50*
Biased and Leading Questions *50*
Monotony *51*
Analysis Required *51*
Scales of Measurement Used in Marketing Research *51*
Nominal Scale *51*
Ordinal Scale *52*
Interval Scale *52*
Ratio Scale *52*
Structured and Unstructured Questionnaires *53*
Structured Questions *53*
Structured Answers *53*
Open-ended and Closed-ended Questions *54*
Disguised *Versus* Undisguised Questions *54*
Types of Questions *55*
Open-ended Question *56*
Dichotomous Questions *56*
Multiple-choice Questions *56*
Ratings or Rankings *56*
Paired Comparisons *57*
Semantic Differential *57*

How to Choose a Scale and Question Type *57*
Transforming Information Needs into a Questionnaire *58*
Example of Information Needs *58*
Double-barrelled Questions *61*
Good Questionnaires and Bad Questionnaires *61*
Reliability and Validity of a Questionnaire *62*
Reliability and Validity *62*
What is a Construct? *62*
Content Validity *62*
Criterion Validity or Predictive Validity *63*
Construct Validity *63*
Reliability of a Scale *64*
Summary *65*
Case Study 1: Tamarind Menswear *68*
Case Study 2: Casual Clothing References of Youth *69*
Case Study 3: Parryware—A Survey on Consumers Perception of Bathroom and Sanitaryware *76*
Assignment Questions *83*

5. Sampling Methods—Theory and Practice 90
Basic Terminology in Sampling *90*
Sampling Element *90*
Population *91*
Sampling Frame *91*
Sampling Unit *91*
The Sample Size Calculation *91*
Formula for Sample Size Calculation when Estimating Means (for Continuous or Interval-scaled Variables) *92*
Formula for Sample Size Calculation when Estimating Proportions *94*
Other Issues that Affect Sample Size Decisions *96*
1. Number of Centres *96*
2. Multiple Questions *96*
3. Cell Size in Analysis *96*
4. Time and Budget Constraints *97*
5. The Role of Experience in Determination of Sample Size *97*
Sampling Techniques *97*
Probability Sampling Techniques *97*
Non-probability Sampling Techniques *104*
Census *Versus* Sample *105*

Types of Errors in Marketing Research *106*
Sampling Error *106*
Non-sampling Error *106*
Total Error *106*
Summary *107*
Assignment Questions *108*

6. Field Procedures **109**
Design of Field Work *109*
Selection of Cities/Centres *109*
Organising Field Work *110*
Quotas *110*
Selection of Respondents *110*
Control Procedures on the Field *111*
Briefing *111*
Debriefing *111*
Summary *111*
Assignment Questions *112*

7. Planning the Data Analysis **114**
Processing of Data with Computer Packages *114*
Statistical and Data Processing Packages *115*
Types of Analysis *115*
Data Processing *115*
Data Input Format *116*
Coding *116*
Variables and Variable Labels *116*
Variable Format *117*
Value Labels *117*
Record Number/Case Number *117*
Missing Data *118*
Statistical Analysis *118*
Hypothesis Testing and Probability Values (p-values) *119*
Approaches to Analysis *119*
Three Types of Analysis *119*
Hypothesis Testing *123*
Summary *127*
Assignment Questions *128*
SPSS Data Input and t-test Commands *128*

Integrated Case Studies for Part I *130*
Case Study 1: Crocin *130*
Case Study 2: Detergents *141*
Case Study 3: BPL *155*

PART 2: DATA ANALYSIS

8. Simple Tabulation and Cross-tabulation **181**
Univariate and Bivariate Analysis *181*
Dependent and Independent Variables *181*
Demographic Variables *182*
First Stage Analysis—Simple Tabulation *182*
Computer Tabulation *183*
Percentages *183*
Simple Tabulation for Ranking Type Questions *184*
Tabulating Ratings *185*
Second Stage Analysis—Cross-Tabulation *186*
Calculating Percentages in a Cross-tabulation *186*
Cross-tabulation of More Than Two Variables *187*
Lack of Causal Inference in Cross-tabulations *187*
The Chi-squared Test for Cross-tabulations *187*
Chi-squared Test: An Illustration *188*
Measures of the Strength of Association Between Variables *190*
Doing More with Data (Transformation of Variables and Use of Part-samples) *191*
Summary *194*
Assignment Questions *195*
SPSS Commands for Frequency Tables, and Cross-tabs with Chi-squared Test *196*
Case Study 1: Chi-square Test for Cross-tabs *197*
Case Study 2: Chi-square Test for Cross-tabs *200*
Case Study 3: Chi-square Test *204*

9. ANOVA and the Design of Experiments **208**
Introduction *208*
Applications *208*
Methods *209*
Variables *209*
Experimental Designs *209*
Completely Randomised Design in a One-way ANOVA *209*
Randomised Block Design *210*

Latin Square Design *212*
Factorial Design with Two or More Factors *212*
Additional Comments *213*
Pairwise Tests *213*
Summary *217*
Assignment Questions *218*
SPSS Commands for ANOVA *220*
Case Study 1: ANOVA *221*
Case Study 2: ANOVA *226*
Case Study 3: ANOVA *235*

10. Correlation and Regression: Explaining Association and Causation 242
Application Areas *242*
Methods *243*
Recommended Usage *244*
Worked Example *244*
Problem *244*
Input Data *245*
Correlation *245*
Regression *245*
Regression Output *245*
Predictions *246*
Forward Stepwise Regression *247*
Backward Stepwise Regression *247*
Additional Comments *248*
Summary *252*
Assignment Questions *253*
SPSS Commands for Correlation and Regression *254*
Case Study 1: Correlation and Regression *256*
Case Study 2: Correlation and Regression *267*
Case Study 3: Correlation and Regression *273*

11. Discriminant Analysis for Classification and Prediction 280
Application Areas *280*
Methods *281*
Variables and Data *281*
Predicting the Group Membership for a New Data Point *281*
Accuracy of Classification *281*
Stepwise/Fixed Model *282*

Relative Importance of Independent Variables *282*
Apriori Probability of Classification into Groups *282*
Worked Example *282*
Problem *282*
Input Data *283*
Interpretation of Computer Output *283*
Additional Comments *285*
Summary *289*
Assignment Questions *290*
SPSS Commands for Discriminant Analysis *292*
Case Study 1: Discriminant Analysis *293*
Case Study 2: Discriminant Analysis *297*
Case Study 3: Discriminant Analysis *302*

12. Logistic Regression for Classification and Prediction 308
Application Areas *308*
Methods *308*
The Algorithm *310*
Logistic Regression *Versus* Linear Discriminant Analysis *310*
Numerical Example with SPSS *310*
Interpretation of Output *311*
Statistical Significance *312*
Predictors *312*
Classification of New Customer *313*
Summary *313*
Assignment Questions *313*
SPSS Commands for Logistic Regression *314*
Case Study 1: Logistic Regression *315*
Case Study 2: Logistic Regression *318*
Case Study 3: Logistic Regression *322*

13. Factor Analysis for Data Reduction 326
Application Areas *326*
Methods *327*
Recommended Usage *328*
Worked Example *328*
Input Data *328*
Interpretation of Computer Output *329*
Additional Comments *330*
Appendix 1 *331*

Summary 334
Assignment Questions 334
SPSS Commands for Factor Analysis 335
Case Study 1: Factor Analysis 336
Case Study 2: Factor Analysis 340
Case Study 3: Factor Analysis 344

14. Cluster Analysis for Market Segmentation 348

Application Areas 348
Methods 349
Data/Scales of Variables 349
Recommended Usage 349
Worked Example 350
Input Data 350
Output and Its Interpretation 350
Stage 1 351
Stage 2 352
Cluster 1 352
Cluster 2 353
Cluster 3 353
Cluster 4 353
ANOVA 353
Additional Comments on Cluster Analysis 354
Objects 354
Scale 354
Statistical Tests 354
Summary 360
Assignment Questions 361
SPSS Commands for Cluster Analysis 362
Case Study 1: Cluster Analysis 363
Case Study 2: Cluster Analysis 374
Case Study 3: Cluster Analysis 382

15. Multidimensional Scaling for Brand Positioning 394

Application Areas 394
Methods 395
Recommended Usage 396
Worked Example 396

Problem 396
Input Data 396
Interpretation of Computer Output 397
3-Dimensional Solution 397
Additional Comments 398
Summary 402
Assignment Questions 403
SPSS Commands for Multidimensional Scaling 404
Case Study 1: Multidimensional Scaling 405
Case Study 2: Multidimensional Scaling 411
Case Study 3: Multidimensional Scaling 416

16. Conjoint Analysis for Product Design 421
Application Areas *421*
Methods *422*
Recommended Usage *423*
Number of Attributes and Levels *423*
Number of Combinations *423*
Worked Example *423*
Ranking *424*
Running Conjoint as a Regression Model *424*
Output and its Interpretation *425*
Utilities Table for Conjoint Analysis *425*
Combination Utilities *426*
Individual Attributes *426*
Additional Comments *426*
Summary 430
Assignment Questions 431
SPSS Commands for Conjoint Analysis 433
Case Study 1: Conjoint Analysis 435
Case Study 2: Conjoint Analysis 440
Case Study 3: Conjoint Analysis 447

17. Attribute-based Perceptual Mapping Using Discriminant Analysis 454
Application Areas *454*
Methods *455*
Worked Example *455*
Problem *455*
Discriminant Analysis Output *456*
Summary of Canonical Discriminant Functions *457*

Unstandardised Coefficients *458*
Putting Variables/Attribute Vectors on the Above Map *459*
Brands and their Association with Attributes/Dimensions *461*
Attribute vs. Non-attribute Based Perceptual Maps: Which are Better? *461*
Summary *461*
Assignment Questions *462*
SPSS Commands for Attribute-based Perceptual Mapping Using Discriminant Analysis *462*
Case Study 1: Attribute-based Perceptual Mapping Using Discriminant Analysis *464*
Case Study 2: Attribute-based Perceptual Mapping Using Discriminant Analysis *471*
Case Study 3: Attribute-based Perceptual Mapping Using Discriminant Analysis *477*

18. Structural Equation Modeling (SEM) for Complex Models (including Confirmatory Factor Analysis) 485

Confirmatory Factor Analysis *485*
Tests Used in CFA and SEM *487*
Goodness-of-fit Tests Comparing the Given Model with an Alternative Model *489*
Chi-square Test in SEM *489*
Confirmatory Factor Analysis *490*
STATISTICA Commands for CFA *490*
SEM Commands Using STATISTICA *491*
Conclusion from SEM Analysis *496*
Summary *496*
Appendix *497*
Assignment Questions *512*

PART 3: APPENDICES

Appendix 1: Industrial Marketing Research 515

Defining the Target Population *515*
Applications *515*
Who Does Industrial Marketing Research? *516*
Internal *versus* External *516*
Technical Qualification of Researcher *517*
Questionnaire Design *517*
Checklists *517*
Use of Secondary Research *518*
Use of Industry Experts *518*

Analysing Government Policy *518*
Forecasting Derived Demand *518*
Assignment 519
Marketing Research for Product Redesign: A Case Study of ABC Ltd. 520

Appendix 2: Careers in Marketing Research 523
Major Companies in Marketing Research 523
Jobs in Marketing Research 524
Research Executive 524
Statistical Analyst 525
Field Supervisors 525
Field Staff 525
Summary of Job Prospects 526
Growth Prospects 526
Getting Business 526
Summary 526
Assignment Questions 527

***References* 528**

***Index* 530**

The McGraw·Hill Companies

PART 1 Fundamentals of Marketing Research

- Introduction, Evolution, and Emerging Issues
- The Marketing Research Process: An Overview
- Research Methods and Design—Additional Inputs
- Questionnaire Design: A Customer-centric Approach
- Sampling Methods—Theory and Practice
- Field Procedures
- Planning the Data Analysis

CHAPTER 1 INTRODUCTION, EVOLUTION, AND EMERGING ISSUES

Learning Objectives

In this chapter, we will

- Define the role of marketing research in making marketing decisions
- Introduce other inputs, such as intelligence, into marketing decisions
- Look at typical marketing research applications at two levels—strategic and tactical
- Discuss when marketing research should be done
- Discuss the limitations of marketing research
- Introduce Primary and Secondary Research
- Ponder over ethical considerations in marketing research, and the consumer's Right to privacy
- Take a brief look at the emerging issues such as the Internet based/online research, and Data Warehousing/Data Mining

ROLE OF MARKETING RESEARCH IN A MARKETING PLAN

Before we go into a description of the process of marketing research, we will try and understand where exactly marketing research fits into the marketing function. Every business works on an explicit or implicit 'business plan', which comprises both the corporate and the competitive strategies of the firm. To implement the above two strategies, there are functional areas, which have their own strategies or plans. The major functional areas of business are marketing, production, finance, and human resource management (information technology is usually an enabler for all the functional areas).

The marketing plan usually follows the marketing strategy. *Marketing strategy can be defined in simple terms as segmentation, target market selection, and positioning*. Marketing research plays an

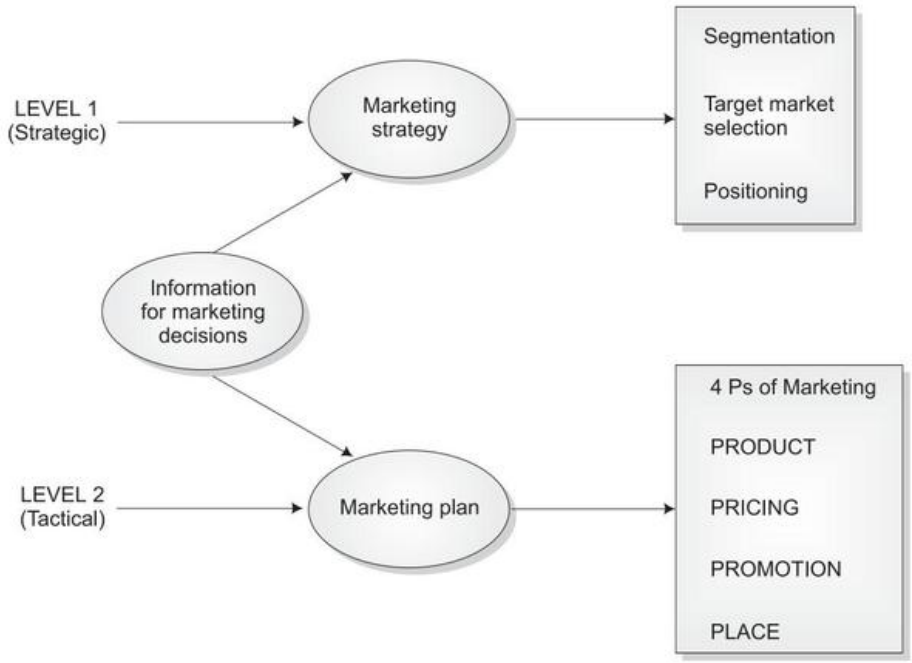


FIGURE 1.1 The Role of Marketing Research

important role in deciding on the marketing strategy by providing information necessary for choosing an appropriate strategy. This could be termed as marketing research at the strategic level of marketing.

Once marketing strategy is in place, the marketing plan at the tactical level proceeds to the arena of the 4 'P's of marketing—Product, Price, Promotion, and Place (Distribution). Readers who have taken a first course in Marketing Management will be familiar with this terminology (4 'P's) and we will not go into the details here. It is sufficient to say that most of the decisions related to marketing plans would be in terms of the product design or packaging, pricing, short-term and long-term promotions, and distribution and logistics (currently, the term supply chain management is also used to denote the incoming and outgoing logistics management). To make a decision regarding any of the above, a marketing manager needs information. This critical role of providing information is fulfilled by marketing research.

MARKETING INTELLIGENCE VERSUS MARKETING RESEARCH

The information, which we have talked of in the above section, necessary for making critical marketing decisions can be obtained in a variety of ways through a combination of sources. First, let us take a look at the two distinct processes that act as information providers to marketing managers. Both together could be termed as the *Marketing Information System*.

The first process, *Marketing Intelligence, could be defined as "an ongoing process of continuously collecting information about the industry in which our company operates, competitors' moves in*

marketing or other functional areas, related industries (e.g. suppliers or substitute products), government policies and actions in areas of export, import, taxation, liberalisation, consumer law enforcement, environmental protection, and so on".

Marketing Research, the second process, could be defined as "a clearly defined search for answers to some questions, which if answered would lead our company to make critical marketing decisions on a strategic or tactical level". This is not a technical definition of the marketing research function, but it serves the purpose of illustrating the differences between itself and marketing intelligence.

TABLE 1.1 Marketing Intelligence versus Marketing Research

Marketing Intelligence	*Marketing Research*
Ongoing process	Project based on information gap
Usually done in-house	Mostly outsourced to M.R. Companies
Not meant for immediate action	Action oriented
General purpose	Very specific answers to questions
Focus on competition, environment	Focus on consumers, influencers, etc.

As we can see, marketing intelligence is a continuous process, whereas marketing research is generally activated when there is a question (or questions) to be answered-in other words, only when there is an 'Information Gap'. Marketing research is generally taken up as a short-term project with a clearly defined time schedule, budget, and a clearly defined output, which should aid marketing decision-making.

To illustrate, let us say a manufacturer of branded coffee is contemplating some critical decisions. The company's market share in a region is declining, and a competitor is making slow and steady inroads. It could initiate a marketing research project to collect data on why its customers are switching brands. This becomes the marketing research 'problem'—to find out why the company is losing market share. Based on the problem, a methodology for conducting research would be worked out. The research could be done internally by company staff or out-sourced to an external market research company (such as ORG-MARG, IMRB in the Indian context). After the study report comes in, marketing managers of the coffee brand would use the findings to redesign their marketing strategy.

We will contrast this example of marketing research with an example of marketing intelligence to further explain the difference between the two. Suppose that this same company, which sells branded coffee, had set up an internal intelligence cell to collect pertinent information on its industry, competitors, consumer lifestyle trends and so forth. Suppose that it was headed by a competent person with a couple of assistants. The way they might operate is as follows.

They may subscribe to business newspapers such as the *Economic Times*, business magazines such as *Business India* or *Business World*, trade/industry journals pertaining to the coffee, tea or related industries, and keep their eyes open to any reports published in the press or aired on television that could affect their coffee sales. They could even be sensitive to changes in government policy on taxes, import, export, auctions, price fixation, and other relevant issues. They could monitor competitors' annual reports and statements of company chairmen in their industry. It could even extend to a study of competing cold beverages like colas, which could be substitutes for tea/coffee in some segments or in

certain seasons. Scanning of websites relevant to the industries or issues mentioned above could be part of their job.

As we can see from the description of marketing intelligence above, it is an ongoing activity with a wide scope. It is quite possible that the marketing intelligence cell could have monitored a move away from its brand strengths to some competitor's brand attributes. For example, consumers may have moved to a healthy lifestyle, and a competitor's brand may be emphasising it more in its advertising and promotion. This could have happened over a period of three to five years. It may take a trained and qualified professional with a keen perception to determine a trend from fragmented pieces of information. Thus how the marketing intelligence issue is handled is a serious concern and it may be best handled in-house.

Who Does the Marketing Research?

In the case of marketing research, it is not necessary to handle it in-house all the time. It could easily be out-sourced. There are many reasons for this. First, it is an intermittent activity, based on an identified information gap or information need. Second, a large number of professional research firms are available to do the needful. They have branches all over the country (in case of India) and some have global associates if research needs to go beyond Indian boundaries. Some of these firms specialise in consumer research, some in industrial research, some in qualitative research, and some in quality of service audits. ORG, a major Indian research firm (it is now A.C. Nielsen ORG-MARG, after a merger between the two companies), is a specialist in retail audits. Thus, it is usually economical for most companies to get their marketing research done from one of the many research firms—IMRB, ORG-MARG, T N Sofres MODE, Gallup-MBA being some of the prominent ones. But in certain large consumer goods companies which have multiple brands, it may be worthwhile having an in-house department consisting of a few qualified researchers. Other reasons for doing research through in-house staff could be to protect confidential information such as new product designs or pricing. But it is generally not viable to do marketing research internally alone, as the cost of hiring and retaining qualified staff may be high, so there could be a mix of internal and out-sourced research.

TYPICAL APPLICATIONS OF MARKETING RESEARCH

Applications of marketing research can be divided into two broad areas:

1. Strategic
2. Tactical

Among the strategic areas, marketing research applications would be demand forecasting, sales forecasting, segmentation studies, identification of target markets for a given product, and positioning strategies identification.

In the second area of tactical applications, we would have applications such as product testing, pricing research, advertising research, promotional research, distribution and logistics related research. In other words, it would include research related to all the 'P's of marketing: how much to price the product, how to distribute it, whether to package it in one way or another, what time to offer a service, consumer satisfaction with respect to the different elements of the marketing mix (product, price,

promotion, distribution), and so on. In general, we would find more tactical applications than strategic applications because these areas can be fine-tuned more easily, based on the marketing research findings. Obviously, strategic changes are likely to be fewer than tactical changes. Therefore, the need for information would be in proportion to the frequency of changes.

The following list is a snapshot of the kind of studies that have actually been done in India.

1. A study of consumer buying habits for detergents—frequency, pack size, effect of promotions, brand loyalty and so forth.
2. To find out the potential demand for ready-to-eat chapatis in Mumbai city.
3. To determine which of the three proposed ingredients—tulsi, coconut oil or neem, the consumer would like to have in a toilet soap
4. To find out what factors would affect the sales of Flue Gas Desulphurisation equipment (industrial pollution control equipment)
5. To find out the effectiveness of the advertising campaign for a car brand
6. To determine brand awareness and brand loyalty for a branded PC (Personal Computer)
7. To determine appropriate product mix, price level, and target market for a new restaurant
8. To find the customer satisfaction level among consumers of an Internet service provider
9. To determine factors which influenced consumers in choosing a brand of cellular phone handset
10. To find out the TV viewing preferences of the target audience in specific time slots in early and late evenings

As the list shows, marketing research tackles a wide variety of subjects. The list is only indicative, and the applications of marketing research in reality can be useful for almost any major decision related to marketing. The next sections discuss some typical application areas.

Concept Research

During a new product launch, there would be several stages—for example, concept development, concept testing, prototype development and testing, test marketing in a designated city or region, estimation of total market size based on the test marketing, and then a national rollout or withdrawal of the product based on the results.

The first stage is the development of a concept and its testing. The concept for a new product may come from several sources—the idea may be from a brain-storming session consisting of company employees, a focus group conducted among consumers, or the brainwave of a top executive. Whatever may be its source, it is generally researched further through what is termed as concept testing, before it goes into prototype or product development stages.

A concept test takes the form of developing a description of the product, its benefits, how to use it, and so on, in about a paragraph, and then asking potential consumers to rate how much they like the concept, how much they would be willing to pay for the product if introduced, and similar questions. As an example, the concept statement for a fabric softener may read as follows:

This fabric softener cum whitener is to be added to the wash cycle in a machine or to the bucket of detergent in which clothes are soaked. Only a few drops of this liquid will be needed per wash to whiten

white clothes and also soften them by eliminating static charge. It will be particularly useful for woollens, undergarments and baby's or children's clothes. It will have a fresh fragrance, and will be sold in handy 200 ml, bottles to last about a month. It can also replace all existing 'blues' with the added benefit of a softener.

This statement can be used to survey existing customers of 'blues' and whiteners, and we could ask customers for their reactions on pack size, pricing, colour of the liquid, ease of use, and whether or not they would buy such a product. More complex concept tests can be done using Conjoint Analysis where specific levels of price or product/service features to be offered are pre-determined and reactions of consumers are in the form of ratings given to each product concept combining various features. This is then used to make predictions about which product concepts would provide the highest utility to the consumer, and to estimate market shares of each concept. The technique of Conjoint Analysis is discussed with an example in Part II of the book.

Product Research

Apart from product concepts, research helps to identify which alternative packaging is most preferred, or what drives a consumer to buy a brand or product category itself, and specifics of satisfaction or dissatisfaction with elements of a product. These days, service elements are as important as product features, because competition is bringing most products on par with each other.

An example of product research would be to find out the reactions of consumers to manual cameras versus automatic cameras. In addition to specific likes or dislikes for each product category, brand preferences within the category could form a part of the research. The objectives may be to find out what type of camera to launch and how strong the brand salience for the sponsor's brand is.

Another example of product research could be to find out from existing users of photocopiers (both commercial and corporate), whether after-sales service is satisfactory, whether spare parts are reasonably priced, and easily available, and any other service improvement ideas—for instance, service contracts, leasing options or buy-backs and trade-ins.

The scope of product research is immense, and includes products or brands at various stages of the product life cycle—introduction, growth, maturity, and decline. One particularly interesting category of research is into the subject of brand positioning. The most commonly used technique for brand positioning studies (though not the only one) is called Multidimensional Scaling. This is covered in more detail with an example and case studies in Part II as a separate chapter.

Pricing Research

Pricing is an important part of the marketing plan. In the late nineties in India, some interesting changes have been tried by marketers of various goods and services. Newer varieties of discounting practices including buy-backs, exchange offers, and straight discounts have been offered by many consumer durable manufacturers—notably AKAI and AIWA brands of TVs. Most FMCG (fast moving consumer goods) manufacturers/marketers of toothpaste, toothbrush, toilet soap, talcum powder have offered a variety of price-offs or premium-based offers which affect the effective consumer price of a product.

Pricing research can delve into questions such as appropriate pricing levels from the customers' point of view, or the dealer's point of view. It could try to find out how the current price of a product is perceived, whether it is a barrier for purchase, how a brand is perceived with respect to its price and relative to other brands' prices (price positioning). Here, it is worth remembering that price has a functional role as well as a psychological role. For instance, a high price may be an indicator of high quality or high esteem value for certain customer segments. Therefore, questions regarding price may need careful framing, and careful interpretation during the analysis.

Associating price with value is a delicate task, which may require indirect methods of research at times. A bland question such as

"Do you think the price of Brand A of refrigerators is appropriate?" may or may not elicit true responses from customers. It is also not easy for a customer to articulate the price he would be willing to pay for convenience of use, easy product availability, good after-sales service, and other elements of the marketing mix. It may require experience of several pricing-related studies before one begins to appreciate the nuances of consumer behaviour related to price as a functional and psychological measure of the value of a product offering.

An interesting area of research into pricing has been determining price elasticity at various price points for a given brand through experiments or simulations. Price framing, or what the consumer compares (frames) price against, is another area of research. For example, one consumer may compare the price of a car against an expensive two-wheeler (his frame of reference), whereas another may compare it with an investment in the stock market or real estate. Another example might be the interest earned from a fixed deposit, which serves as a benchmark for one person before he decides to invest in a mutual fund, whereas for another, the investment may be a substitute for buying gold, which earns no interest. In many cases, therefore, it is the frame of reference used by the customer which determines 'value' for him of a given product. There are tangible as well as intangible (and sometimes not discernible) aspects to a consumer's evaluation of price. Some of the case studies at the end of Part I include pricing or price-related issues as part of the case.

Distribution Research

Traditionally, most marketing research focuses on consumers or buyers. Sometimes this extends to potential buyers or those who were buyers but have switched to other brands. But right now, there is a renewed interest in the entire area of logistics, supply chain, and customer service at dealer locations. There is also increasing standardisation from the point of view of brand building, in displays at the retail level, promotions done at the distribution points. Distribution research focuses on various issues related to the distribution of products including service levels provided by current channels, frequency of salespeople visits to distribution points, routing/transport related issues for deliveries to and from distribution points throughout the channel, testing of new channels, channel displays, linkages between displays and sales performance, and so on. As an example, a biscuit manufacturer wanted to know how it could increase sales of a particular brand of biscuits in cinema theatres. Should it use existing concessionaires selling assorted goods in theatres, or work out some exclusive arrangements? Similarly, a soft drink manufacturer may want to know where to set up vending machines. Potential sites could include roadside stalls, shopping malls, educational institutions, and cinema theatres. Research would help identify factors that would make a particular location a success.

In many service businesses where a customer has to visit the location, it becomes very important to research the location itself. For example, a big hotel or a specialty restaurant may want to know where to locate themselves for better visibility and occupancy rates. Distribution research helps answer many of these questions and thereby make better marketing decisions.

Advertising Research

The two major categories of research in advertising are:

(1) Copy
(2) Media

Copy Testing This is a broad term that includes research into all aspects of advertising—brand awareness, brand recall, copy recall (at various time periods such as day after recall, week after recall), recall of different parts of the advertisement such as the headline for print ads, slogan or jingle for TV ads, the star in an endorsement and so on. Other applications include testing alternative ad copies (copy is the name given to text or words used in the advertisement, and the person in the advertising agency responsible for writing the words is known as the copy writer) for a single ad, alternative layouts (a layout is the way all the elements of the advertisement are laid out in a print advertisement) with the same copy, testing of concepts or storyboards (a storyboard is a scene-by-scene drawing of a TV commercial which is like a rough version before the ad is actually shot on film) of TV commercials to test for positive/negative reactions, and many others. Some of these applications appear in our discussion of Analysis of Variance (ANOVA) in Part II and some case studies elsewhere in the book.

A particular class of advertising research is known as Tracking Studies. When an advertising campaign is running, periodic sample surveys known as tracking studies can be conducted to evaluate the effect of the campaign over a long period of time such as six months or one year, or even longer. This may allow marketers to alter the advertising theme, content, media selection or frequency of airing/releasing advertisements and evaluate the effects. As opposed to a snapshot provided by a one-time survey, tracking studies may provide a continuous or near-continuous monitoring mechanism. But here, one should be careful in assessing the impact of the advertising on sales, because other factors could change along with time. For example, the marketing programmes of the sponsor and the competitors could vary over time. The impact on sales could be due to the combined effect of several factors.

Media Research The major activity under this category is research into viewership of specific television programmes on various TV channels. There are specialised agencies like A.C. Nielsen wordwide which offer viewership data on a syndicated basis (i.e., to anyone who wants to buy the data). In India, both ORG-MARG and IMRB offer this service. They provide peoplemeter data with brand names of TAM and INTAM which is used by advertising agencies when they draw up media plans for their clients. Research could also focus on print media and their readership. Here again, readership surveys such as the National Readership Survey (NRS) and the Indian Readership Survey (IRS) provide syndicated readership data. These surveys are now conducted almost on a continuous basis in India and are helpful to find out circulation and readership figures of major print media.

ABC (Audit Bureau of Circulations) is an autonomous body which provides audited figures on the paid circulation (number of copies printed and sold) of each newspaper and magazine, which is a member of ABC.

Media research can also focus on demographic details of people reached by each medium, and also attempt to correlate consumption habits of these groups with their media preferences. Advertising research is used at all stages of advertising, from conception to release of ads, and thereafter to measure advertising effectiveness based on various parameters. It is a very important area of research for brands that rely a lot on advertising. At the time of writing, TV shows such as *Kaun Banega Crorepati* and *Kyunki Saas Bhi Kabhi Bahu Thi* on STAR PLUS are the top rated shows as measured by various agencies and reported in the business press in India. The other top rated programmes in India are usually cricket matches and film based programmes.

WHEN TO DO MARKETING RESEARCH?

Marketing research can be done when:

1. there is an information gap which can be filled by doing research.
2. the cost of filling the gap through marketing research is less than the cost of taking a wrong decision without doing the research.
3. the time taken for the research does not delay decision-making beyond reasonable limits. A delay can have many undesirable effects, like competitors becoming aware of strategies or tactics being contemplated, consumer opinion changing between the beginning and end of the study, and so forth.

LIMITATIONS OF MARKETING RESEARCH

It must be kept in mind that marketing research, though very useful most of the time, is not the only input for decision-making. For example, many small businesses work without doing marketing research, and some of them are quite successful. It is obviously some other model of informal perceptions about consumer behaviour, needs, and expectations that is at work in such cases. Many businessmen and managers base their work on judgement, intuition, and perceptions rather than numerical data.

There is a famous example in India, where a company commissioned a marketing research company to find out if there was adequate demand for launching a new camera. This was in pre-liberalised India, of the early 1980s. The finding of the research study was that there was no demand, and that the camera would not succeed, if launched. The company went ahead and launched it anyway, and it was a huge success. The camera was Hot Shot. It was able to tap into the need of consumers at that time for an easy-to-use camera at an affordable price.

Thus marketing research is not always the best or only source of information to be used for making decisions. It works best when combined with judgement, intuition, experience, and passion. For instance, even if marketing research were to show there was demand for a certain type of product, it still depends on the design and implementation of the appropriate marketing plans to make it succeed.

Further, competitors could take actions which were not foreseen when marketing research was undertaken. This also leads us to conclude that the time taken for research should be the minimum possible, if we expect the conditions to be dynamic, or fast-changing.

Differences in Methodology

The reader may be familiar with research studies or opinion polls conducted by different agencies showing different results. One of the reasons why results differ is because the methodology followed by each agency is usually different. The sampling method used, the sample size itself, the representativeness of the population, the quality of field staff who conduct interviews, and conceptual skills in design and interpretation all differ from agency to agency. Minor differences are to be expected in sample surveys done by different people, but major differences should be examined for the cause, which will usually lead us to the different methodologies adopted by them. Based on the credibility of the agency doing the research and the appropriateness of the methodology followed, the user decides which result to rely upon. A judgement of which methodology is more appropriate for the research on hand comes from experience of doing a variety of research.

To summarise, it is important to understand the limitations of marketing research, and to use it in such a way that we minimise its limitations.

Complementary Inputs for Decision-making

Along with marketing research, marketing managers may need to look into other information while making a decision. For example, our corporate policy may dictate that a premium image must be maintained in all activities of our company. On the other hand, marketing research may tell us that consumers want a value-for-money product. This creates a dilemma for the basic corporate policy, which has to be balanced with consumer perception as measured by marketing research.

Other inputs for decision-making could be growth strategies for the brand or product, competitors' strategies, and regulatory moves by the government and others. Some of these are available internally—for example, corporate policy and growth plans may be documented internally. Some other inputs may come from a marketing intelligence cell if the company has one. In any case, marketing decisions would be based on many of these complementary inputs, and not on the marketing research results alone.

SECONDARY AND PRIMARY RESEARCH

One of the most basic differentiations is between secondary and primary research. *Secondary research is any information we may use, but which has not been specifically collected for the current marketing research*. This includes published sources of data, periodicals, newspaper reports, and nowadays, the Internet. It is sometimes possible to do a lot of good secondary research and get useful information. But marketing research typically requires a lot of current data which is not available from secondary sources. For example, the customer satisfaction level for a product or brand may not be reported anywhere. The effectiveness of a particular advertisement may be evident from the sales which follow.

But why people liked the advertisement may not be obvious, and can only be ascertained through interviews with consumers. Also, the methodology for the secondary data already collected may be unknown, and therefore we may be unable to judge the reliability and validity of the data.

Primary research is what we will be dealing with throughout this book. *It can be defined as research which involves collecting information specifically for the study on hand, from the actual sources such as consumers, dealers or other entities involved in the research.* The obvious advantages of primary research are that it is timely, focussed, and involves no unnecessary data collection, which could be a wasted effort. The disadvantage could be that it is expensive to collect primary data. But when an information gap exists, the cost could be more than compensated by better decisions, which are taken with the collected data.

ETHICAL CONSIDERATIONS IN MARKETING RESEARCH

It may not appear so to the first time participant in a marketing research project, but there are many ethical considerations involved in doing primary research. In particular, the rights of the respondent are most important. The following are some issues regarding respondent's rights which the marketing researcher would do well to keep in mind while going about his job:

1. Any information collected for the purpose of marketing research from a respondent should not be misused for any other purpose. To ensure this, the field interviewing staff must be selected carefully.
2. Badgering or forcing respondents to answer a questionnaire or certain questions on the questionnaire is not good professional practice. A better approach is to explain why the particular question or questions are necessary, and then leave it to the respondent to decide if he wants to answer them. For example, questions on income of the respondent are always viewed with anxiety, and an explanation may be given before asking the question as to why it is necessary.
3. Confidentiality of the replies given in good faith by respondents should be protected. No one outside the research organisation doing the study and its client company must have access to the information provided by respondents.
4. If questions involved are of a personal nature, which could embarrass the respondent, he/she must be given an opportunity to think about it, and refuse to participate in a study. Trained field staff of appropriate sex may have to be used to reduce the embarrassment caused. For example, if the product is feminine sanitary napkins, it would be a good practice to use female interviewers to talk to respondents.
5. It is the marketing researcher's responsibility to accurately reflect the respondents' replies in his report to the sponsoring organisation (client). The report must not be based on any pre-conceived ideas of the marketing researcher, or twisting of the facts to suit any of the parties involved, including the client organisation. For instance, there used to be an unethical practice in India of a client ordering a marketing feasibility study to get project funding from financial institutions through cooked up 'favourable' research reports. Such practices give the whole industry a bad name, and should be avoided by the conscientious marketing researcher.

CONSUMER'S RIGHT TO PRIVACY

Is it acceptable to disturb a consumer (respondent) in the middle of the night to ask him questions about his buying behaviour or brand preference? Certainly not, we may all agree. But what of other times? When is it OK to do so? When is it not? If a respondent does not want to participate in a research study, what are the options?

Quite often, the field staff or telephonic interviewers (in case this method is used) get carried away by the importance they attach to the study. They do not pause to think about it from the respondent's point of view. He (or she) is really having to sacrifice his time, concentrate on the questions in the interview, and make an honest attempt to answer them. For doing this, he could be taking time away from other important things. Or, it could simply be inconvenient for him to answer questions at that particular place or time.

It is important for the researcher to be aware that a respondent has a right to privacy, and withdraw politely, if he encounters a less than enthusiastic response to his request for cooperation. Interviewing a non-cooperative or non-interested respondent may lead to a completion of the targeted quota of interviews, but could give bad data in the bargain—biased, useless or meaningless for the purpose of the study.

EMERGING ISSUES

Marketing Research in the Internet Era

In the present age of connectivity on the Internet, it is possible to do some of the research work on the net through the use of email or through a website which respondents can access. This is a substitute for methods such as personal interview or telephonic interview. In a country like India, however, the use of the Net for marketing research will remain limited, for the simple reason that penetration of computers and the Internet is low. If only two or three per cent of a given target population have access to the questionnaire sent out on the Net, the sample selection would be biased in their favour. Their opinions on any issues included in our research could be different from people who do not have access to the Net. For the same reason, samples taken from the telephone directory are sometimes considered biased in India. The personal interview is still preferred, and it may continue as the first choice for a few more years. But the telephone is already the preferred mode for many marketing research studies in the U.S., where most of the population has access to one. So, it is to be expected that the Internet would also be a good medium where the sample required fits the profile of the Internet users. It is a medium of the future. Its use will depend on how fast its penetration grows in India.

The advantages of doing research on the Net would be the possibility of showing both text and good graphics, and also the possibility of making the process interactive by asking consumers to create their product designs or pack designs out of given options, and be able to view them at once, before giving their ratings or preferences, and so on. The second major advantage of the Net-based research could be the speed of responses and with proper enabling software, the analysis could be automated and reports (or at least tables) generated very fast.

Of course, there is a chance of hacking by competitors or third parties who may get access to the data unless it is secured. But these problems have been tackled by companies such as banks and other institutions which do business on the Net. So they could be surmounted.

Online Research

We will now discuss some of the methods of doing online research, and their pros and cons. Basically, there are three methods of doing online surveys:

1. Email surveys
2. HTML forms
3. Downloadable interactive survey applications

Email surveys are often the fastest and simplest of the three methods. The advantages are very little setup time and wide reach. However, as email is limited to a flat text format, questionnaires cannot include complex variations in the sequencing of questions, graphics, and so on. But they do give the ability to address anyone who has an email ID. Some Internet service providers may also have segmented address lists available for relevant target populations in some studies.

HTML form surveys can create more complex questionnaires with skip patterns (skipping over questions or asking them in different sequence), grid-style rating questions with graphics and even sound. But they require a lot of programming effort. Setup time required is therefore higher. Typically, respondents are invited through email and are given a website address where the questionnaire can be found and answered.

Downloadable survey applications can be even more interactive. These are in the form of executable files which respondents can download on to their own computers. The advantage of these surveys is that many more features can be added to them, which can run on the respondent's operating system. The disadvantage is that respondent action required is more complex (downloading) and may result in more non-cooperation. That is, the response rate may be low. The programming cost and time is also high. Some respondents may not download for fear of catching a virus through the file. And there are certainly possibilities of hackers posing as market researchers trying this.

Sampling is generally a difficult task in online surveys. Making sure that the population is the one you want is the first difficulty, because someone else has usually compiled the list, and it is almost impossible to crosscheck. There is also the issue of selection bias. Only internet users can be the target population for such research studies. Non-Net users will remain unrepresented.

Qualitative research is possible on the Net in the form of chat sessions which work like focus groups. Transcripts of these can be available almost immediately, unlike in a physical focus group with an audio or manual recording. But facial expressions and emotional inputs cannot be captured online. A video recording in a physical setting is perhaps the best way of doing that. But web cameras can perhaps replace these to a certain extent.

A reasonable length of an online survey should be about 10–20 minutes of respondent time. A longer one will reduce cooperation, as in a physical or telephonic survey.

Data Warehousing and Data Mining

Another emerging technology related in some ways to marketing intelligence and marketing research is that of Data Warehousing and Data Mining. Essentially, it involves the capture of data on a regular basis from several sources. This huge amount of data is stored on computers in a virtual warehouse, and used to find patterns, test hypotheses as in normal research. For example, a credit card company may have card usage data over many years. This can be used to find patterns among things purchased, or used to build profiles of frequent buyers, high value buyers, globetrotters, and so on. This information can then be used for marketing actions such as specially designed promotions, bonuses or rewards, and others. These initiatives are sometimes called CRM initiatives because they can be used for Customer Relationship Management (CRM) activities of a company. The data typically are culled out from company records, or scanner data from retail checkout counters, bought out data, and so on. Survey data could be one of the inputs for a data warehouse, or it can be used as a complement to the data mined from a warehouse built by the company. Some of the techniques of analysis that are used in marketing research are also used to analyse the data mined from warehouses. For example, correlation, regression, cluster analysis, discriminant analysis, etc. can be used in data analysis in both cases. The volume of data is usually very high in a data mining application. The quality of data, since, it is collected at different times and places, may sometimes be suspect in data mining. Handling large amounts of data also requires software with higher capabilities. Analytical skills may involve asking

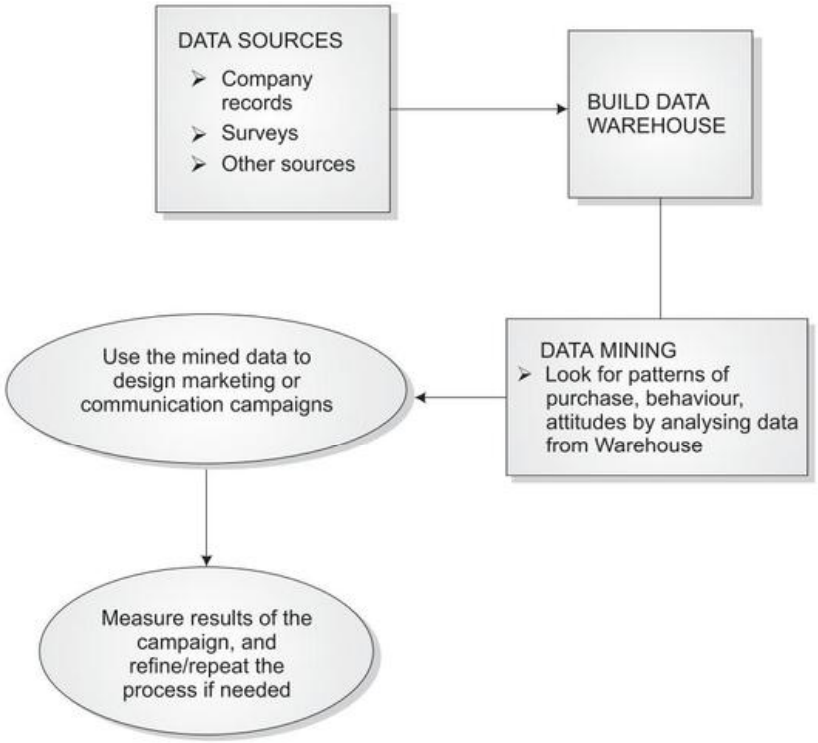


FIGURE 1.2 Using Data Warehouse and Data Mining

creative questions to seek new patterns in existing data. Some Indian companies that have begun to use data mining are ICICI Bank, Foodworld, and Max Hutchison, the cellular phone service provider.

SUMMARY

Marketing research plays the role of providing information for marketing decisions at two levels—strategic, and tactical. It is like a fact-finding mission, which seeks answers to the questions which a marketing manager may have—for example, is there a market for a new brand of talcum powder, why is brand X of a fairness cream the most popular, why a particular product failed in the market, which variant of the advertisement should be used, and so on.

Marketing intelligence is an ongoing activity for gathering information, and complements marketing research. Together, marketing research and marketing intelligence could be called the marketing information system.

Marketing research can either be done by the marketer in-house or it can be out-sourced to a professional marketing research company such as ORG-MARG. Typical applications of marketing research are concept research, product research, pricing research, distribution research and, advertising research.

Marketing research should be done if the cost of doing research is lower than the cost of a wrong decision being taken without research. Also, there should be a clear information gap that the research can fill. Thirdly, the research results should not take too long. That is, results must come in time for taking decisions.

Marketing research complements a marketing manager's judgement, intuition, and understanding, rather than substituting for these. Marketing research also has certain limitations due to a variety of errors that may be a part of the data collection, data handling or sample selection process. Typically, research studies that use different methodologies report differing results. So, an important part of doing and using research is understanding the methodology used, and an assessment of the researcher's credibility.

Secondary research means information that is collected by someone else for purposes other than yours. This includes reports in newspapers, magazines, the Internet, and so on. Primary research is one which collects data for the problem on hand, directly from the source, first-hand. For example, a consumer survey designed and done by your company to find out answers to questions which you have. Primary research is generally more useful but more expensive to do than secondary research.

Marketing researchers must follow some norms and ethics in collecting and using data. For example, a respondent's right to privacy should not be violated. Data should not be used for purposes other than the one it was collected for.

It is possible these days to do marketing research by using the Internet, though there are many limitations to this method. Some new tools such as data warehousing and data mining are being employed by companies as complements or substitutes to traditional marketing research. These techniques involve huge databases built up by companies over the years. These are analysed to find behaviour patterns and purchase probabilities among customers. Marketing actions are then initiated based on the findings.

ASSIGNMENT QUESTIONS

1. What is the role of marketing research at the strategic level of marketing? At the tactical level of marketing?
2. What is the difference between marketing intelligence and marketing research? Can they be used together?
3. Should marketing research by done by a company in-house, or should it be out-sourced to a marketing research agency? Explain your answer.
4. What do you mean by an information gap? Is marketing research the only way to fill an information gap?
5. Why is the methodology of doing marketing research important? What are some elements of research methodology?
6. Is it a good idea to do a given marketing research study through an Internet-based sample? Why or why not?
7. Can secondary research be a substitute for primary research? If yes, under what conditions?
8. What are the limitations of secondary research, which are overcome by primary research?
9. List out some ethical issues in marketing research, which companies/ individuals doing research must grapple with.
10. What is Right to Privacy? Why is it relevant to marketing research?

CHAPTER 2 THE MARKETING RESEARCH PROCESS—AN OVERVIEW

Learning Objectives

In this chapter, we will

- Understand the marketing research process through its different stages
- Define the need for marketing research and its translation into research objectives
- Discuss Exploratory, Descriptive, and Causal research designs
- Describe major research methods like survey, observation, and experimentation
- Understand the need for well-planned sampling, field work, and analysis
- Look at the basics of presentation and report writing to end the M.R. process

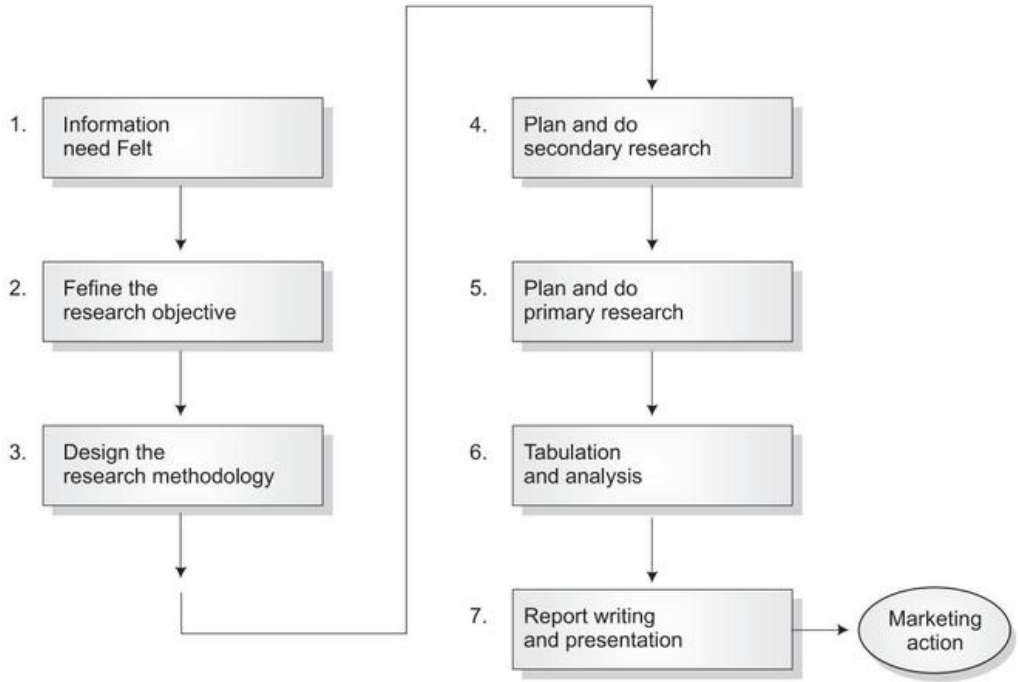
A marketing research project starts with an information need. It ends with an actionable report or presentation or both. In between are various steps to ensure that the marketing research project achieves what it set out to do. In this chapter we will take a look at the marketing research process—how it all happens.

A diagrammatic representation of the marketing research process is shown in the following figure.

We will now consider each of these steps in detail. The finer points of the process and some practical issues involved in the above steps will be the subject of the rest of this chapter.

INFORMATION NEED

This is the need that initiates a marketing research project. Usually, there is a realisation among the marketing managers who have to take a marketing decision—either strategic or tactical—that they need some more information from consumers or dealers or users (if different from buyers). Consider, for example, an expensive advertising campaign which has been running on television for three weeks. It may not have produced the expected jump in sales in some of the major sales territories. The client,



let us assume, is a shaving blades manufacturer. The marketing manager has to decide whether to discontinue the campaign, or change it, or reconfirm that the ad campaign is good. If the ad campaign is good, it may be some other marketing variables such as the price or distribution, or strong competitive promotions that are the reasons for sales not being upto expectations. One way to find out is to do marketing research. Therefore, the marketing manager has identified an information need, and it could be fulfilled by a marketing research study.

There could be a second marketing manager who is considering the launch of a new brand of deodorant in the market. He wants to know how to position the brand in the market, and get a rough estimate of what the market size would be in the chosen segments. He has an information need, which could be filled by doing a consumer survey.

A third marketing manager heads a popular music channel on TV. He wants to know which of his video disc jockeys is the most popular, and which show is the most watched. He could commission a study by an independent marketing research agency to do just that.

The above examples may have made the concept of information need clear to the reader. Of course, as the reader would expect, any need for information must be examined in terms of the cost of obtaining the required information. Also, the cost of not having this information should be estimated. The risk involved in taking a marketing decision with inadequate information, should be weighed against the cost of getting the information, and, hopefully taking a better-informed decision. It is not easy to estimate the risk, which can be considerable in terms of wasted marketing resources. It is generally a good idea to do the marketing research if one has the time and money, rather than shoot blindly. Neither approach, of course, guarantees success in meeting the marketing objectives of a company. Success depends on many factors, and information is only one of them.

DEFINING THE RESEARCH OBJECTIVE

If we do have an information need that can be met by doing marketing research, the next step would be to define the research objective in terms of that information need. For example, a study could have as its objective, "the determination of customer satisfaction with a brand of new frost-free refrigerator launched by our company". This research objective can be met by undertaking a survey of customers who have bought the new brand.

A research objective can be specified broadly, or narrowly. One common pitfall in the field of marketing research is to specify too many objectives for a single marketing research project. The thinking seems to be, "we are doing a study anyway, so let us find out everything we can". Sometimes this strategy backfires, and the data collection task turns unwieldy. It produces a mass of data that is not really needed at that point of time. In general, it is a good idea to be focussed in doing marketing research and have a few, clearly stated objectives. About four or five objectives in most cases are adequate to do a useful marketing research study.

If you are wondering how to decide which objectives to leave out from a long wish list compiled by your department, here is a simple test. You could ask a question—"Will I be able to take a better decision if I include this research objective in the study?" Only if the answer is a firm "yes", it should be included. The reason why this point is so important, is that every objective usually translates into a few questions on a questionnaire, and there is a limit to how many questions a respondent can honestly answer before his interest level goes down.

Sometimes, we call the research objective by another name—the research problem. Broadly, these two terms can be used interchangeably. Whatever the terminology used, the research should end up with useful information that enables a marketing manager or entrepreneur to make a better decision. To repeat what we have touched upon in Chapter 1 earlier, the real test of useful marketing research is how much it helps us in marketing action. If a report is meant to lie on a shelf, it is not really marketing research, but a waste of resources.

RESEARCH DESIGNS: EXPLORATORY, DESCRIPTIVE, AND CAUSAL

Let us look at the three major kinds of research designs to understand some of the basic underlying concepts. A research design provides the framework to be used as a guide in collecting and analysing data. But it is not necessary that a particular research design is always the best. Experience with different research designs will generally provide the researcher with the capability to match a research problem with an appropriate design. For example, in a study for a new English daily newspaper launched in Bangalore in the eighties, it was found that the sales were much below expectations. A survey was proposed. But as a complement to the survey, the author's team at a research agency proposed a Content Analysis of all the major dailies in Bangalore. This method analysed the coverage of various categories of news such as politics, sports, regional, national, city-based news and so on by the client's newspaper and the competitors. This gave vital insights to the publishers of the paper, and over a period, it became successful. This is just an example to show that sometimes unusual research designs do pay off.

Broadly speaking, we can classify research designs into the following three kinds:

1. Exploratory research
2. Descriptive research
3. Causal research

Exploratory Research

This is generally used to clarify thoughts and opinions about the research problem or the respondent population, or to provide insights on how to do more conclusive (causal) research. An example could be a chocolate manufacturer wanting to identify the ten most important variables his consumers use to decide on whether to buy a chocolate brand. The results of this exploratory study could provide him with inputs for a second study using Factor Analysis techniques (discussed in Part II of this book) to reduce the ten variables into a smaller set of factors. Another example of exploratory research is a focus group discussion among housewives to debate the future of convenience foods in India. It may be used to throw up ideas about new products, or suggest modifications to existing products through a free-wheeling discussion.

Often the researcher is new to a problem either because the product (or the brand) is new, or the researcher is studying it for the first time. In such cases, the first few studies tend to be exploratory in nature. Only after some experience with the product category, a researcher's confidence grows to design what can be termed as a 'causal' research study—where a more direct link between cause and effect can be established.

One major application of exploratory research, therefore, is to generate hypotheses for further studies. The methods used in exploratory studies can range from the usual surveys, to focus groups, to consultations with experts in the field, to analysis of selected cases. An example of the last may be to study three of a company's best salespeople, and three of the worst, to try and figure out what drives the sales of the products, and their motivations. This could help in designing a study of customers to find out more from them.

Descriptive Research

Most marketing research is of this type. Typically, descriptive studies are either (1) longitudinal or (2) cross-sectional.

1. Longitudinal Studies This generally takes the form of a sample of respondents who are studied over a period of time—from a few months to a few years. In marketing research, longitudinal (same sample) studies are usually done through a panel. A Panel is a sample of respondents chosen from the defined target population for the study. This sample could be of consumers, retailers or of any other type. A consumer panel could be used to study consumption of products/brands over a period of time. It could also be used to measure viewership of TV shows, or readership of magazines. A retail store audit is a variation of the panel, with data being collected from retail stores on the products/brands being stocked, shelf space allotted, sales and promotions and so on. The most commonly used form of retail audit is about sales of each brand in a given category, which serves as an estimate of market share in the areas covered by the audit.

Panel data has the advantage of enabling comparisons at different points of time. So there are elements that can be checked easily. For example, the effect of a change in price, pack design, or other elements of the marketing mix can be easily measured by comparing the sales or market share before and after the change. This is not so easy to do in typical survey data, because it is cross-sectional in nature, for only one point in time. One other advantage of panels is that if a quick check on something is needed, sample selection time can be saved by approaching panel members. In these days of the Internet it may be possible to get a quick response to a short survey of panel members in a matter of a couple of days.

There is of course a disadvantage to panel data. Panels suffer from a selection bias. Some people are more likely to agree to be on a panel than others, because it needs a commitment in terms of time and effort to regularly record and report data. This is the case even though an incentive or payment is offered to the members of a panel. This selection bias may make panels non-representative of the target population.

2. Cross-sectional Design This is most commonly used in marketing research. This is a one-shot research study at a given point of time, and consists of a sample (cross-section) of the population of interest. The typical market survey is of this type. Its advantages are that it gives a good overall picture of the position at a given time. It can cover many variables of interest, and is not affected by the movement of elements in the sample, because other elements can be substituted for them (at least in consumer research). The disadvantages could be that a cross-sectional study tends to rely too much on numbers, can be affected by poor quality of interviewers or supervisors, and tends to view the population in terms of too many generalisations—the 'average' consumer's views about anything, which may cloud the individuals or segments among the population. To some extent, the last mentioned problem can be overcome with certain techniques of analysis. For example, we can analyse data by town or region or by other segments to prevent unnecessary aggregation which is misleading.

On the whole, though, cross-sectional research appears to be most preferred by market researchers and their clients on account of its simplicity and understandability. It is also quite flexible in nature, and can take care of simple analysis as well as complex statistical methods.

Causal Research Designs

In research, we can never be completely sure that a particular variable (say X) influences another (say Y). But a causal design seeks to establish causation as far as possible, by employing controls and conditions under which we can state with reasonable confidence whether or not Y is affected by X. In addition to X and Y, of course, there may be other variables which could affect the relationship between X and Y. How to treat the other variables during the analysis of the effect of X on Y also forms part of the causal designs.

Causal designs differ from descriptive designs in their greater probability of establishing causality. The reason for this is that causal designs are similar to experiments done in a lab, where we know what goes in, what changes are made, and what results from the changes. Causal designs are also known as Experimental Designs, for this reason.

DESIGNING THE RESEARCH METHODOLOGY

Research methodology depends, to a large extent, on the target population, and how easy or difficult to access it is. The second factor which influences research methodology is, of course, the importance of decisions which will be taken based on the research. The accuracy level required is based on the criticality of the decision, which will follow.

The major parts of the research methodology are:

1. Research method—secondary and primary
2. Sampling plan
3. Questionnaire design (if applicable)
4. Field work plan
5. Analysis plan

As we said earlier, some of these depend on how easy or difficult it is to locate the target population of the study. But usually, the first thing one has to decide is the method to be used for data collection. Every research study will start with some information need. Sometimes, the information required can be collected entirely from published sources or internal records. This is called secondary research.

It is more usual, however, that we will need to collect data from primary sources—customers, buyers, users, dealers or some other respondents. It is possible to collect data from respondents by many different methods. The major methods commonly used are:

1. Survey
2. Observation
3. Experimentation
4. Qualitative techniques
5. Other specialised techniques

Quantitative methods are generally more popular than qualitative techniques in marketing research studies. Also, the survey technique is more popular than other techniques.

Survey

There are different ways a survey can be carried out. It can be done by telephone, by mail, or in person. In present times, it can even be done by email using the Internet. Each of these has its own merits and demerits. For example, personal interviews have the advantage that questions can be explained to respondents, and facial reactions or body language can be observed. Telephonic surveys have the advantage of low cost. But facial reactions cannot be observed. Internet surveys are quite new, but may have the same disadvantages that telephonic and mail surveys have. It is difficult to ensure that all target respondents have an opportunity for selection in the sample. Also, non-response bias may be high. Non-response bias means that some people who are inclined not to participate in such surveys do not get represented in the sample, leading to errors.

For example, every potential respondent for the survey may not be using the email, or even a computer. Therefore, the email survey does not represent a true sample of the target population for

many products or services. To that extent, the results may be wrong, compared to the errors in a door-to-door personal interview done with scientific probability sampling. But if some amount of error is acceptable (a detailed discussion of errors and their relationship with sample size appears in Chapter 6), and speed is of the essence, an email survey or a telephone survey would be excellent methods. A traditional mail survey would be much slower, by comparison.

At present, personal interviews are the preferred method for doing surveys in India. Telephone and mail surveys are used in a minority of cases where they are justified by the target population and the objective of the research.

Observation

Sometimes, however, we may not choose a survey, but some other method which is more appropriate. Observation, or experimentation could be the method of choice. Observation is a technique where the consumer's behaviour is recorded, usually without his knowledge. For example, a video camera in a retail store can be used to record a customer's behaviour while she buys a garment. If it is a full service store, like many Indian stores, she could ask for a particular brand or brands, look for specific colours, or fabric, or prices and so on in a particular sequence. Her facial reactions or eagerness or lack of interest when a piece is displayed to her can be recorded along with the garment. Viewed later, this video tape can be interpreted for the purchase factors, purchase behaviour, brand preference, price and colour preference, and matched with the lady's age and complexion—if she bought for herself.

The obvious advantage of this technique is that it is actual consumer behaviour that gets recorded, rather than their statements of purchase intention. Therefore, we get more accurate information. But a major problem might be that we are not sure if a representative sample of consumers have been chosen, because consumers can usually be recorded only at shops or public places, and we have no control over who shops at a given time. But still, it can be a valid technique to use, if the research objective is to get a fair idea of the kinds of things a consumer does while shopping for a particular product category in a particular kind of shopping environment. If a video recording is too expensive, an audio recording is possible, or even a data collector in person can observe and record his findings on paper.

Experimentation

This is the third major technique in quantitative research. This involves more control over the cause and effect, when compared to a survey. In experiments, we try to measure the effect of one or more variables by changing the level of some variables, and measuring the effects. For example, if an advertisement is released, and we measured the brand awareness of the advertised brand among a sample of target respondents, we would be doing an experiment. In the same way, a product test could be designed as an experiment, with three different variants of the product being tested on three randomly chosen sets of respondents from a target population. The modern method of Simulated Test Marketing (STM), is usually a design which can be termed an experiment.

A detailed discussion of experimental techniques with numerical examples appears in Chapter 9, titled ANOVA. The interested reader may refer to the same.

Qualitative Techniques

Sometimes, the research objective calls for more indirect methods of questioning, either because normal quantitative surveys are inadequate, or inappropriate. In such cases, qualitative methods, which probe the minds of respondents may be used. Here, the emphasis may be on free-wheeling interviews with open-ended, unstructured questions such as "What do you expect from a refrigerator?", "What needs does it fulfill?" or "What do you feel when a friend shoots an envious glance at your car?"

Other methods of qualitative research include the word associations where a respondent is asked to think of a word which comes to mind when he thinks of a brand. Other variations include associating each brand with a person or celebrity, or an animal, and so on.

The major requirement for using qualitative techniques is that we need a behavioural specialist such as a psychologist or sociologist to analyse the findings. The sample sizes in qualitative studies are usually small, and analysis and interpretation is not as easy as it is in quantitative studies. If done by non-experts, qualitative research can be completely misleading. But it is quite useful in probing the minds of the minority among a target population, who cannot reveal themselves fully in a structured survey. Qualitative techniques can also be used in combination with quantitative techniques to gain better insights into consumer mindsets.

An example of qualitative research is a study done by TVS Suzuki, among scooter and moped users in 1989 (cited in The Catalyst, Business Line, July 10, 1989). The research objective was to assess the impact of a newly launched scooterette from Bajaj on the market for TVS mopeds, and to try and find out what people expected TVS to do in response. The study was done only in one centre, and the method used was focus groups who discussed the motivations behind purchase of mopeds and scooters. Projective techniques were also used with respondents being asked to put themselves in place of existing moped brands and talk about themselves as if they were the brands. The concept of a low cost scooterette was then exposed to the participants, and their interest levels appeared high. This research formed one of the bases for TVS to design and launch the SCOOTY.

Specialised Techniques

There are three specialised techniques, used commonly by marketing researchers.

1. A Consumer Panel This is a sample of consumers chosen for keeping a record of what they buy in a given period or what TV shows they watch in a given period. They have to record this information in a diary, usually every week, and hand it over to the marketing research agency for tabulation and analysis. The special feature of this is that the sample remains the same for a year or six months. Usually, there is a payment for being part of such a sample, as it involves extra effort of recording data and maintaining the records.

2. Retail Audit Many companies routinely do a retail audit and publish the results (at least partially). Detailed reports are available for anyone to buy and use. A retail audit measures what brands are sold and their quantity sold in a particular period. It could be done weekly. In India, ORG is a company which routinely performs retail audits. It is a measurement done at the retail level, and if it covers a locality of interest to the marketer, it provides an update of how his brand is doing in

the given locality or territory. It also provides a comparative look at the sales of competing brands, and products in other categories. Both regional and national audits can be done. Usually, such audits are best done by a third party (independent agency), to reduce chances of bias, rather than the marketing company. Sometimes, similar studies are undertaken by the company for its own brands at either consumer level or retail level. Such studies are sometimes known as brand tracking studies—particularly when the tracking of sales or awareness is studied immediately after a new launch.

3. TV Audience Measurements These days, millions of rupees are spent in advertising on TV. It is important for the marketer to know who is watching the TV shows on which he has advertised. Or, to plan for a particular audience profile. There are now commonly used technologies, which record who is watching a given channel and show at any given time, for upto a week. These are called *Peoplemeters*, and are available in India for about Rs. 40,000/- a piece. Indian market research companies such as IMRB and ORG-MARG have already started using them, and their use is likely to grow. The branded names for the peoplemeters in use in India are TAM and INTAM.

Earlier, consumer panels used to be recruited and asked to maintain a diary of what was watched by whom in their household. The method was not very accurate. The new meters have changed the advertising patterns of many TV channels and individual shows after they were introduced in India.

PLAN FOR SAMPLING, FIELD WORK, AND ANALYSIS

The next stage in a marketing research study, after the primary research method has been decided upon, is the plan for

1. Sampling
2. Field work
3. Analysis

These are probably the most important in a study involving primary research, as the credibility and the accuracy of a study is dependent on these stages.

Sampling Plan

This is the statement of what will be the sample composition and size. This is the most critical of all decisions in the marketing research process, because we are usually trying to make a statement about the target population based on our study of the sample. For instance, if we find that 50% of our sample is favourably disposed towards brand A, we are likely to use it as a benchmark for the entire target market, give or take a few percentage points (due to errors). But in order to make the sample representative of the population, a lot of care has to be taken by the researcher.

In general, two precautions should be taken to ensure a good sample.

1. Use a probabilistic sampling technique which is not biased.
2. Try and divide the population to be sampled into segments or strata based on relevant parameters such as users/non-users, or classes based on age, income, and so on. Then, ensure that each segment gets represented adequately in the final sample. This also applies to studies

that are done in multiple cities. If a study is done in twenty cities, and if analysis is required by city (i.e. for each city separately), then the sample size for each city must be adequate for such analysis.

Generally, formulas can be used to determine sample sizes, but they suffer from some limitations. For a more detailed discussion, please refer to Chapter 6, titled "Sampling Methods—Theory and Practice". It is usually a blend of theory, practical limitations, and experience which generates the best sampling plan in any given research situation.

Field Work Plan

This is clearly linked to the sampling plan. Once the sampling centres (cities, towns, etc.) are decided on, and the sample sizes are determined for each, the next step is to plan on the following:

1. Who
2. When

The first question is who will do the field work for collecting data. Field work assumes that we are collecting data from respondents by going to the 'field'—that is, homes, offices, shops, dealerships, and so forth. Before doing field work, whoever is going out in the field needs to have an idea of what is to be collected—and its format of recording. In the traditional format of personal interviews (which is still the most popular format in India for various reasons discussed earlier), a questionnaire is used by the field workers in most cases. Sometimes, a checklist is used instead, if the situation demands it. We will assume here that the questionnaire has been developed. A detailed discussion of how to develop a good questionnaire appears in Chapter 4, titled "Questionnaire Design—A Customer-Centric Approach".

The second question is "when". In many studies carried out nationally, it is not possible always to simultaneously cover all centres, on the same days. There could be logistical problems for supervisors, or there may be difficulties in recruiting adequate field workers, and so on. But it is desirable to have a well-planned schedule so that all field work is completed in an orderly fashion, and cross-checks can be established.

Briefing

For all important studies, the research executive in charge should personally brief the field supervisor (the person who will actually supervise the team of field workers during the data collection). This briefing session is conducted after recruiting field workers, and ends with a practice round of mock interviews and questions from field workers on any special difficulties they may encounter in locating respondents, asking certain questions, and so on.

The mock interviews and the briefing session is designed to explain and clarify to the field workers "how" to go about their data collection task. In most studies, temporary field workers are recruited on a daily wage basis and paid on the basis of a minimum number of complete, usable questionnaires filled up.

The number of field workers required in each centre is usually estimated based on the sample size required, the locations where the sample can be found, the number of supervisors available, and the

time limit for completion of field work. These are communicated by the research executive in charge to the field supervisors in his branch offices, who generally recruit the field workers.

Debriefing

It is important that any problems on the field get reported to the field supervisor or the research executive, and solutions found quickly. These problems may include difficulty in locating target sample units, or non-cooperation in answering some questions, or difficulties in comprehension.

To minimise any problems the field staff may encounter, a *debriefing session* is usually held at the end of the first day's field work in each new centre (location). The field staff reports on the work progress, and problems faced in the field, if any. Solutions are thought of by the research executive or field supervisor, and implemented for the remaining part of the study.

Some of these problems are recognised even earlier if a *pilot study* of a small sample is performed, before starting regular field work. Alternatively, the first day's or half day's field work could be considered as a pilot study, and not included in the survey results.

ANALYSIS PLAN AND EXPECTED OUTCOME

Analysis is based on the answers given to questions. It is important to have an analysis plan in mind even before going to the field with a questionnaire. Regrettably, this is not always given the attention it deserves by the researcher. It is sometimes assumed that it can be done later, or that all possible analyses can be done anyway, so why bother to plan the analysis in advance. But for many reasons, it is vital to do so.

A very powerful reason is that the sample size gets reduced, if the analysis is done on parts of the sample. For instance, in a sample of 200 respondents, there could be 16 combinations of income (4 groups) and age (4 age groups). If analysis is performed for a combination of age and income, we get a 16-celled output matrix. Even assuming a uniform distribution of the sample into these 16 cells, each cell only gets a sample size of 100/16 or 12.5 persons. This may not be good enough to draw conclusions about the given age-income combination. But if it is known in advance that we will analyse the data by this combination, we can increase the sample sizes in each cell to say, 20 or 30 by incurring marginal additional cost. This cannot be done easily at the analysis stage, after all data has been collected and tabulated.

In certain cases, special statistical procedures or tests have to be performed. For example, in a procedure called multidimensional scaling (covered in a later chapter in Part II of this book), the questionnaire has to be constructed in a particular way. Otherwise, it is not possible to do the required analysis.

For these reasons, we must know in advance, at least the types of analyses we want to perform.

There are normally two very basic kinds of analyses in a marketing research study. These are:

1. Simple tabulation
2. Cross-tabulation

Simple Tabulation This involves counting the number of responses in each category for a question, and putting it in a frequency table form. This can be used to compute percentages, by dividing the number of responses by the sample size. Simple tabulation is done for each question in the questionnaire.

Cross-tabulation This is the result of counting simultaneously, answers to two or more different questions on a questionnaire. For example, one question may ask how frequently respondents buy a soap brand. Answers may vary from once a month to thrice a month. Another question on the same questionnaire may ask for their reaction to the fragrance of the soap. We may want to cross tabulate the responses to these two questions. How many of the people who liked the fragrance bought once a month, and how many of them bought twice or thrice a month? Similarly, how many who did not like the fragrance bought it once, twice or thrice a month?

While doing cross-tabulation, it is also necessary that the two questions (variables) that we are cross-tabulating must be related to or associated with each other. For instance, in the above example, it is possible that the frequency of soap purchase is a function of family size, rather than the liking for its fragrance.

It is possible to compute cross-tabulation data for any two questions on a questionnaire—but all of these may not be meaningful. For numerical examples of simple and cross-tabulation, please refer to Chapter 8.

Expected Outcome

One good way to think about expected outcome is to prepare a blank table of output, particularly for any cross-tabulations we may be interested in. This can be done after the questionnaire is designed, but before the field work is done. This helps to anticipate some of the problems in sampling, and corrective action can be taken easily to adjust sample sizes on the field.

Budget and Cost Estimation

There are two or three basic parameters which provide an estimate of how much a study is going to cost.

1. Sample size
2. How difficult it is to find the sampling units (respondents) and their geographical dispersion.
3. Who will do the field work

For example, if hired field workers are doing the field work, a study costs much less per respondent, than if a research executive conducts the interviews. In some industrial marketing research, a qualified research executive may in fact do the field work himself. But in most consumer product or service studies, it is hired temporary field workers who do it. In such cases, sample size is multiplied by the estimated cost per respondent to arrive at a total cost estimate.

This estimate is modified by the number of centres (geographical dispersion) for the study, and the difficulty in locating required respondents. For example, locating a two-wheeler owner for a given brand of two-wheeler (say, a Suzuki or Honda), is much easier than locating an owner of a luxury car—

say, a Mercedes, unless we have the cooperation of the company or its dealer. Additional cities for the survey may entail travel and communication cost for the research executive and supervisory staff in addition to normal cost of field work.

PRESENTATION, REPORT, AND MARKETING ACTION

After the tabulation and analysis is completed, the next step is usually a presentation of the major findings to the sponsor of the study. This includes a presentation of all the major tabulations (frequency tables) and cross-tabulations in percentage terms. It may also include a summary of major findings, and some recommendations. If any additional cross-tabulations are required, the client or sponsor usually requests them at this stage.

A formal report usually follows the presentation. This should normally contain the following:

1. Executive summary
2. Table of contents
3. Introduction
4. Research objectives
5. Research methodology
   - Sample design
   - Field work plan and dates
   - Analysis /expected outcome plan
   - Questionnaire copy (as annexure)
6. Analysis
   - Simple tabulation
   - Cross-tabulation
   - Any special analysis
7. Findings
8. Limitations
9. Recommendations for action
10. Bibliography/List of references

Based on the report, the client normally will take some marketing actions. This is the expected outcome of any marketing research study.

CASE STUDY 1

A recent case study for a cellular phone service provider in Chennai listed its research objectives and methodology (including sampling plan) for a marketing research study as follows:

SKYCELL, A CELLULAR OPERATOR/STUDY ON VALUE ADDED SERVICES LIKE SMS (SHORT MESSAGING SERVICE), VOICE MAIL, AND SO ON

Research Objectives

To find out

- whether people actually use the mobile phone just for talking
- to what extent the mobile phone is used for its VAS (Value Added Services)
- factors influencing choice of service provider
- awareness of Skycell's improved coverage

Locations Covered

Chennai city and the suburbs

Methodology

Primary data:
Through questionnaires

Sample Composition

- Mobile phone users
- Business persons
- Executives
- Youth

Sample size: 75

Age group: 18 – 45 years

The McGraw·Hill Companies

Questions

1. Can you add to the methodolgy section ?
2. Distribute the sample of 75 among the different categories of respondents mentioned under "Sample Composition".

SUMMARY

The marketing research process starts with a need felt for information by decision makers in the marketing, product development or new business development departments. The first step in the process is to define the research objective. This should ideally be a list of questions to which answers will be found at the end of the study. This list can also be called research questions or research problems. Care should be exercised so that the list is not unnecessarily long, because the length of a questionnaire depends on this list.

There are three broad types of research designs that a researcher has to be aware of, to design the research methodology for his study. *Exploratory* research is undertaken when the concept or product being researched is new, or the researcher needs to find out what hypotheses to use in future studies. Generally major decisions are not taken on the basis of exploratory research alone. *Descriptive research* can be either longitudinal, that is, a study spread over a long time, or cross-sectional, done at one point of time or within a short span of time. Longitudinal studies require repeated measurements of the same variables from the same or similar samples over a period. Tracking studies of an advertisement or a brand may fall in this category. But most marketing research studies are of the cross-sectional kind. They are done once, and provide a snapshot of (or describe) things as they are at that point in time. An obvious disadvantage is that variables such as consumer attitude may change with time. But this can be overcome by doing periodic research. On the plus side, cross-sectional research is quite effective in answering a few questions quickly and helping make better-informed decisions. *Causal research designs* try to establish cause and effect relationships between variables being studied. For example, between the amount of advertising and sales or brand awareness. Generally this class of research is called *Experimental* research, because the methods used are similar to laboratory-type experiments.

The most popular method of doing marketing research is a survey, followed by other methods like experimentation and observation. A survey can be done in person, on the telephone, by mail, or on the Internet. Each of these media/methods of execution has its own merits and demerits. In India as of today, personal interviews are the preferred method of doing surveys, at least for common consumers. In the US, telephone surveys are more widely used.

Observation is a technique used when direct questioning is not expected to produce accurate results. Actual purchase behaviour of consumers, for example, can be different from what they state as intended behaviour in a questionnaire. Also, some things like the influence of various point-of-purchase elements such as posters, special offers, or the salesperson's push, can be accurately gauged if a consumer is observed rather than questioned. A method of recording, either a video camera, or a person with a notepad, is required to note down the proceedings. Analysis can follow a structured

pattern of noting what is observed, and a summary of observations made can be prepared. It is a slightly difficult technique to use, but can be of immense help in understanding the way a customer actually behaves.

New product testing, or choice of the best ad copy among three alternatives, and similar problems where three or more variables have to be compared, to find out whether they differ from each other, require experimental methods such as ANOVA. These are covered in detail in Part II of the book.

Qualitative techniques are indirect methods of questioning, or unstructured methods of asking questions to elicit thoughts normally not expressed in structured surveys. There are different qualitative techniques, like depth interviews, focus groups, and sentence completion, brand personality tests, tests of associations, and so on. These are gaining popularity for certain applications. Focus groups, particularly, are being used as a quick method of picking up new ideas from a group of potential consumers or existing consumers, or to gauge reactions to a concept or product idea. The TVS group is one which has effectively used this technique while developing its new brand Scooty.

Doing primary research requires decisions about the sample, field work, and analysis before actually taking the study to the field. A comprehensive statement of intent on the composition of the sample, its size, geographical spread and so on is needed. A timetable for deciding where and when the field work will be carried out is made. Questionnaires or checklists have to be prepared into which data will be recorded. Briefing and debriefing of and by the field staff is a part of the field work execution and control process. The analysis plan must specify any special analyses needed in advance, to aid the questionnaire design. In particular, if analysis will be done on smaller parts of the total sample, this must be specified in advance because it may need a higher-than prescribed total sample size to compensate for the splitting up of the sample during analysis.

The last stage in the marketing research process is presentation and report writing. This is where the communication and analytical skills of the researcher count for an effective presentation to the sponsor of the study, and together with the written report, should lead to some action being taken by the decision-maker.

ASSIGNMENT QUESTIONS

1. Define and describe the first stage and the last stage of the marketing research process.
2. You are the brand manager for a shampoo brand in the market. You wish to conduct a research to find out what is the current perception of consumers about your brand vis-à-vis other brands (competitors). You also want to know what are the most important factors buyers consider while buying a brand of shampoo. The decisions you will take based on the above research (its findings) are
    1. whether to reposition your brand.
    2. whether to launch line extensions of the current shampoo brand (a variant with same brand name).

3. whether to launch a new brand of shampoo.

Write a 'brief' to a leading marketing research agency, describing your proposed research, and asking them for a proposal.

3. Assume the role of the marketing research agency which has received the 'brief' mentioned above in Q. 2. You have to write a proposal for this study, including a general idea of the methodology, sampling plan, time and cost. Attempt writing this proposal.
4. Why does secondary research appear before primary research in the typical marketing research process diagram? Explain briefly.
5. What are the major methods of doing primary research?
6. Why is a survey the most commonly used technique for primary research?
7. What are the different ways of doing a survey? Which of these is most used in India? Why?
8. What is experimentation? Give an example of M.R. which could use experimentation.
9. Why are qualitative techniques used less frequently than quantitative techniques?
10. How are TV audience shares for different channels or different time slots measured?
11. Why should the broad analysis plan be known before the start of a marketing research study?
12. Explain the difference between simple and cross-tabulation with an example.

The McGraw·Hill Companies

CHAPTER 3 RESEARCH METHODS AND DESIGN—ADDITIONAL INPUTS

Learning Objectives

In this chapter, we will

- Introduce the sources of secondary data, and discuss their advantages and disadvantages
- Define exploratory and conclusive research
- Introduce three major qualitative research techniques—depth interview, focus group, and projective techniques
- Discuss test marketing and Simulated Test Marketing (STM) within the class of techniques called Experiments

We have discussed primary research methods used most commonly, such as the survey and some special techniques, in the previous chapter. We will now look at some other issues such as sources of secondary data, their advantages and disadvantages, definitions of exploratory and conclusive research, and explore in some detail a few qualitative research techniques. We will also look at some design issues including validity of research, and look at techniques such as experimentation, test marketing, and simulated test marketing in that context.

SOURCES OF SECONDARY DATA

There are two major sources of secondary data:

1. Internal
2. External

Internal records in the company are generally used as a starting point in any marketing research. This includes information about the product being researched, its history, company background and

history, market share, and competitor information. These types of information are usually maintained by the marketing department, sales department, or a corporate cell for marketing intelligence in the company.

External information sources include syndicated reports such as retail sales data, or market share data, or industry analyses. Some of this information may be available from public sources such as business newspapers or magazines, or industry associations or trade bodies. A prominent source of data on Indian industry is the CMIE or Centre for Monitoring Indian Economy, which publishes monthly reports on various aspects of the Indian economy and industry. *The Hindu*, a prominent daily newspaper, publishes an Annual Survey of Indian Industry, which is a low-priced and useful compilation which deals with industrial goods, infrastructure and core industries, consumer durables' growth prospects and past performance.

Syndicated research studies are studies available to any subscriber or buyer. One such important study is called the National Readership Survey (NRS). This is conducted every three to four years. The latest one was done in 1999. This study covers a large national sample, and measures the readership of newspapers and magazines in great detail. It also covers demographics and consumption patterns of household consumer goods. It is a frequently quoted study in advertisements and news reports. It is one of the sources used by advertisers to select the media in which to place their advertisements. Another such study is called the IRS the Indian Readership Survey. The Audit Bureau of Circulation (ABC) is an autonomous body which certifies the circulation of newspapers and magazines. The Indian Newspapers Society (INS) also publishes a handbook every year with circulation, readership, and advertisement tariffs for various print media in the country.

There are several computer based data sources, which provide on a sale and subscription basis, updated information on financial and sales data on all publicly listed companies. Now, some of this data is available on the Internet, particularly industry analyses.

Usually, a scan of secondary sources gives a lot of background material on the product, industry, and government policy. If the industry we are investigating is known, the most useful way to gather relevant secondary data is to have a cell within the company to monitor and keep cuttings from business magazines such as *Advertising and Marketing, Business India, Business Today,* and *Business World.* This can be supplemented by newspaper reports from *The Economic Times, Business Line* or other business dailies. Over a period of a few years, this method ensures that we can easily look back and know at a glance, the current status of our brands, products, competitors, and so on. This provides a good reference to new employees or trainees who are hired to do their internship or summer projects in the company. The marketing research agency also benefits from a good briefing, and good reference material can add value to the primary research that they may undertake on behalf of our company.

DISADVANTAGES OF SECONDARY DATA

Having looked at its advantages, it is also necessary to keep in mind some disadvantages of secondary data.

1. It may be outdated. We may have cuttings which are two years old, about consumer preferences, and these may have changed over time.

2. It may be done for a different purpose and therefore be slanted or biased. It is important to note who has collected the data, and for what purpose, before making a judgement on its usefulness.
3. The sample or the methodology may be different from, or unrepresentative of, the target population we are studying. For example, the earlier study may have studied only teenagers, whereas we are looking at all adults and teenagers.
4. The units of data aggregation may be different from what we need. For example, we may want to know reactions from different sexes (male and female separately), and these may not be reported separately. Or, only regionwise data may be reported, not centrewise or citywise. Or, the way income groups are formed may be different from what we want to study.

In spite of some obvious limitations, many types of secondary data help in

1. training primary researchers
2. serving as a cross-check for other secondary data
3. provoking thinking about methodology and its impact on results of research.

Used judiciously, secondary research is an appropriate starting point for any marketing research project, mainly because it is much less expensive than primary research. In the age of the Internet, it would certainly be worthwhile to have someone at least download and look at what is available on the product and industry, before venturing out into the field for doing primary research.

EXPLORATORY AND CONCLUSIVE RESEARCH

Exploratory research usually does not directly lead to marketing decisions being made. Conclusive research on the other hand, leads to major marketing decisions being taken.

Exploratory research may be undertaken for knowing a little more about the problem, or the consumer, or the way questions should be formulated, which factors should be included in the study, or in general, to help design a follow-up 'conclusive' research study. As the name indicates, a study which seeks to explore any of these subjects is called an Exploratory Study.

An exploratory study may not have as rigorous a methodology as is used in conclusive studies, and sample sizes may be smaller. But it helps to do the exploratory study as methodically as possible, if it is going to be used for major decisions about the way we are going to conduct our next study.

One of the reasons for conducting an exploratory study is that we do not know enough to even formulate a 'conclusive' study. But if a study is designated as exploratory and treated as such, it must be followed up by another one before any major conclusions or inferences can be drawn.

There is no separate methodology for doing exploratory studies. The same process and methodologies that are available for regular research are also used in exploratory studies.

Conclusive research, as the name indicates, seeks to draw conclusions about marketing or consumer variables or other variables like sales or consumer preferences. This is usually done through a proper research methodology, rigorously used sampling plans and field work, and appropriate analytical techniques. Conclusive research may follow exploratory research in cases where the area of investigation is new. If the field of investigation is not new, it may be a routine activity, repeated every year or half-year or quarter.

Conclusive research is more likely to use statistical tests, advanced analytical techniques, and larger sample sizes, compared with exploratory studies. Conclusive research is also more likely to use quantitative, rather than qualitative techniques. This does not mean that quantitative techniques are necessarily better, but it is a fact that they are more easily understood by the sponsors of most marketing research. Clients of marketing research are comfortable with numbers, in general.

MAJOR QUALITATIVE RESEARCH TECHNIQUES

In addition to the well-known quantitative techniques such as the survey, many qualitative techniques are used for various purposes by marketing researchers. We will look at three of them in some detail. These are:

1. Depth interview
2. Focus group
3. Projective techniques

Depth Interview

This is an unstructured and longish interview on the given subject. The interviewer has to be skilled to be able to get the respondent to give a free-wheeling interview, with a minimum of prodding. Most questions are open-ended, and ask for opinions, anecdotes, feelings about products, occasions of use and so on. The discussion is rich in personal detail, which is individualistic. Compared to a regular structured interview, a depth interview has only minimal instructions for the interviewer, and the respondent is free to respond in any way he likes, not constrained to a set of multiple responses or predetermined categories. But it could also be more difficult for the same reason, for both the interviewer and the interviewee.

The expectation of the respondent from a regular survey is easy to answer, non-intrusive questions, which do not probe too far. It is different with depth interviews. Every selected respondent may not feel comfortable being open with a stranger interviewing him, and this may hinder the process. The interviewer also must have the required training to make a focussed, but unstructured conversation over a period as long as an hour or more.

An example of a depth interview would be to try and probe the feelings of a car owner about his car, what it means to him, how he feels when he is driving it, who generally he takes out with him or who else he allows to drive it, how he perceives other people who drive the same brand, and other brands or models, why he would or would not consider other brands, and so on.

To define it, a depth interview could be called a process of probing for the feelings, associations, reasons for behaviour of a consumer of a product category or brand through a mostly unstructured interview consisting of a lot of open-ended questions, by a trained interviewer.

Like many qualitative techniques, a depth interview tends to be subjective rather than objective, and therefore difficult to interpret. But it is capable of revealing much more about the underlying thought processes and feelings of a consumer about the product or service being researched, compared with traditional structured interviews.

Focus Group

This is essentially a group discussion on a given subject conducted by a trained moderator. The purpose of this is to create a less than formal situation, where people can exchange views, bringing out their opinions, attitudes, feelings about the given subject.

To bring out a fruitful discussion, the subject has to be carefully thought out, and moderated if it veers away from the given subject. The participants have to be called to the venue, and a system of video or audio recording should be used to record the discussion for later analysis. It is possible for the moderator and the 'analyser' to be different persons.

The sample is selected as usual from a target population which is specified by the needs of the study. Usually, a group consists of about 6–10 persons. The length of the discussion can be about an hour to an hour and a half, or until the group has nothing left to add.

This technique is used frequently to check out opinions about new concepts, before a product is launched, and in general, as an exploratory research tool. It is sometimes also used for conclusive research, or in combination with a survey, as a cross-check for the important findings from the survey.

Projective Techniques

There are many different techniques which can be called 'projective'. One popular method is to show a respondent a picture and ask him to describe the persons or objects in the picture. A particular product or brand can be shown being used, or displayed, and the respondent can be asked to guess the type of consumer who would use the product shown. This is essentially a technique which seeks to get indirectly at the underlying motivations, attitudes or emotions of the respondent, which he would not reveal under direct questioning. This method of questioning overcomes some common inhibitions of respondents such as the wish to give socially desirable responses, or giving answers 'acceptable' to the interviewer.

Word Associations Another variation of projective techniques is to ask respondents to associate brands with one word that they can think of when they think of the brand. It could also be a person, a celebrity, or an animal, depending on the interviewer's or the analyst's viewpoint. Interpretation of such association is best left to a psychologist, or a researcher with a psychoanalytical background and experience.

Sentence Completion Another type of projective technique is giving an incomplete sentence to the respondent, and asking him to complete it. For example, "People who use brand B coffee tend to be ……."

This method is similar to word associations, and may result in surprising or unexpected associations. It is equally difficult to interpret, and needs a trained hand to do it.

Indirect methods such as projective techniques have proved themselves useful in many classic research situations, where direct methods proved unsatisfactory. They are used in a minority of the total research projects undertaken.

VALIDITY OF RESEARCH

We have touched upon the issue of validity in the previous chapter. We will again discuss it briefly for some additional insights. Let us assume that we changed the price of a brand of pen and its sales were affected, in the following week. Can we conclude that the price change was responsible for the change in its sales? We cannot be really sure, unless we know what else remained the same and what else changed during the period. An experiment could be designed to draw a 'valid' conclusion that price was a major cause of change in sales. Validity of a result refers to it generalisability and its robustness. Is the result of an experiment occurring merely by chance, or is it due to the intervention of some variables we have no data on or is it a valid relationship between the variables under study?

To obtain a reasonably valid result, a researcher must be aware of all likely variables affecting the variables being studied (let us assume these are price and sales), be able to control or keep constant these variables, and then vary the independent variable (price) to find its impact on the dependent (sales).

EXPERIMENTS

Experiments can be conducted with varying designs and varying amounts of controls or rigour. Laboratory experiments typically have the best controls, and field experiments have the least. Simulations done on a computer can control any variable, which may not be possible when we deal with human beings in a contrived setting in an experiment designed to measure the effect of price, packaging, and promotion on sales. Human or psychological factors such as the effect of brand name, ambience of the simulated store and so on, may affect human respondents participating in an experiment.

Test Marketing

This is the name used for a class of controlled experiments in marketing research. Its objective is to predict sales (either absolute in terms of units, or relative in terms of market share), based on changes in marketing variables such as price, distribution, promotion, advertising and so forth. Although a good method for testing the product in a limited geographical area (one city, or one region) before going for a national launch, test marketing suffers from some problems discussed earlier in the sections on validity and causal designs. For example, novelty of the product being tested may result in high one-time sales due to curiosity. Once having tried the product, there may be no repeat sales of the same magnitude as trial sales. Many marketers have misinterpreted the test market sales due to inadequate analysis of possible causes of consumer behaviour. Another disadvantage commonly cited is that when you are test marketing, your competitors become aware of your product design, and may counter your efforts by introducing a similar product before you. For example, before Procter and Gamble could launch their concentrated detergent Ariel in the Indian market and while they were test marketing it a few years ago, Hindustan Lever launched their brand called Surf Ultra. It was a similar product, a concentrated detergent, and took away the novelty of Ariel. In spite of a good product, Procter and Gamble faced an uphill task in establishing their brand. There have been allegations of an outright

sabotage of test markets by competitors too. For example, they may buy up huge quantities of your brand to give the impression of a huge success, and mislead you into launching a product nationally. It is also a common tactic for a competitor to launch special promotional offers in your test market area to reduce your sales. There is also the question of which centre or centres to use for test marketing, because the wrong choice of centres can affect the generalisability of your interpretation, leading to wrong estimates of national sales.

Some of these disadvantages, along with long lead times, have encouraged marketers to use Simulated Test Marketing (STM). In a simulated test market, consumers are shown product information, are sometimes exposed to commercials (advertisements) for the brand, and then given money or coupons to buy the products made available in a simulated store containing all the major competing brands in the product category. Non-purchasers of the sponsor's brand may be given free samples. After a use period, the users are interviewed to gauge reactions and repeat purchase intention. A computer model is then used to predict real world market share and penetration based on simulated data on many market and product variables. A few years ago, Mahindra and Mahindra, the multi-utility vehicle manufacturer, did a simulated test marketing exercise for their new brand called ARMADA. Of course, this did not include a free vehicle for the respondents. Only after some time had elapsed, their intention to buy the ARMADA was measured, to predict possible market share.

Experimental designs are also classified in terms of number of independent variables used, the number of blocks (extraneous variables which are controlled), and the method of assignment of sample units to specific treatments. These are discussed in greater detail, with numerical examples, in the chapter titled ANOVA in Part II of the book.

CASE STUDY

Consumer Perception of High-end IT Education

This case study of recent origin (2001), illustrates the use of free-response questions which permit respondents to give unstructured answers. The responses are given in the form of excerpted quotes from the study at the end of the case. The entire study was bigger in scope and results. These reported results are only for the purpose of illustration and do not constitute the complete analysis.

BACKGROUND

SSI, a computer education centre, has added **Internet** to its portfolio. Now SSI plans to re-launch its course called **Internet** in its updated form. The course includes ASP, XML, WAP, .NET AND BLUETOOTH, the last one being offered only by SSI's **Internet**.

Research Objectives

To find out

- the deciding factors for taking up a particular High-End I.T. course
- whether the course contents of *Internet* are actually "in demand".
- the strengths and weaknesses of *Internet*.

Methodology

Collecting information through

- questionnaires
- face-to-face interviews
- telephonic interviews
- internet

Sample Composition

Students of SSI as well as from competing computer education providers (NIIT, Aptech, Radiant, Tata Infotech).

Sample size : 80 (25% SSI + 75% others)

Results from Some Free Response Questions Asking for Students' Comments

The following are quotations from some students' comments on the institute, course, and so on.

"Right now the I.T. market in U.S. has gone down. Bluetooth is still in a kind of an infancy stage with no real commercially proven success. There is a lot of investment in the technology. Recently it has hit a few roadblocks—you will see that from the info in the links (viz http://www.bluetooth.com/ and http://www.zdnet.co.uk/news/specials/1999/04/bluetooth/)"

- Computer professional (NJ, USA)

"MS (Micro Soft) has come up with the .NET, which works on the Windows 2000 platform. Anything to do with Internet will be 'hot'. And MS won't leave it halfway".

- Faculty (Radiant).

"I did my GNIIT, now I am doing Java at RADIANT. Did not continue there because I wanted to do only Java; and NIIT, though it is very good, has only long-term courses. Want to get into an I.T. career. From what I have heard, Aptech is not up to the mark. Don't know much about SSI or INTERNET. .NET is the latest course here."

- Student (Radiant)

"I am doing Radiant.NET with C#, ASP.NET, XML, SOAP, and so forth because it is the latest after Java".

- Student (Radiant)

"I joined Radiant because I heard that the course material is very good. Faculty is also good. Finished my Java from there. And I plan to do a post graduate course in I.T. NIIT is too expensive. Cost- wise, I guess SSI and Radiant are comparable. Don't know more about SSI."

- Student (Radiant)

"I did my Java from TCI because I stay close by (Annanagar). Radiant is more expensive. Also TCI gives me a 'Government of India' certificate. I am working as a web page designer. I am being trained in XML and so on by my company itself."

- Ex-Student (TCI)

".NET has not yet come into the market. Hence we do not have the course. We have C#, XML, WAP."

- Counselor (NIIT)

"Of course NIIT is expensive compared to the other institutes. But when one is focussed on one's career, one does not crib about money. After interacting with my faculty, I have a very good knowledge about the I.T. world. Now I would not even think of changing. I have a background in BCA and am doing my Java here."

- Student (NIIT)

"NIIT has got a name that is recognised the world over more than any other institute in India. Hence I prefer to be in NIIT. I plan to work abroad. I am currently doing the E-Commerce course in NIIT, which includes XML, ASP, WAP and so forth."

- Student (NIIT)

"I just know about NIIT. So I am here. Plan to do a short-term course here itself after my GNIIT, which I will finish this year."

- Student (NIIT)

"I have no background in computers, but I do not find any difficulty in doing my INTERNET course. NIIT and APTECH are too expensive."

- Student (SSI)

Question

1. Write down a brief summary of all the answers given above. How does this differ from the analysis of structured-response questions?

SUMMARY

In this chapter, we have looked at some issues concerning secondary data, qualitative research techniques, and discussed validity of research and experimental designs, including simulated test marketing.

Secondary data can be defined as data available from any source including electronic and physical sources, but not specifically collected for the study undertaken by the marketing researcher. Nevertheless, it is of help to a market researcher in background study and planning of his primary data collection effort. A good source of secondary data now is the Internet version of business newspapers and periodicals such as *The Economic Times* (www.economictimes.com), *Business World*, and various other publications. The printed versions of these can be used if net access is not available, or for reasons such as ease of photocopied storage. Supplementing business magazines and periodicals would be reports of organisations such as CMIE (Centre for Monitoring Indian Economy), syndicated surveys conducted earlier, databases to which the researcher's company may subscribe, and various industry associations that may provide data to the researcher. Secondary data should however be double-checked for its authenticity, credibility, and methodology, because it tends to be different in the sampling method, research design, and generalisability from what the researcher may imagine it to be.

In media studies as well as some other marketing studies, data from massive sample surveys such as the NRS(National Readership Survey) and the IRS (Indian Readership Survey) can sometimes be used as pointers for determining sample sizes needed, or for benchmarking or validating purposes. But it must be remembered that in sampling, bigger is not necessarily better. An accurately conducted sample survey requires surprisingly small sample sizes, and a clumsily done large survey has little value.

Exploratory research is a term used for preliminary research conducted for finalising your hypotheses or research questions, generating some more data about the population under study for deciding on the target, or in general, to know more before doing conclusive research. Conclusive research is one on the basis of which marketing (or other) managers take important decisions.

The methods of doing exploratory research could be the same as those used in conclusive research, or different. Frequently, exploratory studies use a lot of qualitative research techniques, and conclusive research tends to use more quantitative techniques.

Three major qualitative research techniques are 1. the depth interview, 2. the focus group, and 3. projective techniques. A depth interview is a long, semi-structured or unstructured discussion with one respondent at a time, designed to find out more than a simple structured interview can. It may be held over multiple sessions, or at one go, but it needs a lot more of time to analyse, than a structured interview.

A focus group is a moderated group discussion conducted at a neutral venue, to which 6–10 respondents of the target population are invited, and to whom an incentive may be given for their effort. The discussion is moderated by a researcher and usually recorded on video for later analysis. Several groups may be conducted as part of the same research study in different cities or based on other stratification variables (for example, males, females, users/non-users, etc.). The method is very popular for checking out new concepts, or for understanding consumer motivations, and in general for exploratory research.

Projective techniques can be of many types, but the idea behind them is to get a feel of the psychology of the consumer through indirect methods. For example, studies to determine brand associations may reveal the consumer's hidden associations or unexpressed motivations for choice of brands, which are not easy to uncover through direct questions. An example could be that consumers associate scooters with femininity, but do not say so. But if you ask them to write a story after showing them a picture of a scooter, the theme they use may suggest that this association exists in their mind, and is a barrier to their using it. The techniques used under the head 'projective' may include sentence completion, writing about the personality of a consumer or brand based on pictures shown, or others.

Validity is an issue covered in greater detail in text books on research methodology. What the marketing researcher needs to be aware of in this context is whether the questions used in his research actually measure what they are supposed to measure, whether the sample is valid (representative of the population targeted), and whether the results are generalisable and robust. This requires a good understanding of all major variables under study, their possible interrelationships, and an effort to operationalise the variables and rule out extraneous variables interfering with the variables under study.

In terms of validity, a class of studies called Experiments scores well, because the variables are controlled tightly, and the effects and causes can therefore be studied with a higher degree of confidence than with some other methods. Two techniques within the general class of experimentation are test marketing and simulated test marketing (STM). Test marketing is used at the trial stage of a product launch, where various marketing mix variables like price, level of promotion, and so on can be varied to measure their impact on sales in a limited geographical area, before deciding on a national launch of a brand. Simulated test marketing is similar to test marketing, but done under simulated store or purchase conditions, instead of the real marketplace. The reasons for using simulations are to prevent competitors from realising your strategy, and to speed up the process of research. Some marketers find STM as useful in decision-making as a real test marketing exercise.

ASSIGNMENT QUESTIONS

1. List out as many different sources of secondary data that you can.
2. Why is some research called exploratory research? How does it differ from conclusive research?
3. Describe (a) depth interviews (b) focus groups.
4. What are the common requirements for qualitative research methods in general?
5. Think of and list down some applications where qualitative methods are better than quantitative methods.
6. What are projective techniques?

CHAPTER 4 QUESTIONNAIRE DESIGN: A CUSTOMER-CENTRIC APPROACH

Learning Objectives

In this chapter, we will

- Introduce the reader to common errors made while designing questionnaires
- Discuss the four major scales used in designing questions
- Define structured and unstructured, disguised and undisguised questions
- Elaborate on various types of questions one can ask, with examples
- Provide an example of transforming information needs into a questionnaire
- Briefly discuss reliability and validity of a questionnaire

If the research objectives of a marketing research study are well-defined, questionnaire design becomes a relatively easy task. But there are many ways to improve upon a questionnaire. Most of it has to do with having a customer-centred, or respondent-centred approach.

DESIGNING QUESTIONNAIRES FOR MARKET RESEARCH

Language

The first and foremost question we have to ask ourselves as a researcher is: *What language is the respondent going to understand and respond in?*

The questionnaire must be designed in such a way that it can be used in any language. This does not necessarily mean it has to be printed in each language in which it has to be administered. For instance, a questionnaire printed in English could be administered to the respondent in the local language he speaks, by a trained interviewer who could translate each question online. The answers can be recorded

in the given English language form if the interviewer is fluent in both the languages. This makes it easier to tabulate.

Alternatively, the numerical codes for the answers can be in usual numbers, and the questionnaire could be translated into any language required for the respondent to understand. But the translation must be as consistent as possible with the original, to keep the data collection process valid and standardised.

Difficulty Level

In addition to the language used to write the questions, the difficulty level of the words used should be a prime concern. It is frequently assumed by the person who formulates a questionnaire, that the respondents can understand what they mean. Even simple words (by their standard) are sometimes not understood. As a rule, it is a good idea to avoid marketing jargon or difficult words unless the respondent is a postgraduate or an experienced executive. In other words, keep the language as simple and straightforward as possible.

Fatigue

Increasingly, consumers are getting weary of answering questions for marketing research. It is one of the researcher's responsibilities to stick to necessary questions, and avoid unnecessary ones. The golden rule is to keep it as short as possible, and the ideal maximum interview time is probably about 20 minutes per interview.

Cooperation with Researcher

The questionnaire must also encourage the respondent to respond. In personal interviews, it is the interviewer's job to introduce the subject of the research and the agency he/she represents, before starting the interview. He must also explicitly ask for the cooperation of the respondent. In questionnaires which are filled by respondents themselves, there must be a two to three line introduction and request for respondent's cooperation at the top of the questionnaire. In mailed questionnaires, a covering letter detailing the purpose of the study and explaining what use its results will be put to, is likely to increase manifold the response rate.

Social Desirability Bias

There is a tendency on the part of respondents to give wrong, but "socially acceptable" answers to even the most ordinary, innocuous questions. For example, the socially desirable answer to the question "Do you read the daily newspaper?" is "yes". It is as likely to be wrong as right. There are many ways to verify the accuracy of responses and to deal with them. Some of the techniques are:

1. repeating the same or similar question in the questionnaire at different places.
2. asking indirect questions.
3. asking follow-up questions to probe if the respondent is really truthful.

For example, the respondent could be asked to state one important headline, or describe one important story he remembers, if he states that he reads the daily newspaper. This could be from the same day's or previous day's newspaper.

4. deliberately introducing non-existent periodicals, or advertisements, and asking the respondent if he/she has seen them.

Ease of Recording

It must be remembered when designing the questionnaire, that it has to be carried on the field, and data may be recorded on it while standing in awkward postures. The questionnaire design should ensure it is easy to carry, visible in different kinds of light, and the distance between different answer categories should be sufficient so that there is no confusion or mistake while placing a tick over the actual response for a given question.

Coding

If the questionnaire is coded before doing the field work (as most questionnaires are likely to be these days), it must be ensured that the field staff knows where to mark the answers—on the code or on the actual answer choice. This should be done during the briefing and mock interview, or pilot survey as described in the earlier discussion on field work in Chapter 2.

Purpose of a Questionnaire

The purpose of a questionnaire is to collect with ease the data required from the target respondents in the marketing research. The questionnaire must be easy to understand, easy to fill, and must fulfill its purpose.

Frequently, a questionnaire contains printed instructions for the interviewer. This includes 'Go To' statements, such as "If respondent is a non-user of brand *X*, then Go To Q.5. If not, Go To Q.9".

Sequencing of Questions

Questions in a questionnaire should appear in a sequence starting from non-threatening or ice-breaking or introductory questions, and then proceed to the main body of questions. Generally, the age, income, occupation, education, and similar demographic questions should appear at the end of a questionnaire, after an interviewer has established a rapport or familiarity with the respondent. If these are asked in the beginning, there is a high likelihood of suspicion and non-cooperation resulting in a wasted effort in many cases.

As far as possible, questions should follow a logical sequence, and must be phrased appropriately.

Biased and Leading Questions

The questions should be carefully worded to avoid bias. It is not a good practice to ask questions such as "Don't you think liberalisation is a good idea?" You could be better off getting an unbiased reply

asking a question like "Some people think liberalisation is a good thing, and some think it is bad. What do you think?"

Monotony

One indicator that a questionnaire is monotonous for the respondent is if he answers "Agree" to every question or "Disagree" to every question, for four to five questions in a row. If this happens, the researcher must find a way to overcome the potential problem, by re-sequencing the questions which force the respondent to think before he answers, or by changing the scale, or by some other method.

If some of the above points are kept in mind, and the questionnaire tested on a small sample before it is cleared for the actual survey, questionnaire design is a fairly straightforward translation of the information needs and research objectives into an instrument for data-gathering.

Analysis Required

One last thing. A questionnaire design is dependent on the analysis required from it. But the most important effect of the analysis required is in the *scale of measurement* that must be used. So we will deal with this topic—the scale of measurement—next.

SCALES OF MEASUREMENT USED IN MARKETING RESEARCH

Broadly, marketing research uses the four major types of scales:

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal Scale

A nominal scale is one in which numbers are only used as labels, and have no numerical sanctity. For example, if we want to categorise male and female respondents, we could use a nominal scale of 1 for male and 2 for female. But 1 and 2 in this case do not represent any order or distance. They are simply used as labels. For instance, we could easily label females as '1' and males as '2', and it could still be a valid nominal scale.

We can use the nominal scale to indicate categories of any variable which is not to be given a numerical significance. For example, demographic variables such as religion, education level, languages spoken, and other variables like magazines read, TV shows watched, user or non-user of a brand, brands bought, can be nominally scaled.

Nominally scaled variables cannot be used to perform many of the statistical computations such as mean, standard deviation and so on, because such statistics do not have any meaning when used with nominal scale variables. However, counting of number of responses in each category and computation

of percentages after division by the sample size is allowed. Also, nominal scale variables can be used to do cross tabulations, one of the most popular methods of routine analysis. The Chi-square test can be performed on a cross-tabulation of nominal scale data. To repeat, simple tabulations (also called frequency tables) and cross-tabulations can be done with nominal scale variables.

Ordinal Scale

Ordinal scale variables are ones which have a meaningful order to them. A typical marketing variable is ranks given to brands by respondents. These ranks are not interchangeable, as nominal scale labels are. This is because rank 1 means it is ranked higher than rank 2. Similarly, rank 2 is higher than rank 3, and so on. Instead of 1, 2, and 3, however, we could use any other numbers which preserve the same *order*. For example, 3, 10, 15 could denote the same ranking order instead of 1, 2, and 3. This is because we do not know for sure what the distance between 1 and 2 is, or what the distance between 2 and 3 is. Ranking simply denotes that 1 is higher than 2, and 2 higher than 3, but higher by *how much* is unknown. For one respondent, 1 and 2 may be close together; for another, they could be far from each other.

The statistics which can be used with the ordinal scale are the median, various percentiles such as the quartile, and the rank correlation. This is in addition to the frequency tables and cross-tabulations, which can also be used. Arithmetic mean (or average) should not be used on the ordinal scale variables. For example, the *average rank* of a set of rankings does not have any meaning. Even though weighted indexes are calculated in practice from rank order data, it is, strictly speaking, not allowed.

Interval Scale

An interval scale variable can be used to compute the commonly used statistical measures such as the average (arithmetic mean), standard deviation, and the Pearson Correlation Coefficient. Many other advanced statistical tests and techniques also require interval-scaled or ratio-scaled data.

Most of the behavioural measurement scales used to measure attitudes of respondents on a scale of 1 to 5 or 1 to 7 or 1 to 10 can be treated as interval scales. These types of scales, also known as *Rating Scales*, are very commonly used in marketing research.

If a consumer is asked for his satisfaction level with a product or service or any other attribute related to it, on a scale of 1 to 10, it is an interval-scaled rating. We could use it to compute the average rating given by all respondents in the sample. Standard deviation can also be computed.

The difference between interval scale and ordinal scale variables is that the distance between 1 and 2 is the same as the distance between 2 and 3, and 3 and 4 in an interval scale. That is, the difference between two successive numerical measures is fixed, whereas in a rank-ordered data set, it is not fixed.

Ratio Scale

All arithmetic operations are possible on a ratio-scaled variable. These include computation of geometric mean, harmonic mean, and all other statistics like the average, standard deviation and Pearson Correlation, and also the tests such as the t-test and the F-test.

In a ratio type scale, there is a unique zero or beginning point. An interval scale does not have a unique zero (it is an arbitrary zero). Also, the ratio of two values of the scale corresponds to the same ratio among the measured values.

For example, distance is a ratio-scaled variable. It has a zero, which is unique. 2 metres is to 1 metre as 2 kilometres is to 1 kilometre. Also, 4 metres to 1 metre, and 30 metres to 7.5 metres. The ratios can be measured at any two points, and they would correctly denote the relationship.

Not many ratio-scaled variables exist in marketing. Some of them are length, height, weight, age and income (measured in rupees, not as an income category).

STRUCTURED AND UNSTRUCTURED QUESTIONNAIRES

We will briefly discuss the issue of structured and unstructured questionnaires now, and then move on to disguised and undisguised questions. Structured questionnaires are those where the questions to be asked are standardised, and no variation is permitted in terms of the wording of the questions between different interviewers. Standardisation in a structured questionnaire usually extends to the answers which are also standardised into a set of permissible ones that the respondent can choose from. In effect, then, we can standardise either (1) questions only, or (2) both questions and answers.

Structured Questions

The reason for asking structured questions is to improve the consistency of the wording used in doing the study at different places, by different people. This increases the reliability of the study, by ensuring that every respondent is asked the same question, word for word. For example, the question "Do you live in Delhi?" may be construed differently from the question "Are you a resident of Delhi?" by some respondents, even though it appears that both questions are asking for the same information. A person who is normally not resident in Delhi but is living there at present on a short visit may answer "yes" to the first question but "no" to the second one. Whatever may be the intention of the researcher in asking this question, it is best served by keeping the question exactly the same (either version 1 or version 2), when asked by different interviewers.

Structured Answers

Structuring or standardising answers which a respondent can choose from in a questionnaire also achieves consistency of form. Additionally, it makes the interpretation of answers, that is, analysis and tabulation, easier than in the case of unstructured answers. Unstructured answers suffer from many problems in a large-scale marketing research study. They become difficult to categorise after the study, and different analysts may interpret them differently—so they may lend themselves to subjective interpretations. Subjectivity by itself is not bad, but it becomes difficult to defend if the sponsors (clients) of the study are quantitatively oriented. Most large-scale studies in marketing research therefore, choose the less risky, and easier to manage, structured answer approach.

Open-ended and Closed-ended Questions

Questions which permit any answer from the respondent in his own words are called open-ended questions. Questions which structure the possible answers beforehand are known as closed-ended questions. An example of an open-ended question is "What do you like about Surf detergent?"

_____

The respondent can say whatever he wants to, in response to this question.

On the other hand, a closed-ended question which elicits similar information could be "What do you like about Surf detergent?"

(a) Its cleaning power
(b) Its price
(c) Its fragrance
(d) It dissolves easily
(e) Its stain-removing ability
(f) Any other, (please specify)_____

Here, options "a" to "e" are predetermined, but "f" provides for anything else the respondent wants to add.

The various kinds of questions with a structured question-answer format are discussed later in the chapter, and range from dichotomous (two choices) to multiple choice, and from simple ranking to different kinds of rating scales. There are also some special types of questions for certain applications, like the paired comparison, a variation of the ranking scale.

Disguised *Versus* Undisguised Questions

There are some pros and cons of asking questions directly versus asking them in a disguised form. For example, we may ask a person if he/she is a good parent. This is a direct question. Or, we may ask for the respondent's opinion on the deficiencies they have observed in how others bring up their children—say, their neighbours, relatives or friends. This is an indirect question, and a qualified analyst can interpret the answers to gauge how good a parent the respondent might be, from the responses given. The problem with the direct question in this case is that most people will not admit to being a bad parent. But they may come out freely with other people's deficiencies, some of which could reflect their own shortcomings.

Another example of this is the question, "Are you afraid of flying?". Almost nobody would actually admit that they are afraid to fly. However, if the question is suitably disguised, say as "Are any of your neighbours/friends afraid to fly?", many more people would be identified as being afraid to fly in a given neigbourhood. An inference can be made about the neighbourhood based on these answers interpreted by an experienced researcher. Many indirect or disguised questions require experts in psychology or sociology to decode the replies and interpret them.

There are other reasons why disguised questions are sometimes needed. It is often found that respondents are biased when they know who is the sponsor of the study. For example, if they know that Pepsi has sponsored a research study, they may be biased towards Pepsi in all their answers. The same could happen to a study sponsored by Coke if the fact is known to the respondent. To get true, unbiased

opinions regarding attitudes towards brands, researchers sometimes do not let on the name of the sponsor. For example, a well-known multinational company making electrical switches for industrial application once did an anonymous survey in Mumbai among its customers (a study done by the author) and found many deficiencies in its products and service which they otherwise may not have found out. Sometimes corporates use students for studies where they do not reveal the sponsoring company's name. Where to draw the line in such misrepresentation depends on the culture, the law, and the definition of ethical standards of the country and the people. If it results in more accurate data without doing any harm to the respondent, it may be a legitimate way to do the study.

Completely disguised or indirect questions probing into the psyche of a person are usually used for qualitative research, as part of depth interviews or projective techniques, and so forth. These are discussed elsewhere in the book.

To summarise, market researchers usually ask structured, undisguised questions in a typical study done on a large sample. Most studies also tend to be of the 'quantitative' type, where numbers (frequencies), percentages, averages or similar summary statistics are computed. These types of analyses are easier to do with structured formats for answers. But in cases where non-routine studies are done which are exploratory in nature or qualitative by design, or both, unstructured questions/ answers may be required.

Even if a study is primarily based on structured responses, a couple of open-ended questions may still be included in it if they are the best suited for the task on hand. One such category of questions is called 'Probing' questions in marketing research terminology. These are used as a follow-up after a structured response question. An example of this use of an open-ended question following a structured question is:

1. Which brand of mosquito mats do you use?
   (a) Good Knight
   (b) Mortein
   (c) Jet
2. Why do you use this particular brand? _____

In this question, the second part is open-ended, while the first part is closed-ended.

TYPES OF QUESTIONS

Having looked at the four major scales and having discussed the broad typology of questions into structured unstructured and disguised undisguised, let us look at the six major types of questions that most questionnaires would generally use:

1. Open-ended
2. Dichotomous (2 choices)
3. Multiple choice
4. Ratings or rankings
5. Paired comparisons
6. Other types such as semantic differential, or other special types of scales.

Open-ended Question

This is one which leaves it to the respondent to answer it as he chooses. An example is "What do you think of the taste of brand *X* of Cola?" No alternatives are suggested. The answer can be in the respondent's own words.

Dichotomous Questions

These are questions which ask the respondent to choose between two given alternatives. The most common example of this is the yes or no type of questions like "Are you a user of brand *X* toilet soap?" Yes or No are the alternatives given. A third choice is sometimes added to dichotomous questions such as "Do you like brand *X* of potato chips?" The choices given are "yes", "no", and "neither like nor dislike". (Sometimes, "any other, please specify _____" is used, instead of "neither like nor dislike".)

Multiple-choice Questions

These are extensions of dichotomous questions, except that the alternatives listed number more than two. A common example is as follows:

Please tick against the factors which made you buy this brand of car:

(a) Reasonable price
(b) Great looks (appearance)
(c) Fuel economy
(d) Easy availability of service
(e) Any other, please specify.

In the above question, more than one category can be chosen. In some other multiple choice questions, only one category is to be chosen. For example, look at the following question.

Please specify your age group:

(a) Below 15
(b) 16–25
(c) 26–40
(d) Above 40

Only one of the above is to be chosen. It must be clear to the respondent and the interviewer whether only one choice is allowed, or more than one are allowed for a multiple choice question.

Ratings or Rankings

This is a question of the type, 'Please rate the following detergent brands on a scale of 1 to 7 in their ability to clean clothes'.

Brand A	1	2	3	4	5	6	7
Brand B	1	2	3	4	5	6	7
Brand X	1	2	3	4	5	6	7

This is an example of rating. Ranking would have looked as follows:

Please rank (1 = best, 2 = next best, etc.) the following detergent brands on their ability to clean clothes.

Brand A -----
Brand B -----
Brand X -----

Paired Comparisons

A special type of question is the paired comparison. This requires the respondent to choose between pairs of choices at a time. For example, there could be six brands of colour TVs, brands A, B, C, D, E, and F. A respondent may be asked to do a paired comparison to say which brand is better, but for only two brands at a time. He is given a table or a card with two brands written on it, and has to choose the better brand, each time. This process has to repeat for as many pairs as exist in the given set of objects or brands.

Some special techniques such as Multidimensional Scaling need data from paired comparisons. This technique is explained later in Part II of this book.

Semantic Differential

Another scale commonly used by marketing researchers is called the semantic differential. This type of question is similar to the rating scale. The only additional feature is that a set of two adjectives forms the two extreme points of the scale. For example, a product is

Easy to use	|-----|-----|-----|-----|	Difficult to use
Expensive	|-----|-----|-----|-----|	Inexpensive
Easily available	|-----|-----|-----|-----|	Not easily available
Convenient	|-----|-----|-----|-----|	Inconvenient

There may be several intermediate points between the two extreme values of the scale. These could be coded 1 to 5 or 1 to 7 or whatever the number of points is. A commonly used 5-point scale is from *completely agree* to *completely disagree*.

There may be questions based on other scales which are standard or specially constructed. Some scales like the Likert Scale or Thurston Scale are named after people who invented them.

How to Choose a Scale and Question Type

The researcher must decide on the scale and type of question based on the following factors:

1. Information need
2. Output format desired
3. Ease of tabulation

4. Ease of interpretation
5. Ease of statistical analysis
6. Reduction of various errors in understanding or use by respondents and field workers

TRANSFORMING INFORMATION NEEDS INTO A QUESTIONNAIRE

There are many finer points to designing an appropriate questionnaire. We can only discuss them if we have a reference point for the discussion. We will illustrate by developing a complete questionnaire for a given situation and then discussing the merits and demerits of the individual questions.

Example of Information Needs

A soft drink concentrate manufacturer (for example, Rasna's manufacturer) wants to know the following:

1. Demographic profile of users versus non-users of soft drink concentrates
2. Among users,
   (a) the preference for liquid concentrate versus powder
   (b) preference for powder with sugar added, versus powder with no added sugar
   (c) occasions of use by self
   (d) whether served to guests
   (e) rating on convenience, taste, price, and availability
   (f) brand preferred among soft drink concentrates
3. Among non-users,
   (a) reasons for not using soft drink concentrate
   (b) substitute product usage, if any, and reasons for using or consuming them

Let us attempt to develop a questionnaire for the above information needs. A possible questionnaire is shown below.

Questionnaire for Soft Drink Concentrate Study

Date _____ Q. No. _____

Centre _____

Dear Sir/Madam,

We are doing a brief survey to find out more about consumer preferences regarding soft drink concentrate. We would be grateful if you could spare a few minutes to participate in it. Thank you for your cooperation.

1. Do you use soft drink concentrate to make your own soft drinks at home?
   Yes ☐ No ☐
   If Yes, continue with Q.2. If No, Go To Q.9.

2. Do you use liquid or powdered concentrate? (Tick only one)
   Liquid ☐
   Powder ☐
   Both ☐
3. Which type of concentrate do you prefer out of the following?
   Concentrate with sugar added ☐
   Concentrate without sugar added ☐
4. What are the occasions when you use soft drink concentrate to make soft drinks? (Tick only one)
   Regularly, all year round ☐
   Regularly, only in summer ☐
   Occasionally, all year round ☐
   Occasionally, only in summer ☐
5. Do you serve it to guests?
   Yes ☐ No ☐ Depends on the guest ☐
6. Which brand do you use?
   Rasna ☐ Brand X ☐ Brand Y ☐
   Any other (please specify) _____
7. Please rate the brand you use on the following attributes, on a scale of 1 to 7 (7 = Very Good, 1 = Very poor).

	1	2	3	4	5	6	7
Availability	|	|	|	|	|	|	|
Taste	|	|	|	|	|	|	|
Convenience	|	|	|	|	|	|	|
Price	|	|	|	|	|	|	|

8. Any other comments on the brand you use ?

   _____

   _____

   _____

After Q. 8, Go To Demographics Q.11.

Non-Users

9. Do you consume any of the following regularly? (You may tick more than one)

	Yes	No
Fruit juice	☐	☐
Squash	☐	☐

Bottled soft drinks	☐	☐
Tea	☐	☐
Coffee	☐	☐
Nimbu pani	☐	☐
Buttermilk	☐	☐

10. What are the reasons for not using soft drink concentrate? (You may tick more than one)

Does not taste good	☐
Expensive	☐
Chemical additives	☐
Does not contain natural fruit juice	☐
Not available easily	☐
No nutritional value	☐

Any other (please specify) _____

_____

Demographics Please let us know a little more about yourself.

11. Your age group

Less than 25	☐
26 – 40	☐
41 – 50	☐
Over 50	☐

12. Your monthly household income

Less than 5000 rupees/month	☐
5001 to 10,000 rupees/month	☐
10,001 to 15,000 rupees/month	☐
Over 15,000 rupees/month	☐

13. Address _____

_____

As a practice exercise, you can critically examine the questionnaire above to suggest improvements in any of the questions or the scales or the choices given in the multiple choice questions. Use as a guideline the principles of questionnaire design we have discussed, the information needs given, and the kind of audience who will respond to the questionnaire. In fact, we should define the target audience/respondents before we develop the questionnaire.

Some Hints for Discussing the Merits and Demerits of the above Questionnaire

1. Are the income and age categories adequate for analysis of the data? (Questions 11 and 12)
2. Is the 7-point scale used in Question 7 easy to understand? Is it appropriate? Adequate?
3. Should there be an open-ended question number 8? Why?

4. Have we left out anything? Such as who decides on the brand to buy (for users)? Who decides to buy/use substitutes (for non-users)?
5. Should we also ask which family members drink the soft drink (for users) made from concentrate?
6. Should we ask the convenience and price questions separately (Question 7) and differently? What exactly do we want to know from respondents regarding price? Are we getting the answer?

Double-barrelled Questions

Inexperienced questionnaire designers have a tendency to combine two questions into a single question, such as:

Are you happy with the price and quality of brand Y?

Yes ☐ No ☐

This is not a good question to ask, because the answer will be ambiguous, whether it is yes or no. It would not be clear whether the respondent has said yes for price alone, quality alone, or for both. The same problem exists for a 'no' answer.

It is better to rephrase the question and provide for different answer categories for each attribute, or ask two separate questions, one for price and one about quality. Then the interpretation of answers becomes far easier.

Good Questionnaires and Bad Questionnaires

A final comment is in order, on what is a 'good questionnaire'. In general, a questionnaire is good if it measures what it set out to measure (i.e., it is VALID) and does it in an efficient manner. Usually, a questionnaire goes through various stages before it is used in the field. Listing of information needs, conversion into questions with suitable scales of measurement, sequencing of questions into a logical order, trying it out in a pre-test on a handful of respondents in a convenience sample or a field sample, modifications in the wording, scale or sequence as a result of the pre-test, and then the final draft for the actual study, are the steps involved. There are also discussions between the researcher and the sponsor company, if these two drafts (first and final) are not the same. Thus, there are many stages in questionnaire construction, and most faults would be ironed out in this process, if followed meticulously.

From the author's experience, though, the problems most frequently occurring in a typical study are from the lack of sufficient thought given to the analysis required in advance. The solution for this is to prepare blank output formats for each question on the questionnaire, and think about other types of analysis desired after the field work. If this can be done before doing the field work, there would be hardly any problems. In many cases, the value of the research increases manifold by slightly modifying the scale or wording of the questions asked. You must remember that it is cheaper to modify the questionnaire in advance than think about what could have been done after the study is over.

RELIABILITY AND VALIDITY OF A QUESTIONNAIRE

Reliability is the property by which consistent results are achieved when we repeat the measurement of something. A questionnaire used on a similar population that produces similar results can be termed as reliable. Consistency of form and manner of asking questions (their exact wording, the amount of structuring, etc.) generally ensures reliability. Proper training given to interviewers in a study also improves reliability, by reducing variation in the way they ask questions and record answers.

Validity is the property by which a questionnaire measures what it is supposed to measure. If we want to measure attitudes towards brands of washing machines in terms of service and product features, then that is what the critical questions in the questionnaire should measure. The validity of questions on a questionnaire can be checked by comparing it with previously used items (questions) measuring the same thing, and also trying out different questions to find out which one seems to measure what we intended to measure. A certain amount of judgement which comes with experience is of great help in framing 'valid' questions. It is also possible to consult experts in research methodology, or the subject on hand to check that a given set of questions is 'valid'.

RELIABILITY AND VALIDITY

This section is likely to be especially useful to academic researchers who work with multi-item scales measuring certain constructs.

What is a Construct?

A construct represents a researcher's conception of some concept to be operationalised and measured. Service Quality for example, is a construct, and needs to be operationalised into a set of measures through a set of questions or statements that will form a measurement scale. There could be any number of *items* (individual questions) which form a scale to measure a construct. There may be parts of the overall construct. For example, Service Quality was shown by Parasuraman et al to have five different dimensions such as reliability, assurance and so on.

The challenge in measurement is, after devising a measure in the form of some questions/ statements, to show that they measure what we think they measure. This is the basic validity issue. But in practice, there are several types of validity and the researcher may have to try and prove validity of his construct through various methods.

Content Validity

This is the representativeness of what is measured, in drawing conclusions about the property (say, service quality), we wanted to measure. From the universe of content regarding service quality, if what we measure is proved to be adequately representative, we have achieved content validity. How is it done?

Mostly, it is the judgement of the researcher, backed by other experts or researchers /past authors in the field. Some form of judgemental assessment of the items (questions) being used as being

appropriate for measuring the property we wish to measure has to be made and used to justify the content validity of a measurement scale containing these items.

Criterion Validity or Predictive Validity

To what extent can our measurement be validated by an external or different measure? This is essentially the meaning of criterion validity. For example, the admission test of an institute may correlate highly with the final grades (measured by a Grade Point Average at the end of an academic program). Here, the emphasis is on a high correlation between our original measurement (the admission test score) and the later measurement (the Grade Point Average). If this is high, this is a validation that our admission test is good at "predicting" the performance of incoming students. Or, to use the service quality example, our measures of input service quality in a restaurant (staff quality, infrastructure quality etc) may be validated if we get high scores on customer ratings of quality.

Note that what is contained in the test (that is, the content validity) is not very important, if the criterion validity is good. That is, it may be possible to have criterion or predictive validity even if the content of our measurement is suspect.

Construct Validity

Perhaps the most important type of validity for researchers, this is the strongest evidence that our measurements are appropriate. The major reason is that there is a strong link between theory and empirical measurement (actual or proposed) that this type of validity seeks to establish. Testing of hypotheses can be a follow up or a part of proving construct validity. If we prove that the constructs we have defined are valid, we can test further hypotheses related to them. In proving construct validity itself, there are some hypotheses being tested. For example, we may test the hypotheses that—

1. the construct we are testing is different from another (discriminant validity)
2. different measurements of the same construct lead to a similar interpretation of the construct's meaning (convergent validity)

and thus prove construct validity.

There are many ways of approaching construct validation. The three major ways are—

1. The multi-trait multi-method matrix (Campbell and Fiske, 1959)
2. Item to total score correlation in a scale
3. Factor Analysis

We can look at these three methods in some detail.

The Multi-trait Multi-method Matrix This is a somewhat complex way of proving construct validity and essentially tries to approach the evidence through proving convergent and discriminant validity. To use this, it is essential that at least two methods of measurement are available for a construct and two constructs are measured using the two methods. The following table illustrates what we mean. The constructs here are Dimension A and Dimension B, each measured by two methods—1 and 2.

TABLE 4.1

		Method 1	*Method 1*	*Method 2*	*Method 2*
		A	*B*	*A*	*B*
Method 1	Dimension A				
Method 1	Dimension B	Low			
Method 2	Dimension A	High			
Method 2	Dimension B		High	Low	

If the correlations in the above matrix after the respective measurements are made are as indicated above, there is proof of construct validity. For instance, the two dimensions (constructs) could be Reliability and Assurance in service quality. If there is substantial agreement in the two methods measuring A and B (indicated by the two High correlations in the matrix), this proves the two measures are measuring the same constructs (A and B in this case).

Also, if using any of the two methods (1 or 2), the correlation between A and B is low (indicated by the Low values in the above table), it indicates good discrimination between A and B. Therefore, construct validity is indicated for both constructs A and B.

In practice, it may be difficult to find two different methods and measure the same constructs twice. But this method is useful if construct validity is critical to further research.

Item to Total Correlation In this method, we assume that the total score for a scale is valid. Then, the validity of individual items in the scale can be tested by measuring the correlation between that item and the total score. If this correlation is high, then it is assumed that the item is valid. If this correlation is low, we could drop the item from the scale. Practically, this can be done through options on the Reliability measurement menu.

Reliability of a Scale

Mainly, reliability is a measure of how a scale can be relied on to produce similar measurements every time we use the scale. SPSS offers the usual measures of reliability, like Cronbach's alpha. To perform Reliability analysis of a scale that consists of 5 items/questions/variables (say, Var 1 to Var 5 in a dataset), select these variables on your screen after choosing **Analyze, Scale, Reliability** from the main menu. Then choose **Alpha** as the model and from the **Statistics** option on the same screen, choose the *Descriptives for Scale* and *Descriptives for scale if item deleted*. Also choose *Inter-item correlation.* Then click OK to get the output.

If the alpha value for the scale is 0.7 or more, it is usually considered a good scale. If the item to total correlation is low for an item, you can consider dropping the item from the scale. You can also take this decision based on a look at the alpha value after dropping an item. If the alpha value is high even after dropping a particular item, you can drop the item.

In the hypothetical output tables shown below, we are examining the reliability of a scale that contains six items (v2, v3... to v7). What we find is that the scale reliability is .708, which is good. If we

look at item to total correlation (correlation between one item of the scale and the complete scale), we see that **v6** has the lowest value in the third column of the second table below. This item with an item to total correlation of .364 can be considered for being dropped from the scale if an item has to be reduced. The alpha value of the scale if we drop this item from the scale will be .690, as mentioned in the last column. This is a decision based on the researcher's requirements. You may decide not to drop any item if you think that it serves no purpose.

TABLE 4.2 Reliability Statistics

Cronbach's Alpha	*Cronbach's Alpha Based on Standardized Items*	*N of Items*
.706	.708	6

TABLE 4.3 Item-Total Statistics

	Scale Mean if Item Deleted	*Scale Variance if Item Deleted*	*Corrected Item-Total Correlation*	*Squared Multiple Correlation*	*Cronbach's Alpha if Item Deleted*
V2	25.89	22.655	.431	.231	.669
V3	25.76	22.641	.523	.380	.641
V4	25.69	22.561	.480	.315	.653
V5	25.90	24.268	.425	.185	.671
V6	26.14	23.715	.364	.171	.690
V7	26.28	23.273	.411	.182	.675

Factor Analysis This is a third option that is often used to prove validity. Here, the hypothesized subscales or dimensions of the total construct should represent themselves as different dimensions or "factors" at the end of factor analysis. For details on how to perform factor analysis, refer to the Factor Analysis chapter in the book.

SUMMARY

Questionnaire design is aimed at eliciting accurate answers from respondents and makes it possible to analyse the responses to be used in making decisions. Being the primary instrument of data collection in marketing research, it is vitally important to the usefulness of the study. But frequently, questionnaires are not designed carefully, or tested inadequately. One of the major drawbacks of questionnaires is that they force a lot of respondents to think unnaturally, or respond in ways that are unfamiliar to them as a consumer. This chapter has spelt out a few rules which should help those who design questionnaires.

One of the foremost things to remember in questionnaire design is that the respondent must understand what the question means, and therefore the language used must be simple enough for the housewife or man on the street or the least educated respondent to answer. If the questionnaire will be translated formally or informally while asking questions, it must be amenable to such translation. Management jargon or any other difficult words should be avoided unless there is an appropriate target respondent.

Biases of all types, including leading questions must be avoided. A questionnaire must be easy to code, should specify whether multiple responses for a question are permissible, and should allow for unexpected or unusual responses instead of forcing respondents to necessarily pick one of the predetermined options which the researcher has decided.

Above all, a questionnaire must not be too long to answer in terms of time, should be interesting if possible, and should serve its purpose.

Variables of interest or research questions are converted into questions. Questions can be scaled on one of the four scales—nominal, ordinal, interval or ratio, depending on the nature of the variable. In certain cases, the same variable can be measured using two different types of scales. Here, the researcher must decide on which scale is to be used. For example, age can be measured in years, or in categories of years (making it a nominal variable). The requirements of analytical techniques also decide on the scale of variables needed. Most parametric statistical require that the variables are either interval-scaled or ratio-scaled.

Questions can be either closed-ended or open-ended. These are also known as structured and unstructured questions. Quantitative studies normally use structured or closed-ended questions where all responses fall into pre-determined categories. Qualitative research techniques use mostly unstructured questions or items on a checklist. Questions can also be either direct (undisguised) or disguised. It may be necessary to ask disguised questions if we feel that there will be problems getting accurate answers from respondents. Some methods of disguising questions are discussed in the chapter.

An open-ended question leaves it to the respondent to answer as he chooses. There are no pre-determined alternatives. Dichotomous questions have two categories such as Yes or No. Multiple choice or single choice out of multiple alternatives (polytomous) questions are an extension of the dichotomous question. Marketing research frequently requires respondents to rank brands or product attributes or benefits. These are done through rating or ranking type questions. These ratings or rankings can sometimes be combined during the analysis stage. For example, we can combine ranks 1 and 2 given by respondents, and count such respondents for each brand. There are other types of scales such as the Semantic Differential, with or without numbered stops in between. Special scales such as paired comparisons may be required by certain multivariate analytical techniques. Here a respondent is asked to rate the better and worse among a pair of objects such as brands, and goes through a series of such paired comparisons.

Reliability and validity are two tests of a good questionnaire. A questionnaire is valid if it measures what it had set out to measure. A questionnaire is reliable if it provides consistent results.

Questionnaire design is an art, but there are certain common sense rules that can help, as we have discussed throughout this chapter. Scales to be used should be decided on by the researcher in

consultation with the study sponsor, keeping in mind the kind of output formats or tables required for decision-making. Validity and reliability issues are of particular importance if the subject of the study is new or the researcher is inexperienced. Practice with designing questionnaires is the best way to perfect the art. Please do test the questionnaire on a small sample, and modify it if necessary, before going full steam ahead.

CASE STUDY 1

Tamarind Menswear

Given below is a preliminary questionnaire for retailers and consumers of a recently launched menswear brand. Can you list down the research objectives for both questionnaires? Can you modify the given questionnaires to a final draft?

TAMARIND QUESTIONNAIRE FOR RETAILERS

1. Do you have Tamarind? Yes/No
2. What do you think about it?
3. Is there place in the market for one more readymade garment company?
4. What kind of products does Tamarind have? Are they good?
5. Is it a threat to any existing brand? If yes, which one ?
6. If it is not available, what is your view about advertising so heavily before the product is launched?
7. Are people coming and asking for Tamarind?
8. The range of clothes with the retailer.
9. Price range.
10. Name of shop and so on.

TAMARIND QUESTIONNAIRE FOR CONSUMERS

1. Which ads do you recall?
2. Which garment ads do you recall?
3. Have you seen the Tamarind ad?
4. What do you remember from the ad?
5. Do you like the ad? Why?
6. What is the main message?
7. What kind of clothes are Tamarind?
8. What do you think will be the price range?
9. Will you buy it? Why?

CASE STUDY 2

Casual Clothing Preferences of Youth

BACKGROUND

Given below is a preliminary questionnaire from a recent study by a leading advertising agency in Bangalore which wanted to study the youth in that city to find out their casual clothing preferences. The results were to provide inputs for the communication strategy for one of their clients who marketed a range of garments. Some of the questions were modified before the study was done.

CASUAL CLOTHING QUESTIONNAIRE

Quantitative

1. Do you wear casual clothes?
   ____ Yes ____ No
2. How often do you wear casuals?
   ____ Everyday
   ____ Almost everyday
   ____ Often
   ____ Rarely
3. What kinds of casuals are these?
   ____ Jeans
   ____ Cargos
   ____ T-shirts
   ____ Casual shirts
   ____ Cotton trousers
   ____ Any other
4. What are the first 5 brands of casuals that come to your mind?
5. Which brands of casuals do you have/own?
   ____ Jeans
   ____ Cargos
   ____ T-shirts
   ____ Casual shirts
   ____ Cotton trousers
   ____ Any other

6. Which is your preferred brand?
   ____ Jeans
   ____ Cargos
   ____ T-shirts
   ____ Casual shirts
   ____ Cotton trousers
   ____ Any other
7. What is it that you like about your brand?
8. What does your brand give you?
   ____ Recognition
   ____ Satisfaction
   ____ Value for money
   ____ Praise from friends
   ____ Social acceptability
   ____ Any other
9. Which advertisements do you remember about your brand of casuals?
10. What did you like most about that ad?
11. How often do you purchase casuals?
    ____ Once a month
    ____ Once in 6 months
    ____ On impulse
    ____ During sales, discounts and so on
    ____ Any other
12. Do you have clothes specific to occasions/locations?
    ____Yes ____ No
13. On which occasion do you take extra care to dress smartly and which brand do you keep reserved for this occasion (1 Brand)?
    ____ Festivals
    ____ Parties
    ____ Dates
    ____ Discotheques
    ____ Hanging out
    ____ Birth Day
    ____ Any other
    _____ Brand which I reserve
14. How important is the brand?
    ____ Very important
    ____ Important

____ Not important at all
____ Any other

15. On what basis do you select a brand?
    ____ Popularity
    ____ Brand name
    ____ Image
    ____ Current trends
    ____ Availability
    ____ Any other
16. Do you go in for repurchases or do you go in for different brands?
    ____ Yes, I purchase the same brands again (If "Yes", go to Q. 18)
    ____ No, I go in for different brands.
17. Why do you look for a different brand?
    ____ Variety
    ____ Current fashion
    ____ Not satisfied with previous brand
    ____ Qualities of new brand
    ____ On impulse
    ____ New in the market
    ____ Bored with previous brand
    ____ Any other
18. Do you shop... ?
    ____ Alone
    ____ With your friends
    ____ With your family
    ____ With members of the opposite sex
19. Why?
    ____ Financial support
    ____ Gives you a second opinion
    ____ Their choice would be the best for me
    ____ I don't like going alone/For company
    ____ Any other
20. What influences you the most when you make a purchase decision?
    ____ Family
    ____ Friends
    ____ Price
    ____ Quality

____ Colours, fabric, and so on.
____ Any other

21. How much would you pay maximum for your brand of casuals?
    Jeans: ____
    Cargos: ____
    T-shirts: ____
    Cotton trousers: ____
    Casual shirts: ____
    Any other: ____
22. Suppose your pocket money is Rs. 1000, how much would you allocate to each?
    Movies ____
    Dates ____
    Snooker and so on ____
    Clothes ____
    Junk food ____
    Any other ____
23. Do discounts/free gifts and so on, affect your purchase decision?
    ____ Yes ____ No
24. What are the latest trends in clothes presently?
25. Who do you think sets trends in clothing today?
    ____ Filmstars
    ____ Fashion models
    ____ Sportstars
    ____ Business celebrities
    ____ Politicians
    ____ Any other
26. Where do you come to know of latest trends?
    ____ Television
    ____ Films
    ____ Friends
    ____ Just by looking around
    ____ While shopping/window shopping
    ____ Advertisements
    ____ Any other
27. Is there anybody you look up to for cues in fashion?
    _____

28. What are the key motivators when you purchase casuals?
(Rank on a scale of 1–10; 1–lowest; 10–highest)
____ Product
____ Style
____ Design
____ Colours
____ Brand name
____ Price
____ International fashion
____ Fabric quality
____ Comfort
____ Availability
____ During sales, discount season, and so on
29. What kind of guys do girls look for nowadays?
____ Humorous
____ Caring
____ Well dressed
____ Rich
____ Intelligent
____ Any other

Qualitative

1. I wear the kind of clothes I wear because
____ I want to be seen as part of the happening crowd who is in tune with the latest trends.
____ I feel comfortable, don't bother about others, easy to slip into, shows my laid-back life style.
____ I'm fashionable, am the trend setter, want to be the pioneer in making fashion statements.
____ looks sleek, makes me noticed, makes me different, and attractive to the opposite sex.
2. I want to be seen as:
____ the most happening guy in college
____ the laid-back, take it easy kind
____ the most fashionable
____ the 1st guy to try and do anything
____ a career conscious, ambitious guy
____ the no-nonsense kind
____ jolly and fun loving

____ helpful and benevolent
____ any other

3. What would you do to be seen as stated above?
____ Change your style accordingly
____ Make minimal changes
____ Stay the way you are
____ Won't care
____ Go to any lengths
4. Name the guy who's most popular amongst you _____
5. Why ?
____ Dresses well
____ Humorous
____ Helpful
____ Flexible
____ Macho
____ Any other
6. Miscellaneous:
Your favorite bike _____
Your favorite movie star _____
Your favorite sports star _____
Your kind of music _____
Your favorite game _____
Guy you would like to dress as _____
Do you think Net surfing is cool any more? Yes/No
How do you spend your leisure time? _____

Demographics

1. Name:
2. Age:
3. Address:
4. Occupation/Designation:
5. Income group:
____ Less than Rs. 5000
____ Rs. 5000 – Rs. 10000
____ More than Rs. 10000

Questions for the Reader

1. Comment on the casual clothing questionnaires.
2. Improve upon the questionnaires. (You can change individual question format, scale, wording or change the sequencing, or even add or delete questions).
3. Why do you think two questionnaires (quantitative and qualitative) are used?
4. What would happen if you used only the quantitative study questionnaire?
5. Would you recommend using the qualitative questionnaire alone? If yes, under what conditions?

CASE STUDY 3

Parryware—A Survey on Consumers Perception of Bathroom and Sanitaryware

BACKGROUND

Parryware is a range of premium sanitaryware produced and marketed by EID Parry, a part of the Rs. 3500 crore Murugappa Group based in Chennai. It is the leading player in ceramic sanitaryware business in India. In the recent past, it has increased its share of sanitaryware business through acquisition of Johnson Peddar (I) Ltd. Parryware was the first to tap into the growing 'home consciousness' of the Indian consumers and attempted to bring sanitaryware into the realm of conscious brand choice.

However over the last few years there have been many brands of sanitaryware and tiles that have also begun marketing their products in a similar manner. The sanitary ware market is also getting highly competitive with new entrants into the scene. There are brands like Cera and Hindware, which occupy the mass end of the market being competitive in price. Also expected is the entry of up market MNC brands flooding the top of the market.

The sanitaryware market can be basically divided into two categories. These are the

- Premium category
- Secondary category

To keep up with the present competitive market, the Premium category is further subdivided into two categories:

- Leaders such as Parryware, Hindware, Cera
- Laggards such as Johnson Peddar, Raasi, Neycer

On the other hand, the Secondary category is further sub-divided into categories having the following characteristics:

- **Secondary A:** Higher quality standards
- Cost of production and selling too are high
- **Secondary B:** This category tends to follow low cost, low price model of manufacturing and sales—the brand is incidental.

In this scenario, Parryware felt a need to re-examine what consumers feel about bathroom and sanitaryware. The study was designed to examine the significance of the bathroom in a consumer's mind and psychographically profile the users, competitive users, and non-users based on their attitude, lifestyle, occupation, and significance attached to their bathroom. It was to also try and ascertain whether there has been any significant change in the perception of bathrooms and bathroom fittings in the last decade. This research was to help Parryware to give a new dimension to its thinking about the product category in line with consumer thought processes and fine-tune its positioning and marketing strategy.

A description of the study follows, starting with a statement of the research objective, and going on to sample design and ending with a detailed questionnaire.

Research Objective

- The major objective of this research is to understand the significance of a bathroom in the consumer's mind. To find out what rational and emotional values are attached with the bathroom and to determine whether there has been any significant shift in the consumer's perception of bathrooms and sanitaryware with the passage of time.
- The study will try to reveal the difference in perception, attitude, and lifestyle of North Indians and South Indians so far as bathrooms and sanitaryware are concerned.
- The survey will also be conducted to find out the
  (a) Present scenario of sanitaryware market
  (b) Key demand drivers in sanitaryware market
  (c) Key players in sanitaryware market by understanding the consumer's
    (i) Awareness level
    (ii) Perception, attitude to the brands
    (iii) Perceived strength and weakness
    (iv) Values associated with the brand
    (v) Image associated with the brand
- To explore an alternative route to position Parryware/ways and means to communicate the current route in a more relevant and competitive and unique manner.

Methodology

The methodology that will be followed will include both

- ☐ Primary data collection
- ☐ Secondary data collection

Primary data collection will include both in-depth interview and simple questionnaire.

In-depth interview will be conducted with North Indian and South Indian couples who either own a house/flat or stay in a rented house/flat, and those who are planning to build/renovate.

Such interviews will help us to ascertain the difference in perception and attitude towards bathroom and sanitaryware between North Indian and South Indian couples.

Couple interviews will provide detailed information needed to profile the consumers based on their needs. These interviews will be held at their homes and hence an observation of the bathroom will also be made simultaneously.

The survey will also incorporate a simple questionnaire along with the in-depth interview for those who are planning to renovate their house or bathroom.

To further understand the process of brand choice of sanitaryware, the research will also identify the possible influencers (e.g. architects, contractors) for intenders to build/renovate a house, bathroom in particular with the help of a simple questionnaire.

Secondary Data Collection will be incorporated to find out the present scenario of sanitaryware market and key demand-drivers in sanitaryware business.

Target Audience

Category: Couples

- North Indians—28
- South Indians—28

A. Owners of house/flat having a car
- Businessmen—8
- Professional—8

B. Owners of house/flat having a two-wheeler
- Businessmen—8
- Professional—8

C. Staying in a rented house/flat having a car or two-wheeler
- Businessmen—8
- Professional—8

D. Intended owners, that is, those planning to build or renovate—8

Age group: 25–45 year-old couples

Usership: Owners of Parryware and other competitive sanitaryware including the local or other unbranded ones.

AREA TO BE COVERED: CHENNAI

Consumers Discussion Guide—Parryware

Psychographics

1. **Daily Routine:**
   - What is your daily routine on weekdays and weekends?
     Actual and Ideal
     Activities enjoyed and activities not enjoyed
2. **Relationships: Parents, In laws, Spouse, Friends, Colleagues**
   - Describe each relationship, what role does each of them play in your life and how significant is each relationship to you?
   - Do you think that the definition of each of this relationship has changed over the years…. if so ….why?
   - When do you interact…what are the subjects that you discuss ……why?
   - On what occasions does your husband normally take the decision……why?

- Cite any instance when your husband took a decision, which you disliked… How did you feel?
- What are the situations in which you would turn towards this person ……when would you hesitate…when would you never approach to them?

3. **Occupation: Level of involvement and attitude**
   - If you had to describe your profession to someone how would you explain it…why?
   - What would you describe as the most enjoyable aspect of work and unenjoyable aspect of work?

     Satisfaction, Remuneration, Experience, Opportunity for development, Ambience
4. **Leisure Activities**
   - What do you normally do in your spare time?
   - What kind of magazines, newspapers do you normally browse through?
   - TV—which programmes do you watch, how much time, how often?
   - What are your favourite programmes?
   - How often do you go for a tour?
   - What are your holiday destinations?
   - Who normally takes the decision of going for a tour?
   - On tours what kind of hotels do you generally stay in?
   - What do you think about discos, pubs, coffee pubs.
   - What is your opinion about shopping malls with respect to time and money……..Why?
5. **Home furnishing and Decor**
   - What are the areas in your life where fashion is relevant…. why? (Home, furniture, furnishing, accessories, durables, sanitaryware)
   - List out all the brands that you consider *Fashionable* and *Unfashionable*…why? What makes one brand fashionable and the other not?
   - Which of the durables and non-durables do you own….which do you consider either as *Necessity* or *Luxury*…which do you consider as *High cost* or *Affordable*
   - In this day and age what does the concept of 'Home Furnishing and Décor' mean to you?
   - What kind of brands do you generally look for while selecting home accessories and furnishings for your home?
   - What style (contemporary, ethnic) of sofa, chairs, curtains, upholstery, do you go for, for enhancing the look of your house?
   - Where do you generally go for shopping of home accessories and furnishings …and with whom?
   - When you go for shopping what do you observe as being the critical attitude to shopping? How has this changed in the last 10 years…why?
   - On what occasions do you think people are willing to spend money and when they are resistant to spending it?
   - What are the three things you bought recently to enhance the look of your house?

- Of all the purchases that you have mentioned what could be classified as planned purchase, impulse purchase, splurging.
- Who normally takes the decision in the purchase of the following items…. kitchenware, sanitaryware, technology driven products…why?
- Imagine that, as a part of your yearly planning you are asked to draw up the budget plan and allocate funds for the year, what purchases would you allocate for quarterly allocation and monthly allocation.

6. **Decision Making: Related to the purchase of Kitchenware, Sanitaryware, and Technology driven products**
   - Imagine that you have decided to purchase a product, where would you get information from…why?
   - Magazines, Advertisements, Dealers, the Internet, Showrooms
   - Having information about the product what would you do next…why? (Visit showrooms, scouring the market, taking advice of the people)
   - Who would you go for advice/suggestions for each of the product category?
   - Whom would you rely on the most and the least?
   - In today's world whom (*Friends or Relatives*) do you consider as the prime influencer in the purchase of the following items?
   - What are the factors in their suggestions that would be critical in swaying your decision…….why? What would be lower order priorities?
   - Who in the family takes the final decision…why? Is there anyway they can be persuaded otherwise…how?
7. **Psycho drawing: Everyday life vs. Ideal life**
   - What aspects of your everyday life do you *Like* and *Not Like* to see in your ideal life?
   - If you were allowed to design a desirable lifestyle what would it include—housing, clothes worn, vehicle owned, food consumed, type of family, durables owned ?
   - How would it be different from your current lifestyle?
8. **Perception of Bathrooms: Past and Present**
   - Are you the owner of the house?
   - If no, are you satisfied with the size of the bathroom?
   - Which size of a bathroom would you prefer to have? 20 sq ft/40 sq ft/60 sq ft
   - How many bathrooms would you prefer to have?
   - What do you think are the benefits of having attached bathrooms?
   - On a scale of 1–5 (1—most important, 5—least important) how would you rank different rooms (bedroom, kitchen, drawing room, bathroom) on basis of *Lifestyle and Privacy*?
   - How would you define 'Bathroom'? How do you think the definition of bathroom has changed with the passage of time?
   - How can you relate *Romance* with the bathroom?
   - Besides taking bath what other activities do you like to do in the bathroom?

- Can you recall any special moment in the bathroom?
- How much time do you spend per day in a bathroom?
- How do you normally take a bath…if you were given a choice how would you prefer to take it in the near future?
- Where do you prefer to dress up…what will you like to have in your bathroom in order to dress up?
- When you go to hotel how do you feel when you enter their bathroom?
- Do you prefer to go for such bathrooms…if Yes, why?…if No…what do you feel about those who go for such bathrooms?
- What will you like to have in your *Dream Bathroom*?
- How old are your kids…how much time do they spend in the bathroom…how do you feel when they take long hours in the bathroom?
- What was the level of involvement of each of your family members, friends, and colleagues in deciding how your bathroom should look like?
- How often will you like to go for remodelling of the bathroom? (once in a year/twice in a year/more than twice a year)?
- What constraints did/will you face for remodelling of the bathroom?
- What triggered you/will trigger you to go for remodelling of the bathroom?
  Hobby/Recent trend/In-laws/Friends/Parents/Status
- If you think of remodelling it once again what will you like to have in your bathroom…how will it be different from the present bathroom…why?

9. **Sanitaryware:**
   - If you had to elicit all the factors that you consider while buying/ if you were to buy …what would they be…could you please list them in order of their importance and reasons behind each of them.
   - How will you/did you go about deciding on the brand. Why?
   - What do you think is the striking feature of the sanitaryware that you own?
   - In the near future if you think of replacing it what other features will you look for…What other brands will you opt for….why?
   - From the list of words can you tell me which of these is applied to sanitaryware and which cannot be applied to them at all.
     - Economical
     - Good appearance
     - Compact design
     - Variety of colors
     - Sculpted shapes
     - Easy to use
     - Long-term purchase
     - Source of envy

- Enhance the role of a care giver
- Time bound
- Adds to your glamour
- Symbol of upward mobility

- If you had to identify any way in which sanitaryware can be improved and made better …what would be your suggestions?
- What effect will this have on your current sanitaryware?

10. **Triggers /Barriers**
    - Mukesh and Suresh are friends …Mukesh is extremely keen on purchasing a 'Premium Brand of sanitaryware'…what could his reasons be…where would Mukesh have got this information from…what are the critical factors that have convinced him…what can you say about his occupation/lifestyle/personality …what Premium brands do you think he has opted for….why?
    - On the other hand Suresh is resistant to purchasing a Premium brand of sanitaryware although he has every opportunity. Where does his hesitation stem from..why? What can you suggest about his occupation/lifestyle/personality. What brands do you think he will go for?
11. **Brands:**
    - When you think of any sanitaryware which brand comes to your mind?
    - When you first saw/hear of this brand what kind of impression did you form….why?
    - What other brands of sanitaryware are you aware of? If you were asked to classify these brands on the following parameters, how will you do so?

Parameter	**Brand A**	**Brand B**	**Brand C**
Frequent promotion vs. Infrequent promotion			
Affordable vs. Expensive			
For all vs. For few			
Reflects Lifestyle vs. Does not necessarily reflects lifestyle			
One-time purchase vs. All-time purchase			

- Taking one brand at a time what are its benefits and drawbacks?
- Imagine that this brand turned into a person …what kind of a person would this be…age, lifestyle, occupation.
- If you had to associate this brand with the following….how will you do so…..why?

  A restaurant, a brand of car, a brand of watch, a brand of clothes, a holiday destination.
- Which brand of sanitaryware do you own…why ….how is it different from other brands…in the near future if you think of replacing it which brand will you go for …why?
- When you first saw/heard of this brand what kind of impression did you form….why?

Thank you

Questions for the Reader

1. What differences do you find in a questionnaire for a depth interview such as the one in this case and the one used in routine surveys?
2. Comment on the above questionnaire in terms of appropriateness of the questions for the research objectives, length of the questionnaire, logical flow of questions, and so on.
3. Analyse each section of the questionnaire separately and try to improve on it.
4. Can some other approaches be used to gather similar information from the target group? (This may require knowledge of advanced statistical methods covered in Part II of the book).
5. How would you go about analysing this study after finishing data collection? Explain the approach and details of expected formats for reporting findings, relevant to this study.

ASSIGNMENT QUESTIONS

1. Look at the questionnaire given below.

SHIMOGA CO-OPERATIVE MILK PRODUCERS SOCIETIES UNION LTD.

Urban Household Survey

Questionnaire

A. Household Identification

1. Milk Union [7|4]
2. City [|]
3. Zone [|]
4. Investigator [|]
5. Interview No. (for each ward/zone) [| |]

Name and address .....

.....

.....

B. Household Particulars

1. a) Respondent's name ..... (Name)

   b) His relationship with head of the household ..... (Relationship)

   c) Education and occupation of main Income Earning Member of family

Education

Education	Code
No response	1
Uneducated	2
Primary school educated	3

Occupation

Occupation	Code
No response	1
Govt service	2
Private service	3

Middle school educated	4
Higher Secondary/SSC/SSLC	5
Graduet diploma course	6
Post graduate/Higher studies	7
Professional training	8
Others	9

Private practice	4
Petty trade/shop owner self-employed	5
Business	6
Labour/daily paid worker	7
Farming	8
Not working/unemployed	9

2. Please provide details of your family

Below 7 Yrs		7–15 Yrs		16–25 Yrs		Over 25 Yrs		Total	

3. Do you own milch animals? Yes: 1 No: 2 ☐
   If yes, go to E & then return to fill C & D

C. Milk Purchases & End Use

1. (a) Do you purchase liquid milk daily? Yes: 1 Not daily: 2 Never: 3 ☐

   If no, please indicate reason for not buying liquid milk daily or never buying ☐

Reason for not buying milk	Code
Cannot afford	1
Milk Producer	2
Good quality not available	3
Buying milk powder	4
Bachelor/eating outside	5
Any other	6

(b) No. of days the household purchases milk in each month, if not daily ☐☐

(c) Whether buying loose or packed? Loose: 1 Packed: 2 Both: 3 ☐

(d) If home delivery of packed milk then what are the delivery charges Rs/1 pack (500 ml)

Rs Paise

Rs/ 1 month

2. Details of last days' liquid milk purchases/retained:

Source	Quantity (ltrs)				Total Qty (ltr)	Total value (Rs)
	Milk Type	Morning	Afternoon	Evening/ Night		

Source of milk supply	Code	Type of milk	Code
Nandini	1	Skim milk/Seprated milk	1
Laxmi	2	Toned milk	2
Surabhi	3	Double toned milk	3
Krishna	4	Standardised milk	4
Milk Producer	5	Whole milk	5
Others:	6	Cow milk	6
Milkman/Vendor	7	Buffalo milk	7
Milk shop	8	Goat/Loose (unpacked) milk	8
		Do not know the type of pack	9

3. Why are you prefering/patronizing particular sources(s) of milk?

Source ☐☐☐

1st reason ☐☐☐

2nd reason ☐☐☐

Reasons	Code
Additional emergency requirements	1
Time of milk supply convenient	2
Immediately available/no waiting time	3
Reasonable price	4
Quality is Good	5
Suitable mode of payment	6
Proximity of house	7
Home delivery	8
No other source available	9

4. Are you satisfied with regard to the type of milk you buy?

Source	Milk type	Satis code	Reason

	Code
Satisfied	1
Not Satisfied	2
No comment	3

Reason for satisfaction:

Reason	Code
Available on all days	1
Comparatively lower price	2
No advance payment	3
Timely supply	4
Good quality (Taste)	5
Good quality (Cream/Malai)	6
Good quality (No sourage)	7
Good quality (No adulteration)	8
Any other (name).....	9

Reason for dissatisfaction:

Reason	Code
Irregular supply	1
Higher price	2
Bad quality (taste)	3
Bad quality (less cream/malai)	4
Bad quality (frequent sourage)	5
Bad quality (often adulteration)	6
Insist on advance payment	7
No time schedule for supply	8
Any other (name).....	9

5. (a) How the milk bought/produced last day was used? (ltrs)

Drinking
a) Children (below 7 yrs)
b) Children (7–15 yrs)
c) Other members

Other uses
—Making tea/coffee
—Dahi/curd making
—Converted into channa/paneer
—Other uses
Total quantity

(b) Use of home-made dahi:

Prime use	Second use

Use	Code
Eating	1
Making butter/ghee	2
Others	3

(c) Do you make ghee/butter from malai? Yes: 1 No: 2 [ ]

D. Milk Products

Quantity of fresh products & milk powder purchased during last month

Product code		Qty unit code	Qty Purchased			Price Rs/kg or Rs/ltr			Branded—1 Unbranded—2	Preferred Pack size code

Pack size	Code
100 gms	1
200 gms	2
250 gms	3
500 gms	4
1 kg	5
5 kg	6

Unit	Code
Kg	1
Ltr	2

Product Name	Code
Peda	1
Butter	2
Ghee	3
Skim milk powder	4
Curd in packs	5
Lassi	6
Butter milk	7
Paneer	8

Products to be decided based on the MPU's requirements.
Print names of the products in the questionnaire

E. Milch Animal Holding and Milk Production

1. How many adult female animals do you have? (Nos)

Cattle

Buffalo

Goat

2. What was the total milk production yesterday (Qty in ltr)

Cattle				
Buffalo				
Goat				
Total				

3. How was the milk produced yesterday used ?

a) Retained (ltr)				
b) Sale Qty (ltr)				
Sale Price (Rs/ltr)				
Sold to a)				
b)				

Sold to	Code
Neighbour	1
Dudhia	2
Shop/Pvt. Dairy	3
Coop Dairy	4

F. Monthly Expenditure & Income

1. Monthly expenditure on:

	Food items		Non-food items		Total
Rs		+	Rs.	= Rs	

Liquid Milk (Rs)	
Pure ghee/butter (Rs)	
Milk Products (Rs)	
Oil/Vanaspati Ghee (Rs)	
Fringe Food (fast food) (Rs)	

Details of monthly expenditure

Food items	(Rs/month)	Non-food items	(Rs month)
1. Cereals (Rice/ Wheat, etc.)		1. House Rent	
2. Pulses		2. Clothing	
3. Vegetables		3. Laundry/Washing	
4. Liquid Milk		4. Education (fees, books/notebooks, etc.)	
5. Pure ghee		5. Education (travelling by bus, auto, taxi)	
6. Butter		6. Travelling (petrol, bus/auto/taxi)	
7. Milk Products		7. House servant's salary	
8. Edible oils/vanaspati		8. Entertainment (going to garden, theatre, etc.)	
9. Fruits		9. a) Electricity/Telephone	
10. Others Name a) Gas		b)	
b)		10. Others (miscellaneous)	

Remarks .....

Checked by (Name) .....

Signature .....

Date.....

(a) What could be the research objectives of this study?

(b) How would you improve upon the questionnaire, keeping the objectives the same?

2. Under what conditions would you use a self-administered questionnaire? A mailed questionnaire?
3. How would you design a questionnaire for a telephonic survey?
4. What points do you need to keep in mind when you design a questionnaire for a mail survey?
5. If you had a study that required you to do a survey of email users and you had to design a questionnaire for this survey, is there anything special you would do? What are the decisions you need to take regarding design and to ensure a high rate of cooperation from respondents?
6. Design a questionnaire for a pharmaceutical manufacturer who has launched a new Hepatitis-B vaccine in the market six months ago. The survey is intended for doctors, is to be done nationally in India, and is required to test the doctors' awareness of the brand *X* vaccine launched by this company. There are two other brands *Y* and *Z* in the market, made by competitors. The company also wants to know doctors' perception of the price, efficacy and side-effects of the three vaccine brands. If possible, the company wants to know the consumers' perceptions about the brand and/or company (through doctors).
7. A second-tier consumer goods company had launched a fairness cream in the market two years ago. It competed against the leading brand 'Fair & Lovely', and a couple of other brands. This company's brand was called 'Fairy Tales', and was launched with a campaign based on adaptations of some popular fairy tales. Design a questionnaire to test consumer reactions to the advertising campaign, and the themes used. The consumer's demographics are of particular interest to the company. Factors such as age, education level, family income, hobbies are to be included. Consumer's awareness and usage status of any of the brands of fairness cream in the market, and reactions to the new brand 'Fairy Tales' are to be tested.
8. Pepsi has commissioned you to do a survey to find the public perception of its advertisement featuring actor Shah Rukh Khan in which there is a dig at another actor Hrithik Roshan. The respondent has to be a cola drinker, should have seen the ad, and has to express his feelings about it. Design a brief questionnaire to do the job, which should takes about 6–8 minutes to complete (per questionnaire).

The McGraw·Hill Companies

CHAPTER 5 SAMPLING METHODS—THEORY AND PRACTICE

Learning Objectives

In this chapter, we will

- Introduce the terminology in sampling such as sampling unit, sampling element, population, sampling frame
- Discuss the formulas for calculating sample size
- Discuss other issues which affect sample size decisions
- Elaborate on specific probability and non-probability sampling techniques
- Introduce the ideas of sampling error, non-sampling error, and total error in a research study, and discuss their impact on sample size decisions.

BASIC TERMINOLOGY IN SAMPLING

We will start by defining some basic sampling terminology and then move on to the issue of how to calculate the sample size required to fulfill the needs of a particular research study.

Sampling Element

This is the unit about which information is sought by the marketing researcher for further analysis and action. The most common sampling element in marketing research is a human respondent who could be a consumer, a potential consumer, a dealer or a person exposed to an advertisement, and so on. But some other possible elements for a study could be companies, families or households, retail stores and so on.

Population

This is not the entire population of a given geographical area, but the predefined set of potential respondents (elements) in a geographical area. For example, a population may be defined as "all mothers who buy branded baby food in a given area" or "all teenagers who watch MTV in the country" or "all adult males who have heard about or use the AQUAFRESH brand of toothpaste" or similar definitions in line with the study being done.

Sampling Frame

Even though it is a relatively easy task to define a target population for a study, it is much more difficult to identify or list every member of such a population. In most cases, therefore, we decide on a sampling frame, a subset of the defined target population, from which we can realistically select a sample for our research. For example, we may use a telephone directory of Mumbai as a sampling frame to represent the target population defined as "the adult residents of Mumbai". Obviously, there would be a number of elements (people) who fit our population definition, but do not figure in the telephone directory. Similarly, some who have moved out of Mumbai recently would still be listed. Thus, a sampling frame is usually a practical listing of the population, or a definition of the elements or areas which can be used for the sampling exercise. For reasons of budget or time constraints, certain areas of the country may be excluded from the sampling frame of a study, even though they would be a part of the defined population.

Sampling Unit

If individual respondents form the sample elements, and if we directly select some individuals in a single step, the sampling unit is also the element. That is, both the unit and the element are the same. But in most marketing research, there is a multi-stage selection. For example, we may first select areas or blocks in a city or town. These form the first stage sampling units. Then, we may select specific streets within a block or area, and these are called second stage sampling units. Then we may select apartments or houses—the third stage sampling units. At the last stage, we reach the individual *sampling element*—the respondent we wanted to meet.

THE SAMPLE SIZE CALCULATION

Students of marketing research are most interested in the subject of how sample sizes are determined. The first thing that should be said about the subject is that it is not a formula alone that determines sample size in actual marketing research. Sampling in practice is based on science, but is also an art. The basic assumptions made while computing sample sizes through the use of formulae are sometimes not met in practice. At other times, there are other factors which are influential in increasing or decreasing sample sizes obtained through the use of formulae. However, we will first discuss the use of the two commonly used formulae for sample size determination, and later move on to a discussion of their limitations, and how the limitations are overcome in practice. For now, the reader would do well to remember that sample size determination is a blend of using formulae, experience of similar studies,

time and budget constraints and a few other elements such as output or analysis requirements, number of segments of the target population, number of centres where the study is conducted, and so on.

Formula for Sample Size Calculation when Estimating Means (for Continuous or Interval-scaled Variables)

The formula for computing '*n*', the sample size required to do the study, is:

$$n = \left(\frac{Zs}{e}\right)^2$$

Let us examine one by one what the quantities '*Z*', '*s*', and '*e*' represent. We will then apply the same to an example to see how it works in practice.

***Z*:** The '*Z*' value represents the *Z* score from the standard normal distribution for the confidence level desired by the researcher. For example, a 95 per cent confidence level would indicate (from a standard normal distribution for a two-sided probability value of 0.95) a *Z* score of 1.96. Similarly, if the researcher desires a 90 per cent confidence level, the corresponding *Z* score would be 1.645 (again, from the standard normal distribution, for a two-sided probability of 0.90).

Generally, 90 or 95 per cent confidence is adequate for most marketing research studies. A 100 per cent confidence level is not practical, as it means we have to take a census of the entire population, instead of using a sample.

We will use $Z = 1.96$, equivalent to a 95 per cent confidence level, in our example.

***s*:** The '*s*' represents the population standard deviation for the variable which we are trying to measure from the study. By definition, this is an unknown quantity, since we have not taken a sample yet. So, the question of knowing the value of '*s*', the population standard deviation, does not arise.

However, we can use a rough estimate of the population standard deviation for the variable being measured. This estimate can be obtained in the following ways:

1. If past studies have measured this variable, we can use the standard deviation of the variable from one of the studies from the recent past. It serves as a good approximation.
2. A very small sample can be taken as a test or pilot sample, only for the purpose of roughly estimating the population standard deviation of the concerned variable.
3. If the minimum and maximum values of the variable can be estimated, then the range of the variable's values is known. Range = maximum value – minimum value. Assuming that in practically all variables, 99.7 per cent of the values of the variables would lie within ± 3 standard deviations of the mean, we could get an approximate value of the standard deviation by dividing the range by 6.

   The logic of this is, that range is equal to 6 standard deviations for most variables. Therefore, range, when divided by 6, should give a fairly good estimate of the standard deviation.

***e*:** The third value required for calculating the sample size required for the study is '*e*', called tolerable error in estimating the variable in question. This can be decided only by the researcher or his sponsor for the study. The lower the tolerance, the higher will be the sample size. The higher the tolerable error, the smaller will be the sample size required.

Now, let us take an example of the use of the above formula, to see how it works.

Let us assume we are doing a customer satisfaction study for a washing machine. We are measuring satisfaction on a scale of 1 to 10. 1 represents 'Not at all satisfied', and 10 represents 'Completely Satisfied'. The scale would look like this on a questionnaire:

1 2 3 4 5 6 7 8 9 10

Customer Satisfaction Scale

For a moment, we will assume that the questionnaire consists only of 7–8 questions, all of them using this 10-point scale. Therefore, the variable we are trying to measure or estimate through the survey, is customer satisfaction, which is being measured on a 10-point interval scale.

We will apply the formula discussed for sample size calculation, and check for its usefulness.

$n = \left(\frac{Zs}{e}\right)^2$ is the formula, for variables which are continuous, or interval scaled.

Z Let us assume we want a 95 per cent confidence level in our estimate of customer satisfaction level from the study. Then, from the standard normal distribution tables, (for a two-sided probability value of 0.95), the *Z* value is 1.96.

s Let us assume that such a customer satisfaction study was not conducted in the past by us. We have no idea of the standard deviation of the variable 'Customer Satisfaction'. We can then use the rough approximation of range divided by 6 to estimate the sample standard deviation.

In this case, the lowest value of customer satisfaction is 1, and the highest value is 10. Thus, the range of values for this variable is 10 – 1 = 9. Therefore, the estimated sample standard deviation becomes 9/6 = 1.5. We will use this value of 1.5, as '*s*' in our formula.

e The tolerable error is expressed in the same units as the variable being measured or estimated by the study. Thus, we have to decide how much error (on a scale of 1 to 10) we can tolerate in the estimate of average customer satisfaction. Let us say, we put the value at + 0.5. That means we are putting the value of '*e*' as 0.5. This means, we would like our estimate of customer satisfaction to be within 0.5 of the actual value, with a confidence level of 95 percent (decided earlier while setting the '*z*' value).

Now, we have all 3 values required for calculating '*n*', the sample size. So let us calculate '*n*'.

$$n = \left(\frac{Zs}{e}\right)^2 \left(\frac{1.95 \times 1.5}{0.5}\right)^2$$

$$= (1.96 \times 3)\ 2 = 34.57 \text{ or } 35 \text{ (approximately)}$$

Therefore, a sample size of 35 would give us an estimate of customer satisfaction measured on a 1 to 10-point scale, with 95 per cent confidence level, and error level maintained within ± 0.5 of the actual value.

If we were to tighten our tolerance level of error (e) to ± 0.25 instead of ± 0.5, we would have to take a sample of higher size.

'n' would then be equal to

$$\left(\frac{1.96 \times 1.5}{0.25}\right)^2 = (1.96 \times 6)\,2 = 138.3$$

$$= 138 \text{ (approximately)}$$

Similarly, for any change in the estimate of 's' or the value of 'Z' we choose to set, the value of 'n', the sample size, would change.

In general, sample size would increase if

1. standard deviation 's' is higher
2. confidence level required is higher
3. error tolerance 'e' is lower

The major things to remember in the above formula are that

1. 'Z' value is set based on the confidence level we desire.
2. 's' value is estimated from past studies involving the same variable, or from the approximate formula of $\left(\frac{\text{Range}}{6}\right)$, if we can estimate the range of values for the variable in question.
3. 'e' value is also set by us.

The reason for reiterating these points is to emphasise the fact that sample size to a large extent is determined by the researcher's priorities and assumptions, and at least two of the values (Z and e) are set by the researcher himself. If one changes the assumptions, the formula would give different sample size recommendations. Even apart from this, there are some other limitations of the formula, which we will discuss later. First, we will look at one more commonly used formula—for estimating proportions.

Formula for Sample Size Calculation when Estimating Proportions

In cases where the variable being estimated is not continuous, but a proportion or a percentage, a variation of the formula mentioned earlier should be used. Such variables are typically found in questions that have a dichotomous scale, with only two choices for an answer. For example, regular users versus non-users. If we are estimating the proportion of respondents who are regular users of our brand of toothpaste, say, we might use following formula to determine sample size.

Here, the formula is

$$n = \boldsymbol{pq}\left(\frac{\boldsymbol{z}}{\boldsymbol{e}}\right)^2$$

Let us look at the meaning of each of the terms on the right hand side of the formula.

'*p*' is the frequency of occurrence of something expressed as a proportion. For example, if the number of users you would expect to find in a sample is 1 out of every 4 respondents, '*p*' would be ¼ or 0.25. '*q*' is simply the frequency of non-occurrence of the same event, and is calculated as (1-*p*). In other words, '*p*' and '*q*' always add up to 1. Here again, it should be noted that we are actually trying to determine '*p*' or estimate '*p*' by doing our survey. So, the estimate of '*p*' that we use to compute '*n*' in the formula is either a very rough guess based on prior studies, or on some other data. It is used only to calculate the sample size '*n*'. Only after doing the study will we have our true estimate of '*p*', the proportion of users in the population. It is similar to the problem mentioned earlier (in the estimation of means for continuous variables) when we used an estimate of '*s*' before doing the actual study, only for the purpose of computing sample size.

***Z*:** '*Z*' is the confidence level-related value of the standard normal variable, as discussed in the earlier section. It is equal to 1.645 for 90 per cent confidence level, and 1.96 for 95 per cent confidence level (from the standard normal distribution table).

***e*:** '*e*' is once again, the tolerable level of error in estimating '*p*' that the researcher has to decide. If we decide that we can tolerate only a 3 per cent error, '*e*' has to be expressed in terms of the same units as '*p*'. So, a 3 per cent tolerable error would translate into $e = 0.03$ because '*p*' is a proportion, with values ranging from 0 to 1 only. '*q*' is also a proportion, with the same range of values, and $p + q$ is equal to 1.

Example of Use of Formula for Proportions

Let us plug in some numbers to see how the formula works. Assuming we are trying to estimate the proportion of the population who use our toothpaste brand AQUA, let us assume that we want a confidence level of 95 per cent in our results (which means $Z = 1.96$), and '*e*' is 0.03, as discussed above. '*p*', from previous studies or from prior knowledge, is estimated as 0.25 for the purpose of sample size determination.

Then,
$$n = pq\left(\frac{z}{e}\right)^2$$

which is equal to $(0.25)\,(0.75)\left(\frac{1.96}{0.03}\right)^2$

or
$$n = (0.25)\,(0.75)\,(4268.4)$$
$$= 800$$

Therefore, we need a sample size of 800 respondents to estimate the true value of '*p*', with a 95 per cent confidence level, and with an error tolerance of ± 0.03 from the true value.

Here again, the sample size is higher if

1. the confidence level is higher
2. the error tolerance is lower

But, the relationship between sample size and estimated 'p' is somewhat different. The sample size increases as 'p' increases from 0 to 0.5, but decreases thereafter, as 'p' increases from 0.5 to 1. Thus, other things being equal, sample size required is maximum if 'p' is equal to 0.5. This is because the formula also contains 'q' which is equal to (1-p). The product of 'p' and 'q' is maximum when $p = 0.5$, $q = 0.5$ ($0.5 \times 0.5 = 0.25$). At all other 'p' values, the product of 'p' and 'q' is less than 0.25. Therefore, the sample size formula gives the highest value when $p = 0.5$.

This also gives us an easy way out of estimating the value of 'p', if past information is not available. We can simply set the value of 'p' to 0.5, because that will give us the maximum sample size. This could be an overestimated sample size, but it can never underestimate sample size.

OTHER ISSUES THAT AFFECT SAMPLE SIZE DECISIONS

Even though we use formulae in some situations, particularly when doing a study for the first time, we have to recognise the several limitations in using them. Some of the important issues and limitations are discussed below.

1. Number of Centres

One major practical constraint in determining appropriate sample size is the assumption in the above formulae that we are dealing with one set of respondents.

Actually, there are several segments one deals with. For example, most studies deal with multiple locations spread across the country. If the data is to be analysed separately for each geographical segment, the overall sample size obtained from the formula has to be split into these geographical centres or segments. In such cases, we may intervene, and ensure that a minimum sample size is made compulsory for each centre/city.

2. Multiple Questions

Different varieties and scales of variables are used in a questionnaire. Our assumption in using the above formulae was that we have only one major type of variable in the questionnaire—either a continuous variable or a proportion.

Actually, we have many different types of variables in any commonly used questionnaire. This may require formulas to be used for each different scale/type of variable. Then, one may have to reconcile the different sample sizes arrived at for each different variable type. Usually, the easy way out in such cases is to take the maximum sample size which is calculated, for any major variable in the questionnaire.

3. Cell Size in Analysis

Just as there are segments in geographical terms, one may want to analyse data by other segments, one or two segments at a time. For example, we may be interested in analysing the combined effect of income and age on some variable of interest. There may be 5 income categories among our

respondents, and 4 age categories. This creates a table with 5 × 4, or 20 cells. Now, even though the overall sample size was adequate for simple analysis, the sample size in some of these 20 cells may not be adequate. There are various rules-of-thumb used to overcome or prevent such problems. One says that each cell must have a minimum of 10 entries for us to do any analysis using that cell. Such problems can be overcome more easily if we know in advance what types of analysis we are likely to do. In other words, blank formats of output tables can be specified before doing the study.

4. Time and Budget Constraints

Many a time, a study has to be done quickly to aid decision-making, or to prevent competitors from learning too much about possible marketing strategy changes. There may also be budget constraints, because more money has been spent in product development, or in promotions, and so on. Sampling design has to keep in mind both the time and budget constraints for the study, before finalising a sampling plan.

5. The Role of Experience in Determination of Sample Size

Given the many limitations in using formulae to determine the 'right' sample size, it is important to recognise that past experience of conducting marketing research studies is a quick substitute, at times, for the use of formulae. Even otherwise, experience of similar research can be used to moderate or adjust the numbers crunched out by the formulae.

We will next discuss some of the commonly used sampling techniques, their merits and demerits.

SAMPLING TECHNIQUES

These can be classified under two major types—probability and non-probability.

Probability Sampling Techniques

These are techniques where each sampling unit (usually a household or individual in a marketing research study) has a known probability of being included in the sample. The probability of inclusion need not be equal for every sampling unit. In some methods, it is equal, and in some others, it is unequal. But it should be a known probability, for it to be classified as a probability sampling method.

The other major distinguishing feature of probability sampling methods is that they are unbiased. The scheme of selection of units from the target population is pre-specified, and then the sample is selected according to the scheme and not according to any biases or preferences of the researcher. In practice, however some small compromises do take place, but it is a largely unbiased set of techniques when used in the field.

There are quite a few practical difficulties in using the probability sampling methods, whichever method we may want to use. In such cases, the best feasible theoretical method with minor modifications may be used. The major types of probability sampling techniques are:

1. Simple random sampling
2. Stratified random sampling
3. Cluster sampling
4. Systematic sampling
5. Multi-stage or Combination sampling

We shall briefly look at each one, its merits and demerits.

1. Simple Random Sampling This technique is conceptually the easiest to understand, but quite difficult to implement in a realistic marketing research project. To illustrate what it is, assume that we wish to estimate the average income level of 100 employees of a company. We do not have access to their income levels, so we have to interview them and find out their income level. We have a time constraint, and we just need a quick estimate. Assume that we have decided we would be happy with a sample of 5, randomly selected from the 100. How do we select the sample?

If we wish to use simple random sampling we could make a list of all 100 employeees. Then, a number could be allotted to each employee. We could then write these 100 numbers on small pieces of paper, one number on each. Shuffling these folded pieces of paper, we can draw 5 pieces out of the 100, and use these employees as our sample.

Practical Issues This appears very easy to do when there is a relatively small number of people to pick from. But when we deal with typical marketing research problems, the numbers are quite large, and more importantly, the exact numbers are not known. This creates a very practical difficulty for the marketing researcher who wishes to use simple random sampling. Imagine trying to procure a list of all Indian consumers of toilet soap, for a study into their brand preferences. It is an impossible task, and therefore, simple random sampling, strictly speaking, is not feasible.

But it is possible to use modifications of the basic technique, with reasonable checks and balances to keep the method unbiased in practice.

2. Stratified Random Sampling In this technique, the total target population is divided into strata or segments on the basis of some important variables. For example, a consumer population may be divided into age brackets of below 25, 25–40, and above 40 years. Then, a sample is taken from each of the strata defined earlier.

Practically, the overall sample size is first calculated, using a formula of the type discussed earlier, or based on judgement and experience. This overall sample is then divided into sub-samples for each stratum or segment. There are two ways of doing this—called proportionate stratification, and disproportionate stratification. We will illustrate, based on our example of the three age-based strata.

Total Sample Size for Proportionate Stratified Sample First, to compute the overall sample size for a proportionate stratified sample, we have to use a modified formula,

$$\left(\frac{Z}{e}\right)^2 \Sigma W_i S_i^2$$

instead of the earlier formula discussed at the beginning of this chapter. The pre-condition for using this formula is that we need to know the standard deviation (estimated) of the concerned variable for each of the strata S_1, S_2, S_3, and so on. We also have to assign a weight to each stratum, which is W_i in the formula above. W_i is generally calculated as a proportion of number of people in stratum 'i', to the number of people in all the strata. In other words,

$W_i = \left(\frac{N_i}{N}\right)$, where N_i is the population of stratum 'i', and 'N' is the total population targeted for the study.

For calculating the weights, therefore, we must have at least an estimate of the distribution of our target population among the strata. We also need S_i, the standard deviation of the variable being estimated, for each stratum. These are not always easy to get.

However, we will illustrate, assuming we are trying to gather data for a customer satisfaction study for a TV channel. Let us assume we want to know the overall customer satisfaction level among three age groups—below 25, 25 to 40, and above 40, for an entertainment channel such as Sony. We want to determine the customer satisfaction on a 7-point scale, 1 being low satisfaction level, and 7 being high satisfaction level.

Our formula for total sample size, we recall, is

$$n = \left(\frac{Z}{e}\right)^2 \Sigma\, W_i S_i^2$$

We will now assume that

Z = 1.96 (assuming 95 per cent confidence level)

e = 0.05 (tolerable error on the 7-point scale)

We will assume that for the three age-based strata, the weights and standard deviations are known or can be calculated. A rough estimate of the standard deviation 's' (overall) is given by the formula (Range ÷ 6). Range is 7 – 1 = 6 because the maximum value of the rating can be 7, and minimum can be 1.

Therefore, $\frac{\text{Range}}{6} = \frac{6}{6} = 1$

We will now assume that S_1, S_2, S_3, the standard deviations of customer satisfaction are 1.2, 0.9, and 0.7 for the three age-based strata we have described. Also, let us assume that 40 per cent of the target population of TV watchers is in the 40 plus age group, 30 per cent is in the 25–40 age group, and 30 per cent is in the below 25 age group. The weights for the age groups W_1, W_2, W_3 will then be (from the lower age group to the higher), 0.3, 0.3, and 0.4. The values are written again below:

S_1 = 1.2	W_1 = 0.3
S_2 = 0.9	W_2 = 0.3
S_3 = 0.7	W_3 = 0.4

Now, applying the formula,

$$n = \left(\frac{Z}{e}\right)^2 \Sigma W_i S_i^2, \text{ we get}$$

$$n = \left(\frac{1.96}{0.05}\right)^2 [(0.3)(1.2)^2 + (0.3)(0.9)^2 + (0.4)(0.7)^2]$$

$$= 1536\ [0.871] = 1338 \text{ (approx.)}$$

This is the total sample size required. (Note that if we had used the formula for simple random sampling discussed earlier, sample size n would have been (using $s = 1$ as estimated above) equal to 1536. So, stratified sampling has led to a smaller sample size of 1338 for the same z and e values.) To split this total sample of 1338 into proportionately stratified sub-samples, we simply use the same weights as determined earlier. Thus, the sample size for stratum 1 (below 25 age group) would be

$$1338 \times W_1 = 1338 \times 0.3 = 401$$

For stratum 2(25 – 40 age group), it would be

$$1338 \times W_2 = 1338 \times 0.3 = 401$$

For stratum 3 (above 40 age group), it would be

$$1338 \times W_3 = 1338 \times 0.4 = 536 \text{ (approx.)}$$

Thus, we would take a sample of 401, 401, and 536 from each of the three strata. The total sample size is maintained at 1338.

Disproportionate Stratified Sampling One of the keys to effective sampling is to take a sample as *large* or as *small* as required. Not too high and not too low. But in practice, we need to know the variability of the population to be able achieve an accurate sampling plan. As we know intuitively, the higher the variability among the population (of the variable we are measuring or estimating), the higher the sample size required from the population. As an illustration (though exaggerated), if we know that all the population is of exactly the same characteristics, then a sample size of 1 is enough to tell us the characteristics of the entire population. At the other extreme, if the population is extremely variable, each unit having its own different characteristics, we would need a very large sample to accurately represent the population. Most populations do not fall into extreme zones, and generally *strata* or segments consist of units that are similar to each other.

But coming back to our problem of deciding on sample sizes for stratified sampling, we would probably go for disproportionate stratified samples if the variability of the variable being estimated is different from segment to segment. If the variability had been the same, we could have taken a proportionate stratified sample. We measure variability by the standard deviation of the population stratum or segment.

The formula for the total sample size calculation is (for disproportionate sampling)

$$n = \left(\frac{Z}{e}\right)^2 (\Sigma\, W_i\, S_i)^2$$

This is slightly different from the formula used in case of proportionate stratified sampling.

To illustrate, let us use the same example of three age-based strata, and check how to use a disproportionate sample in the same.

$$n = \left(\frac{Z}{e}\right)^2 (\Sigma\, W_i\, S_i)^2$$

$$n = \left(\frac{1.96}{0.05}\right)^2 [(0.3)\,(1.2) + (0.3)\,(0.9) + (0.4)\,(0.7)]^2$$

$$= (1536)\,(0.8281)$$

$$= 1272 \text{ (approx.)}$$

Thus, we see that compared to the proportionate stratified sample, we have got a lower sample size, for the same level of tolerable error (e) and Z (1.96, 95 per cent confidence level). In general, we will note that disproportionate stratified samples tend to be more efficient (lower sample sizes are obtained), than proportionate stratified samples, because we allocate sample size according to the variability in the strata.

We have yet to allocate the sub-samples to the strata. We will now do that. The criterion for doing so would be to do it in proportion to the variation in a given stratum, compared to the total variation in all strata.

In other words,

$$n_i = \frac{(N_i\, S_i)}{(\Sigma\, N_i S_i)}\, n$$

In our three strata,

n_i = sample size for stratum 'i'

n = total sample size = 1272 (calculated above)

N_i = proportion of population belonging to stratum 'i'

S_i = standard deviation of the variable (customer satisfaction) in stratum 'i'

We have assumed

$N_1 = 0.3$	$S_1 = 1.2$
$N_2 = 0.3$	$S_2 = 0.9$
$N_3 = 0.4$	$S_3 = 0.7$

n = 1272 from our calculation

Therefore, the sample size in stratum 1 (age group below 25),

$$n_1 = \frac{(0.3)(1.2)}{(0.3)(1.2) + (0.3)(0.9) + (0.4)(0.7)}(1272)$$

$$= \frac{(0.36)(1272)}{0.91}$$

$$= 503$$

Similarly,

$$n_2 = \frac{(0.3)(0.9)}{0.91} \times 1272$$

$$= \frac{0.27}{0.91} \times 1272$$

$$= 377$$

and,

$$n_3 = \frac{(0.4)(0.7)}{0.91} \times 1272$$

$$= \frac{0.28}{0.91} \times 1272$$

$$= 391$$

Thus, the sample is divided into the three age groups in proportion to the variation in customer satisfaction, and not in proportion to the number of respondents in each stratum. For example, the below 25 segment has the largest sample size of 503, even though it has only 0.3 or 30 per cent of the population. If we had gone for proportionate stratified sampling, this segment would have got a sample size of $0.3 \times 1272 = 382$ only. This would have been under-representative for this segment.

We have discussed the pros and cons of proportionate and disproportionate sampling in these two sections. The reason for such an extensive discussion is because many of the questions about sampling efficiency get answered when we think about the need for stratification. It has also been researched and proven that if feasible, stratified sampling is the most efficient method of probabilistic sampling. That is, for a given sample size, it produces less sampling error than either simple random sampling or cluster sampling.

We now move on to a discussion of cluster sampling.

3. Cluster Sampling/Area Sampling A major difference between previously discussed methods of sampling and cluster sampling is that a group of objects/units for sampling is selected in cluster sampling. A cluster is a group of sampling units or elements, which can be identified, listed, and a sample of which can be chosen. Theoretically, a cluster could be on the basis of any criterion.

But in practice, clusters tend to be found either in terms of geographical areas, or membership of some groups such as a church, a club, or a social organisation. In practice, marketing researchers use clusters of households located in a city or block, as a first stage in sampling. When the clusters are selected on the basis of geographical area, it is also called area sampling.

If cluster sampling is only a single stage procedure, then

1. a list of all available clusters should be prepared.
2. all clusters should be numbered.
3. a sample of clusters (number to be decided by researcher) should be randomly drawn.
4. all sampling units/elements such as households in the selected clusters should be chosen to be a part of the sample.

Practically, what happens most of the time is that two or more stages of sampling takes place. Out of the clusters selected in the first stage, a sample of units (households) is generally taken, because the number of people in a cluster is usually too large for sampling purposes.

One problem with cluster sampling is that the members of a cluster tend to be similar—for example, people living in a block or neighbourhood come from the same socio-economic background; have similar tastes, buying behaviour, and so on. In general, cluster sampling is statistically inferior to simple random sampling and stratified random sampling. Its sample tends to be less representative than the other two methods. In other words, it produces more sampling error for the same sample size, when compared to the other two methods. But on the positive side, the cost of cluster sampling is also usually lower. So, the researcher may be able to justify using this technique on the grounds of low cost and convenience.

4. Systematic Sampling Systematic sampling is very similar to simple random sampling, and easier to practice. Just as we do in a simple random sample, we start with a list of all sampling units or respondents in the population. We compute the sample size required, based on a formula or by judgement. Once the sample size (n) is decided, we divide the total population into ($N \div n$) parts, where 'n' is the sample size required. From the first part of sampling units, we pick one at random. Thereafter, we pick every ($N \div n$) th item from the remaining parts.

To illustrate, say we have a population of 300 students, for some research. We need a sample of 15 out of these. The sampling fraction is 15/300 which means 1 out of every 20 students will be selected, on an average. We divide the list into 300/15 = 20 parts. Out of the first 20 students, we choose any one at random. Let us say, we choose student number 7 (all students are listed). Thereafter, we choose student numbers 7 + 20, 7 + 20 + 20, 7 + 20 + 20 + 20 and so on in a systematic sampling plan. Therefore, the selected students will be numbers 7, 27, 47, 67, 87, 107, 127, 147, 167, 187, 217, 237, 257, 277, and 297. All these 15 students will comprise our total sample for the study.

In an ordered list according to the criterion of interest, systematic sampling produces a more representative sample than simple random sampling. For example, if all students were arranged in ascending order of age, a systematic sample would produce a sample consisting of all age groups. However, a potential drawback also exists. If the list is drawn up such that every 20th student were similar on the characteristic we are estimating, either by chance or design, then systematic samples can go very wrong. So a list should be examined to see that there is no cyclicality which coincides with our sampling interval.

5. Multi-stage or Combination Sampling As the name indicates, in this type of sampling, we do not choose the final sample in one stage. We combine two or more stages, and sometimes two or more different methods of probability sampling.

We have already talked about two-stage area samples while discussing cluster sampling. Usually, multi-stage methods have to be used when doing research on a national scale. We may divide the national level target population for our survey into clusters or some such units. For example, we may divide India into 5 metro clusters, 20 class A towns, 200 class B towns, and take our first stage sample as 1 metro, 3 class A towns, and 10 class B towns, based on our sampling plan. In the second stage, we may choose a stratified sample based on household income and age of respondent. In such a case, we are using a two stage sampling plan, which is a combination of cluster sampling, and stratified random sampling.

If we go on sampling by geographical area-based clusters in all the stages, it could be a three or four stage cluster sample.

Such combination sampling plans are frequently used in many marketing research studies and National Opinion Polls. A detailed discussion of these is beyond our scope, except to say that these techniques can be used effectively to get a representative sample from diverse groups of respondents.

Non-probability Sampling Techniques

We have so far discussed probability sampling techniques. In reality, because of various difficulties involved in obtaining reliable lists of the desired target population, it is difficult to use a textbook probability sampling prescription. Therefore, some compromises could be made, or approximately probability-type of sampling procedures may be used. Some of the non-probabilistic techniques may also be used explicitly in cases where it is not feasible to use probability based methods. The major difference is that in non-probability techniques, the extent of bias in selecting a sample is not known. This makes it difficult to say anything about the representativeness or accuracy of the sample. Nevertheless, if done conscientiously, some of these are good approximations for the probability sampling techniques.

There are four major non-probability sampling techniques. These are:

1. Quota sampling
2. Judgement sampling
3. Convenience sampling
4. Snowball sampling

1. Quota Sampling The first method, quota sampling, is very similar to stratified random sampling. The first step of deciding on the strata, or segments which the population is divided into, is actually the same. The second step, of calculating a total sample size, and allocating it to the various strata, is also the same. The major difference is that, random selection of respondents is not strictly adhered to. More liberty is given to the field worker to select enough respondents to complete the segmentwise quota.

In practice, unless there are untrained field workers, or the field supervision is lax, the results produced by a quota sample could be very similar to the one produced by a stratified random sample.

But there is no guarantee that it would be similar. In practice, many researchers use quota sampling, because it saves time, compared with stratified random sampling. For example, if a household is locked, a quota sample would permit the field worker to use a substitute household in the same apartment block. But with a stratified random sample, he would be expected to make a second or third attempt at different times of the day to contact the same locked household. This would increase the time taken to complete the required 'quota'.

2. Judgement Sampling This is not practiced often, as it is difficult to justify. The method relies only on the judgement of the researcher as to who should be in the sample. It obviously suffers from a researcher bias. If a different researcher were to do the same study, he is likely to select an entirely different kind of sample.

3. Convenience Sampling This is employed many times in pre-testing of questionnaires. It involves picking any available set of respondents convenient for the researcher to use. For example, students could be used as a sample by a marketing researcher who lives in a college town. They (the students) need not be representative of the target population for the study, for the product being researched.

Other examples of convenience sampling includes on-the-street interviews, or any other meetings, or from employees of one office block or factory. Another common example of convenience sampling is the one by TV reporters who catch any person passing by and interview him on the street. Unless a sample is chosen with a known and explicit, replicable methodology, the danger is that a sampling unit may be one-of-a-kind, or non-representative of the population.

4. Snowball Sampling This technique is used when the population being sought is a small one, and chances of finding them by traditional means are low. For example, to find owners of Mercedes Benz cars in a city, we may go to one or two, and ask them if they know anyone else who owns one. Or, to find golf players, one can ask each of five golf players if they know one or two others.

One respondent being used to generate names of others is called snowballing, and it can be done again on the second set of respondents. It could also be called 'networking' to find respondents.

As a general rule, this method is useful for niche markets, but is not very reliable for estimating sampling error.

Census *Versus* Sample

It would appear from our discussion of sampling in this chapter that it is not possible to do a census in marketing research. Strictly speaking, it is possible to do one if the population size is small. For example, if 200 solar cooker owners exist in a town, it may be possible to meet all of them, if their addresses were available, or could be obtained.

In some cases, like a survey of distributors or dealers, or even industrial buyers, it may make sense to do a census if it is feasible. Particularly if opinions or buying behaviour of respondents in a small population are likely to be widely divergent.

But in most cases, if populations are reasonably large or very large, it makes little sense to do a census. One major reason is that it may simply take too long. Data may arrive too late for decision-making. Inaccuracies also are likely to exist due to the volume of data collected. We will discuss these in the next section under the subject "Sampling and Non-sampling Errors".

TYPES OF ERRORS IN MARKETING RESEARCH

Any research study has an error margin associated with it. No method is foolproof, as we will see, including a census. This is because there are two major types of errors associated with a research study. These are called:

(a) Sampling error or random error
(b) Non-sampling or Human error

Sampling Error

This is the error which occurs due to the selection of some units and non-selection of other units into the sample. It is controllable if the selection of sample is done in a random, unbiased way. In other words, if a probability sampling technique is used, it is possible to control this error. In general, this error reduces as sample size increases.

Non-sampling Error

This is the effect of various errors in doing the study, by the interviewer, data entry operator or the researcher himself. Handling a large quantity of data is not an easy job, and errors may creep in at any stage of the research. The data entry person may interchange the column of 'yes' and 'no' responses while entering or compiling data, or the interviewer may cheat by not filling up the questionnaire in the field, and instead, fudge the data. Or, the respondent may say one thing, but another may be recorded by mistake. These errors are usually a function of the sample size. That is, the larger the sample size, the larger the non-sampling error. Also, it is difficult to estimate the size of non-sampling error.

Total Error

This is the total of sampling error + non-sampling error. Out of this, the sampling error can be estimated in the case of probability samples, but not in the case of non-probability samples. Non-sampling errors can be controlled through hiring better field workers, qualified data entry persons, and good control procedures throughout the project.

One important outcome of this discussion of errors is that the total error is usually unknown. But, we may have to live with higher non-sampling error in our attempt to reduce sampling error by increasing the sample size of the study, not to mention the higher cost of a larger sample. Therefore, it is worthwhile to optimise total error by optimising the sample size, rather than going blindly for the largest posssible sample size.

SUMMARY

Sampling is the process of making a selection of sampling elements from a defined set of elements called a Population. A population is defined as per the marketing researcher's objective and research questions. Usually practical M.R. uses a multi-stage selection of sampling units, until the respondents are selected in the final stage of the sampling process.

The formulas for sample size calculation (for interval-scaled variables or for proportions) can be used as one of the inputs for deciding on the final sample size. But other considerations that the researcher must also look into include—the number of cities or centres used, the criticality of the various questions in the questionnaire (assuming that one is used), cell size during the analysis stage (this comes from subsets of the total sample, which will be separately analysed), time and budget constraints, and other issues based on the researcher's experience (for example, the chances of questionnaires getting rejected during scrutiny at the tabulation or data entry stage, which may require further additions to calculated sample sizes). The important thing to remember in sample size determination is that it is not completely a deterministic process, but involves several assumptions and judgements on the part of the researcher, as detailed in the chapter.

Probability sampling techniques are those in which the researcher knows the probability of a sampling unit getting selected in the sample. In non-probability sampling methods, no such knowledge of the selection probability is possible. The biggest problem with real-life populations is that there is rarely an accurate listing available of the sampling units or the sampling elements. This makes it difficult to use probability sampling techniques in practice. Therefore, in real life, researchers tend to use a combination of probability and non-probability techniques, or an adapted form of probability sampling techniques. The popular probability sampling techniques are simple random sampling, systematic sampling, stratified sampling, and cluster sampling. The non-probability techniques include quota sampling, judgement sampling, convenience sampling, and snowball sampling.

In general, the higher the variability in a population on the variable of interest, the higher would be the sample size needed. Disproportionately high sample size is used when the variation in a segment is high, and lower sample size can be used if the variation is low.

A sample is usually more accurate than a census of a large population due to the high non-sampling errors that occur in doing large measurements. Sampling errors are those which occur due to the selection process, and can be estimated and controlled by using probability sampling methods. Non-sampling errors include human errors in the way of asking questions (which could be inconsistent from one interviewer to the next), counting, data entry or tabulation, and so on. These cannot be accurately estimated, but they increase with sample size. Non-sampling errors can be minimised by careful selection and training of interviewers, field control procedures, and avoiding known sources of such errors. Total error in a research project is the sum of the sampling and non-sampling errors. The researcher should aim to minimise total error by balancing the sampling and non-sampling errors.

ASSIGNMENT QUESTIONS

1. What does a 100 per cent confidence level mean, in the context of sampling and sample size?
2. What is the relationship between sample size and
   (a) confidence level needed
   (b) standard deviation of the variable
   (c) tolerable error

   for a continuous variable such as income (in Rs.)? Describe the answer in three statements like the following:

   "The higher the confidence level required, the lower/higher the sample size".
3. Calculate the sample size, from the formula, for a question, which is to be answered on a 7-point scale. Assume a 95 per cent confidence level, an error tolerance of 0.15 points, and an unknown standard deviation. Use the approximate rule-of-thumb for the estimation of standard deviation.
4. In the above problem, change the tolerable error to 0.05 and calculate the sample size required. Now, increase the tolerable error to 0.30 and recalculate the sample size. What conclusions can you draw from this exercise?
5. What do 'p' and 'q' represent in the sample size formula for proportions? Is 'q' related to 'p'? If so, what is the relationship?
6. What is the range of values that p can have? Which value of p would give us the maximum sample size? Why?
7. For the same study, would you vary the sample size if the number of centres (cities) was changed from two to ten? Why or why not?
8. What do we mean by a representative sample? Discuss?
9. How do we ensure that a sample represents its population?
10. Comment on the impact of output table cell size on sample size. Explain the effect with an example.
11. What should be our approach to the sample size calculation when we have some continuous, and some dichotomous (proportion or percentage) type of variables in a single questionnaire?
12. Explain the basic difference between probability and non-probability sampling techniques.
13. Among the three techniques of simple random sampling, stratified random sampling, and cluster sampling, which produces the least sampling error for a given sample size? In other words, which technique is the most efficient?
14. Under what conditions would you use snowball sampling?
15. Can we estimate sampling error? Non-sampling error?
16. How can we reduce sampling error? Non-sampling error?

The McGraw·Hill Companies

CHAPTER 6 FIELD PROCEDURES

Learning Objectives

In this chapter, we will

- Introduce the concept of selection of centres for field work
- Explain the concept of quotas and targets on the field
- Explain how respondents are actually selected
- Explain how the field supervisor controls the field staff to minimise cheating and non-sampling error
- Describe the need for briefing of the field staff and debriefing by them

Having looked at questionnaire design and sampling methods in the two earlier chapters, we will take a brief look at field procedures in this chapter.

DESIGN OF FIELD WORK

We are dealing with primary data collection for most part of this book. Field work is a very important component of it. In India usually, field work is done physically by interviewing people at homes, offices, or on the streets. The sampling method determines which of these places will yield a sample with the required characteristics. The sampling method also dictates if a random sample has to be chosen, and by what method.

SELECTION OF CITIES/CENTRES

In actual practice, a sample of cities is usually chosen, or the client's instructions are followed (assuming a marketing research agency is doing the research on behalf of a corporate client). For

example, a client may want the four metros of Mumbai, Delhi, Kolkata and Chennai, to be covered, for strategic reasons. Since these also represent the four different regions of India (West, North, East, and South respectively), it makes sense to include all four in a national sample. In addition, a client may want Bangalore, Hyderabad, Kanpur, Bhopal, Ahmedabad, and Chandigarh to be included, because he feels a potential market exists in these cities for his brand. These may all be included due to his requirements. Minor changes could be made if the research agency feels some other centres are more likely to yield accurate estimates of whatever information the client wants to collect. For example, Pune may be used instead of Ahmedabad, if Maharashtra is an area of focus for the client.

If the client has not specified any cities, a national sample of cities may be chosen randomly or based on the research agency's experience of what would be the most representative cities for the target population. For example, if a cosmopolitan, multi-linguistic, well-educated population is required, the bigger cities may be chosen. If smaller, class A or class B towns are to be targeted, based on their population levels, a sample may be chosen from a listing of such towns according to census data or other sources.

ORGANISING FIELD WORK

Once the centres for field work are finalised, it has to be organised in each of these places. The research agency may or may not have its own offices in each of the centres. If it has an office, a field supervisor from the office is sent a written 'brief' and a copy of the questionnaire, and asked to recruit a field force and conduct a briefing for them. The written brief explains the necessary details like the client, the purpose of the study, and most importantly, the target population and how the sample is to be selected.

QUOTAS

Most large consumer marketing research studies have quotas for demographics like age, income, sex of the respondent. This is because the output has to be analysed by these characteristics. It is the job of the field supervisor to see that these quotas are achieved. In practice, these quotas are achieved by selecting residential areas whose resident profile is known, particularly the income profile. Most of the time, some extra interviews (other than the sample size planned) are conducted to achieve the required quotas in terms of age, income, and so on. This is because a few questionnaires may get rejected at a later stage, during tabulation or data entry, due to inconsistencies or incompleteness of answers.

SELECTION OF RESPONDENTS

The field supervisor actually leads the team of field workers on the field, and instructs them on how to select a household. For example, they may be told to select every third apartment in a block of 10 apartments. If the respondent is not of the required characteristics, or is not available, an alternative is given to the field worker. He may be permitted to try the neighbour's door, for example, in such a case. The field worker has a tendency, usually, to overdo things by selecting too many similar respondents from the same block, street, or area. The field supervisor has to control this tendency, because this may lead to an over-representation of one type of respondent, and under-representation of other types.

CONTROL PROCEDURES ON THE FIELD

To ensure that a field worker is doing his job, the field supervisor can randomly go back to a few addresses and talk to the respondents to ensure that they were interviewed accurately. This is known as a call-back, and is one of the most commonly used control procedures on the field. It also ensures that there is no cheating by interviewers. Otherwise, there may be an interviewer who is not honest, and may do only a few interviews, and fill up other questionnaires without doing the interviews. The call-back serves the dual purpose of minimising cheating and also verifying the accuracy of the answers by re-asking some of the important questions. Field control procedures reduce non-sampling errors.

Of course, there is a chance that the respondent may get irritated by having to answer the questions again. But an experienced field supervisor would handle the situation properly, by first explaining why he is calling back.

BRIEFING

Before the field workers are sent on the field to do interviews, they are given a thorough briefing by the field supervisor. At this time, they generally go through a couple of mock interviews to ensure that they understand the questions, the answer categories, and the sequence. The field workers can also clarify any doubts they may have regarding the sample selection process, and the quotas for income, age, or any other variables. What to do in case of contingencies is also discussed. A target for the day in terms of filled-in questionnaires is also set, for each field worker.

It is after the briefing session that the field force starts work on data collection.

DEBRIEFING

After returning from field work on Day One of the study in a given centre, there is usually a debriefing session where any problems in the field are discussed, and solutions found by the supervisor. It may also be desirable to have a debriefing session at the end of the survey in a city, to summarise the main findings, and discuss any special comments or answers given by respondents in a city. These can be noted down and sent along with the filled-in questionnaires to the research executive in-charge of the study, who may be at the head office or the branch office where the study originated.

As mentioned earlier, field work is the backbone of primary data collection. It has to be carefully planned and supervised to ensure that errors are minimised, and accuracy levels maintained.

SUMMARY

The major decisions pertaining to field work are:

1. to determine the centres where the field work has to be done, based on the target population, specific client requirements, and the researcher's own experience.
2. to organise manpower under the control of a field supervisor in each centre.

This includes selection and briefing of interviewers or field staff in every centre by the field supervisor appointed for the study in each centre. The research executive in charge of the project may also be present at the briefing, which normally includes a mock interview and a question and answer session.

The field supervisor also explains the requirements for the selection of respondents, screening of respondents, and handling of contingencies like "respondent not available" and so on, while conducting the field work. The venue for starting the field work for the following day is decided. Questionnaire forms are distributed and logistics are discussed. This forms the planning part of field work.

Actual field work is carried out under the direct supervision of the supervisor in the pre-selected areas, with a random cross-checking of the work by the supervisor. This cross-check on the field is called a 'call-back' and is done to verify that respondents were actually met and their answers recorded correctly. This reduces some non-sampling errors.

After the field work in a centre is over on day one, there is usually a debriefing session to take stock of progress in achieving the day's targeted quota, and to resolve any problems and take corrective measures. This may be continued on all days when field work is on.

There is also a debriefing note by the field supervisor at each centre along with the filled-in questionnaires, which goes to the research executive in charge. It may include a summary of observations and problems on the field in that centre. There could be 5–10% extra questionnaires filled in if it is anticipated that some questionnaires will get rejected at the time of tabulation because they are incomplete or appear to be filled inaccurately.

ASSIGNMENT QUESTIONS

1. What are likely to be the target populations and centres for a study of
   (a) South Indian preferences for a new brand of coffee?
   (b) service experiences of a business class air traveller?
   (c) satisfaction with product and service, for buyers of Toyota Qualis, one year after its launch?
   (d) satisfaction among consumers of a brand of fertiliser?
   (e) success or otherwise of a new ATM facility launched by a private consumer bank in Delhi?
   (f) usage pattern and ease of use of an Internet banking facility offered by a multinational bank in India?
2. Suppose a study requires you to interview housewives who only speak the native language of the state, in ten different linguistic states. How would you deal with the practical problem of question translation and recording of answers? Discuss.
3. What are the issues relevant to selection of appropriate field staff to interview respondents in a survey? Does it depend on the product, type of respondent, time of survey, or other factors? Explain.

4. What do we mean by control procedures on the field? What do they achieve?
5. Explain what is meant by briefing and debriefing in marketing research.

CHAPTER 7 PLANNING THE DATA ANALYSIS

Learning Objectives

In this chapter, we will

- Introduce concepts like data coding, data processing, and analysis
- Define Univariate, Bivariate, and Multivariate analysis
- Illustrate data input formats, Variable Labels, and Value Labels
- Discuss selection of an analytical technique appropriate for the research design and type of variables (metric/non-metric), and their relationships (independent, dependent, etc.)
- Introduce hypothesis testing as a concept
- Discuss manual versus computer based hypothesis testing
- Demonstrate how to do the independent sample t-test and the paired sample t-test on the SPSS package

PROCESSING OF DATA WITH COMPUTER PACKAGES

We will begin this chapter with a brief description of data processing and analysis packages for computerised analysis. We will also look into common rules for adapting data for computerised analysis, including coding. Then, we will cover some analytical approaches for univariate, bivariate and multivariate analysis. The 3 factors which determine the analytical technique to be selected for a problem will be discussed. Finally, we will introduce the concept of hypothesis testing and look at how to perform a 't' test using the computer.

STATISTICAL AND DATA PROCESSING PACKAGES

Once data has been collected in the form of filled up questionnaires, the next step is to process it. This can be done either manually or with the help of a computer. Until the late eighties, mostly manual methods were used in India for data processing. But now, it is rarely done manually. In almost all cases, the computer is used. Most students of management are familiar with simple data processing packages like Excel and FoxPro, which are essentially spreadsheets and database management packages. But for managing the types and quantum of data generated by a field survey, there is another set of packages available, and the student can choose from several which are commercially available. Most of these have been developed in the U.S., and are now available either directly from the respective company's marketing office in India, or through their dealers.

Some of these packages are called SPSS, SAS, STATISTICA, and SYSTAT. There are several others also available, but these four are among the more popular and widely available. The names of these packages are registered trademarks of these companies. Usually, the package name is an abbreviation of its function. For example, SPSS stands for Statistical Package for the Social Sciences. The older versions of these packages were MSDOS-based, while the new versions are usually WINDOWS-based. The new versions have many user-friendly features, and can be learnt fairly easily.

Types of Analysis

The above packages (or similar ones from other companies) can be used for two major types of applications in Marketing Research.

1. Data processing—General
2. Statistical analysis—Specialised (Univariate, Bivariate, and Multivariate)

Data Processing

This application is for coding and entering data for all respondents, for all questions on a questionnaire. For example, there may be a question which asks for the education level of a participant. The choices may be 12th or below, Graduate, Post-Graduate, and any other. The objective of data processing is to assign a code for each of the options—for instance, **1** for 12th or below, **2** for Graduate, **3** for Post-Graduate, and **4** for any other. Next, depending on the option ticked for each respondent, to enter the respective code against his row (usually, the data for one respondent is entered in a row assigned to him in the data set), in the column assigned to the question.

The end result of data processing for this question would be to be able to tell the researcher how many of the sample of respondents were of education level 12th or below (Code 1), how many were Graduates (Code 2), how many Post-Graduates (Code 3), and how many were in any other category (Code 4). For example, it could be that out of a sample of 500 respondents, 100 were in Code 1 category, 200 in Code 2, 150 in Code 3, and 50 in Code 4 (any other).

Similarly, all other questions on the questionnaire are processed, and totals for each category of answers can be computed. The menu commands used for such data processing are called FREQUENCIES, SUMMARY STATISTICS, DESCRIPTIVE STATISTICS, TABLES and so on. depending on the software package used.

DATA INPUT FORMAT

Most of the packages mentioned earlier have a format similar to spreadsheet packages for data entry. Readers familiar with any spreadsheet package like Excel can easily handle the data entry (input) part of these statistical packages.

The input follows a matrix format, where the variable appears on the column heading, and data for one person (respondent or record, also called a **case** in statistical terminology) is entered in one row. For example, the data for respondent no. 1 is entered in row 1. The answer given by respondent no.1 to Question 1 is entered in Row 1 and Column 1. The answer given by respondent no.1 to Question 2 is entered in Row 1 and Column 2. The input matrix looks like the following:

	Var 1	Var 2	Var 3…….	Var k
Respondent 1	*x*	*x*	*x*	*x*
Respondent 2	*x*	*x*	*x*	*x*
Respondent 3				
…				
…				
…	*x*	*x*	*x*	*x*
Respondent *n*	*x*	*x*	*x*	*x*

Coding

One limitation of doing analysis on the computer with these statistical packages is that all data must be converted into numerical form. Otherwise, it cannot be counted or manipulated for analysis. So, all data must be coded and converted to numbers, if it is non-numerical. We saw one example of coding in the previous section, where we gave numerical codes of 1, 2, 3, and 4 to the education level of the respondent. Similarly, any non-numerical data can be converted into numbers. Usually, nominal scale variables (categorical variables) need to be coded and entered into the packages.

An important aspect of coding is to remember which code stands for what. Most software packages have a facility called definition of Value Labels for each variable, which should be used to define the codes for every value of a variable. This is illustrated in a section labelled ‘value labels’ a little later.

VARIABLES AND VARIABLE LABELS

Please note that usually a question on the questionnaire represents a variable in the package. This is not always the case, because sometimes we may create more than one variables out of answers to a question. For example, it could be a ranking question which requires respondents to rank five brands 1 to 5. We may define ranking given to brand *X* as variable 10, and ranks given to it could be any number from 1 to 5. Similarly, ranking of brand *Y* could be defined as variable 11, and again, the responses could be from 1 to 5. Therefore, we may end up with five variables from that single ranking question on the questionnaire. It all depends on how we want the output to look like, and how we want to analyse it.

One very useful provision that all the packages have is the variable name. For instance, if the particular question (variable) represents the respondent's income, then the variable name can be INCOME on the column representing this variable. There is a provision to give a longer name to each variable if required (usually called Variable Label) in each one of the packages.

Variable Format

There is a provision by which the user can define in these packages the type of variable (numeric or non-numeric), and the number of digits it will have. A non-numeric variable can be defined, but no mathematical calculations can be performed with it. For a numerical variable, you can usually also define the number of decimal points (if applicable).

In SPSS, to define variable labels and formats, you can double click on the column heading of a variable, and fill up the label and format in the dialogue box which opens up.

VALUE LABELS

Sometimes, the different values taken by the variable are continuous numbers. But sometimes, they are categories. For example, income categories could be:

1. below 5,000 per month
2. 5,001 to 10,000 per month
3. 10,001 to 20,000 per month
4. more than 20,000 per month

Each of these could be coded as 1, 2, 3, or 4. While entering data into the computer, these codes are entered, depending on the response of a particular respondent. But later on, when the computer prints the output, we tend to lose track of what the codes represent. To avoid this problem, we could make use of a feature that allows us to use 'Value Labels'. We can use the feature and label 1 as 'Below Rs. 5,000 p.m.', 2 as 'Rs. 5001 to 10,000 p.m.', 3 as 'Rs. 10,001 to 20,000 p.m.', and 4 as 'More than Rs. 20,000 p.m.'. The words used in quotes are called value labels, and can be defined for each variable separately.

This process simplifies the problems while interpreting the output. The value labels are generally printed along with the codes when a table is printed involving the given variables (for example, income).

Record Number/Case Number

Every row is called a 'case' or 'record', and represents data for one respondent. In rare cases, the respondent may occupy two rows, if the number of variables is too large to be accommodated in one row. We may not encounter such cases in our examples, but these are sometimes encountered in commercial applications of marketing research. The manual for the package being used (SPSS, SAS, SYSTAT, etc.) can be referred to for an explanation of how to use two or more rows for representing a single case (respondent).

If a respondent is represented by one row, usually the row number and the serial number of respondent become identical.

In other words, the number of rows will add up to the sample size.

Missing Data

Frequently, respondents do not answer all the questions asked. This leaves some blanks on the questionnaire. There are two approaches for handling this problem.

1. Pairwise deletion: The computer can be asked to use the pairwise deletion, which means that if one respondent's data is missing for one question, then the package simply treats the sample size as one less than the given number of respondents for that question alone, and computes the information asked for. All other questions are treated as usual.
2. Listwise deletion: This instruction to the computer results in the entire row of data being deleted, even if there is one missing (blank) piece of data in the questionnaire. This may result in a large reduction in sample size, if there is a lot of missing data on different questions.

STATISTICAL ANALYSIS

We have so for discussed general data processing applications of statistical packages. But that is not all that these packages are capable of. They can help us to do a lot of statistical tests, like the Chi-squared, the *t*-test and the *F*-test. They can also be used to perform analyses such as Correlation and Regression Analysis, ANOVA or Analysis of Variance, Factor Analysis, Cluster Analysis, Discriminant Analysis, Multidimensional Scaling, Conjoint Analysis, and many other advanced statistical analyses. The packages we have mentioned (SPSS, SAS, SYSTAT) generally perform most of these analyses. In addition, the statistical packages also have varying graphical capabilities for drawing graphs.

Some of the packages require a large amount of computer memory to operate some of the advanced multivariate statistical techniques, particularly if the data size is large.

Most of the important multivariate statistical analysis techniques typically used by a marketing researcher are described in detail from the next chapter onwords. The exact commands used will vary depending on which statistical package is used by the reader. But in most of the current packages, a pull-down menu is used, and a Help feature is available online, so a user can easily perform most of these analyses if he is slightly familiar with WINDOWS operating system and general data entry into packages like EXCEL. For details, the manual for whichever package is being used should be consulted.

The chapters which follow guide even the inexperienced users with a detailed example of how to use each major statistical technique. A description of a problem is accompanied by the input data, and the exact output of the computer for the analysis being described. It is desirable for the user to have access to one of the statistical packages, which can perform these analyses, but it is possible to understand the essence of these methods even if one has no access to a computer package.

The McGraw·Hill Companies

HYPOTHESIS TESTING AND PROBABILITY VALUES (P-VALUES)

In manual forms of hypothesis testing, we generally compute the value of a statistic (the Z, the t, or the F statistic, for example), and compare it with a table value of the same statistic for a given constraint (sample size, degrees of freedom, etc.). But in the computer output for any analysis involving a statistical test, a more convenient way is to interpret the p-value printed for a particular test. For example, if we are conducting a hypothesis, we only need to decide on the confidence level (statistical) for the test before the computerised analysis. Suppose we decide that we want a confidence level of 95 per cent for the test (assume it is a t-test). Suppose now that the computer gives an output that shows the p-value as 0.067 for the t-test we requested. This value being more than 0.05 (100—confidence level of 95%), the null hypothesis cannot be rejected. If the p-value had been less than 0.05, we would have rejected the null hypothesis. But what is a null hypothesis? In general, a null hypothesis is the opposite of any statistical relationship between variables that we expect to prove. In other words, if we want to check if variables x and y are related to each other, the null hypothesis would be that there is no significant relationship between x and y.

This method of proving or disproving a hypothesis is very simple to understand and use, in the context of computers doing the testing. This is what we will use throughout this book.

APPROACHES TO ANALYSIS

What is Analysis?

Analysis of data is the process by which data is converted into useful information. Raw data as collected from questionnaires cannot be used unless it is processed in some way to make it amenable to drawing conclusions. Various techniques of data analysis are available, and it is sometimes difficult to choose one that will be the most appropriate for the research problems on hand. We have earlier mentioned the fact that analysis should be planned at the time of designing the questionnaire. This is true particularly when special kinds of analysis are needed, requiring specific forms or scale of data.

Three Types of Analysis

Broadly, we can classify analysis into three types—

1. Univariate, involving a single variable at a time,
2. Bivariate, involving two variables at a time, and
3. Multivariate, involving three or more variables simultaneously.

The choice of which of the above types of data analysis to use depends on at least three factors —1) the scale of measurement of the data, 2) the research design, and 3) assumptions about the test statistic being used, if one is used. We will briefly discuss these factors and their implications with some illustrations.

1. Scale of Data If the variables being measured are nominally scaled (involving categories such as age, income, gender, education level, etc.), many of the advanced multivariate analysis techniques

such as factor analysis, discriminant analysis, etc. cannot be used. Even common statistics such as the average (mean) or standard deviation have no meaning for nominally scaled data.

The permissible analysis is of univariate, or some specific types of bivariate analysis. For example, counting (frequencies) and percentage calculation for the nominal variable categories is possible. For the Age variable, we can say that 50% of the sample were below 25 years old, 30% between 25 and 40, and 20% over 40. This is a univariate analysis called a frequency distribution.

Bivariate analysis of certain types can also be done with nominally scaled variables. For example, the chi-squared test of association between two variables can be performed with two nominally scaled variables.This is explained in the next chapter in the section on cross-tabulation.

Ordinally scaled data also has severe limitations on the usage of multivariate statistics. Mostly, univariate or at the most bivariate analysis can be used on ordinal data. For example, a ranking of 5 brands of audio systems by a sample of consumers may produce ordinal scale data consisting of these ranks. We cannot comput an "average" rank for each brand, because averages are not meaningful for ordinal level data. But univariate analysis can be done to make statements such as "70 percent of the sample ranked Brand A (say, Aiwa) as no.1", or "20 percent of the sample ranked Brand B (say, Philips), as no.1". Similarly, numbers and percentages can be calculated for ranks 2, 3, 4 and 5.

We can also do some types of bivariate analysis such as a chi-squared test of association between say, "the brand ranked as no. 1" and say, "the income group to which the respondent belongs" (a nominal variable). This would tell us if a significant association exists between these variables.

This may be understood further by reading the section on crosstabs in the next chapter. The crosstabs in this case may look as follows—

Brand Ranked 1	*Income Grp.1*	*Inc. Grp.2*	*Inc. Grp.3*	*Inc. Grp. 4*
Brand A	x	x	x	x
Brand B	x	x	x	x
Brand C	x	x	x	x
Brand D	x	x	x	x
Brand E	x	x	x	x

The *x* values in the above table represent the number of respondents in each cell.

Nominal and ordinal scale data are also called *non-metric* data, and generally various *non-parametric tests* are used on non-metric data. Interval scaled or ratio scaled data are also called metric data, and many more statistical techniques, including univariate, bivariate and multivariate, can be used for their analysis. Part II of this book deals with many of these techniques.

2. Research Design The second determinant of the analysis technique is the Research Design. For example, whether one sample is taken or two, and whether one set of measurements is independent of the other or dependent on the other determine the analysis technique.

Let us consider an example of Attitude towards a Brand, measured from Buyers and Non-buyers of the brand. These two are independent samples, and a '*t*' test for independent samples can be used to measure if the "mean attitude" is different among the users and non-users, if the attitude is measured with an interval scale.

As an example of dependent samples, assume that a group of respondents is given a new product to try. Before and after trial, their opinion about the product is measured, using an interval scale. This is a set of dependent samples, and a different type of '*t*' test called the paired difference '*t*' test, is used in this case to find out if there is a significant difference in their opinion before and after the trial.

In the earlier example of independent samples, if there had been three means (averages) to compare instead of two (for example, Heavy Users' Attitude, Medium Users' Attitude and Non-users'Attitude), we would use a method known as ANOVA (Analysis of Variance) instead of the '*t*' test.

In general, if the number of variables associated with or causing simultaneous change in another variable are two or more, we need a multivariate technique of analysis, rather than a univariate or bivariate tehnique. This is because multivariate techniques assume that many variables simultaneously affect a given variable. For example, we can measure the simultaneous effect of advertising, pricing and sales promotion on sales through a multivariate technique called Multiple Regression.

3. Assumptions About the Test Statistic or Technique The third factor affecting the choice of analytical technique is the set of assumptions made while using a particular test statistic. For example, the independent samples '*t*' test assumes that the two populations from which the samples are drawn is independent. In addition, it assumes that the populations are normally distributed and that they have equal variances. When these assumptions are violated, the test's efficacy is reduced, or sometimes, totally lost. Another type of assumption is related to the scale of the variable. For example, chi-squared test assumes the data are nominally scaled simple counts, whereas the techniques of factor analysis and cluster analysis assume the data to be interval scaled.

Figure 7.1 lists out the various options available to the analyst who wants to do univariate or bivariate analysis. Figure 7.2 lists out a roadmap for selecting appropriate multivariate analysis techniques.

The next chapter describes how simple tabulation and crosstabulation of data can be done. These two are the most widely used analysis techniques in survey research. A detailed coverage of the non-parametric techniques mentioned on the left side of Fig. 7.1 is beyond the scope of this book. Out of these non-parametric tests, we will discuss only the chi-squared test for crosstabulations in the next chapter, because that is the most popular in practice. Readers interested in other non-parametric tests may refer to the chapter on Nonparametric Methods in "Statistics for Management" by Richard Levin and David Rubin (Prentice-Hall India), or a similar book.

For the univariate and bivariate analysis of **metric data** (interval scale or ratio scale), we use ***'t'* tests** of different types, or the **Z test**. We will illustrate the use of two types of '*t*' tests, which are shown in the right half of Fig. 7.1. These are

1. The independent sample '*t*' test and
2. The paired sample '*t*' test

These two are the most likely tests which a marketing researcher would encounter, if the data collected are metric in nature, for univariate/bivariate testing of hypotheses.

The major focus of this book will be on simple and crosstabulations for univariate and bivariate analysis (used mainly for non-metric data), and a variety of multivariate analysis techniques for special applications (using primarily metric data, with a few exceptions).

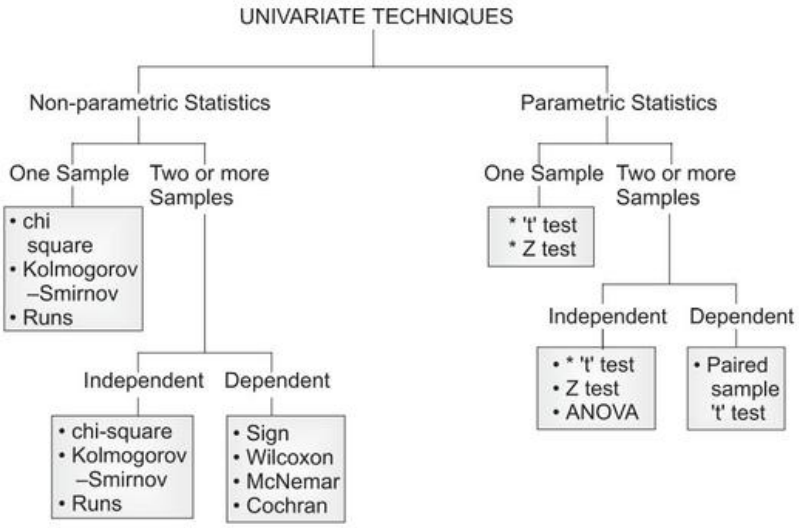


FIGURE 7.1 Univariate Techniques

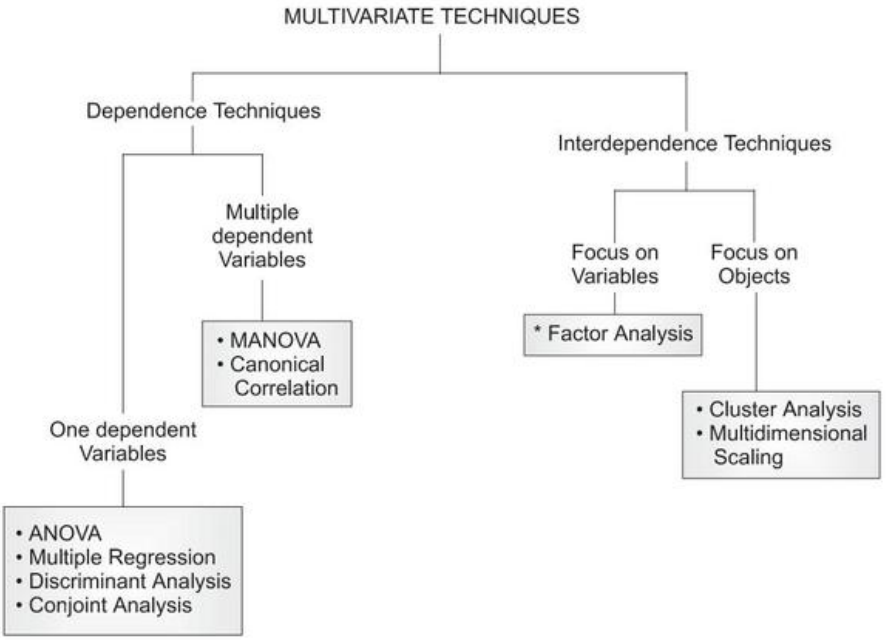


FIGURE 7.2 Multivariate Techniques

HYPOTHESIS TESTING

Before we illustrate the use of the independent sample '*t*' test and the paired sample '*t*' test, we need to know a little more about the concept of hypothesis testing. This is a brief introduction to the concept, in the context of the '*t*' test.

Suppose, as marketers of a brand of jeans, we wanted to find out whether a set of customers in Delhi and a set of customers in Mumbai thought of our brand in the same way or not. Suppose we conducted a small survey in both cities and got Ratings on an interval scale (assume it was a seven point scale with ratings 1 to 7) from our customers. We now want to do a statistical test to find out if the two sets of Ratings are "significantly different" from each other or not. We have to now set a level of "statistical significance" and select a suitable test. We also need to specify a null hypothesis.

The 'null hypothesis' represents a statement to be used to perform a statistical test to prove or to disprove (reject) the statement. In the above example, the null hypothesis for the '*t*' test would be "There is no significant difference in the ratings given by customers in Mumbai and Delhi". In other words, the null hypothesis states that the mean (average) rating from these two places is the same.

Now, we have to set a level of significance for the test. This represents the chance that we may be making a mistake of a certain type. It can also be set as (100 minus confidence level desired in the test, divided by 100). For example, if we desire that the confidence level for the test should be 95, then (100-95)/100, or .05, becomes the significance level. We can think of it as a .05 probability that we are making a certain type of error (called Type I error) in our decision-making process. Type I error is the error of rejecting the null hypothesis (wrongly, of course) when it is true. Commonly used values of significance used in marketing research are .05 (corresponding to a confidence level of 95 percent) or 0.10 (corresponding to a confidence level of 90 percent). But there is no hard and fast rule, and the significance level can be set at a different level if necessary. Let us assume that we take the conventional value of .05 for our test.

Now, a suitable test for the problem discussed above has to be found. In this case, from Fig. 8.1, we know that the independent sample '*t*' test is required.

What do we expect to achieve from this test? We will either reject the null hypothesis (that is, prove that the Delhi and Mumbai ratings are significantly different), or fail to reject it (conclude that there is no difference between the Delhi and Mumbai ratings).

1. The independent sample 't' test Let us proceed with the same example and set up an independent sample '*t*' test as discussed above, at a significance level of .05. Table 7.1 presents the input data (assumed) for the test. This assumes that 15 customers of our brand each in Mumbai and Delhi were asked to rate our brand on a 7 point scale. The responses of all the 30 customers are in column labelled 'Ratings' in the table. The column labelled City indicates the city from which the ratings came, with a code of 1 for Mumbai and 2 for Delhi.

Table 7.2 presents the output from the independent sample '*t*' test performed on the above data. The decision rule for the test (for any computerised output which gives a '*p*' value for the test) at .05 significance level is this—

If the '*p*' value is less than the significance level set up by us for the test, we reject the null hypothesis. Otherwise, we accept the null hypothesis. In this case, we find that the '*p*' value for the '*t*'

TABLE 7.1 Input Data for Independent Sample 't' test

SERIAL No.	*RATINGS*	*CITY*
1	2	1
2	3	1
3	3	1
4	4	1
5	5	1
6	4	1
7	4	1
8	5	1
9	3	1
10	4	1
11	5	1
12	4	1
13	3	1
14	3	1
15	4	1
16	3	2
17	4	2
18	5	2
19	6	2
20	5	2
21	5	2
22	5	2
23	4	2
24	3	2
25	3	2
26	5	2
27	6	2
28	6	2
29	6	2
30	5	2

TABLE 7.2 't' tests for Independent Samples of CITY

Variable	*No. of Cases*	*Mean*	*Std. Deviation*
Mumbai	15	3.7333	0.884
Delhi	15	4.7333	1.100
Mean Difference = – 1.0000			
Levene's test for Equality of Variances: F = .727 p = .401			
't' test for Equality of Means			
Variances	*t-value*	*df*	*2-tail Significance*
Equal	– 2.75	28	0.010
Unequal	– 2.75	26.76	0.011

test is .011 assuming unequal variances in two populations. This value of .011 being less than our significance level of .05, *we reject the null hypothesis* and conclude that the Ratings of Mumbai and Delhi are **different**. If the '*p*' value had been larger than .05, we would have accepted the null hypothesis that there was no difference between the two ratings.

Manual Versus Computer-based Hypothesis Testing: Please note that conventional hypothesis testing would have done a manual computation of the *t* value from the data, compared it with a value from the '*t*' tables and arrived at the same kind of conclusion that we did. The advantage of using the computer is that the test is performed by the package automatically, and we get the '*p*' value for the test in the computer output. We are going to use this approach (computerised testing) throughout this book for all the tests and analytical procedures. This removes the need for tedious manual calculations, and leaves the student to do managerial jobs like interpreting computer outputs rather than waste time in manual computation.

2. Paired Sample 't' test In some cases, we may not have independent samples, but the same sample could be used to do a research study involving two measurements. For instance, we may measure somebody's attitude towards a brand before it is advertised, and after it is advertised, to try and find out if their attitude has changed due to the ad campaign. In such cases, a paired sample '*t*' test is the appropriate statistical test.

We will illustrate using the example mentioned above. Assume that we used a sample of 18 respondents whom we asked to rate on a 10 point interval scale, their attitude towards say, Tamarind brand of garments, before and after an ad campaign was released for this brand. A rating of 1 represents "Brand is Highly Disliked" and a rating of 10 represents "Brand is Highly Liked", with other ratings having appropriate meanings.

The assumed data are in Table 7.3. The first column contains ratings given by respondents Before they saw the ad campaign, and the second column represents their ratings After they saw the ad campaign.

Table 7.4 contains the resultant computer output for a paired sample '*t*' test. Assume that we had set the significance level at .05, and that the null hypothesis is that "there is no difference in the ratings given by respondents before and after they saw the ad campaign".

TABLE 7.3 Input Data for Paired Sample '*t*' test

SERIAL No.	*BEFORE*	*AFTER*
1	3	5
2	4	6
3	2	6
4	5	7
5	3	8
6	4	4
7	5	6
8	3	7
9	4	5
10	2	4
11	2	6
12	4	7
13	1	4
14	3	6
15	6	8
16	3	4
17	2	5
18	3	6

TABLE 7.4 '*t*' Tests for Paired Samples

		Mean	*Std. Deviation*
AFTER	Ratings after Ad Campaign	5.7778	1.309
BEFORE	Ratings before Ad Campaign	3.2778	1.274

Paired Differences				
Mean Difference	Std. Deviation	t value	df	2-tail Significance
2.5000	1.295	8.19	17	0.000

The output table shows that the 2-tailed significance of the test is .000, from the last column titled "2-tail Significance". This is the '*p*' value, and it is less than the level of .05 we had set. Therefore, as per our decision rule specified in the earlier example, we have to reject the null hypothesis at a significance level of .05, and conclude that there is a significant difference in the ratings given by

respondents Before and After their exposure to the ad campaign. The mean ratings after the ad campaign is 5.7778 and before the campaign, it is 3.2778, and the difference of 2.5 is statistically significant.

Large sample sizes: If we have a sample size larger than 30 for the independent sample *'t'* test, we can use the '*Z*' test instead of the *'t'* test. The statement of null hypothesis etc. will remain the same in the case of a *'Z'* test also.

Proportions: Even though we have tested for differences in mean values of variables in this section, we could also test in the same way for differences in Proportions. The procedure is the same, and a *'Z'* test or a *'t'* test is used, depending on whether the sample size is more than or less than 30.

SUMMARY

A computer has made data processing and analysis very simple. But to make full use of the computer's power, a few preliminary steps have to be taken. One is coding of all data into numerical form so that it can be fed into a computer data file in any package we intend using for data processing and analysis. This step can be done either at the time of questionnaire design or later, depending on whether the questionnaire is structured or unstructured. If structured, it is advisable to code it before the survey.

SPSS, SAS, STATISTICA, and SYSTAT are some computer packages capable of data processing and statistical analysis too. The data entry into these packages is similar to that in an EXCEL data file, and these packages are WINDOWS based with a help manual, which should be referred to if needed. All data should be in the form of variables which must be given Variable Labels. Individual values of a variable should be given a Value Label, for which a provision exists in all these statistical packages. For example, in SPSS, double click on the column heading of a variable (column) for defining the variable and value labels.

Once data have been entered and saved as a data file, a variety of analytical techniques can be used to analyse or summarise the data. Counts (frequencies) of data can be computed, graphs or charts can be drawn, tests of varying complexity can be run, and so on. What analysis is performed depends on the nature of data and the research design. There are certain restrictions on non-metric data for which even means (averages) cannot be computed, but only counting is permitted. Mean or mode may be used as permitted. The simplest type of test on metric data is the '*t*' or '*Z*' test for small or large samples respectively. If there are two independent measurements, we can use the independent sample *t*-test to test the hypothesis that the means of the two samples are equal. This test is illustrated in the chapter with computer-based output from the SPSS package.

If there are two related samples such as before-after measurements, we use the paired sample *t*-test. This is also illustrated in the chapter. This chapter illustrates the concept of computer-based testing of hypotheses (as opposed to manual), which is the modern and easier way to do it, and one which we will use throughout this book. From this chapter onwards, it is recommended that the reader use a statistical package to try out the problems illustrated in the text on a computer, so that his learning is faster. In case access to a computer is difficult, however, the chapter can be understood from the computer outputs provided in it. Please consult any good text on business statistics for further illustrations of the

t-test. This is a basic topic in most Statistics books, and therefore only a couple of illustrations are provided in this chapter. The same is true of non-parametric tests used for non-metric data. The next chapter, however, illustrates the most commonly used non-parametric test in marketing research, called the Chi-squared test.

ASSIGNMENT QUESTIONS

1. Find out from the Internet, through a search engine, the names of statistical packages, which can perform multivariate statistical analysis. If possible, list down the features and the cost of each one of these.
2. What is the data input format for typical survey data?
3. What is coding? Why is it needed?
4. What is a variable? How is it related to a question in a survey questionnaire?
5. What is a variable label?
6. What is a value label? Explain with an example.
7. What is a 'case' or a 'record' in a survey?
8. What is missing data? How should it be handled?
9. Do we always need to do multivariate analysis of survey data? Is bivariate analysis adequate in some surveys?
10. How difficult is it to pick up multivariate analysis skills on a computer? What are the prerequisites for doing multivariate analysis involving large amounts of data?

SPSS DATA INPUT AND t-TEST COMMANDS

DATA in SPSS

When you start the SPSS program, you will get a blank screen like a blank EXCEL spreadsheet.

1. You can type in your data for the problem (or data from a survey which has to be processed) in this file. Data should be numerical (coded if nominal scale).
2. To define the data format, variable labels, and value labels for each variable, double-click on the heading of the respective column. Fill the details in the relevant boxes/cells.
3. Save this file with a FILE SAVE command.

t-tests (Independent sample and Paired sample) After the input data has been typed along with variable labels and value labels in an SPSS file, to get the *t*-test output for an independent sample *t*-test for comparing the means of two metric variables as described in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYSIS).

2. Click on COMPARE MEANS, followed by 'Independent sample *t*-test'.
3. Select the test variable for which this test is to be done, by clicking on the arrow after highlighting the appropriate variable to transfer it from left to right.
4. Select the GROUPING VARIABLE in the same way, and transfer it to the right side box. This variable defines the codes for segregating the test variable into two groups.
5. Then define the codes for the two groups by clicking on DEFINE GROUPS just below the GROUPING VARIABLE and typing in the codes (1, 2 for example, or whichever codes are used in your problem).
6. Click OK to get the output for an independent sample *t*-test.

For the paired sample *t*-test,

1. repeat step 1 above, after your data is typed and labels are defined.
2. click on COMPARE MEANS, followed by 'Paired sample *t*-test'.
3. Select two variables from the variable list appearing on the left side. These should be transferred to the box on the right by clicking on the arrow.
4. Click OK to get the desired output.

Note: In both these tests, you can set a confidence level by clicking on OPTIONS in the dialogue box, and choosing the desired confidence level for the *t*-test. The default value would generally be 95% if you do not choose any.

INTEGRATED CASE STUDIES FOR PART 1

CASE STUDY 1

Crocin*

BACKGROUND

Crocin—the safe drug for all. Our group chose Crocin, a product of SmithKline Beecham as a subject for market research. Of late, Crocin, an Over-the-Counter (OTC) drug, has been advertising on television. The company claims that sales of Crocin have increased by 10% due to the advertising. On the other hand, there have also been concerns among some that advertisements for drugs such as Crocin have a tendency to promote self-medication, and this is a cause for worry, especially among the medical community. There have also been concerns, that such commercials may affect the 'prescription style' of doctors. The general feeling is that doctors have, in fact, stopped prescribing Crocin after the widespread airing of the commercial.

Keeping in mind these diverse views, a market research was conducted in order to get a fair idea about the general effect that the commercials for Crocin has had on its sales. This is the primary objective of this research.

Research Objectives

The objectives of our study are as follows:

- The primary objective of our research was to study the impact of Crocin commercials on the sale of Crocin. The company had claimed that the sales of Crocin had increased by l0% after the advertisement. We wanted to test this claim at least within our sample in the Harihar and Davangere region.
- Another major focus of the study was to glean if the commercial for Crocin had in any way affected the consumer and/ or the influencer (i.e. the doctor) and/or the distribution channel (i.e. the chemist)
- We wanted to study if the commercial for Crocin had in any way encouraged self-medication.

Methodology

The basic methodology that we followed was the questionnaire method. To serve our purposes, we designed three separate questionnaires for the consumer, the retailer (i.e. the chemist), and the doctor. Each questionnaire was designed in a manner so as to gain the maximum relevant information from the respondent taking minimum of their time. The questionnaires that were designed are presented below:

* Prepared by Aditi Sood, Mukesh, Vikram Rathi, and Vinita Nair.

Questionnaire for the Consumer

1. If you have a fever what would you take?
   - ☐ Crocin
   - ☐ Calpol
   - ☐ Paracetamol (generic)
   - ☐ Others (please specify) ...
2. Who influences your choice of medication?
   - ☐ Doctor
   - ☐ Chemist
   - ☐ Family/friends
   - ☐ Self-medication
3. What influences you to take self-medication?
   - ☐ Past experience
   - ☐ Advertisement
   - ☐ Lack of time to visit a doctor
   - ☐ Others (please specify) ...
4. Have you seen the advertisement for Crocin?
   - ☐ Yes
   - ☐ No
5. If yes, where?
   - ☐ Newspapers/Magazines
   - ☐ Chemist shop
   - ☐ Television
   - ☐ Others {please specify}
6. On the basis of the advertisement have you prescribed Crocin to your family/friends?
   - ☐ Yes
   - ☐ No
7. Are you aware that Crocin is available in syrup form specifica1ly for children?
   - ☐ Yes
   - ☐ No
8. If yes, how did you become aware of it? (Please specify)

   .....

Questionnaire for Doctors

1. What drug do you prefer to prescribe for fever?
   - ☐ Crocin
   - ☐ Calpol
   - ☐ Others (please specify) ...

2. Reasons for prescribing the above drug ... (tick as many as applicable)
   - ☐ Medical representative
   - ☐ Experience
   - ☐ Minimum side-effect
   - ☐ Others
3. Have you seen the TV commercial for Crocin ?
   - ☐ Yes
   - ☐ No
4. If yes, what are your comments on the commercia1?
5. In your opinion are commercials regarding drugs appropriate for the general public?
6. Has the commercial in any way changed your attitude towards prescribing Crocin for your patients?

Questionnaire for Retailer

1. Which is the largest selling brand of paracetamol?
   - ☐ Crocin
   - ☐ Calpol
   - ☐ Others (please specify) ...
2. Have you seen the Crocin commercial?
   - ☐ Yes
   - ☐ No
3. If yes, how has the commercial affected the sale?
   - ☐ Increased sales
   - ☐ Decreased sales
   - ☐ No change in sales
4. Among the customers of Crocin, what percentage carries a prescription?
   - ☐ less than 10%
   - ☐ 10–50 %
   - ☐ more than 50%

Apart from the questionnaire method, we also tried to gather primary information through informal conversations with the respondents, especially the doctors and the chemists.

The market research was carried out in the towns of Harihar and Davangere. We attempted to cover as wide an area as possible, and we were successful to a great extent.

Sample Composition

Our sample comprised three categories of respondents:

1. Doctors
2. Chemists/Retailers
3. Consumers

Sample size: Our sample size for survey comprised 35 persons, the break up being as under:

Doctors (from Davangere)—15

Chemists (from Davangere and Harihar)—10

Consumers (from Davangere)—10

Experiences and Difficulties Faced in the Field

The market research that we conducted provided us an excellent opportunity to implement all that we had learnt in our class room sessions in the practical outfield.

After having designed the questionnaire and having made a few minor changes, we were ready to go out into the field to carry out the survey. As mentioned earlier, we covered the Harihar and Davangere areas during our research. We found that the respondents were very forthcoming and cooperated with us for the most part.

Nonetheless, we did face a few minor problems while conducting the survey. These are as under:

- LANGUAGE PROBLEM: One of the main problems faced was the language. Communication posed to be a problem in some cases, we were mistaken to be sellers of Crocin! However, we were able to overcome the communication problem to some extent as one of the group members was relatively conversant in Kannada.
- SAMPLE: While conducting the survey among consumers, we found that the locality that we were visiting was mainly inhabited by doctors. This maybe because Davangere is home to a very large medical community. However, this problem was solved by conducting the survey in a totally different area of Davangere.
- LIMITATION: In the opinion of our group, the sample taken for the study is too small and not representative enough to draw very relevant conclusions that may hold good in the larger scheme of things. This, we feel, is a major limitation of the study.
- NEW DOSAGE FORM: SmithKline Beecham has introduced a new dosage form for paracetomol (Crocin—1000 mg tablet) used in cases of Osteo Arthiritis. This dosage form is a prescription drug. However, our focus was only on the 500 mg dosage of Crocin which is a Non-Steroidal Anti Inflammatory Drug (NSAID), used as an antipyretic and in some cases an analgesic. This limited the scope of our study.

Analysis

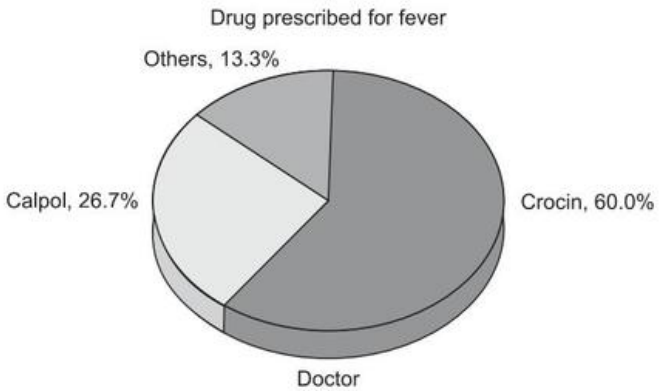
Our survey comprises three different questionnaires for the three different categories of respondents. Therefore, our analysis is also divided into three distinct parts, which we briefly present here:

Doctor's Questionnaire While analysing the data collected from the doctors, we basically identified three important questions:

1. What is the drug you prescribe in case of fever?
2. What is the reason for the prescription of the above drug?
3. Have you seen the commercial for Crocin?

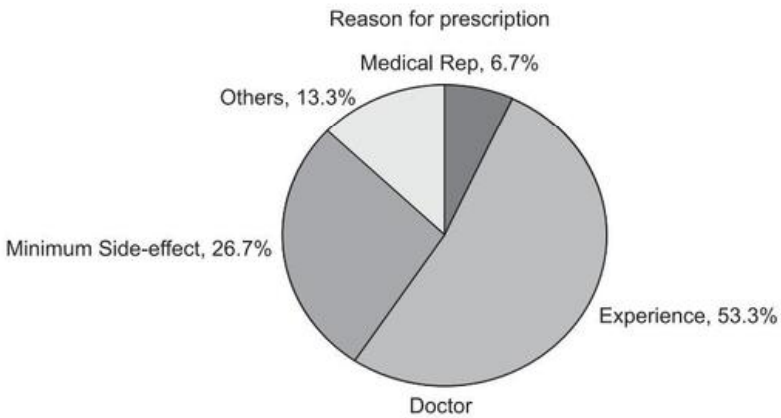
In response to the first question, 9 out of 15 doctors interviewed stated that they prescribed Crocin in the case of fever. 4 of the 15 doctors prescribe Calpol for fever, while the remaining 2 doctors prescribe any other formulation such as Paracetamol.

Thus, it is clear that in the case of our sample, Crocin is the preferred drug prescribed by 60% of the doctors surveyed by us.



In response to the second question, regarding the reason for the prescription of the above drug, 8 out of 15 doctors interviewed stated **past experience** as the major reason for the prescription of the afore mentioned drug. This was followed by the reason of **minimum side effect** caused by the prescribed drug, with 4 of 15 doctors choosing this alternative. Surprisingly, only 1 of the respondents chose **medical representative** as the reason for prescribing the drug. This is surprising because it is common knowledge that medical representatives do in fact play a very major role in promoting any drug and persuading the doctor to prescribe it. Our sample, however, did not reflect this trend. Only 2 of the 15 doctors interviewed gave other reasons for the prescription of the drug.

Thus, past experience is the overwhelming reason for the prescription of the preferred drug in the case of the sample covered, with 53.3% of respondents choosing this option.



The third important question in the case of the doctor's questionnaire, was if they had seen the commercial for Crocin. In response to this question, we were very surprised to find that only 4 of the 15 doctors interviewed had actually seen the commercial for Crocin. The remaining 11 doctors had never seen the commercial for Crocin on TV.

Therefore, in the case of our sample 60% of the doctors interviewed had not seen the commercial for Crocin.

Major Findings (Doctor's Questionnaire)

From an analysis of the data pertaining to the Doctor's questionnaire, we arrived at the following conclusions:

- As is evident from the above analysis, Crocin is definitely the preferred drug prescribed by doctors surveyed in our sample study. 60% prescribed Crocin, which is followed by Calpol, prescribed by 26.7% of the doctors interviewed.
- Crocin being the preferred drug for prescription, the main reason for this is overwhelmingly past experience, with 53.3% of our sample choosing this option.
- It was found that the Crocin commercial had not had much of an impact on the 'prescription style' of doctors. We could infer this from the informal conversations that we were able to conduct with the doctors interviewed.
- As was expected, the survey showed that doctors were against self-medication. This was gathered from the questionnaire as well as from the conversations with the doctors.

Recommendations (Doctor's Questionnaire)

On the basis of the responses and an analysis of the same, we would make the following recommendations and suggestions:

- It was observed from the responses, that only 6.7% of the doctors interviewed stated medical representatives as the main reason for prescribing Crocin. Therefore, it is suggested that promotion by medical representatives must be increased to a great extent.

- It is also observed from the market research that the brand awareness of Crocin is extremely high. Also, most doctors prescribe Crocin due to their past experience. Thus, it may be recommended that Crocin may further leverage its brand name in order to further enhance its current market.

Consumer's Questionnaire

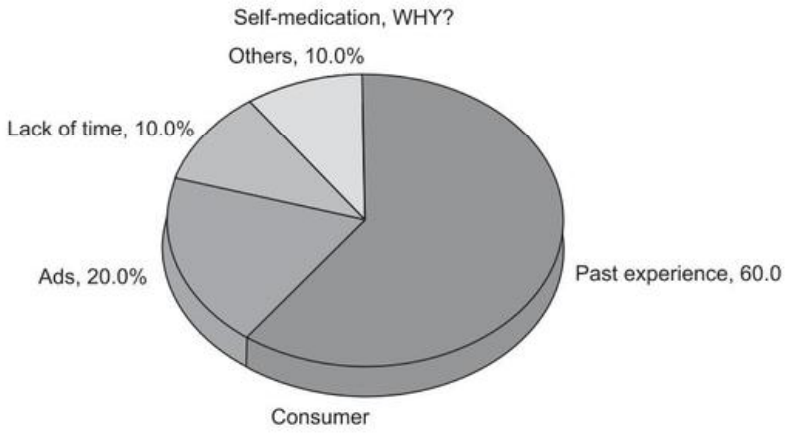
In the case of the consumer's questionnaire, we took into consideration four main questions:

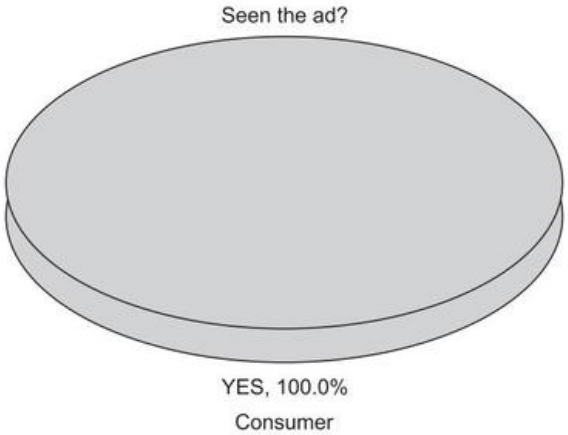
1. What influences you to take self-medication?
2. Have you seen the Crocin commercial?
3. If yes, where have you seen the commercial?
4. On the basis of the advertisement, have you prescribed Crocin to your family or friends?

In reply to the question on the main influencer for the consumer to take self-medication, 6 of 10 respondents stated that **past experience** was the major reason for self-medication. This meant that after having tried a certain drug which had been prescribed by a doctor, the respondent then simply used that particular drug in the future without reverting back to the doctor. Only 2 of 10 respondents stated that they had resorted to self-medication on the basis of the advertisement. 1 respondent stated the **lack of time to visit a doctor** as the reason for self medication, while the remaining 1 respondent gave **other reasons**.

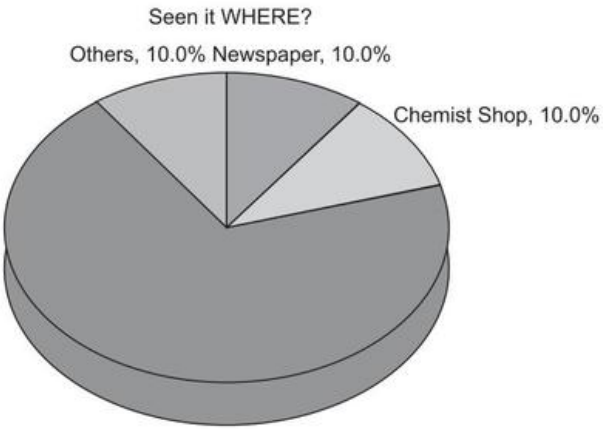
On being asked if they had seen the Crocin commercial, every single respondent replied in the affirmative. All 10 consumers interviewed had seen the commercial for Crocin, showing that the commercial had a very effective penetration in case of the consumer.

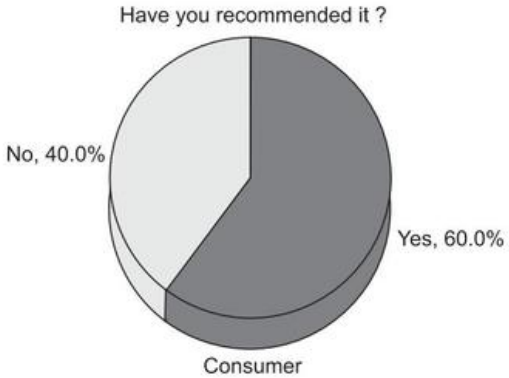
On being asked where they had seen the advertisement, 7 out of 10 consumers responded that they had seen the advertisement on television, once again reiterating the fact that the **television** commercial for Crocin had had a very good penetration in the case of the consumer. Only one respondent each stated that they had seen the advertisement in the **print media,** in **chemist shops,** and from **other sources** respectively.





Finally, on being asked if they had **prescribed Crocin** to their relatives or friends, in 6 cases the respondent, replied in the affirmative. Only in 4 cases the consumer did not recommend Crocin to relatives and friends. This shows that the reliability perception and brand name factor is of importance in the case of the consumer. This emphasises the fact that Crocin would stand to gain if it further exploited its strong brand image. A substantial 60% of respondents had recommended the drug to friends and relatives.





Major Findings (Consumer's Questionnaire)

On the basis of the analysis of the questionnaires filled in by the consumers, we may conclude the following:

- The preferred drug for fever is Crocin. Most consumers covered in our survey stated Crocin as the preferred drug in case of fever, thus showing that Crocin has a major advantage of being the preferred drug for prescription by doctors as well as the drug of choice for the consumers.

- Most consumers are influenced by doctors. And as seen earlier Crocin is the preferred drug for prescription. Thus, this finding follows from the previous ones.
- 50% take self-medication from past experience and 20% are influenced by ads.
- Most of them have seen the TV commercial and have recommended Crocin to others.

Recommendations (Consumer's Questionnaire)

The major recommendations that can be made on the basis of the responses obtained from the consumer's questionnaire are as follows:

- SmithKline Beecham should continue to advertise for Crocin, as it is found from our study that most consumers have recommended Crocin to others on the basis of the advertisement, and 20%of the consumers are influenced by the commercial to take self-medication.
- The company must also try to promote prescription sales in order to increase sales. This is evident as Crocin is the preferred drug for prescription and therefore sales can be raised by promoting prescription sales.

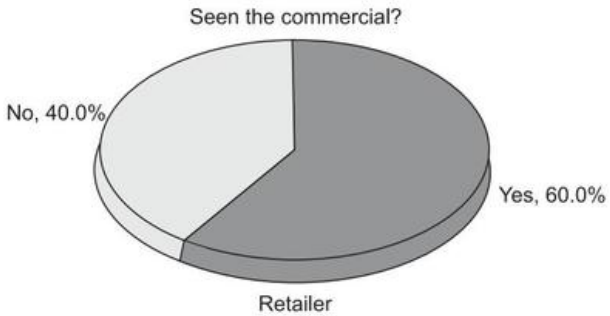
Retailer's Questionnaire

The first question that we asked in the case of the retailers is—what is the largest selling brand of paracetamol? In response, 70% of retailers responded that Crocin was the largest selling brand, while 30% responded that Calpol sold the most.

In the case of the questionnaire for retailers, we identified three basic questions for our analysis. These are:

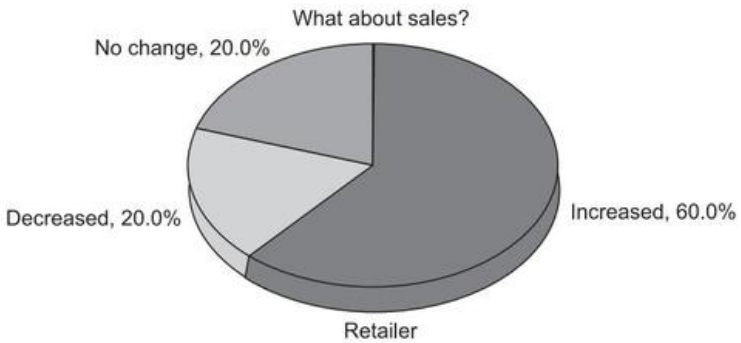
1. Have you seen the TV commercial for Crocin?
2. Has there been any change in the sales of Crocin after the commercial?
3. What is the percentage of customers for Crocin who carry a prescription?

In response to the first question, we found that 6 of 10 chemists had **seen the commercial** for Crocin, while only 4 responded that they had not seen the commercial under question. **Thus,** a **majority of 60% of the chemists had been exposed to the commercial.**



The next question was aimed at our primary objective of finding out the **impact of the commercial for Crocin on its sales.** 80% of the retailers responded that the sales of Crocin had **increased** by 12%–20% after the advertisement, thus showing that the advertisement had a major positive impact on the sales of Crocin in the case of our sample. Only one retailer stated that the sales of Crocin had in fact **reduced** after the advertisement. One respondent stated that there had been no **change** in sales after the advertisement.

Thus, it may be concluded that the commercial for Crocin has had a very positive impact on its sales.



The next question of importance was to ascertain the percentage of customers of Crocin who carry a prescription. We found that majority of retailers stating that less than 10% of customers for Crocin carry a prescription. The remaining 20% retailers stated that 10%–50% of the customers for Crocin carry a prescription.

Thus, it is evident that the prescription sales of Crocin are not very high.

Major Findings (Retailer's Questionnaire)

On the basis of the data gathered from the retailer's questionnaire, we may draw the following conclusions:

- The largest selling brand of paracetamol is Crocin. This is evident from the fact that 70% of the retailers stated that Crocin was the largest selling brand followed by Calpol.
- Sales have increased after the TV ads, which is clear from the responses of the retailers. Once again, this emphasises the positive impact that the advertisement has had in the case of the consumers.
- Majority of its sales is OTC, as is evident from the fact that most of the customers of Crocin do not carry a prescription.

Recommendations (Retailer's Questionnaire)

Till recently, Crocin was a prescription drug. It is observed from our questionnaire that Crocin is still the preferred drug for prescription by doctors. Through our informal conversations with retailers, we were able to infer that doctors had in fact reduced the prescription of Crocin after the commercial. Therefore, it is evident that if Crocin were to revert back to being a prescription drug, it would be able to recapture that part of the market that it has currently lost out to its closest rival, Calpol which is a prescription drug.

Thus, we recommend that Crocin may revert back to being a prescription drug so as to recapture the market that it has lost out to Calpol.

CASE STUDY 2

Detergents*

BACKGROUND

This project was done in Harihar and Davangere. The data are real, but they are not generalisable to other towns and cities in India.

Research Objectives

To find out:

- which is the most commonly used detergent in the market.
- what influences people to buy a particular brand?
- what is the penetration level of Surf in the market?
- to identify customer needs.

Methodology

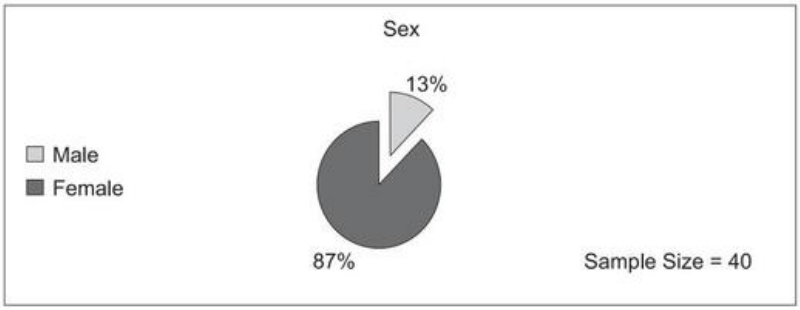
- Research objectives were clearly stated before designing the questionnaire.
- Information was collected from a sample size of 40 respondents, which included both males and females. Demographic details are provided later in the report.
- Information obtained from the respondents was analysed and interpreted with the help of the SPSS software.
- Simple tabulations (annexed in the report) were calculated for each question and cross-tabs were provided as and where necessary.
- Findings revealed by the tabulations were listed in a summarised form as **Recommended Actions.**

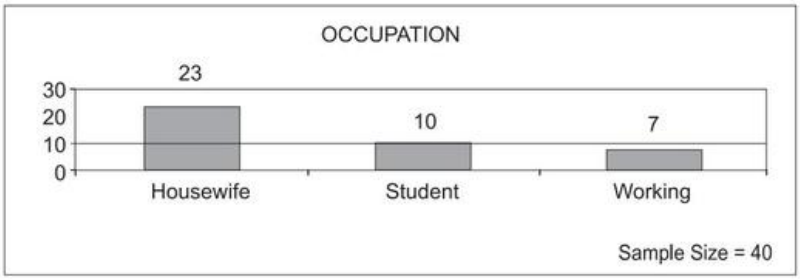
* Prepared by Harakuni Rajiv, Jitender Singh, Siddharth Agarwala, and Vidya Thekke Cherupilli under the supervision of Dr R Nargundkar, faculty member at Kirloskar Institute of Advanced Management Studies. This is a part of a course project for the Marketing Research course.

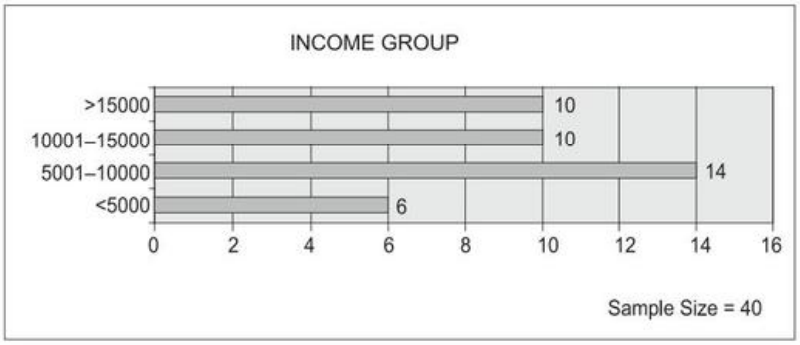
Sample Composition

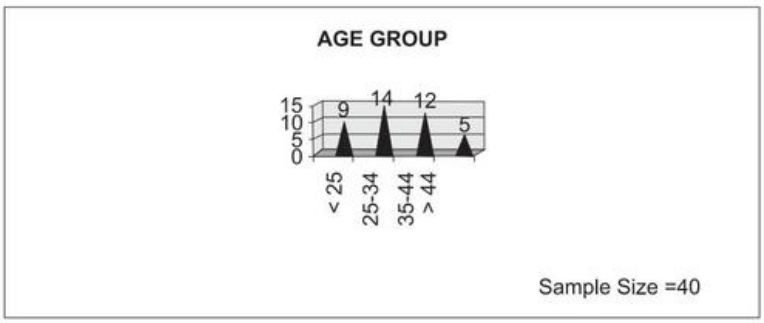
In all, our group members as a part of our survey visited 52 households. 12 of them revealed that they were entirely dependent on local washermen or launderettes. Therefore these respondents were not considered for answering the questionnaire. The remaining **40** thereby formed the **sample size** of our survey.

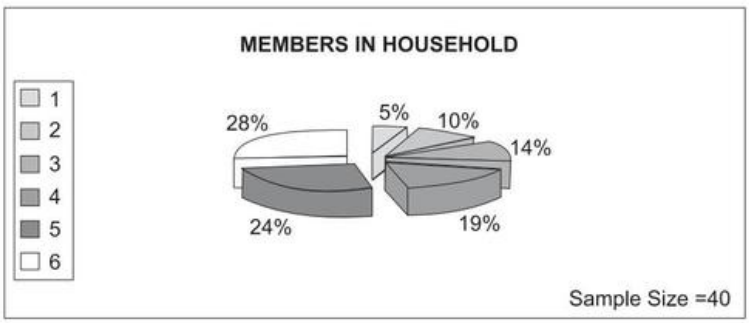
The respondents were from Davangere district (Local residential areas of Davangere and MKL Colony). The following charts and graphs depict the distributions of the people interviewed.

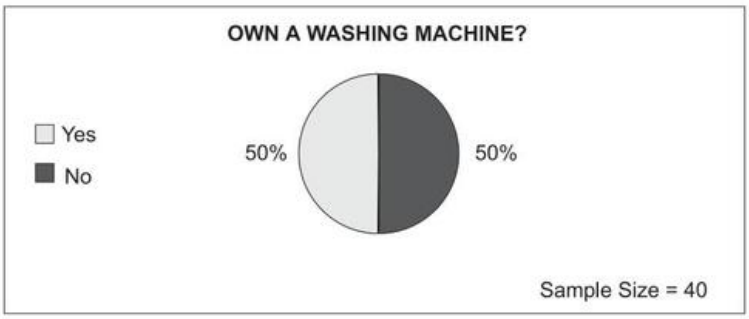


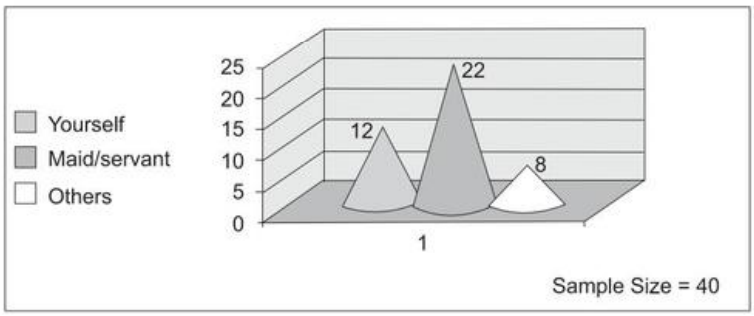





AGE GROUP
15
10
5
0
9
14
12
5
< 25
25-34
35-44
> 44
Sample Size =40


MEMBERS IN HOUSEHOLD
1
2
3
4
5
6
28%
5%
10%
14%
24%
19%
Sample Size =40


OWN A WASHING MACHINE?
Yes
No
50%
50%
Sample Size = 40


Yourself
Maid/servant
Others
25
20
15
10
5
0
12
22
8
1
Sample Size = 40

Experiences/Difficulties Encountered in the Field

The survey conducted as a part of our curriculum provided loads of experiences to each member of the group. However the fact that some difficulties were encountered while performing this task cannot be overlooked. These are listed below:

Difficulties:

- Our group consisted of four members. We were divided into two subgroups of two each. Since we had only one Kannada speaking member in the group, one of the subgroups had problems communicating with respondents in some areas of Harihar and Davangere.
- Not much of importance was attached to this task by some of the households. We were greeted with either reluctance or nonchalance. In some cases, we were given contradictory answers, which seemed to show that they were trying to get over with the interview as soon as possible.

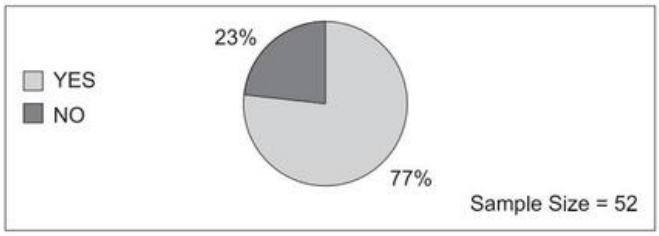
Experiences:

- We got an understanding of how the door-to-door salesmen might be feeling when they are turned away from the gate itself while marketing their products.
- The exercise provided us with first-hand knowledge of how real life surveys were conducted by organisations involved in market research.
- It gave us a hands-on experience on understanding the mentality of consumers varying from rural areas to urban ones. It also gave us an understanding as to the effect of income and personal experiences while indulging in a purchase.
- We also found that in many cases, brand loyalty exists irrespective of income if the product satisfies the desired needs of the consumer. For example, Rin being used by families of high income while Surf (which is comparatively costlier) was used by some of the households falling under the lower income category.
- In many cases it was felt that when a consumer was confused as to a particular response, he or she usually responded with the same answer that the surveyor hinted at.
- Although it was thought that only women would be enthusiastic participants, due to the product being a woman-oriented one, to the great surprise and pleasure of the surveyors it was found that men were equally knowledgeable and willing to answer the questionnaire.

Findings of our Survey

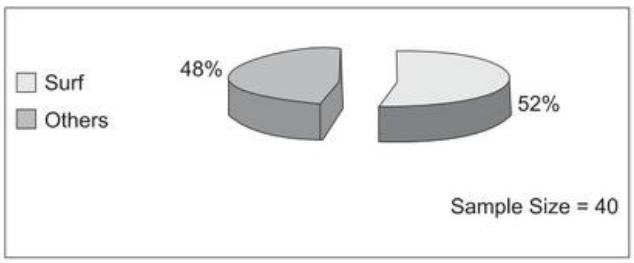
(Simple Tabulation)

1. **People who wash clothes at home.**



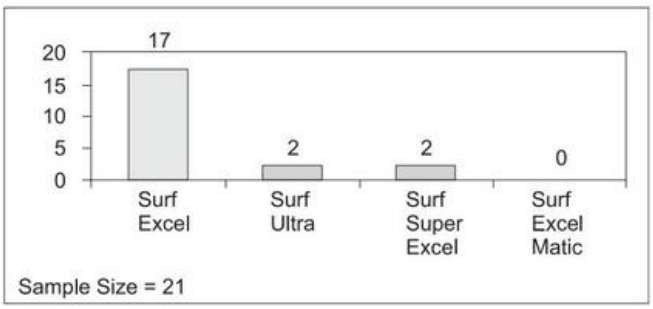
As a part of our survey, we visited 52 houses. It was found that 12 households gave all their clothes to launderettes, while 40 households washed their clothes at home. Since the objective of our survey was to find out which detergent is popular in the households, we did not take into consideration the 12 who depended entirely on launderettes.

2. People using Surf.



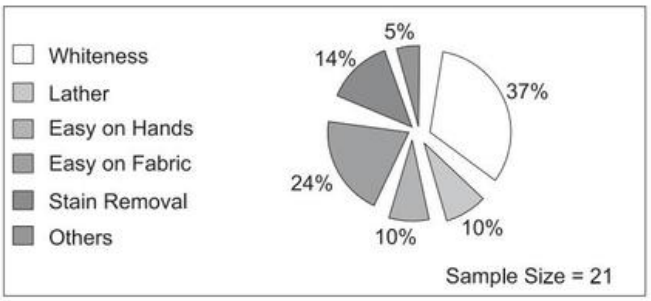
Of the 40 people interviewed, we found that 21 households used Surf, while 19 of them washed their clothes with other detergents. This is a clear indicator of the popularity and the penetration of this particular brand in the consumer's mind and in the market.

3. Sub-brands of Surf used.



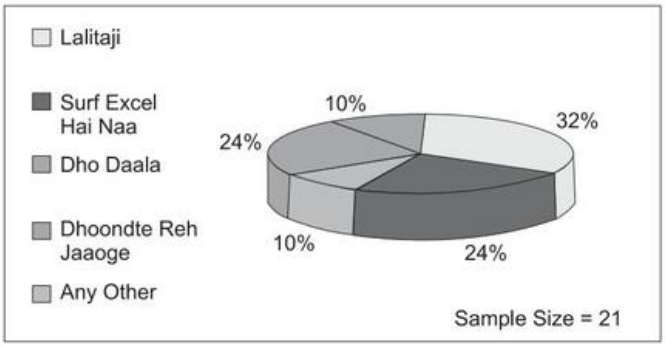
Of the 21 consumers using Surf, it was found that **Surf Excel** as a sub-brand was the most commonly used, with 17 consumers stating it as their preference. This was followed by **Surf Ultra** and **Surf Super Excel** with 2 consumers each. However, no users could be detected for **Surf Excel Matic**.

4. Influential factors while buying Surf.



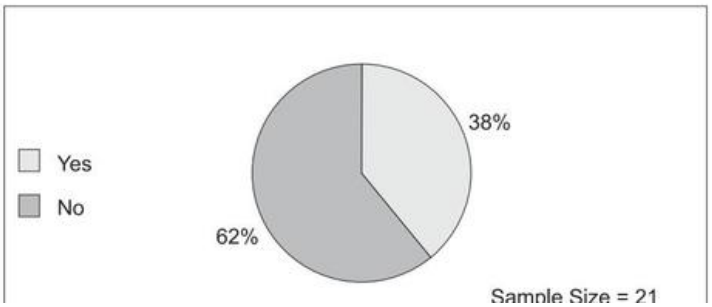
As indicated above, **whiteness** that the detergent provides, say 8 of the consumers, is one of the most potent influences while buying the detergent. The second most important influence is the fact that it is **easy on the fabric**, say 5 of them. Other influential factors are its **gentleness on hands** and its **good stain removing capacity** (*Daag dhoondte reh jaaoge*).

5. **Consumer awareness with respect to the advertising campaign of Surf.**



Of the consumers surveyed, awareness with respect to advertising by Surf was cent per cent-that is, all consumers using Surf were aware of its promotional campaigns and all had seen Surf ads at one point of time or the other. Of the different types of ads aired by the media, the **Lalitaji** ad held the greatest retention power and linking, with 7 out of the 21 consumers liking it the most, followed by the ad for Surf Excel and *Dhoondthe Reh Jaaoge*, with a fan following of 5 consumers each.

6. **Persuasive powers of various schemes.**



Schemes, which are launched by Surf to promote sales, are generally not THE major criteria when the consumer goes in for a purchase. This is also reflected by the survey in which 13 out of 21 of the consumers supported the fact. Only 8 were those who were affected by the schemes propagated by Surf.

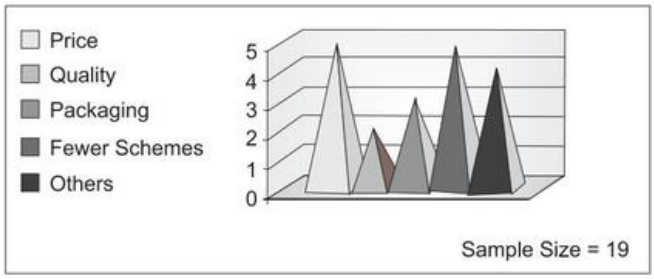
7. Suggestions provided by the consumers.

Suggestions for Improvements	*No.*
Different quantities available	2
Style of packaging	4
More schemes	5
Price	5
Any other	3
No comment	2
Total	**21**

The following changes were suggested in the Any other category:

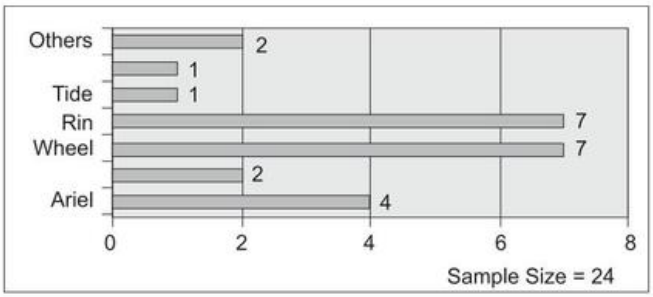
- Demands for a measuring scale so as to avoid wastage of powder
- Change in the colour of the detergent powder.
- Fragrance of the detergent.

8. Reasons for not using Surf.



Price of the detergent and association with fewer schemes were the two primary reasons for which consumers preferred other brands to Surf. Price was a factor for users of cheaper washing powders such as Nirma, Rin, and Wheel. Users of Henko, Tide and Ariel insisted that the quality of their detergent was superior to that of Surf.

9. Detergents (other than Surf) frequently used by consumers.



Amongst many existing brands available (excluding Surf) in the market, the most frequently used ones are **Rin** and **Wheel** followed by **Ariel** with **others** (Local) constituting the rest of the market. In the chart indicated above, 19 were nonusers of Surf, while 5 of them also preferred an additional detergent besides Surf.

10. Major influencers while making a purchase.

Factors	*No.*
Friends	3
Neighbors	4
Advertisements	13
Self-experience	19
Others	1
Total	**40**

While conducting the survey, personal experience of using the product along with many others over a period was major influence while indulging in the purchase. Apart from this, effective advertising was a close runner–up and was largely responsible in influencing people while buying their preferred brand.

11. Quantity usually purchased.

Quantity	*No.*
Less than 1 kg.	10
1 – 2 kg.	20
2 – 3 kg.	6
3 – 4 kg.	3
More than 4 kg.	1
Total	**40**

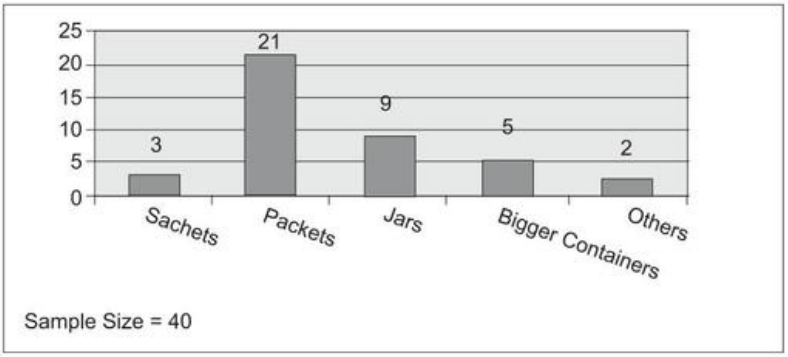
As is predictable, due to the fact that the survey was done in an area which was a middle – class one, the housewives usually went in for the 1–2 kg pack and the frequency of purchase was once in a month, which is depicted in the chart above.

12. Frequency of purchase.

Frequency	*No.*
Once a week	3
Once a fortnight	11
Once a month	23
Once in two months	3
Total	**40**

Usually households preferred to buy their stock of detergent once in a month, as is mostly the case with all stock being ordered along with the ration that comes monthly. But still many households also buy it fortnightly.

13. Packaging Preference.



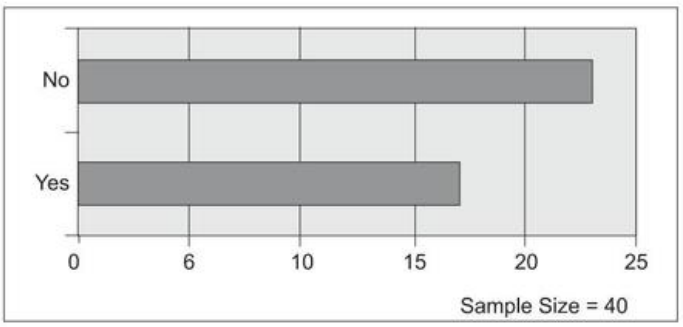
Packets (500 gm., 1 kg., 2 kg.) are the outright winners in this section with more than 50% consumers in this category preferring this particular style of packaging. However, Jars were also preferred because of their multi-utility purpose after using the primary product.

14. Alternative brand of detergent.

First choice		*Second Choice*	
Brands	*No.*	*Brands*	*No.*
Rin	9	Rin	11
Wheel	8	Wheel	12
Surf	10	Surf	5
Tide	1	Tide	6
Ariel	7	Ariel	3
Henko	4	Henko	1
Nirma	1	Nirma	2
Total	40	Total	40

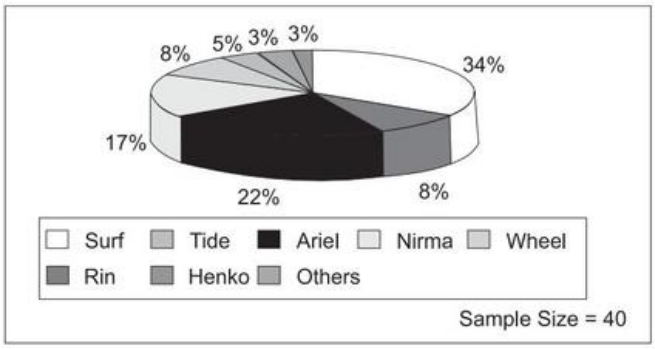
There were 10 non-users of Surf who preferred it as their first choice of purchase in case of non-availability of their preferred brand. Users of Surf voted for Ariel, Rin, and Wheel as their first choice, given the same situation.

15. Stock of detergents.



More than 50% of the households did not keep a stock of detergents at home and resorted to purchase only when the need arose.

16. Preferred detergent amongst acquaintances of consumers.



The general impression that we get after conducting the survey is that Surf rules the market because it was revealed that amongst the acquaintances also Surf was the most popular brand followed by Ariel and Nirma.

Summary of Findings

(Based on Simple & Cross-tabulations)

- In Q. No.7, eight respondents stated that they would like more schemes to be associated with Surf. However, when they were asked that what change would they suggest in their detergent (Q. No.8), only 5 of them suggested more schemes.
- Surf Excel (17/21) is preferred by the consumers because of its extraordinary whiteness (8/21) and the fact that it is easy on the fabric (5/21).
- When it comes to housewives the verdict is almost equal with 12 saying 'Yes' and 11 saying that they do not use Surf but when it comes to students, Surf is the clear winner with 6 out of 7 favoring the product.
- Of the 40 consumers surveyed, 21 used Surf and of those 21, 19 were women as Surf is more a product that homemakers use. Of the 19 non-users, 16 again were women with the rest being men who had genuine knowledge about the product and who had used it at one moment of time or another.
- Of the users of Surf, all of them were more or less equally distributed when categorised according to the income group with the higher income group categories preferring Surf a little more as Surf is costlier than most of the other brands (13/21).
- An interesting fact is revealed, 4 users stated that some of their clothes were washed either by themselves or by their maids, however the expensive clothes were given to launderettes.

  We also find that the trend of people who are in the different categories is almost the same with almost an equal number in each category.

- One interesting observation may be possible. It is seen that the less than 25 age group of users are more inclined to use Surf and as the age group increases the number of users decrease, this may be due to the new positioning that Surf is using where it is targeting the younger generation too, through its advertisements.
- Surf is popular with acquaintances of both the users and the non-users. In the acquaintances of users section Ariel follows (5/21) while in the alternative category Nirma (6/19) and Ariel (4/19) are preferred widely.

Recommendations

- Of the sub-brands, Surf Excel was the most recognised one, so the company ought to take some measures to make the consumer aware about other sub-brands.
- If possible, pricing should be reviewed, with many consumers citing it as a negative factor.
- Surf being viewed as a premium product could come up with a lower priced sub-brand for more rural market penetration to compete with Wheel, Nirma, and so on
- More schemes should be introduced to attract non-users.
- Advertising standards should be maintained, if possible improved, as advertisements have contributed immensely to the awareness level and usage of the product.

Questionnaire

Dear Sir/Madam.

The students of Kirloskar Institute of Advanced Management Studies, Harihar, are conducting this survey, as a part of their project in the field of Market Research. The purpose of this activity is to measure the penetration of Surf in Davangere district.

1. Do you wash your clothes at home?
   - ☐ Yes
   - ☐ No
2. Do you use Surf?
   - ☐ Yes
   - ☐ No

   (If no. then go to Q. No.9.)
3. If Surf, which sub-brand do you use?
   - ☐ Surf Excel
   - ☐ Surf Ultra
   - ☐ Surf Super Excel
   - ☐ Surf Excel Matic
4. What influences your decision while buying Surf? (Tick as many as applicable.)
   - ☐ Whiteness
   - ☐ Lather
   - ☐ Easy on hands

- ☐ Easy on fabric
- ☐ Stain removal
- ☐ Any other (please specify)

5. Have you seen any promotional campaign of Surf?
   - ☐ Yes
   - ☐ No

   If yes, which one do you like the most?
   - ☐ Lalitaji
   - ☐ Surf Excel hai na
   - ☐ Dho daala
   - ☐ Dhoondthe reh jaaoge
   - ☐ Any Other (please specify) _____
6. Do the various schemes associated with Surf affect your purchase?
   - ☐ Yes
   - ☐ No
7. Would you suggest any changes for Surf in the following fields?
   - ☐ Availability in different quantities
   - ☐ Style of packaging
   - ☐ More schemes to be associated with the brand
   - ☐ Pricing
   - ☐ Any Other (Please specify) _____
8. Why not Surf ?
   - ☐ Price
   - ☐ Quality
   - ☐ Packaging
   - ☐ Fewer schemes as compared to other brands
   - ☐ Any Other(Please specify) _____
9. Which detergent do you most frequently use? (Tick as many applicable)
   - ☐ Ariel
   - ☐ Nirma
   - ☐ Wheel
   - ☐ Rin
   - ☐ Tide
   - ☐ Henko
   - ☐ Any other (Please specify) _____
10. What influences you to buy your preferred brand?
    - ☐ Friends
    - ☐ Neighbours

- ☐ Advertisements
- ☐ Self Experience
- ☐ Any other (please specify) _____

11. While purchasing a detergent, what quantity do you usually go for?
    - ☐ Less than 1 Kg
    - ☐ 1–2 Kg
    - ☐ 2–3 Kg
    - ☐ 3–4 Kg
    - ☐ More than 4 Kg
12. How frequently do you purchase detergents?
    - ☐ Once a week
    - ☐ Once a fortnight
    - ☐ Once a month
    - ☐ Once in two months
13. You prefer your detergent in:
    - ☐ Sachets (10 gm, 20 gm, 50 gm etc.)
    - ☐ Packets
    - ☐ Jars
    - ☐ Bigger containers
    - ☐ Any other (Please Specify) _____
14. If your preferred detergent is not available, you go for:

    First Choice _____

    Second Choice _____
15. Do you keep a stock of detergents in your home?
    - ☐ Yes
    - ☐ No
16. Most preferred detergent among people you know
    - ☐ Surf
    - ☐ Ariel
    - ☐ Nirma
    - ☐ Wheel
    - ☐ Rin
    - ☐ Tide
    - ☐ Henko
    - ☐ Any other (Please specify) _____

Something about you

Name : Mr./Mrs./Ms _____

Age Group : Kindly tick whichever is applicable

- ☐ < 25
- ☐ 25–34
- ☐ 35–44
- ☐ 45 and above

Address:

Occupation:

Do you own a washing machine?

- ☐ Yes
- ☐ No

Who washes the clothes in your house?

- ☐ Yourself
- ☐ Maid
- ☐ Any other (Please specify) _____

How many members are there in your household?

_____

Income Group: (Tick whicher is applicable.)

- ☐ < 5,000
- ☐ 5,001–10,000
- ☐ 10,001–15,000
- ☐ 15,001 and above

Thank you

CASE STUDY 3

BPL*

BACKGROUND

British Physical Laboratories (BPL) is the flagship company of BPL group, India's largest consumer electronics company. The company was incorporated as a private limited company in 1963 under the leadership of TPG Nambiar. It started operations by manufacturing hermetically sealed panels at a manufacturing plant in Palakkad, Kerala.

In 1979, the company started manufacturing plain paper copiers. In 1982, the company diversified in the area of consumer electronics. In 1992, the company became a public company making a public issue in March 1994. In 1996, the company integrated its operation backward by the takeover of Uptron's colour picture tube manufacturing facility. In 1997, the company successfully implemented its alkaline battery project at a cost of Rs. 1.2 billion in Karnataka.

The company in the recent past entered into a technical collaboration with the world's best: Sanyo Japan, Toshiba Corporation, France Telecom, Media One, Harris Communications, Octel, Nokia, and so on.

The BPL group at present has 12 regional, 35 branch offices, and a 4000 strong dealer network which enables it to have direct and constant access to the most important markets across the country. It has a 13000 strong skilled workforce. The group has 30 factories, which covers an area of approximately 5 million sq. ft.

Business areas now include consumer electronics, telecommunications, consumer durables, soft energy, professional and medical products, power, and components covering over 230 products and services.

The **highlights** of the financial year 1999–2000 were as follows:

- The market capitalisation of BPL was Rs. 5955 million as on 31st March 2000.
- The turnover for the group stood at Rs. 20146.68 million and profit after tax was Rs. 1071.28 million—4.5 per cent improvement over the previous year (1998–1999).

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala under the support and guidance of Dr Rajendra Nargundkar, for which they are sincerely grateful.

- BPL's return on capital employed was 15.72 per cent.
- Investment of Rs. 24 crore in the area of Oracle Financing, Data Warehousing, and Net enabling.
- The company's contribution to the exchequer amounted to Rs 410.93 crores in the form of duties and taxes.

Research Objectives

1) To study how the customer perceives the BPL range of CTV.
2) To identify the factors that influence the buyers to buy and non-buyers not to buy BPL CTV.
3) To examine the effectiveness of various promotional activities.

Methodology

Sources of data

- The data is basically primary in nature.
- It was obtained from the employees of Harihar Polyfibres company, and households and dealers in Harihar and Davanagere.

Methods

- Our communication approach was basically structured questioning, that is, personal interview with the aid of printed questionnaires.

Sample size

- Convenience sampling: No sampling technique was employed in arriving at a sample size. it is a convenience sampling suiting our use.
- Consumer sample size—30
- Dealer sample size—6

Limitations

1. This being a convenience sample, the analysis may not be a true picture of the target population.
2. Language problem, especially with the household respondents of Harihar.
3. Prejudice of some of the respondents.

   For instance, one of the respondents had a bitter experience with respect to BPL VCR. For this reason, he is totally against the brand name BPL and could not consider even a single factor in favour of the BPL CTV.
4. Low sample size of the dealers-that is, we could not find more than 6 who were able to answer our questions in the areas surveyed.

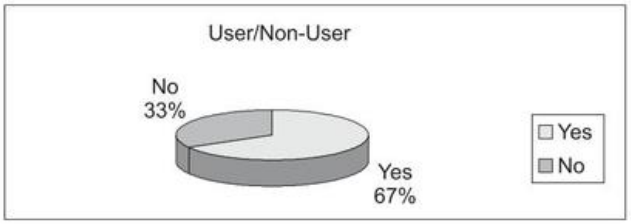
ANALYSIS OF SIMPLE TABULATIONS

Customer's Questionnaire

1. Are you a user of BPL CTV?

Sample Size 30

	Frequency	*Percentage*
Yes	20	33.3
No	10	66.7

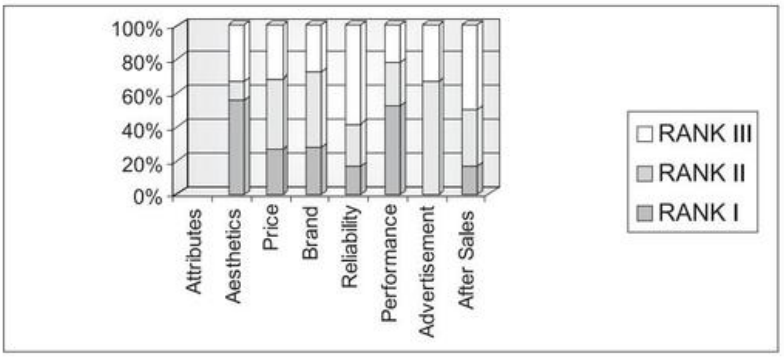


67% of the respondents are users of BPL CTV and 33% are non users.

2. What influences your decision to purchase a Colour TV?

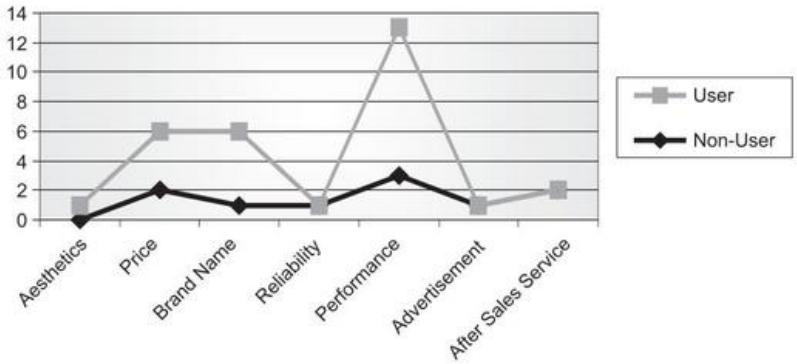
Sample Size 30

Ranks	*I*	*II*	*III*
Attributes			
Aesthetics	5	1	3
Price	5	8	6
Brand	5	8	5
Reliability	2	3	7
Performance	12	6	5
Advertisement	0	2	1
After sales service	1	2	3



The respondents ranked performance as the most important attribute influencing their purchasing decision. The next important attributes are price brand, and aesthetics.

3. What influenced your decision to purchase/not to purchase BPL CTV?



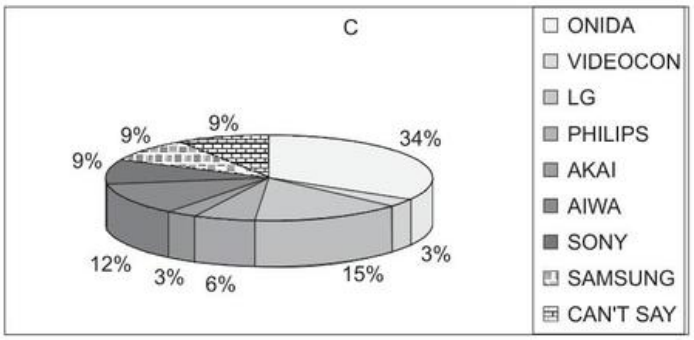
It is found that the non-users have not purchased BPL CTV because of dissatisfaction with its price and performance.

Surprisingly, price and performance are also found to be the factors that have influenced the purchase of BPL CTV.

4. Which brand do you think is the toughest competitor to BPL CTV?

Sample Size 30

	FREQUENCY	*PERCENT*
ONIDA	11	36.7
VIDEOCON	1	3.3
LG	5	16.7
PHILIPS	2	6.7
AKAI	1	3.3
AIWA	4	13.3
SONY	3	10
SAMSUNG	3	10
CAN'T SAY	3	10

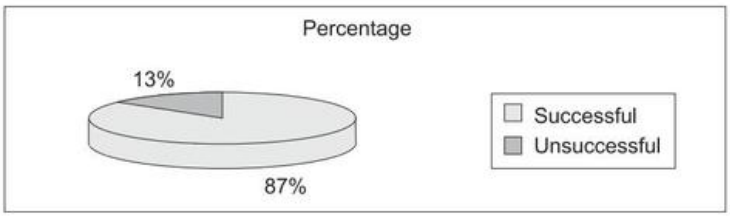


From the table it can be clearly inferred that Onida is the toughest competitor to BPL according to the perception of the consumers.

5. Do you think BPL CTV is a successful/unsuccessful brand?

Sample Size 30

	Frequency	*Percentage*
Successful	26	86.7
Unsuccessful	4	13.3

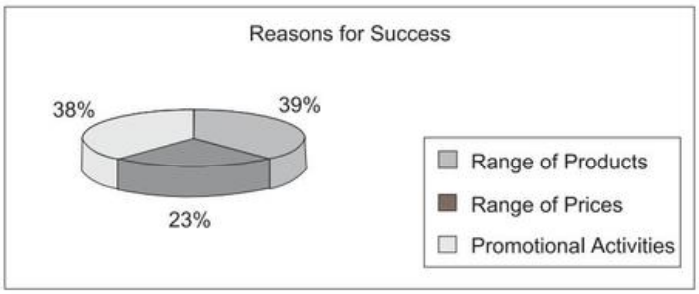


87% of the respondents feel that BPL CTV is a successful brand. Rest 13% find it to be unsuccessful.

6. Why do you think BPL has been a successful/unsuccessful brand?

Sample size :30

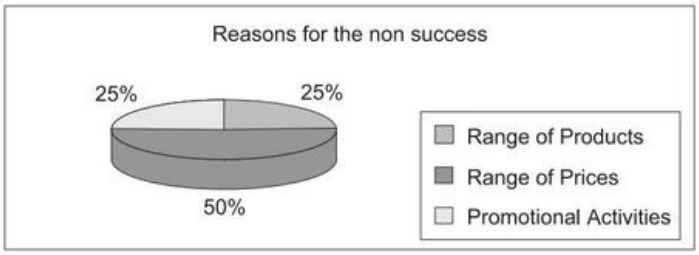
Successful	*Frequency*	*Percentage*
Range of Products	10	38.5
Range of Prices	6	23.1
Promotional Activities	10	38.4



As far as the reasons for BPL CTV being a successful brand is concerned, 39% of the consumers feel that the success is due to the range of products offered by BPL, 38%feel its due to the range of prices and the rest 23% feel that its due to the promotional activities carried out by BPL.

Sample Size 30

Unsuccessful	*Frequency*	*Percentage*
Range of Products	1	25
Range of Prices	2	50
Promotional Activities	1	25

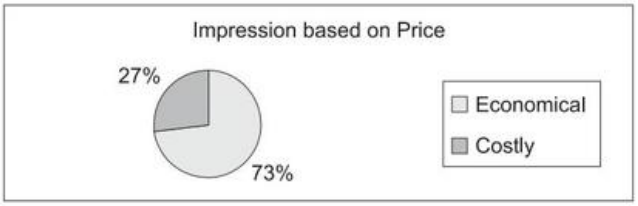


Among the people who find BPL an unsuccessful brand, 50% feel its due to the range of prices offered by BPL, rest 25% each feel its due to the range of prices and promotional activities respectively.

7. Based on the price range what impression do you carry of BPL CTV?

Sample Size 30

	Frequency	*Percentage*
Economical	22	73.3
Costly	8	26.7

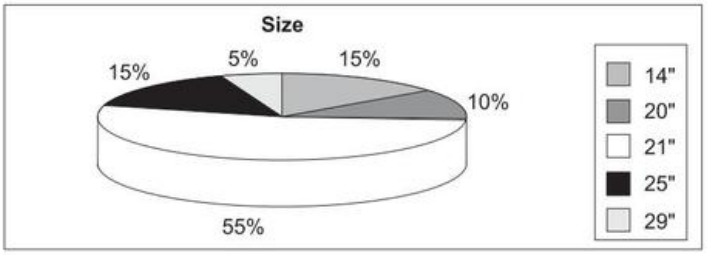


73% find BPL to be an economical brand and the rest 27% find it costly. Thus the overall impression of BPL in CTV segment is that of an economical one.

8. Give the size of the BPL CTV you own (User's Question).

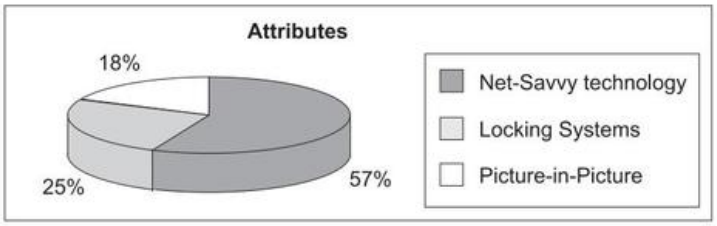
Sample size : 20

Size	*Frequency*	*Percentage*
14"	3	15
20"	2	10
21"	11	55
25"	3	15
29"	1	5



We found that the most preferred size by 55 % respondents was 21", followed by 14" and 25", with 15% each.

9. Which of the following additional attributes do you perceive as most important in your CTV?



Sample size : 28

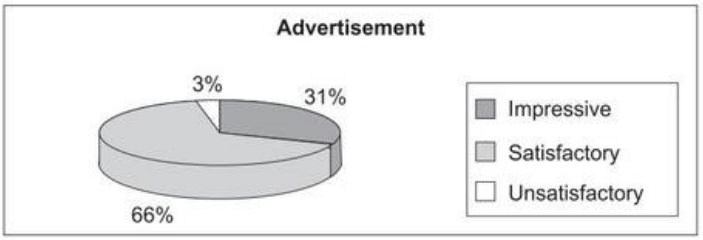
	Frequency	*Percentage*
Net-Savvy technology	16	57.1
Locking Systems	7	25
Picture-in-Picture	5	17.9

57.1% of the people prefer to have Net Savvy Technology as an additional attribute in their CTVs. Following this are locking systems with 25% and Picture in Picture with 17.9%.

10. What do you think of the advertisement of BPL CTV?

Sample Size 29

	Frequency	*Percentage*
Impressive	9	31
Satisfactory	19	65.5
Unsatisfactory	1	3.4

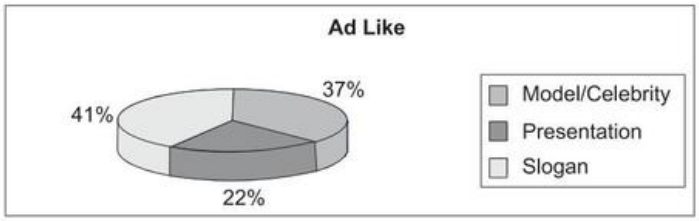


65.5 % of the respondents find the advertisements satisfactory while 31% and 3% find it impressive and unsatisfactory respectively.

11. You like/dislike the advertisement because of

Sample Size 30

Like	*Frequency*	*Percentage*
Model/Celebrity	10	35.7
Presentation	6	21.4
Slogan	11	39.3



41% of the respondents who liked the advertisement like it because of the slogan 'Believe in the Best'.

37% like it because of the model or celebrity, that is, i.e. undoubtedly Amitabh Bachchan.

Sample Size: 30

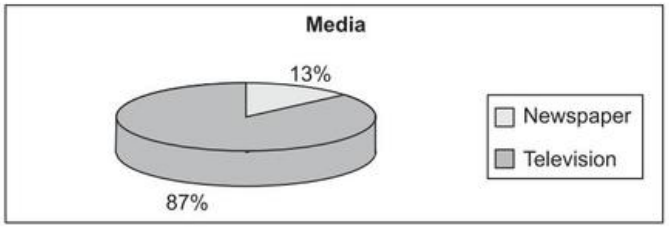
Dislike	*Frequency*	*Percentage*
Model/Celebrity	0	0
Presentation	3	100
Slogan	0	0

Out of the sample size of 30, only 3 people disliked the ad and all of them disliked it because of the presentation (no ad was specifically mentioned).

12. Which do you think is the most striking media to advertise BPL CTV?

Sample Size : 30

	Frequency	Percentage
Newspaper	4	13.3
Television	26	86.7

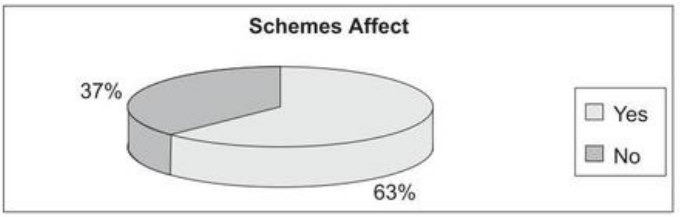


87% of the respondents asserted that television is the most striking media to advertise BPL CTV while the rest 13% felt that the newspaper is the best media.

13. Do the various schemes/promotional activities affect you purchase plans?

Sample Size 30

	Frequency	*Percentage*
Yes	19	63.3
No	11	36.7

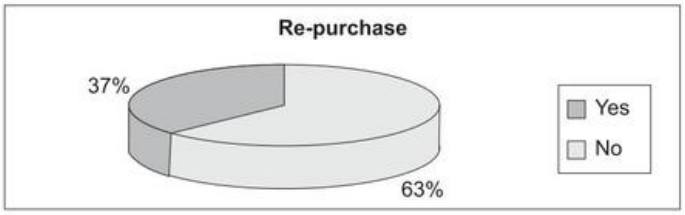


63% of the respondents felt that the various schemes offered by the company would affect their purchase decision, while the rest 37% did not feel so.

14. If you go for re-purchase of TV, will you prefer BPL CTV?

Sample Size 30

	Frequency	*Percentage*
Yes	19	63.3
No	11	36.7

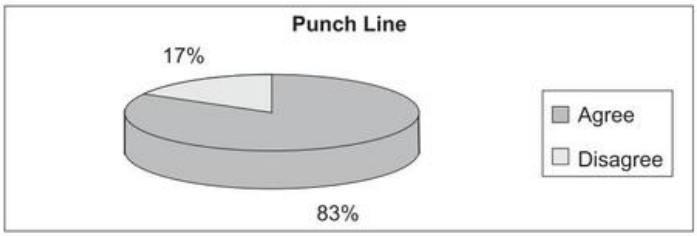


63.3% of the respondents said that they would go for BPL CTV as their re-purchase choice while the rest 37% felt otherwise.

15. Do you agree/disagree with the Punch line of BPL—'Believe in the Best'?

Sample Size 30

	Frequency	*Percentage*
Agree	25	83.3
Disagree	5	16.7

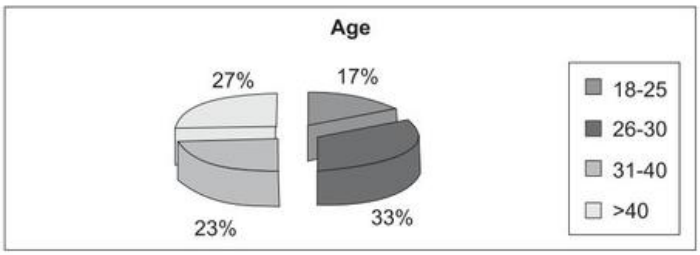


83% of the respondents agree with what the punch line of BPL has to say and the rest 17% disagree with the statement that BPL is the best.

Demographics

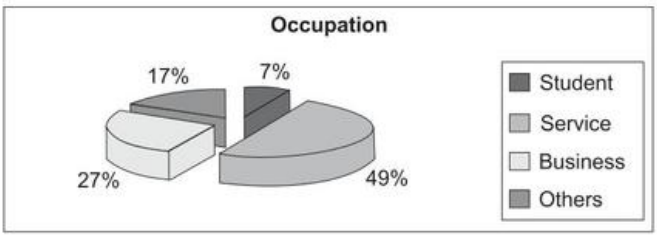
Age

Frequency	*Age Group*	*Percentage*
5	18–25	16.70%
10	26–30	33.30%
7	31–40	23.30%
8	> 40	26.70%



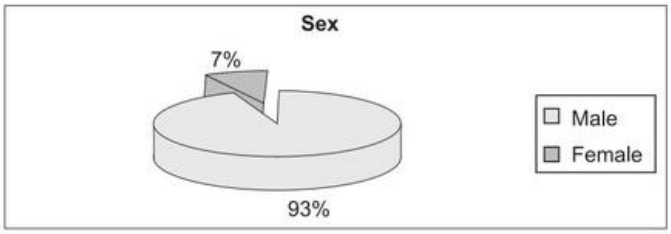
Occupation

Occupation	*Frequency*	*Percentage*
Student	2	6.70%
Service	15	50%
Business	8	26.70%
Others	5	16.70%



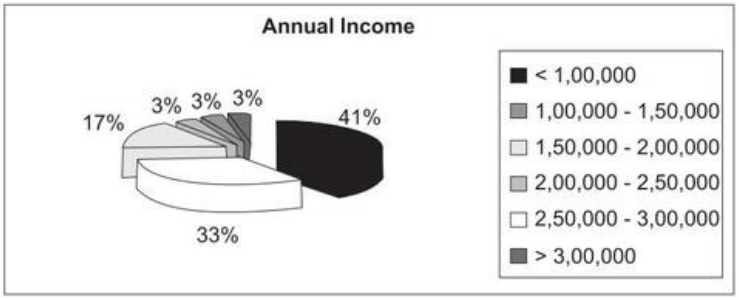
Sex

Sex	*Frequency*	*Percentage*
Male	28	93.30%
Female	2	6.70%



Annual Income

Annual Income	*Frequency*	*Percentage*
< 1,00,000	12	40%
1,00,000–1,50,000	10	33.30%
1,50,000–2,00,000	5	16.70%
2,00,000–2,50,000	1	3.30%
2,50,000–3,00,000	1	3.30%
> 3,00,000	1	3.30%

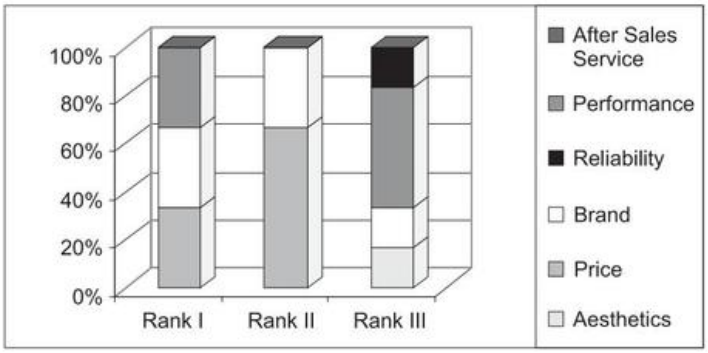


Dealer's Questionnaire

1. **What influences your customer's decision to purchase a BPL CTV?**

Sample Size 6

Ranks	*I*	*II*	*III*
Aesthetics	…	…	1 (16.7%)
Price	2 (33.3%)	4 (66.7%)	…
Brand	2 (33.3%)	2 (33.3%)	1 (16.7%)
Reliability	…	…	…
Performance	2 (33.3%)	…	3 (50%)
After sales service	…	…	1 (16.7%)

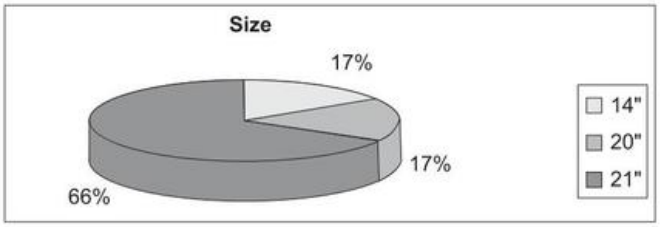


According to the dealers, the top three attributes that influence their customer's purchase decision are: performance, brand name, and price.

2. Which size of BPL CTV do the customers purchase the most?

Sample Size 6

Size	*Frequency*	*Percentage*
14"	1	16.7
20"	1	16.7
21"	4	66.7
25"	0	0
29"	0	0

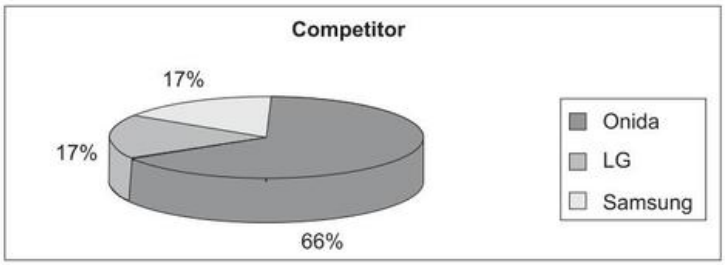


21" is the most preferred size of the BPL CTV. It is accepted by 66% of the dealers.

3. Which do you think is the toughest competitor to BPL?

Sample Size 6

Competitors	*Frequency*	*Percentage*
Onida	4	66.7
LG	1	16.7
Samsung	1	16.7

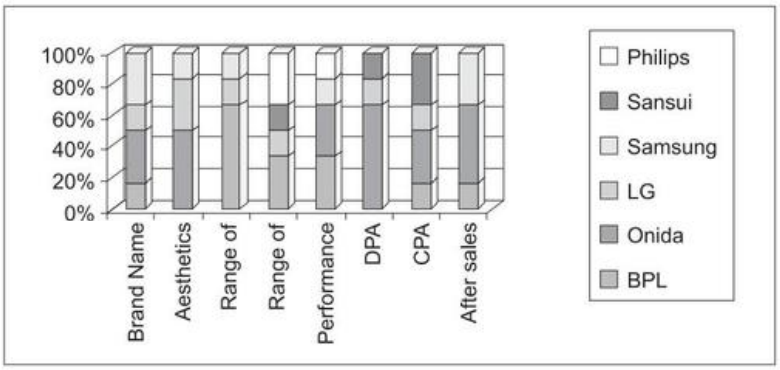


Onida is found to be the toughest competitor to BPL CTV. This is a true picture of the national scenario. Next comes LG and then Samsung as felt by 17% of the Dealers.

4. Where do you rank the following CTVs in terms of

Sample Size 6

	BPL	*Onida*	*LG*	*Samsung*	*Sansui*	*Philips*
Brand Name	1	2	1	2		
Aesthetics		3	2	1		
Range of Products	4		1	1		
Range of Prices	2		1		1	2
Performance	2	2		1		1
DPA		4	1		1	
CPA	1	2	1		2	
After sales service	1	3		2		



The best in terms of

Brand name: *BPL, Samsung*

Aesthetics: *Onida*

Range of products: *BPL*

Range of prices: *BPL, Philips*

Performance: *BPL, Onida*

Dealers promotional activities: *Onida*

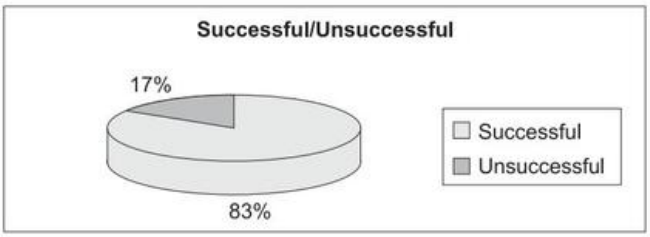
Consumers promotional activities: *Onida, Sansui*

After sales service: *Onida*

5. Do you think BPL CTV is a successful/unsuccessful Brand?

Sample Size 6

	Frequency	*Percentage*
Successful	5	83.3
Unsuccessful	1	16.7

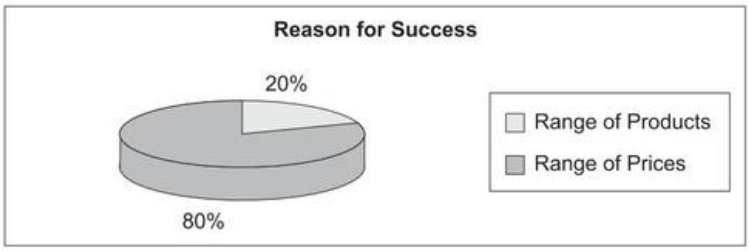


83% of the dealers accepted that it is a successful brand and 17% felt it is an unsuccessful brand.

6. Why do you think BPL CTV has been a successful/unsuccessful Brand?

Sample Size 5

Successful	*Frequency*	*Percentage*
Range of Products	1	20
Range of Prices	4	80
Promotional Activities	0	0
Distribution Network	0	0



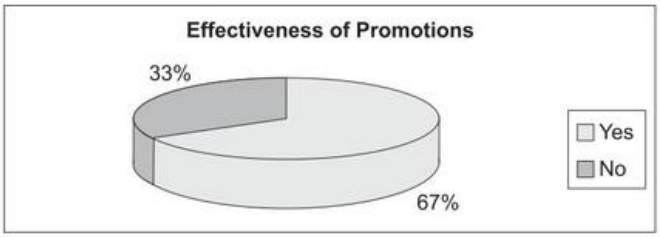
80% of the dealers felt it is successful because of its range of prices. 20% felt it is because of the range of products. Promotional activities and distribution network have nothing to do with the success of BPL CTV.

Only one dealer found the brand unsuccessful because of its range of products.

7. Are the consumer sales promotional activities of BPL effective?

Sample Size 6

	Frequency	*Percent*
Yes	4	66.7
No	2	33.3

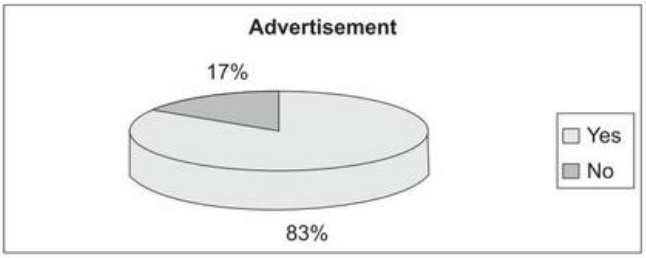


67% of the dealers are happy with the effectiveness of sales promotional activities and 33% are not.

8. Are you satisfied with the current media used by BPL for advertisement?

Sample Size 6

	Frequency	*Percent*
Yes	5	83.3
No	1	16.7

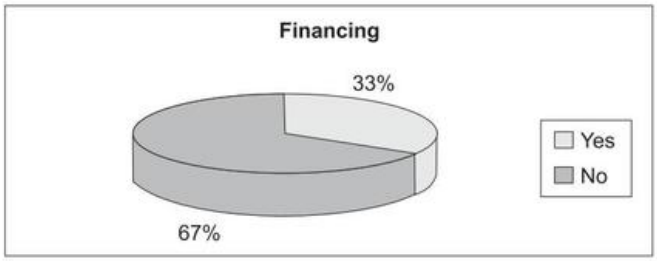


83% are happy with the current media used for advertisement and 17% are not happy.

9. Do you think BPL should go for financing the purchase of the Colour TV?

Sample Size 6

	Frequency	*Percent*
Yes	4	33.3
No	2	66.7



67% of the dealers wanted BPL to go for financing the purchase of the CTV.

10. Are you satisfied with the dealer incentive schemes provided to you by BPL vis-à-vis the competitors?

Sample Size 6

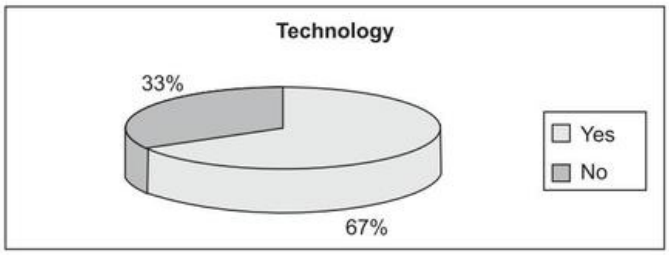
	Frequency	*Percent*
Yes	4	66.7
No	2	33.3

It was found that 67% of the dealers were satisfied with the incentive schemes provided to them by the company. But 33% of the dealers weren't satisfied.

11. Do you think the latest technological changes introduced in the BPL CTV will result in the boost of it's sales?

Sample Size 6

	Frequency	*Percent*
Yes	4	66.7
No	2	33.3

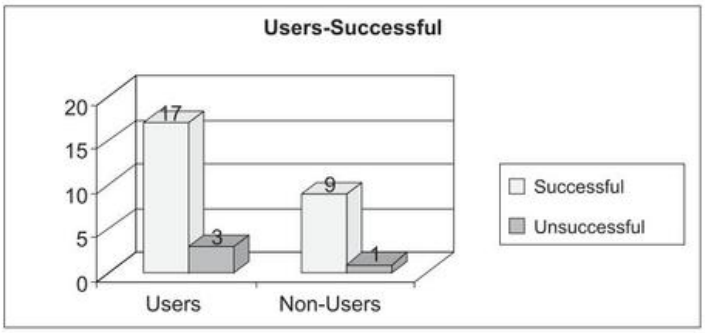


67% of the dealers thought that the technological advances would really affect the sale of the CTV. There were 33% of the dealers who felt that it would not affect the sale.

Cross-tabulation

1. **The objective of this tabulation is to find the break-up between the percentage of users/ non-users perceiving BPL as a successful brand.**

	Successful	*Unsuccessful*	
Users	17	3	20 (66.7%)
Non-users	9	1	10 (33.33%)
	26 (86.7%)	4 (13.3%)	

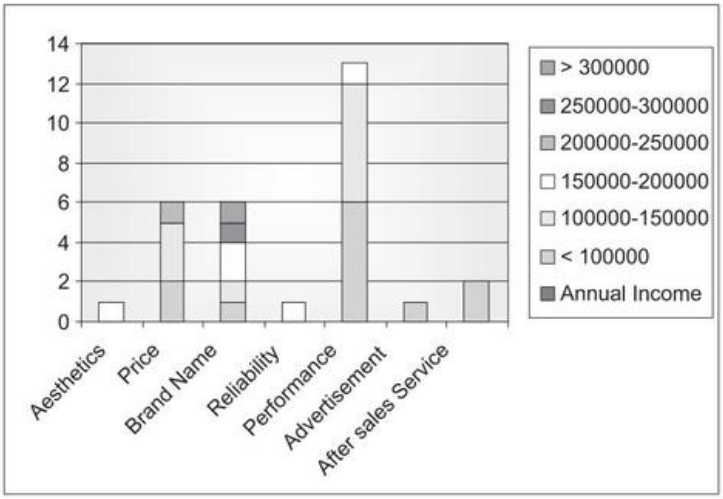


Finding We see that out of the ten non-users questioned, nine of them perceive the brand to be successful but still they have not gone for it. During our research process we found out that though majority of the people find the brand to be successful they have not actually bought it, because of the following reasons:

- at the time of their purchase either the other brands had come out with better promotional schemes, or
- while in the showroom they made on the spot changes because they found the range of products offered by its competitors more appealing.

2. **The main objective of this cross-tabulation is to show how the choice of attributes of CTV vary with income levels**

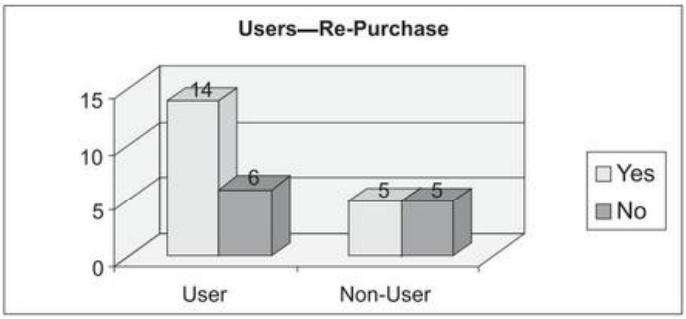
	Aesthetics	*Price*	*Brand*	*Reliability name*	*Performance*	*Advertisement*	*After sales*	*service*
Annual Income								
< 100000		2	1		6	1	2	40%
100000–150000		3	1		6			33.3%
150000–200000	1		2	1	1			16.7%
200000–250000		1						3.30%
250000–300000			1					3.30%
> 300000			1					3.30%
	3.30%	20%	20%	3.30%	43.30%	3.30%	6.70%	



Finding It is very clear from the table and the graph above that the lower income groups (i.e., < 1,00,000 and 1,00,000 – 1,50,000) give more preference to performance. Whereas the higher income groups (i.e., 2,50,000 – 3,00,000 and > 3,00,000) give more preference to brand name.

3. **The objective of this cross-tabulation is to know the break-up between the number of users/non-users going for re-purchase of BPL CTV.**

	Yes	*No*	
User	14	6	20 (66.7%)
Non-user	5	5	10 (33.3%)
	19 (63.3%)	11 (36.7%)	



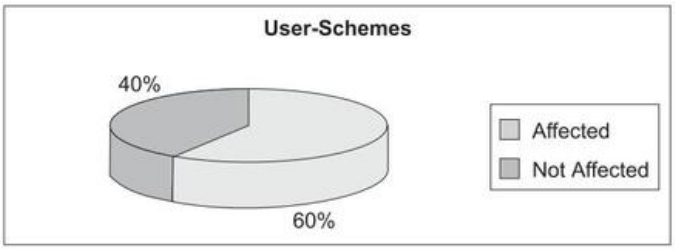
Finding We find that out of 20, 6 of the users don't want to go for re-purchase of BPL CTV .The reasons are:

- The lower income group people find the range of prices and promotional schemes offered by other brands more attractive vis-a-vis BPL.
- The higher income groups people feel that the quality of BPL is not up to the expected level of standards. The trust associated with the brand name is falling.

- Surprisingly, half of the non-users are willing to go for BPL as their re-purchase choice. This shows the privilege enjoyed by BPL as the market leader in the CTV segment and also the minds of the people.

4. **The objective of this cross-tabulation is to know how many of the users have really been affected by the various purchase schemes offered by the company.**

	Schemes	
	Affected	Not Affected
Users	12	8



Finding From the above table it is very clear that 60% of the users are affected by the schemes and 40% are not.

RECOMMENDATIONS

Brand Image Should be Maintained

BPL enjoys a good brand image, especially due to its reputation of conforming to quality in the consumer electronic goods market. This story dates back to the time when there were not many brands in the market and BPL was the only brand, which got hugely popular especially due to its quality levels. But lately it has been noticed that

- Their quality is going down
- No prompt after sales service
- Less incentive schemes to dealers compared to competitors.

So our recommendation to BPL is that if they want to retain their market share it becomes vital for them to maintain the brand image and the trust which they have built over the years.

Better Advertisements

Another finding of our survey was that people were in a very ambiguous state when asked as to what they thought of the advertisements. Even on being told what they were, the respondents were not able to recall them exactly. The only ad which could create a considerable impression was the one featuring Amitabh Bachchan. Thus, it is necessary to come out with more attractive ads. Also, the number and the frequency of the ads has to be increased on the television channels.

Quality Checks

In the past, BPL could boast of its best quality. But today we find that both the customers and dealers have too many complaints with the quality of the product. Thus we think BPL should go for more stringent quality checks.

Attractive Schemes for Customers and Dealers in this Region

Particularly in the areas of Harihar and Davangere, we found that the customer promotional and the dealer incentive schemes offered by the BPL CTV are almost nil, compared to its competitors. Thus, the company should give due weightage to this aspect to survive as the undisputed leader in the market.

questionnaire for Consumers

Date: Q. No.

Dear sir/madam,

We the students of KIAMS are conducting a market research to know your brand perception in the colour TV segment. Kindly extend your cooperation in filling this questionnaire and enable us in doing the research successfully.

1. Are you a user of BPL colour television?
   - ☐ Yes.
   - ☐ No.

   If no, specify which brand _____
2. What influenced your decision to purchase a Colour TV?
   (Rank the following in the order of preference) 1—Best, 2—next best, and so on.
   - __ Aesthetics/Appearance
   - __ Price
   - __ Brand name
   - __ Reliability
   - __ Performance
   - __ Advertisement
   - __ After sales service
3. What influenced your decision to purchase/not to purchase BPL Colour TV?
   (Please select any two from above)
   - i. _____
   - ii. _____
4. Which Brand do you think is the toughest competitor to BPL?
   - ☐ Onida
   - ☐ Videocon

- ☐ LG
- ☐ Philips
- ☐ Akai
- ☐ Aiwa
- ☐ Sony
- ☐ Samsung

Why? _____

5. Do you think BPL Colour TV is a
   - ☐ Successful brand
   - ☐ Unsuccessful brand
6. Why do you think BPL Colour TV has been a successful/ unsuccessful brand?
   - ☐ Range of products
   - ☐ Range of prices
   - ☐ Promotional activities
7. Based on the price range what impression do you carry of BPL CTV:
   - ☐ Economical
   - ☐ Costly
8. Tick the size of BPL Colour TV you own (user's question)

   ( ) 14" ( ) 20" ( ) 21" ( ) 25" ( ) 29"
9. Which of the following additional attributes do you perceive as most important in your Colour TV?
   - ☐ Net savvy technology (e.g.: Web TV, Multi-media TV, etc.)
   - ☐ Locking systems (e.g.: Child Lock, Volume Lock)
   - ☐ Picture in Picture

   Any others, specify_____
10. What do you think of the advertisement of BPL?
    - ☐ Impressive
    - ☐ Satisfactory
    - ☐ Unsatisfactory.
11. You like/dislike the advertisement because of:

	Like	Dislike
☐ Model/Celebrity	_____	_____
Specify the celebrity	_____	_____
☐ Presentation,	_____	_____
What you liked/disliked	_____	_____
☐ Slogan	_____	_____
Any other (specify)	_____	_____

12. Which do you think is the most striking media to advertise BPL CTV?
    - ☐ Newspaper.
    - ☐ Magazines.
    - ☐ TV.
    - ☐ The Internet.
    - ☐ Hoarding.
    - ☐ Radio.
13. Do the various schemes/promotional activities affect your purchase plans?
    - ☐ Yes
    - ☐ No
14. If you go for re-purchase of TV, will you prefer BPL Colour TV?
    - ☐ Yes
    - ☐ No
15. Do you agree or disagree with the punch line of BPL—***'Believe in the Best'***.

    _____

A word about yourself

Name:

Address:

Age: (a) 18–25 (b) 26–30 (c) 31–40 (d) above 40

Occupation: Student/Service/Business/Any other (specify)_____

Sex: Male/Female

Annual Income:
1) < 100000
2) 100000–150000
3) 150000–200000
4) 200000–250000
5) 250000–300000
6) >300000

Thank you

Questionnaire for Dealers

Date: Q. No.

Dear Sir/Madam,

We, the students of KIAMS are conducting a market research to know the consumer perception and sales promotion effectiveness of BPL CTV.

Kindly extend your cooperation in filling this questionnaire and enable us to conduct the research successfully.

1. What influences your customer's decision to purchase a BPL Colour TV? (Rank the following in the order of preference) 1—best, 2—next best, etc.
   __Appearance/Aesthetics
   __Price
   __Brand name
   __Reliability
   __Performance
   __After sales service
2. Which size of BPL CTV do the customers purchase the most?
   ( )14" ( ) 20" ( ) 21" ( ) 25" ( ) 29"
3. Which do you think is the toughest competitor to BPL?
   ☐ Onida
   ☐ Videocon
   ☐ LG
   ☐ Philips
   ☐ Samsung
   ☐ Akai
   ☐ Aiwa
   ☐ Thompson
   ☐ Sony
   Why? _____
4. Where do you rank the following colour TVs in terms of (Rank the top 3 for each)

	BPL	*Onida*	*LG*	*Videocon*	*AIWA*	*AKAI*	*SAMSUNG*
Brand Name							
Aesthetics							
Range of Products							
Range of Prices							
Performance							
Dealer Promotional Activities							
Consumer Promotional Activities							
After Sales Service							

5. Do you think BPL Colour TV is a
   ☐ Successful brand
   ☐ Unsuccessful brand

6. Why do you think BPL Colour TV has been a successful/unsuccessful brand?
   - ☐ Range of products
   - ☐ Range of prices
   - ☐ Distribution network
   - ☐ Promotional activities
7. Are the consumer sales promotion activities of BPL effective?
   - ☐ Yes
   - ☐ No

   If No, why? _____
8. Are you satisfied with the current media used by BPL for advertisement?
   - ☐ Yes
   - ☐ No
9. If No, which do you think is the most striking media for BPL to advertise?
   - ☐ Newspaper
   - ☐ Magazines
   - ☐ TV
   - ☐ The Internet
   - ☐ Hoarding
   - ☐ Radio
10. Do you think BPL should go for financing the purchase of the Colour TV?
    - ☐ Yes
    - ☐ No
11. Are you satisfied with the dealer incentive schemes provided to you by BPL vis-à-vis the competitors?
    - ☐ Yes
    - ☐ No
12. Do you think the latest technological changes introduced in the BPL CTV will result a boost in it's sales?
    - ☐ Yes
    - ☐ No

Shop name:

Address:

Thank you

The McGraw·Hill Companies

PART 2 Data Analysis

- Simple Tabulation and Cross-tabulation
- ANOVA and the Design of Experiments
- Correlation and Regression: Explaining Association and Causation
- Discriminant Analysis for Classification and Prediction
- Logistic Regression for Classification and Prediction
- Factor Analysis for Data Reduction
- Cluster Analysis for Market Segmentation
- Multidimensional Scaling for Brand Positioning
- Conjoint Analysis for Product Design
- Attribute-based Perceptual Mapping Using Discriminant Analysis
- Structural Equation Modeling (SEM) for Complex Models (including Confirmatory Factor Analysis)

The McGraw·Hill Companies

CHAPTER 8 SIMPLE TABULATION AND CROSS-TABULATION

Learning Objectives

In this chapter, we will

- Illustrate First Stage Analysis—counting of frequencies
- Illustrate the concepts of simple and cross-tabulation
- Discuss the use of percentages in simple and cross-tabulations
- Discuss the rule for direction of percentage calculation in a cross-tabulation
- Demonstrate the use of a Chi-squared test for testing if two variables in a cross-tab are significantly associated with each other.
- Discuss some measures of the strength of association between two variables in a cross-tab, known as indexes of agreement, such as the Contingency Coefficient and Goodman and Kruskal's Lambda (asymmetric)

UNIVARIATE AND BIVARIATE ANALYSIS

Univariate analysis is a single-variable analysis. In a questionnaire-based marketing research project, each question usually represents a variable under study. Bivariate analysis involves two variables at a time. Two different questions in a questionnaire may represent two variables. If these two variables are analysed together, it is an example of bivariate analysis.

In other words, simple tabulation involving a single variable constitutes univariate analysis, and cross-tabulation of two variables at a time constitutes bivariate analysis.

Dependent and Independent Variables

If two or more variables are analysed together, it may be necessary to spell out the relationship between the two variables. The concept of dependent and independent variables is useful in spelling out the

relationship. Two variables are called independent variables if a change in one does not influence or cause a change in the other. But if a change in one variable causes a change in the other, the first one is called an independent variable, and the second one is called a dependent variable (dependent on the first).

A common example of a dependent variable in marketing is 'sales'. Annual sales of a brand usually depend on several factors or variables. One of the independent variables on which annual sales depend could be the quantum of advertising (in rupees) done for the brand. A second variable on which sales may depend could be the number of retailers stocking the brand.

In a consumer research questionnaire, the dependent variable could be satisfaction with the brand, which may depend on taste (if it is a food brand), and availability. Another example is the quantity of a product bought, a dependent variable, which depends on family size and household income.

Demographic Variables

Many demographic variables such as age, location, income, occupation, sex, education are generally independent variables for the purposes of most marketing studies. This is because other variables *depend* on them. Of course, sometimes there could be a relationship of dependence between two demographic variables themselves. For instance, income may depend on the level of education, or age of the respondent, in some cases. But most of the time in marketing studies, we are interested in demographic variables for their impact on other marketing variables such as purchase behaviour of consumers. Hence, we treat them as independent variables.

The purchase decision, or the brand purchased, or intention to buy, are usually treated as dependent variables in many marketing studies. For a marketing researcher, these variables or similar ones, are the real variables of interest, as they help in arriving at strategies for increasing sales or market share.

The other major types of independent variables are the elements of four 'P's of marketing. The marketing effort of a company can be measured in terms of its promotional efforts, price variations, and distribution changes. It can also be gauged from new product launches, or repositioning or repackaging of existing brands. Therefore, we could measure sales as the dependent variable with any of the marketing 'P's as independent variables. This could be done even without a consumer survey.

FIRST STAGE ANALYSIS—SIMPLE TABULATION

In a questionnaire-based survey, the first stage of analysis is called simple tabulation. This consists of every question being treated separately. For every question, the number of responses in each category of answers is counted. Assuming the sample size is 500, and all 500 have answered the question, the simple tabulation of a respondent's gender may look like the following:

1. Male	–	300
2. Female	–	200
Total		500

The simple tabulation for another question on the questionnaire may look like this:

1.	Regular users of brand *X*	–	200
2.	Occasional users of brand *X*	–	150
3.	Non-users of Brand *X*	–	150
	Total		500

A title can be included for each table, and on the top of each column, to explain the variable name through a label. For example, the above simple table can be titled, **Frequency of Usage,** or **Number of Users and Non-users of Brand *X*.**

Computer Tabulation

If codes were used to input the data into the computer for tabulation, the numbers 1, 2, and 3 could have also been the numerical codes for the three categories of responses to the above question. The descriptions 'Regular users of brand *X*', 'Occasional users of brand *X*', and 'Non-users of brand *X*' are called value labels in most of the computer packages, and can be defined by the user. They will appear on the table whenever the table is printed as output. The variable label is usually the title of a column of data in the package. In this case, the column could have been labeled with a variable label, 'Usage of brand *X*' or some similar title.

Percentages

In addition to the number of respondents who fall into each category, we usually compute percentage of the respondents also. This appears as one more column on the table, and is automatically printed out in most computer packages when you request a table to be printed. For example, in the above table, it would look like the following, with percentages added:

	Usage of brand X		*Number*	*Percentage*
1.	Regular users of brand *X*	–	200	(40)
2.	Occasional users of brand *X*	–	150	(30)
3.	Non-users of brand *X*	–	150	(30)
	Total		500	(100)

Please note that the percentage is based on the total number of respondents who answered this question. If in a questionnaire, the number of respondents is different for some of the questions, the percentage will be calculated with respect to the total number of respondents for the respective questions. For example, in the above example, there may be a question for non-users only, after the above question has identified them. Since there are only 150 non-users of brand *X*, the sample size of respondents for the question will be 150. Another question for users (both occasional and regular) may have 200 + 150 = 350 as the number of total respondents. So, the percentages will be calculated on different totals for these two subsequent questions.

It is worth remembering that if the categories of answers to a question are such that multiple choices can be ticked by respondents, the percentages may not add up to 100. For example, the question may ask respondents which brand or brands of toothpaste they have used before, and the answer categories may be—

1. Colgate
2. Pepsodent
3. Close Up
4. Promise
5. Any other (specify)

In such a case, people may tick more than one brands. Therefore, the percentages may add up to more than 100. For example, 30% of the respondents may choose Colgate, 40% may say Pepsodent, 50% may say Close Up, 10% may tick Promise, and 20% may pick other brands. The total percentage would then add up to 30 + 40 + 50 + 10 + 20, or 150. This total percentage is not a meaningful percentage if multiple options can be ticked by respondents. But the individual percentages for each brand do hold meaning.

These types of simple tables are also known as Frequency Tables. Many computer packages provide graphics capabilities to print out a variety of graphs and charts to represent the data in addition to the tables. One popular chart is the Bar Chart. Another is a Pie Chart, in the form of a circle with segments representing different categories of answers to a question.

SIMPLE TABULATION FOR RANKING TYPE QUESTIONS

Suppose we had ordinally-scaled questions in our questionnaire. Then, we may have a complex answer to tabulate. For example, the question could have been—

Q. Rank the five brands of refrigerators shown below on a scale of 1 to 5 (1 = Best and 5 = Worst), according to your opinion.

BRAND	*RANK*
Whirlpool	—
Kelvinator	—
Godrej	—
Samsung	—
Videocon	—

The tabulation of this question will end up with an output table that looks like this—

BRAND	*RANK 1*	*RANK 2*	*RANK 3*	*RANK 4*	*RANK 5*
Whirlpool	*x*	*x*	*x*	*x*	*x*
Kelvinator	*x*	*x*	*x*	*x*	*x*
Godrej	*x*	*x*	*x*	*x*	*x*
Samsung	x	x	x	x	x
Videocon	*x*	*x*	*x*	*x*	*x*

The *x* values in the table represent the number of respondents who gave a particular rank to each brand. This is actually a bivariate table, because Brand of refrigerator and Rank are the two variables. If we want to construct univariate tables out of the above data, we can take up one column at a time from Table 1 and do separate frequency tables or charts. If we assume some numbers, one of the univariate tables may look as follows–

BRAND	*No. of People who Ranked it No.1*
Whirlpool	90
Kelvinator	60
Godrej	70
Samsung	32
Videocon	45
Total	297

This is a univariate table, and if we wish to, we can calculate the percentages on a total for each brand. For example, 90/297 works out to .303 or 30.3% who ranked Whirlpool as no.1. Similar calculations can be done for other brands in the column.

We can construct similar tables for Ranks 2, 3, 4 and 5 if we want to look at the frequencies of people who gave those ranks to the brands separately. But the overall picture is already available from Table 1.

Tabulating Ratings

Commonly used rating scales are of the following type—

Q. Rate the following attributes of LIRIL soap on a scale of 1 to 5 (1 = Very unsatisfactory, 2 = Unsatisfactory, 3 = Neither satisfactory nor unsatisfactory, 4 = Satisfactory, 5 = Very satisfactory).

Lather	1	2	3	4	5
Fragrance	1	2	3	4	5

For each attribute, the number of people who rated it as 1, 2, 3, 4, or 5 can be tabulated in separate tables, one of which will look as follows—

Rating	*Lather*
1	30
2	25
3	50
4	76
5	22
Total	203

Alternatively, we can tabulate ratings for all attributes in one table as follows—

Rating	*Lather*	*Fragrance*	*ATR. 3*	*ATR. 4*	*ATR. 5*
1	x	x	x	x	x
2	x	x	x	x	x
3	x	x	x	x	x
4	x	x	x	x	x
5	x	x	x	x	x

SECOND STAGE ANALYSIS—CROSS-TABULATION

After the simple frequency and percentage tabulation for every question on the questionnaire comes the second stage—the cross-tabulations. A cross-tabulation can be done by combining any two of the questions and tabulating the data together. This is a 2-variable cross-tabulation. An example could be a cross-tabulation between brand preference for brands of tea and region to which respondent belongs. Assuming we have the data on these two variables from a study, the cross tabulation may look like this:

Brand	*Regionwise buyers (No.)*				
	North	*South*	*East*	*West*	*Total*
Brooke Bond	25	20	20	15	80
Lipton	10	15	20	5	50
Tata	15	15	10	30	70
Total	50	50	50	50	200

This is a cross tabulation of two variables. An extension of this could be adding percentages.

Calculating Percentages in a Cross-tabulation

For computing percentages in a cross-tabulation, however, there is a problem which needs to be addressed. There are two or three different ways percentages can be calculated. For example, in the above example, we can compute percentages row-wise, column-wise, or on the total sample of 200.

The interpretation of percentages is different in each of the three cases. So which way is right?

The general rule for percentage calculation is to calculate it across the dependent variable. In the above example, we may assume that brand preference depends on the region to which respondents belong. In other words, 'brand' is the dependent variable, and 'Region' is the independent variable. The rule says that percentages must be calculated across brand categories—that is, column-wise. This appears to be the better interpretation, because the interpretation is—'Out of 50 respondents from the northern region, 50% buy Brooke Bond, 20% buy Lipton, and 30% buy Tata tea. All these percentages can be displayed in a table form separately, or in brackets along with number of respondents. The table of percentages along with numbers will look like this—

BRAND		*Regionwise Buyers—Numbers and Percentage*				
		North	*South*	*East*	*West*	*Total*
Brooke Bond	(No.)	25	20	20	15	80
	(%)	(50%)	(40%)	(40%)	(30%)	(40%)
Lipton	(No.)	10	15	20	5	50
	(%)	(20%)	(30%)	(40%)	(10%)	(25%)
Tata	(No.)	15	15	10	30	70
	(%)	(30%)	(30%)	(20%)	(60%)	(35%)
Total	(No.)	50	50	50	50	200
	(%)	(100%)	(100%)	(100%)	(100%)	(100%)

The above table can be interpreted according to the column (region) we are looking at. The first four columns represent findings for each region, and the fifth column (Total) represents overall findings for all the regions on an average. For example, from column 4, 30% of buyers in the west prefer Brooke Bond, 10% Lipton, and 60% prefer Tata tea. From column 5, out of the total 200 respondents, across all regions, 40% prefer Brooke Bond, 25% Lipton, and 35% Tata tea.

Cross-tabulation of More than Two Variables

It is possible to have cross-tabulations of three or more variables in a table. But most people find it difficult to assimilate information contained in three-variable cross-tabulations. For the purpose of drawing bivariate analysis conclusions, a two-variable cross-tabulation is quite adequate. A series of two-variable cross-tabulations can be performed on the important variables in the questionnaire. It is for the researcher to decide which variables need to be cross-tabulated. It is very easy to overdo the cross-tabulations, and too many of these may end up confusing the researcher or his client. It is a good idea to do only those cross-tabulations which are likely to help in the analysis and to draw useful conclusions.

Lack of Causal Inference in Cross-tabulations

It must be mentioned here that any two variables can be cross-tabulated. Even if the cross-tabulation shows a significant association between the two variables, it does not necessarily mean that one of them (the independent) causes the other (the dependent). Causality or direct effect is more of an assumption made by the researcher based on his expectation or experience. The mere existence of a statistically significant association does not necessarily imply a cause-and-effect relationship between the (presumed) independent and the (presumed) dependent variable.

The Chi-squared Test for Cross-tabulations

In the case of cross-tabulations featuring two variables, a test of significance called the Chi-squared test can be used to test if the two variables are statistically associated with each other significantly. The

user who is analysing the data on the computer and using a statistical package, can request a Chi-squared test along with any cross-tabulations. Commands such as CROSSTABS or CROSSTABULATION on most statistical packages have the option of doing a Chi-squared test.

In the manual technique, a Chi-squared statistic had to be calculated from the numbers in the cross tabulation. This had to be compared with the Chi-squared value from the Chi-squared tables for the given degrees of freedom, and a given confidence level. But in the computer user's case, none of these manual steps are needed. An illustration will explain how to perform and interpret the Chi-squared test on cross-taulations using a computer.

Chi-squared Test: An Illustration

Let us assume that we have conducted a consumer survey for a brand of detergent. One of the questions dealt with income category of the respondent. Another asked the respondent to rate his purchase intention. These two variables are listed in Table 8.1. Both variables are coded. Income codes and their equivalent incomes are:

Code	*Income in Rs. per month*
1	Less than 5000
2	5001 to 10,000
3	10,001 to 20,000
4	Above 20,000

Purchase intention codes are as follows:

Code	*Explanation (value labels for the variable)*
1	None—No intention to buy
2	Low—Low intention to buy
3	High—High intention
4	Very high—Very high intention
5	Certain—Certain to buy

These two variables were cross-tabulated from a sample of 20 respondents for the sake of this illustration. A cross-tabulation with a Chi-squared test was requested from the computer package. The output is shown in Table 8.2.

The cross-tabulation shows the number of respondents falling into each cell (a cell is the combination of one INCOME category with one PURCHASE INTENTION category). For example, 2 respondents with income less than Rs. 5,000 per month said they had no intention of buying this detergent brand.

Is there a Significant Association Between Respondent Income and Purchase Intention?

The Chi-squared test basically answers the above question. At the lower part of Table 8.2, we have the results of the Chi-squared test. The first line of the Chi-squared test reads a significance level of 0.09690. This means the Chi-squared test is showing a significant association between these two variables at a 90 per cent confidence level (equivalent to 100 – 90 ÷ 100 or 0.10 significance level).

TABLE 8.1

S. No.	Income	Code	Intent	Intcode
1	Less Than 5000	1	None	1
2	Less Than 5000	1	Low	2
3	Less Than 5000	1	Low	2
4	Less Than 5000	1	None	1
5	Less Than 5000	1	High	3
6	5001–10000	2	Low	2
7	5001–10000	2	High	3
8	5001–10000	2	Very High	4
9	5001–10000	2	High	3
10	5001–10000	2	Low	2
11	10001–20000	3	High	3
12	10001–20000	3	Very High	4
13	10001–20000	3	Certain	5
14	10001–20000	3	High	3
15	10001–20000	3	Very High	4
16	Above 20000	4	High	3
17	Above 20000	4	Certain	5
18	Above 20000	4	Very High	4
19	Above 20000	4	Certain	5
20	Above 20000	4	Certain	5

TABLE 8.2 Income Per Month by Purchase Intention

INCOME PER MONTH in Rs.

Purchase Intent	*Code*	*Less than 5000*	*5000–10000*	*10000–20000*	*Above 20000*	*Total*
NONE	1	2	0	0	0	2
LOW	2	2	2	0	0	4
HIGH	3	1	2	2	1	6
V. HIGH	4	0	1	2	1	4
CERTAIN	5	0	0	1	3	4
TOTAL		5	5	5	5	20

Chi-Square	Value	DF	Significance
Pearson	18.66667	12	.09690

Thus, we conclude that at 90 per cent confidence level, PURCHASE INTENTION and INCOME are associated significantly with each other. This may lead us to conclude that the price of the detergent is important in its purchase.

Like we said earlier, it is possible to do a cross tabulation (and a chi-squared test) for any two nominal variables in the survey. But it is a good idea to use the cross tabulation sparingly only for those variables where a link or association makes some sense theoretically.

Measures of the Strength of Association Between Variables

In our discussion of the Chi-squared test so far, we have only looked at the statistical significance by looking at the *p*-value (probability value) reported on the computer output. This does not tell us the strength of the association between the two variables in the cross tabulation. If we want a measure of the strength, we have to request the package to give us one of the following (these measures are called the indexes of agreement):

1. Contingency Coefficient C
2. Cramer's V
3. The Phi Correlation Coefficient
4. Goodman and Kruskal's Lambda Asymmetric Coefficient

We will briefly discuss these ***indexes of agreement***, as these measures are known.

The Contingency Coefficient lies between 0 and 1, and can be used for any cross-tabulation with any number of rows (R) and any number of columns (C), provided R and C are equal (symmetric cross-tabulation). However, it cannot attain the maximum value of 1. The maximum value of the Contingency Coefficient depends on the number of rows and columns in the cross-tabulation. For instance, it can be a maximum of .707 in a 2×2 table, and a maximum of .87 in a 4×4 table.

Cramer's V is a variation of the Phi Correlation Coefficient, but it is not restricted to 2×2 tables. It can have a maximum value of 1.

Phi Correlation Coefficient is used mainly for 2×2 contingency tables (cross-tabulations) because otherwise its value can go beyond the 0-1 range, which becomes difficult to interpret.

Lambda Asymmetric Coefficient measures the error reduction in predicting the value (category) of one variable (say, the column variable), if we know the category (or value) of the other (say, row) variable. Thus, if Lambda (for the row variable, given the column variable), is 0.43, the reduction in error in predicting the row variable value, given the column variable value is 0.43, or 43 per cent. Similarly, we could compute Lambda Asymmetric for the row variable, given knowledge of the column variable. Also, Lambda Symmetric could also be computed as a weighted average of the above two Lambda Asymmetric values (for the row and the column variables).

All these indexes of agreement can be requested on SPSS or other computer packages. Generally one or two of them are sufficient to find out if the association between the row and column variable in the cross-tabulation is weak (close to 0) or strong (close to 1).

DOING MORE WITH DATA

TRANSFORMATION OF VARIABLES AND USE OF PART-SAMPLES

Using Rows and Columns in Different Ways

So far, we have assumed that all the data (rows and columns) that are typed into SPSS (or imported from an EXCEL data file) are used in our analysis. But this need not be the case. We may actually want to use parts of the dataset for some of our analyses. Or, after completing the analysis of the full dataset, we may want to look at some parts of the data for a more complete understanding. This can lead to some rows of the data being selected for analysis (for example, people from one city, or one income group, or one usage category).

It is also possible that we need to make changes in the variables (columns in a data matrix) after we enter the raw data from a survey questionnaire. For example, we may want to combine two variables into a new one or construct a new variable which is a function of some of the old ones. We have an easy way of doing this using the TRANSFORM/COMPUTE command in SPSS (explained at the end of the chapter). Essentially, this creates a new data column (or columns) containing the transformed variable based on the earlier raw data columns.

An example would be calculating a new variable which is the product of two existing variables, or one divided by another. In financial data, for example, income and assets could be two variables, and ROA could be the new variable, found as INCOME divided by ASSETS. In Sales, productivity could be calculated as the Sales achieved, divided by number of Target Customers met. To do these kinds of transformation, a new variable can be computed using the COMPUTE command (available under the TRANSFORM menu) on SPSS.

Why Recoding

A situation where you need to recode the existing data can arise from the need to relook at the scale of measurement. For example, having collected data on a 7 point scale, we may want to look at only the top end of the scale (say, Scores of 6 and 7) for a certain analysis. We can then RECODE the existing variable into a new variable, which will list out only the top scores of 6-7 on the old variable. All other scores on the old variable (1-5) will be coded separately. For instance, we could specify, for a 7 point-scaled old variable called Customer Satisfaction (CUSTSAT), a new variable called NEWSAT with only 2 values as follows

Old Values (CUSTSAT)	New Variable Values (NEWSAT)
1 – 5	1
6 – 7	2

This would create separately a new variable (in a column with the new label) with only two values, for which an easy procedure called RECODE is available in SPSS (its use is explained in detail below).

Transforming Variables

Two important commands under the TRANSFORM menu are COMPUTE and RECODE. We will explain how to use both.

Compute

Example 1: Let us say, we have 4 variables, V1, V2, V3, V4 in our database. We now want to create a new variable, which is the Average or Mean of all these.

1. Go to **TRANSFORM** on the main menu, and choose **COMPUTE**.
2. In the dialog box that opens, type in the name of the **TARGET VARIABLE** (the new variable you want to create. This will appear as an additional column after it is computed). For example AVGCOST could be the new variable's name (average cost to be computed from 4 variables, say, V1, V2, V3 and V4—these four could be the distribution cost, advertising cost, sales promotion cost and administrative cost).
3. Go to the '**Function Group**' on the dialog box and choose **STATISTICAL**. In the options that open up below, choose **MEAN.**
4. Click on the ↑ **Up** arrow to move the '**Mean**' into the **Numeric Expression** box at the top.
5. In the **Mean [? ?]** that appears in the *Numeric Expression* box, fill the parentheses by selecting the variables V1, V2, V3, V4 by clicking on each, then on the → (Right Arrow) to move them into the parentheses, so that it looks like this – **Mean [V1,V2,V3,V4].**
6. Click on **OK**. Your new variable AVGCOST will appear in the last column of the dataset.

Recode If you want to recast a variable, say, from an interval scale to nominal scale or create a new categorical variable by collapsing / combining the categories of the existing variable, you can do it using the RECODE command.

1. Go to **TRANSFORM** and click on **RECODE**.
2. Choose **'Into Different Variable'**. This basically means you will create a new column for the recoded variable, and not overwrite the existing variable (column).
3. In the dialogue box that opens up, select an old variable you want to recode. Assume that we are trying to change **V5**, a numerical variable, into 2 categories **Low** and **High**. The original variable V5 has values from 1 to 7. We want to Recode it as **1-4 = Low**, **5-7 = High** and will use 1 as a code for Low and 2 as a code for High: Essentially, any value of V5 from 1 to 4 will be recoded as 1, and any value of V5, from 5-7 will be recoded as 2. This new variable will be called **NEWV5**.
4. Select V5 and move it into the top box titled **'Input Variable' => 'Output Variable'**.
5. Define (name) the Output Variable by giving it a Name (for example NEWV5) and Label (same or different from Name). Click '**Change**' below the box.
6. Click on '**Old and New Values**' in the middle of the dialogue box. A new dialogue box opens up.
7. Fill '**Old Values Range**' by choosing RANGE and fill in **1 through 4**.

8. On the right side, type in the **New Value of the variable** (1 in this case). This means that the NEWV5 =1 when V5 has any value from 1 to 4.
9. Click **Add**.
10. Type in the **old value RANGE 5 through 7, and the New Value of 2**. This means that the old values of V5 from 5 to 7 will be recoded into NEWV5=2.
11. Click **Add**.
12. Click **CONTINUE**. This takes you back.
13. Click **OK** to make the changes. Now, NEWV5 can be used just like any other variable. The variable NEWV5 appears in the last column of your data file.

Using a Part (Sub-sample) of the Data Collected

When we want to perform certain analyses on a sub-sample of the data file, two useful commands on the DATA menu are **Select Cases** and **Split File**. Their use is discussed below. In both these, we are working for selecting certain rows of data to create sub-samples of the data collected.

Multi Centre Survey: Analysis by City

For example, we may have done a survey in four centres—Mumbai, Bangalore, Delhi & Hyderabad.

Let us assume these cities are entered as one variable, coded as 1, 2, 3, 4 for each city. The variable name could be CITY. If we want to do a particular analysis (say, Tabulation or Regression), on the data from Mumbai alone, we can make use of the SELECT CASES option from the DATA menu in SPSS. It enables you to filter out all the data except Mumbai (Code=1). This process can be used for any segmented analysis of the data set.

This is followed by the analysis (say, a frequency tabulation or any DESCRIPTIVES or regression) we want to perform. Only the Mumbai sub-sample will be analysed.

If we want to do the analysis for all the 4 cities (in the above example), we can either repeat the SELECT CASES option for each city and do the analysis, or make use of the SPLIT FILE option along with 'ORGANISE OUTPUT BY GROUPS' within that. This enables the analysis by groups on the grouping variable (in this case, CITY is the grouping variable. It could also be any other segmentation variable, say Gender—Male / Female, or Frequent Users / Infrequent Users, etc). If we want a quick comparison of Means of a given variable by demographic groups, this command is quite useful.

Another Example of the SELECT CASES Command—Selecting a Subset of the Cases in a Data File Using a Condition

If you want to select a part of the data set for analysis based on a condition, you can use this command to select cases based on any condition that can be specified using an **If** statement, as follows—

1. Go to the **DATA** Menu.
2. Choose '***Select Cases***' .

3. Specify the condition for selection of cases by choosing '***If condition is Satisfied***' and then specifying the condition in the '**If**' dialog box. For example, you can specify a condition like "**If Variable 2 > 14**", by choosing **Variable 2** from the variable list, **' >'** from the list of available functions, and '**14**' from the list of available digits.
4. Click on **CONTINUE** after you specify the condition.
5. Click **OK** to activate your choice. This will filter out all cases, that do not satisfy the '**If**' statement you specified, and only these cases which satisfy the condition will be used in any ANALYSIS (regression, ANOVA or any other) that you perform.
6. To go back to the entire data set, again choose **SELECT CASES** from the DATA Menu, and choose '**All Cases**' from the dialog box and click 'OK'. This will make all the cases available for analysis.

Note that you can specify any condition here, in terms of any existing variable. You do not need to have variables coded, like you would for the Split File command.

Same Analysis to be Repeated on Sub-Groups of Data: An Example of Using the SPLIT FILE Command

This feature can be used when an analysis is to be done for many sub-samples (not just one) simultaneously, and the grouping variable in the data set is already coded. Suppose Variable 8 has four groups 1,2,3,4. These could be Income groups. If you want some tables (or any other analysis) to be done by these groups (separate computations for each income group), you can follow the instructions below:

1. From the **DATA** menu, choose **SPLIT FILE**.
2. Choose '**Organize output by Groups**'.
3. Select a variable and enter it into the box titled '**Groups Based on**'. This is the variable whose groups you want to analyse separately (say, INCOME)
4. Choose '**Sort the file by grouping variable**'.
5. Click OK.

Now, you can go through with any analysis. It will be done separately for each group of the chosen variable (say, Variable 8, which has 4 groups of income). That is, four separate outputs will appear, one for each income group.

SUMMARY

Analysis of survey based data starts with simply tabulating the collected data. Before we do this, data is assumed to be coded if it is nominal scaled (categorical). If we are using SPSS, value labels for nominal data variables must also be input and saved.

A table can be generated for each question on our survey questionnaire by using the FREQUENCIES command as described below. Usually, percentages are also generated by these tables, and are easier to handle than numbers. This is the first level of analysis, from which many conclusions can be drawn.

After this is done, we can examine the variables to see if we need any two-variable analysis. This can be to check if there is a relationship between two variables. For example, is the *Purchase intention* different according to the respondent's *Age*? If we want to examine this kind of hypothesis, we can do a cross-tabulation of these two variables, and perform a chi-squared test to check the hypothesis. If the chi-squared test is significant, there is a relationship between the two variables in the cross-tab.

In addition to the chi-squared test, there are measures of the strength of association between two variables which can be requested from the SPSS package. Some of these are the contingency coefficient C, Cramer's V and Lambda. These measures tell us whether a strong relationship exists among pairs of variables where chi-squared test has shown a significant relationship.

ASSIGNMENT QUESTIONS

1. Explain the concept of dependent and independent variables.
2. For the following pairs of variables, indicate which one is independent and which is dependent according to you.
   (a) Sales and number of salespeople
   (b) Product quality and sales
   (c) Service quality and satisfaction
   (d) Age and income
   (e) Brand preference and gender
   (f) Frequency of purchase and family size
3. What are demographic variables? Why do we need to study them in marketing research?
4. Why are percentages used in univariate tables? Do they explain the findings better than numbers alone?
5. Why should each table in an analysis mention the sample size?
6. What is a cross-tabulation? Give an example of a 2-variable cross-tabulation.
7. What do we mean by causal inference?
8. Explain the rule for computing percentages in a cross-tabulation.
9. What does a Chi-squared test do in a given cross-tabulation of two variables?
10. Explain how to interpret from the computer output of a Chi-squared test whether there is a significant association between the two variables in a cross-tabulation. If the significance level of a Chi-squared test is 0.02 on an output, is the test significant at a confidence level of 95 per cent? Explain.

SPSS COMMANDS FOR FREQUENCY TABLES, AND CROSS-TABS WITH CHI-SQUARED TEST

After the input data has been typed along with variable labels and value labels in an SPSS data file, to get the frequency tables output for a problem similar to that described in Chapter 8 of the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on DESCRIPTIVE STATISTICS, followed by FREQUENCIES.
3. on the dialogue box which appears, select the variables for which FREQUENCY TABLES are required, by clicking on the right arrow to transfer them from the variable list on the left to the VARIABLES box on the right.
4. click OK to get the tables with counts and percentages, for each of the selected variables.

Note: Charts can be requested by clicking on CHARTS on the main dialogue box, selecting the required type of charts, and clicking CONTINUE before step 4 above. Similarly, if the variables are interval-scaled, you can click STATISTICS on the dialogue box and request Means, Standard Deviations, and so on for each variable.

CROSS-TABS and Chi-squared Test

After the input data has been typed along with variable labels and value labels in an SPSS data file, to get the CROSS-TABULATIONS and Chi-squared test output for a problem similar to that described in Chapter 8 of the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on DESCRIPTIVE STATISTICS, followed by CROSS-TABS.
3. select the row variable for a cross-tabulation by highlighting it in the variable list on the left side and clicking on the arrow leading to the row variable box. Similarly, select the variable you wish to be the column variable in the cross-tabulation.
4. click on STATISTICS in the main dialogue box. Then click on ‘Chi-Square’. In the box titled ‘Nominal’, click on ‘Contingency coefficient’, ‘Phi and Cramer’s V’, and ‘Lambda’ to give you these statistics associated which measure the strength of the association in a cross-tab. Click CONTINUE to return to the main dialogue box.
5. click on CELLS in the main dialogue box. Under 'Percentages’, select either 'ROW' or 'COLUMN' depending on which is desired, as per the discussion and rule given in the text. Click CONTINUE to return to the main dialogue box.
6. click OK to get the output containing the required cross-tab, along with the Chi-squared test and the measures of association like Lambda and Contingency Coefficients.

Note: The Chi-squared test requires counts to be in the cross-tables, and not percentages. Original data should have counts when using this test.

CASE STUDY 1

Chi-square Test for Cross-tabs*

Problem

In this case study, we observed the association between educational background (independent variable) of PGDM students and their performance in terms of grade (dependent variable) secured.

A bivariate cross-tabulation has been done by combining the above two variables and tabulating the data together. Though it is not necessarily a fact that the independent variable (educational background) causes a change in the dependent variable (grade secured), direct effect is an assumption made by our group based on information extracted from the database (performance) of B-schools. We wanted to test at 90% and 95% confidence level, what is the level of significance of association between **EDUCATIONAL BACKGROUND OF PGDM STUDENTS and THEIR PERFORMANCE IN TERMS OF GRADE.**

Further, both the variables are coded as 'code' and 'grade code' representing educational background and grade obtained. Educational backgrounds and their equivalent codes are as follows:

Educational Background	*Code*
B.Com.	1
B.E.	2
B.Sc.	3
B.B.A	4
B.A.	5

Grade codes are as follows:

Grade Obtained	*Grade Code*
A	1
B	2
C	3

These two variables were cross-tabulated for twenty-five observations. A cross-tabulation with a Chi-squared test was requested from the computer SPSS package, the input and output of which are shown in the following tables.

*Prepared by Biswajit Mishra, Imran Ahmed, Vikrant Sharma, and Sharmistha Kundu

Input Data Table

	rollno	*background*	*code*	*grade*	*grdcode*
1	1	B.Com	1	B	2
2	2	B.Com	1	C	3
3	3	B.Com	1	A	1
4	4	B.Com	1	C	3
5	5	B.Com	1	B	2
6	6	B.E.	2	A	1
7	7	B.E.	2	A	1
8	8	B.E.	2	A	1
9	9	B.E.	2	B	2
10	10	B.E.	2	A	1
11	11	B.Sc.	3	B	2
12	12	B.Sc.	3	B	2
13	13	B.Sc.	3	C	3
14	14	B.Sc.	3	C	3
15	15	B.Sc.	3	C	3
16	16	BBA	4	A	1
17	17	BBA	4	B	2
18	18	BBA	4	C	3
19	19	BBA	4	C	3
20	20	BBA	4	B	2
21	21	B.A.	5	C	3
22	22	B.A.	5	C	3
23	23	B.A.	5	C	3
24	24	B.A.	5	C	3
25	25	B.A.	5	B	2

Output Table: Grades Versus Entry Qualification

		B.Com *1*	*B.Engg.* *2*	*B.Sc.* *3*	*BBA* *4*	*B.A.* *5*	*Row Total*
Grade A	1	1 20.0%	4 80.0%	0	1 20.0%	0	6 24.0%
Grade B	2	2 40.0%	1 20.0%	2 40.0%	2 40.0%	1 20.0%	8 32.0%
Grade C	3	2 40.0%	0	3 60.0%	2 40.0%	4 80.0%	11 44.0%
	Column Total	5 20.0%	5 20.0%	5 20.0%	5 20.0%	5 20.0%	25 100.0%

Chi-Square	*Value*	*DF*	*Significance*
Pearson	13.75000	8	.088
Likelihood Ratio	15.58135	8	.048
Linear-by-Linear Association	3.63000	1	.056

Minimum Expected Frequency – 1.200

Cells with Expected Frequency < 5 – 15 of 15 (100.0%)

Statistic		*Value*	*ASE 1*	*Val/ASE 0*	*Approximate Significance*
Contingency Coefficient		.59568			.0885
Lambda:					
symmetric		.26471	.15211	1.59699	.1102
with GRDCODE	dependent	.28571	.20912	1.18678	.2353
with CODE	dependent	.25000	.14361	1.58114	.1138
Goodman & Kruskal Tau:					
with GRDCODE	dependent	.25743	.09603		.1360
with CODE	dependent	.13750	.06503		.1051

Number of Missing Observations: 0

Conclusion and Analysis

The Chi-square test revealed the significant association between the educational background of the students and their performance in terms of grade. From the Chi-square test output table we see that a significance level of 0.08852 (Pearson's) has been achieved. This means the Chi-square test is showing a significant association between the above two variables at 91.15% confidence level (100 – 8.85).

Thus we conclude that at 90% confidence level, EDUCATIONAL BACKGROUND OF PGDM STUDENTS and THEIR PERFORMANCE IN TERMS OF GRADE are associated significantly with each other, whereas this is not significant at the 95% confidence level. From the obtained contingency coefficient (C) of 0.59568, it can be inferred that the association between the dependent and independent variable is significant, as the value 0.59568 is closer to 1 than to 0. Also from the lambda asymmetric value (with GRD CODE dependent) of 0.28571, we conclude that there is a moderate level of association between the above two variables. This lambda value tells us that there is a 28.5% reduction in predicting the grade of a student when we know his educational background.

This leads us to conclude that educational background plays a vital role in the performance of the students of PGDM course.

CASE STUDY 2

Chi-square Test for Cross-tabs*

Methodology

The data generated for this assignment has been taken from an earlier survey** on detergents (Surf). In this report, we have attempted to inter-relate two demographic details of our respondents, which formed a part of our survey. The basic objective of this assignment is to find out whether or not, the income of the household plays a determining factor in deciding the person responsible for washing clothes, that is, whether the clothes are washed by:

- themselves
- maids
- any other (e.g.: a combination of the above, launderettes etc.)

The following steps were followed while formulating this report:

- The data for the assignment was collected as stated above.
- The various categories under each variable (dependent and independent) were coded. This data was incorporated from SPSS (Statistical Package for Social Sciences) software, which was used by our group in our earlier report.
- A cross-tabulation was done taking the two variables as stated above. This was done with the help of SPSS. The cross-tab generated the output that showed the number and percentage of respondents in each category.
- This was followed by a Pearson's Chi-square test, which showed the significance of the inter-relationship between the two variables.
- The relationship thus obtained was analysed and the relevant conclusions were noted.

Input Data

We are attempting to find out if there is any significant relationship between two variables, namely income and person(s) washing the clothes at home. This is in reference to an earlier suvey on detergents (Surf) conducted by our group. The following data has been taken from the same.

*Prepared by Harakuni Rajiv, Jitender Singh, Siddharth Agarwala, and Vidya Thekke Cherupilli
**By the same group

S. No.	A	B	S. No.	A	B
1	2	1	21	4	3
2	4	2	22	4	2
3	3	1	23	3	2
4	2	1	24	2	2
5	2	1	25	4	2
6	4	2	26	2	2
7	1	2	27	1	2
8	1	1	28	2	2
9	2	2	29	2	1
10	3	1	30	1	1
11	4	3	31	3	2
12	4	2	32	3	3
13	3	1	33	4	2
14	2	2	34	2	2
15	2	2	35	4	3
16	3	2	36	1	1
17	2	1	37	4	2
18	1	1	38	4	2
19	2	2	39	3	3
20	3	2	40	3	3

A→Income Code

1: < than 5,000

2: 5,001–10,000

3: 10,001–15,000

4: > than 15,000.

B→Person washing clothes at home

1: Yourself

2: Maid

3: Any other

Explanation of Output

1. Explanation of Cross-tabulation

Person washing clothes at home	*Income Group*				
	< than 5,000	*5,001 to 10,000*	*10,001 to 15,000*	*> than 15,000*	*Total*
Yourself (No.) (%)	4 66.67%	5 35.70%	3 30%	0 0%	12 30%
Maid (No.) (%)	2 33.33%	9 64.30%	4 40%	7 70%	22 55%
Any other (No.) (%)	0 0%	0 0%	3 30%	3 30%	6 15%
Total (No.) (%)	6 100%	14 100%	10 100%	10 100%	40 100%

In the above cross-tabulation, we have tried to find the inter-relationship between the income groups of the respondents (independent variable) and the person washing clothes at home (dependent variable). The reason for income group being the independent variable is because it has been found that generally income of the household determines whether domestic help can be employed or the same need can be met by the respondents themselves.

We observe from the table that as income rises, percentage of people employing maids also increases. Another interesting observation made was that some households had their inexpensive and daily wear clothes washed by maids while the expensive ones were taken care either by themselves or given to launderettes.

2. Explanation of Pearson's Chi-square

Chi-square	*Value*	*DF*	*Significance*
Pearson	13.39105	6	0.03723

Is there a significant association between the respondent's income and the person(s) washing clothes at home?

The small value of Pearson's Chi-square test clearly states that there exists a significant inter-relationship between the dependent and independent variables.

The Chi-squared test is carried out at a 90 per cent confidence level (equivalent to 100 – 90 divided by 100 or 0.10 significance level).

3. Explanation of Contingency Coefficient and Lambda

Contingency coefficient

Statistic	*Value*	*Approximate Significance*
Contingency Coefficient	0.50081	0.03723 (*1)

*1: Pearson's Chi-square probability

The contingency coefficient gives us the measure of strength of the output. If the values close to 0, there is no strong correlation between the two variables. However, if the value ranges between 0.5 and 1, there exists a strong correlation.

From the above table, we can therefore conclude that there exists a correlation between the independent variable (income group of the respondent) and the dependent variable (person washing clothes at home).

Lambda

Statistic	*Value*	*ASE1*	*Val/ASE0*
Symmetric	0.11364	0.06088	1.72774
Dependent variable	0.11111	0.12830	0.82339

Lambda is a measure of reduction in error in measuring the association between the two variables. For example, if lambda = 0.3, it implies that it is leading to a 30% reduction in error in estimating or predicting one variable from the other. In our case, from the table, the value of lambda is .1111, which means there is a 11.11 per cent error reduction. This is a moderate or small value, and therefore we conclude that there is a moderate relationship between the two variables, but that is statistically significant (from the earlier result of chi-squared test).

CASE STUDY 3

Chi-square Test*

Methodology

1. A fictitious data set consisting of thirty respondents was created. The data was mainly constructed to find the relationship between the dependent and independent variable. Age was taken as the independent variable and choice of a drink as dependent variable. Six brands of soft drinks were considered as the different choices for the respondents.
2. The age group was coded into six categories as 1 to 6 and the brands of soft drinks were coded into six categories and the codings are as follows:

   (a) Independent variable

Age	*Coding*
<15	1
16–25	2
26–35	3
36–45	4
46–55	5
>55	6

   (b) Dependent variable

Different brands	*Coding*
Coke	1
Pepsi	2
Mirinda	3
Sprite	4
Slice	5
Fruit Juice	6

3. Chi-square test has been used to cross-tabulate and to understand the relationship between the independent and the dependent variable
4. Calculation of contingency coefficient and the lambda asymmetric coefficient is done to find the strength of the association between the two variables.
5. Sample size is taken as thirty.

*Prepared by Gangambika Deshnur, Mallika Ratnam, Piyush Verma, Raji A, and Shilpa George

6. Analysis of cross-tabulation.
7. SPSS software package for the cross tabulation analysis.

Problem

This is a bivariate problem. The basic intention of the problem is to understand the relationship between AGE and BRAND PREFERENCE of different brands of soft drinks.

Input Data Table

Serial No.	*AGE*	*AGECODE*	*SOFT DRINK*	*DRINKCODE*
1	< 15	1	FRUIT JUICE	6
2	< 15	1	SPRITE	4
3	< 15	1	MIRINDA	3
4	< 15	1	PEPSI	2
5	< 15	1	FRUIT JUICE	6
6	16–25	2	COKE	1
7	16–25	2	SLICE	5
8	16–25	2	COKE	1
9	16–25	2	PEPSI	2
10	16–25	2	MIRINDA	3
11	26–35	3	SLICE	5
12	26–35	3	SPRITE	4
13	26–35	3	FRUIT JUICE	6
14	26–35	3	PEPSI	2
15	26–35	3	SLICE	5
16	36–45	4	MIRINDA	3
17	36–45	4	FRUIT JUICE	6
18	36–45	4	FRUIT JUICE	6
19	36–45	4	SLICE	5
20	36–45	4	PEPSI	2
21	46–55	5	COKE	1
22	46–55	5	SPRITE	4
23	46–55	5	SLICE	5
24	46–55	5	FRUIT JUICE	6
25	46–55	5	SLICE	5
26	> 55	6	MIRINDA	3
27	> 55	6	COKE	1
28	> 55	6	COKE	1
29	> 55	6	PEPSI	2
30	> 55	6	FRUIT JUICE	6

Output Data

Age by Drink Preference

Age ⟶

Drink Preference	*Code*	*<15*	*16–25*	*26–35*	*36–45*	*46–55*	*>55*	*Total*
Coke	1	0	2 33.32%	0	0	1 20%	2 40%	5 16.67%
Pepsi	2	1 20%	1 16.67%	1 25%	1 20%	0	1 20%	5 16.67%
Mirinda	3	1 20%	1 16.67%	0	1 20%	0	1 20%	4 13.33%
Sprite	4	1 20%	0	1 25%	0	1 20%	0	3 10%
Slice	5	0	1 16.67%	2 50%	1 20%	2 40%	0	6 20%
Fruit Juice	6	2 40%	1 16.67%	0	2 40%	1 20%	1 20%	7 23.33%
Total		5 100%	6 100%	4 100%	5 100%	5 100%	5 100%	30 100%

Chi-Square	*Value*	*DF*	*Significance*
Pearson	18.22857	25	.08325
Likelihood Ratio	25.52646	25	.04332
Mantel-Haenszel test for linear association	.13961	1	.07086

Minimum Expected Frequency—.500

Cells with Expected Frequency < 5—36 of 36 (100.0%)

Approximate Statistic	*Value*	*ASE 1*	*VAL/ASE 0*	*Significance*
Contingency Coefficient	.61479			.08325*1
Lambda:				
Symmetric	.18750	.08892	1.99754	
With 'DRINK CODE' dependent	.21739	.12757	1.56813	
With 'AGE CODE' dependent	.16000	.07332	2.14834	
Goodman & Kruskal Tau:				
With 'DRINK CODE' dependent	.12432	.03912		.08412*2
With 'AGE CODE' dependent	.12152	.02580		.08580*2

*1 Pearson Chi-square probability

*2 Based on Chi-square approximation

Number of Missing Observations: 0

Analysis

In a Chi-square test, for a 90 per cent confidence level, if the significance level is greater than or equal to 0.1, it signifies that there is no association between the two variables in the cross-tabulation and if significance level is less than 0.1, then it signifies that there is a significant relationship between the selected variables.

The result of the cross-tabulation

From the output tables, the Chi-square test read a significance level of 0.08325 at 90 per cent confidence level. For 90 per cent, significance level is 0.1, that is, (1–0.9), so the above result shows that at 0.08 (which is less than 0.1), there is a significant relationship between the two variables. At 95 per cent confidence level, significance level being 0.05, and the above output giving a significance level of 0.08 which is greater than 0.05, there is no relationship between the variables:

If contingency coefficient value is greater than + 0.5 then the variables are strongly associated. In the above case the contingency coefficient value being 0.6 which is greater than 0.5, hence the variables are strongly associated.

The asymmetric lambda value (with DRINKCODE dependent) 0.21739 means that 21.7% of error is reduced in predicting brand preference when age is known.

From the above result we can conclude that there is a significant relationship between **AGE** (independent variable) and **BRAND PREFERENCE** (dependent variable), of the respondents.

Thus we can conclude that the age of the respondent plays an important role in the purchasing intention of a particular brand of soft drink.

CHAPTER 9 ANOVA AND THE DESIGN OF EXPERIMENTS

Learning Objectives

In this chapter, we will

- Introduce the reader to experimental designs like the Completely Randomised, Randomised Block, Latin Square and the Factorial Design
- Introduce Analysis of Variance or ANOVA as a technique for analysing marketing research problems
- Illustrate the use of ANOVA and the testing of hypotheses using the three different types of experimental designs
- provide three case studies with fictitious data at the end of the chapter to further enhance understanding of the applications of ANOVA in marketing research

INTRODUCTION

In most marketing research applications, a survey of some sort is the method used, whether it is conducted through mail, a personal interview, over the phone, or more recently, on the Internet. There are however, other classes of study available, one of which is observation. The other widely used class of study is known as experimentation. Just like in a laboratory, we manipulate certain variables (usually marketing related ones in marketing research), and observe changes in other variables (for example: sales, or consumer behaviour, or attitude).

APPLICATIONS

The application areas for experiments in marketing research are wide. Whenever a marketing mix variable (independent variable) such as price, a specific promotion, or type of distribution, even

specific elements like shelf space, or colour of packaging, and so on is changed, we would want to know its effect. Under proper conditions, an experiment can tell us the effects of specific variations in one or more elements of the marketing mix. Therefore, the potential application areas are quite wide-ranging. The experiment can be done either with only one independent variable (factor) or with multiple independent variables.

METHODS

A one-independent variable experiment is called one-way ANOVA. ANOVA stands for Analysis of Variance, the generic name given to a set of techniques for studying the cause-and-effect of one or more factors on a single dependent variable. When more than one dependent variable is studied, the technique called MANOVA or Multivariate Analysis of Variance is used. However, we will limit ourselves to the discussion of three major types of ANOVA in this chapter.

VARIABLES

The Analysis of Variance technique is used when the independent variables are of nominal scale (categorical) and the dependent variable is metric (continuous), or at least interval scaled.

EXPERIMENTAL DESIGNS

The design of the experiment is the most critical when performing any experiment to be analysed through the technique of ANOVA. There are four major types of designs, of which three frequently used types will be illustrated with a worked example for each. These four major types are:

1. Completely Randomised Design in a One-Way ANOVA (single factor)
2. Randomised Block Design (single blocking factor)
3. Latin Square Design (two blocking factors)
4. Factorial Design with 2 or more factors.

We will discuss in detail the first two, and the fourth.

Completely Randomised Design in a One-way ANOVA

This particular design is used when there is only one categorical independent variable, and one dependent (metric) variable. Each category of an independent variable is called a level. The independent variable may be different levels of prices, or different pack sizes, or different product colours, and the effect (dependent variable) could be the sales of the product.

Worked Example In this worked example, we assume that three different versions of advertising copy have been created by an advertising agency for a campaign. Let us call these versions of copy AD COPY 1, 2, and 3. Now, the ad agency wants to test which of these three versions of the advertising copy is preferred by its target population, before they launch the campaign.

A sample of 18 respondents is selected from the target population in the nearby areas of the city. At random, these 18 respondents are assigned to the 3 versions of the ad copy. Each version of the ad copy is thus shown to six of the respondents. The respondents are asked to rate their liking for the ad copy shown to them on a scale of 1 to 10 (1 = not liked at all, 10 = liked a lot). The ratings given by the 18 respondents are tabulated.

Input Data Table 9.1 shows the input data for the 18 respondents. In column 1, the 'ad copy', is coded (1, 2, 3) indicating three different versions of the ad. The second column, 'rating', is the rating given by a respondent. Thus, six respondents have rated each ad. Please note, that these eighteen respondents were randomly assigned to each of the three ad versions. This random assignment is called a completely randomised assignment or design.

The input data in Table 9.1 is now used for performing a One-way ANOVA, because we have only 1 categorical factor (ad copy) at 3 levels—1, 2, 3, and 1 dependent variable—Rating. We are testing the null hypothesis that there is no difference in the Rating for the three ads.

Output The above input data were fed into an SPSS data file (Please refer to the appendices at the end of previous chapters for SPSS data input commands). Then, SPSS commands were used to perform a One-way ANOVA (these commands are described at the end of this chapter).

The output of the computerised One-way ANOVA algorithm is shown in Table 9.2. The first column is titled 'Source of Variation'. Under this, labelled Main Effects, is the single independent variable called AD COPY.

We then go to the last column, where the significance of the F-test is given. It is .203 in this case, for the AD COPY. This indicates that at the confidence level of 95 per cent, (corresponding to significance level of 0.05), the F-test proves the model is not significant. In other words, the ratings given to the three ad copy versions are not significantly different from each other. The ANOVA has thus told us what we may not have been able to gauge if we had simply looked at the mean ratings for each ad copy. For example, the ratings for the ad copy version 1 are 6, 7, 5, 8, 8, 8, and the mean rating is (6 + 7 + 5 + 8 + 8 + 8)/6, or 42/6 = 7. Similarly, the mean rating of ad copy version 2 is (4 + 4 + 5 + 7 + 7 + 6)/6, or 33/6 = 5.5. The mean rating for ad copy version 3 is (5 + 5 + 4 + 7 + 8 + 7)/6, or 36/6 = 6. At a glance, the three mean ratings appear to be different—7, 5.5, and 6. But the ANOVA tells us that this difference is not statistically significant at the 95 per cent confidence level. This it does by performing an F-test. The null hypothesis for this F-test is that there is no significant difference in the mean ratings for the three ad copy versions.

(H_0: $M_1 = M_2 = M_3$ where M_1, M_2, and M_3 are the mean ratings for the three versions of the ad copy). Thus, in this case, we have accepted the null hypothesis (or failed to reject the null hypothesis), at the 95 per cent confidence level. If the significance of F in the last column of Table 9.2 had been less than 0.05, we would have rejected the null hypothesis. In that case, we would have concluded that significant differences exist between mean ratings given to the three ad copy versions.

Randomised Block Design

Let us continue with the same input data as in Table 9.1 with one more column added to it as shown in Table 9.3.

We have made a slightly different assumption in this case. We assume that the three versions of the ad copy were each used in 6 different magazines. These six magazines are coded 1, 2, 3, 4, 5, 6 and appear in the column titled 'magazine'. Out of the people who saw these ads, 18 randomly chosen respondents are picked each of whom has seen a particular version of the ad. Thus, we finally have one respondent who has seen a given version of the ad in a given magazine. In other words, we have one respondent for every combination of magazine and ad copy.

Hypothesis The assignment of our sample of 18 in the above manner assumes that the magazine in which the version of the ad copy appears may have an impact on the ratings. We can test this hypothesis—in fact, two hypotheses—by doing an ANOVA with a randomised block design. For this purpose, we use the variable 'rating' as the dependent variable, and the 'ad copy' as the factor, and 'magazine' as the block. A block is defined as some variable which could affect the relationship between the independent factor and the dependent variable under study in an ANOVA. In our example, the magazine in which the advertisement appears could influence the rating given to the ad copy by the respondents. We are trying to remove the effect of the magazine used, by 'blocking' its effect, or treating the block separately. If we do not block a variable, its effect gets included with the error (residual) term. This may lead to wrong conclusions about the relationship between the independent and dependent variables. In that sense, a randomised block design is more 'powerful' than a simple one-way ANOVA, if the block effect is significantly influencing the relationship.

Output The computer output for this problem using a randomised block design is shown in Table 9.4. This table is similar to the output table of the one-way ANOVA we got earlier (Table 9.2), except that there is an additional source of variation called 'magazine' in the first column of Table 9.4. This is the 'block' we have used, to test the null hypotheses.

1. The first null hypothesis is that the mean rating of the ad copy is the same for all 3 versions. This is the same as the null hypothesis we had used earlier for the one-way ANOVA.
2. The second null hypothesis is that the 'block' used (magazine in this case) has no effect on the mean ratings given to the AD COPY versions by respondents.

To test if the null hypotheses are rejected or not, we turn to the last column of Table 9.4, which gives the result of an F-test for any assumed confidence level. Suppose we wanted to test these hypotheses at the 95 per cent confidence level.

We know that the significance level of F in the last column should be less than 0.05 for the null hypothesis to be rejected. We see that for both the rows labelled AD COPY and MAGAZINE, the significance of F is less than .05. It is .005 for AD COPY and .000 for MAGAZINE. This means that both the null hypotheses are rejected.

We conclude that the mean ratings given to the 3 versions of the ad copy are significantly different, and also that the magazine in which the ad copy appears has an impact on its rating. Please note that this conclusion is different from our earlier conclusion when we performed a completely randomised test. This is because the Randomised Block Design has been more efficient in isolating the variance due to the 'magazine' (the block variable). In general, the Randomised Block Design should be used when we suspect that a Blocking variable is affecting the relationship between the independent and the dependent variables.

Latin Square Design

The Latin Square Design is an extension of the Randomised Block Design. It consists of one independent variable (factor) and two blocks, instead of one which we saw in the Randomised Block Design. It has no special significance in marketing research, so we will move on to the more general case of a factorial design where any number of factors can be tested simultaneously for their effects on the dependent variable.

Factorial Design with Two or more Factors

This type of design is employed when we have Two or more independent variables or factors. The major advantage of this design is that multiple factors can be simultaneously tested. There are two kinds of effects that we can test. One is called the Main Effect. The second is called the Interaction Effect. To illustrate, we will take up an example.

Worked Example In this example, we assume that we are testing for a toilet soap brand, the effect of two factors (independent variables)—pack design and price—on sales (dependent variable). We would like to know (1) if each of the factors independently affects sales (called the main effect), and (2) if there is a combined effect of pack design and price (called the interaction effect) on sales.

Incidentally, if there are 3 factors in a study, then we could test for 2-way interactions effects and 3-way interaction effects in addition to the main effects of the individual factors. But the concept of main effects and interaction effects remains the same.

To continue with our example, the experiment is conducted in a simulated environment on 18 randomly selected respondents. There are 3 levels of price—Rs. 8, Rs. 11, and Rs. 14 and 3 levels of Pack Design—designated by the main colours used— Blue, Red, and Green. The coding of these variables is 1, 2, 3 respectively for Rs. 8, 11, and 14 and 1, 2, 3 for Blue, Red, and Green in the case of pack design.

Input Data The input dataset is shown in Table 9.5. Column 1 is 'sales', column 2 is 'pack design' and column 3 is 'price'. Please note that even though price is a continuous metric variable, for the purpose of ANOVA, being an independent variable, it has to be treated as a categorical variable. Hence the coding for price.

Also note from Table 9.5 that each combination of price and pack design appears twice in the dataset. For example, pack design = 1 and price = 1 appears in row 1 and also in row 10. This is known as a replication in design of experiments. This is similar to having a higher sample size in a survey, to represent certain segments in the population. Depending on the number of factors and the number of levels of each factor, the minimum sample size required for ANOVA may go up. In such cases, multiple observations or replications become necessary. In general, replications reduce chances of random error affecting the results of ANOVA experiments, similar to the effects of increasing sample size in surveys.

Output The output data for our factorial experiment is presented in Table 9.6. Let us first look at 'sources of variation' listed in the first column. The last source of variation listed is the residual or error term. It is the first three, which are of interest to us.

In this case, we had three hypotheses:

1. The mean level of sales remains the same for all 3 levels of pack design (main effect 1).
2. The mean level of sales remains the same for all 3 levels of price (main effect 2).
3. The mean level of sales remains the same for all combinations of pack design and price (interaction effect).

To check if these hypotheses are to be accepted or rejected, we set the significance level. Assuming 0.05 level of significance, we check whether for each of the rows corresponding to the above hypotheses, the significance of F is below 0.05 in the last column of Table 9.6.

We find that the significance of F values are

pack design—.248 (main effect 1)

price—.000 (main effect 2)

pack design by price—.646 (interaction effect)

Therefore, only the price effect, one of the two main effects, is significant statistically, at 95 per cent confidence level. This means that hypothesis no. 2 is rejected.

Hypotheses 1 and 3 cannot be rejected, as the significance of F values are greater than .05 in both cases—.248 and .646 respectively).

Thus, we conclude that price alone has an impact on sales. Neither pack design alone nor the combination of pack design with price have any significant impact on sales of the toilet soap.

ADDITIONAL COMMENTS

Experiments are today widely used in many ways in marketing research. For example, test marketing of new concepts, products or prototypes is usually done through procedures explained above, or similar to these. It is not possible to cover exhaustively in this chapter all possible variations of ANOVA applications. But the examples in the chapter should provide interested students enough grasp of the concepts of ANOVA and the potential uses in marketing—which are many.

STM or Simulated Test Marketing procedures are extensions of the basic ANOVA type experiments, with the added tools of forecasting based on the results of experiments conducted. Separate software packages are now available for many specialised applications such as STM.

The mean values of sales for each value of price design and price are presented in Table 9.7. In Table 9.7 (a) and (b) the mean sales values for all combinations of pack design and price are presented. This is a part of the computer output. But it can also be computed from the input data manually.

PAIRWISE TESTS

If any main effect turns out significant, and has more than two levels, there is one additional test required to check for pairwise differences in the means. For instance, in our example of one-way ANOVA, if the mean ratings had turned out to be significantly different at the 95 per cent confidence level, we still would not know whether only one of the pairs (say, ad copy 1 and ad copy 2) are significantly different from each other, or if the remaining pairs (ad copy 1 and 3, and ad copy 2 and 3)

are also significantly different. To find out, we can use tests such as Tukey's Test, Duncan's Test, or Scheffe's Test. These can be requested while doing the ANOVA on most computer packages. These tests give us a pairwise test result of significant difference among means. These are meaningful only if the *F*-test value for a main effect is significant, leading to the rejection of the relevant null hypothesis.

TABLE 9.1 Input Data

S. No.	*Ad copy*	*Rating*
1	1	6.00
2	1	7.00
3	1	5.00
4	1	8.00
5	1	8.00
6	1	8.00
7	2	4.00
8	2	4.00
9	2	5.00
10	2	7.00
11	2	7.00
12	2	6.00
13	3	5.00
14	3	5.00
15	3	4.00
16	3	7.00
17	3	8.00
18	3	7.00

TABLE 9.2 Output for Completely Randomised Design

Source of Variation	*Sum of Squares*	*DF*	*Mean Square*	*F*	*Sig. of F*
Main Effects	7.000	2	3.500	1.780	.203
Ad copy	7.000	2	3.500	1.780	.203
Explained	7.000	2	3.500	1.780	.203
Residual	29.500	15	1.967		
Total	36.500	17	2.147		

18 cases were processed.
0 cases (.0 pct) were missing.

TABLE 9.3 Input Data for Randomised Block Design

S. No.	*Ad copy*	*Rating*	*Magazine*
1	1	6.00	1
2	1	7.00	2
3	1	5.00	3
4	1	8.00	4
5	1	8.00	5
6	1	8.00	6
7	2	4.00	1
8	2	4.00	2
9	2	5.00	3
10	2	7.00	4
11	2	7.00	5
12	2	6.00	6
13	3	5.00	1
14	3	5.00	2
15	3	4.00	3
16	3	7.00	4
17	3	8.00	5
18	3	7.00	6

TABLE 9.4 Tests of Significance for Rating Using Unique Sums of Squares. (Randomised Block Design)

Source of Variation	*SS*	*DF*	*MS*	*F*	*Sig of F*
Residual	3.67	10	.37		
Ad copy	7.00	2	3.50	9.55	.005
Magazine	25.83	5	5.17	14.09	.000
(Model)	32.83	7	4.69	12.79	.000
(Total)	36.50	17	2.15		

R-Squared = .900
Adjusted R-Squared = .829

TABLE 9.5 Input Data for Factorial Design

S. No.	*Sales*	*Pack design*	*Price*
1	500	1	1
2	440	2	1
3	360	3	1
4	300	1	2
5	280	2	2
6	250	3	2
7	200	1	3
8	150	2	3
9	250	3	3
10	600	1	1
11	450	2	1
12	510	3	1
13	400	1	2
14	350	2	2
15	300	3	2
16	250	1	3
17	275	2	3
18	220	3	3

TABLE 9.6 Output for Factorial Design

Source of Variation	*Sum of Squares*	*DF*	*Mean Square*	*F*	*Sig of F*
Main effects	209305.556	4	52326.389	13.645	.001
Pack design	12536.111	2	6268.056	1.635	.248
Price	196769.444	2	98384.722	25.656	.000
2-Way interactions	9838.889	4	2459.722	.641	.646
Pack design price	9838.889	4	2459.722	.641	.646
Explained	219144.444	8	27393.056	7.143	.004
Residual	34512.500	9	3834.722		
Total	253656.944	17	14920.997		

TABLE 9.7 (a) Mean Values of Sales (Sample Size in Brackets)

	Total population 338.06 (18)		
Pack design			
	1	2	3
	375.00	324.17	315.00
	(6)	(6)	(6)
Price			
	1	2	3
	476.67	313.33	224.17
	(6)	(6)	(6)

TABLE 9.7 (b) Mean Values of Sales (Sample Size in Brackets)

	Price		
Pack design	1	2	3
1	550.00	350.00	225.00
	(2)	(2)	(2)
2	445.00	315.00	212.50
	(2)	(2)	(2)
3	435.00	275.00	235.00
	(2)	(2)	(2)

SUMMARY

ANOVA stands for Analysis of Variance, the generic name given to a set of techniques for studying cause-and-effect of one or more factors (independent variables) on a single dependent variable. The Analysis of Variance technique is used when the independent variables are of nominal scale (categorical) and the dependent variable is metric (continuous). The application areas in marketing research for experiments using ANOVA as the analytical method are wide. Whenever a marketing mix variable (independent variable) such as price, a specific promotion or type of distribution, even specific elements like shelf space, or colour of packaging and so on is changed, we would want to know its effect. Under proper conditions, an experiment can tell us the effects of specific variations in one or more elements of the marketing mix.

One-Way ANOVA: This particular design is used when there is only one categorical independent variable, and one dependent (metric) variable. Each category of an independent variable is called a level. The independent variable may be different levels of prices, or different pack sizes, or different

product colours, and the effect (dependent variable) could be sales of the product. In this type of design, we randomly allocate the various sampling elements to the different levels of the independent variable, and measure the resulting dependent variable. Then, we conduct an *F*-test under the ANOVA to test the null hypothesis that the mean values of the dependent variable are not significantly different from each other, at different levels of the independent variable. If the computer output from the *F*-test shows a significance level (*p*-value) of less than .05 on the ANOVA table, the null hypothesis is rejected. If the *p*-value from the *F*-test is greater than or equal to .05, the null hypothesis is accepted. Thus, if the *p*-value from the *F*-test is less than .05, it proves at the 95 per cent confidence level that variation in the independent variable is able to cause significant variation in the dependent variable. That is, one variable depends on the other, assuming other variables are not in the picture.

A **randomised block design** is used, if there is an additional variable (called the block) which has an impact on the relationship between the independent and dependent variables. This variable is accounted for in the design of a randomised block design by explicitly changing the levels of the block and testing if that has an impact on the relationship between the independent and dependent variables. An example of this is the **day of the week** having an impact on the relationship between type of display in a store (independent variable) and sales (dependent variable). Another example is the magazine in which an advertisement appears (block) affecting the impact (dependent variable) of different versions of the ad (independent variable).

If two or more independent variables are to be tested through an ANOVA, we use a **factorial design,** because each independent variable in ANOVA is also known as a factor. The factorial design can accommodate several factors (independent variables) at several levels (categories) each. The major difference in analysing factorial design with two or more factors is that interactions of two or three factors among themselves form a separate effect. These **interaction effects** need to be tested along with the **main effects** of individual factors. This is quite easy to do on a computer, but care should be taken to assign sampling units or elements to each combination of the factors being tested for their effect. The interaction effects should first be looked at to see if they are significant (*p*-value less than .05 on the ANOVA table), which would show if there is a significant combined effect of the factors. If yes, the analysis may stop there. If not, we must also look at the results of the individual factor *F*-tests (called the main effects) to check if the factors individually have an effect on the dependent variable.

If any of the main effects are statistically significant, there are further pairwise tests to find out which levels of each factor have a significant impact on the dependent. These are called pairwise tests, and the common options available in most packages are Tukey's test, Duncan's test, and Scheffe's test.

The chapter-end case studies provide more applied examples of the material on ANOVA. SPSS commands for doing the various types of ANOVA are listed after the Assignment Questions.

ASSIGNMENT QUESTIONS

1. What is the difference between a completely randomised design and a randomised block design in an ANOVA?
2. What is a 'factor' and what are 'levels' of a factor in an ANOVA? Illustrate with an example.

3. a) What should be the scale of the dependent variable in an ANOVA?
   b) What should be the scale of the independent variables in an ANOVA?
4. What is confidence level of a test?
5. Which test is used to check for statistical significance in an ANOVA?
6. What does the column 'Significance of *F*' or 'Sig. of *F*' on the ANOVA output table represent? Explain with respect to the confidence level or significance level) of the test.
7. What is a null hypothesis and what is an alternate hypothesis in an ANOVA? Can these be interchanged? Explain.
8. How many null hypotheses are there in a 2-factor ANOVA with tests for the main effects and the 2-way interaction? Explain with an illustration.
9. What are the commonly occurring dependent and independent variables in marketing applications of ANOVA?
10. Which other techniques of multivariate analysis are similar to ANOVA? Discuss the similarities between ANOVA and multiple regression, and the differences between the two techniques. Also discuss the similarities and differences between discriminant analysis and ANOVA.
11. A retailer wants to measure the effects of two variables on the sales of a particular brand of detergent. The variables are—position on the shelf and promotion. To check out the effects, he conducts an experiment in his retail store. He changes every week, the position of the detergent brand on his shelf. He first places the detergent on the left side of the shelf for a week, then on the right side of the shelf for another week, and in the third week, he places the detergent in the middle of the shelf. He records the weekly sales in all 3 weeks, without running any promotion. For weeks 4, 5, and 6, he runs a sales promotion, and repeats the three positions on the shelf—left, right, and middle. Weekly sales for all six weeks of the detergent are tabulated. He then repeats the whole experiment 2 times for twelve more weeks—6 weeks with a promotion, and 6 weeks without a promotion.

The data in kgs of detergent sales for the 18 weeks are as follows:

Week	*Detergent Sales*	*Promotion On/Off*	*Shelf Position*
1	60	Off	Left
2	52	Off	Right
3	38	Off	Middle
4	100	On	Left
5	86	On	Right
6	95	On	Middle
7	70	Off	Left
8	45	Off	Right
9	41	Off	Middle
10	92	On	Left
11	75	On	Right

12	84	On	Middle
13	75	Off	Left
14	65	Off	Right
15	55	Off	Middle
16	88	On	Left
17	76	On	Right
18	80	On	Middle

With the help of an ANOVA, advise the retailer whether (a) the position of the detergent on the shelf alone, (b) the promotion alone, or (c) the interaction of the shelf position and the promotion have any impact on the sales of the detergent or not.

SPSS COMMANDS FOR ANOVA

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the first output for a **One-way ANOVA problem** described in the chapter on ANOVA in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYSIS).
2. click on COMPARE MEANS.
3. click on ONE-WAY ANOVA.
4. in the dialogue box that appears, select one appropriate variable as the DEPENDENT by highlighting it in the left hand side box and clicking on the arrow towards the DEPENDENT box. Then select another appropriate variable as a FACTOR (independent variable) from the list of variable labels that appears on the left side of the box and click on the arrow directing it to the FACTOR box. The variables should get transferred to the right hand side boxes after the selection.
5. Click OK to get the output for the one-way ANOVA.

Randomised Block Design

1. Click on **Analyze**.
2. Click on **General Linear Model**.
3. Click on **Univariate**.
4. In the dialogue box that appears, select the dependent variable, and then the independent variable as one FACTOR, and the block variable as the second FACTOR.
5. Click OK to get the output of the Randomised Block Design.

Factorial Experiment with Interactions

Repeat steps 1 to 3 above (used in the Randomised Block Design).

Then,

1. Specify the dependent variable and all the FACTORS (independent variables).
2. Click OK to get the output which contains the **Main effects** and all the **interaction effects**.

CASE STUDY 1

ANOVA*

Methodology

To do the ANOVA analysis we have taken a hypothetical situation of the impact of shelf space and the level of merchandising on the consumer offtake from the retail outlets. Shelf-space has been categorised in three types:

- POP-SHELF = 1 (Point of purchase)
- WINDOWS = 2
- SHELF RACKS = 3

Merchandising and outlet decorations are considered in general to be quite significant to the level of offtake from the respective outlets in consideration and the classification of the merchandising activities has been made on the basis of the variety used in the process and the budget involved in it. So the three types of merchandising that we have taken are as follows:

- Full set (danglers, suspenders, stickers, posters, and POP stand set) = 1
- Retail pack (danglers, stickers, and posters) = 2
- No-display pack (posters and stickers) = 3

We shall go about envisaging the impact of the different types of shelf space acquired and that of the merchandising packages on the volume of sales from the respective retail outlets taken into consideration, through the analysis of variance (ANOVA).

Input Data: Problem (I) We are trying to find out the impact of shelf space on the volume of sales through one-way ANOVA. The input data for that is given as follows:

S. No.	*Sales*	*Shelf Space*
1	550	1
2	500	2
3	450	3
4	570	1

* Prepared by Subhro Kanti Banerjee, Pushpak Sengupta, Bhaskar Sinha, P Gopi Shankar, and Chandrashekar Akki

5	530	2
6	490	3
7	560	1
8	510	2
9	500	3
10	530	1
11	500	2
12	470	3
13	590	1
14	530	2
15	480	3
16	600	1
17	540	2
18	480	3
19	580	1
20	520	2
21	500	3
22	530	1
23	500	2
24	460	3

Problem (2) We shall proceed with the problem now adding another column to the table, that of the merchandising. We shall test two hypotheses by doing an ANOVA with a randomised block design. We have used 'sales' as the dependent variable and 'shelf space' as the factor and 'merchandising' as the block. These are used to test the null hypotheses:

1. The first null hypothesis is that the sales are the same for all the three categories of 'shelf space'.
2. The second null hypothesis is that the block (merchandising) used has no effect on the sales.

The input data for Problem (2) are in the following table.

S. No	*Sales*	*Shelf Space*	*Merchandising*
1	550	1	1
2	500	2	1
3	450	3	1
4	570	1	2
5	530	2	2
6	490	3	2
7	560	1	3

8	510	2	3
9	500	3	3
10	530	1	1
11	500	2	1
12	470	3	1
13	590	1	2
14	530	2	2
15	480	3	2
16	600	1	3
17	540	2	3
18	480	3	3
19	580	1	1
20	520	2	1
21	500	3	1
22	530	1	2
23	500	2	2
24	460	3	2

Problem (3) For this, we have used a modified approach to the previous problem. The same procedure for problem (2) is followed but in the model set up we have chosen Full Factorial instead of Custom. The Error term is taken to be (within + residual). The input data are in the following table.

S. No.	*Sales*	*Shelf Space*	*Merchandising*
1	550	1	1
2	500	2	1
3	450	3	1
4	570	1	2
5	530	2	2
6	490	3	2
7	560	1	3
8	510	2	3
9	500	3	3
10	530	1	1
11	500	2	1
12	470	3	1
13	590	1	2

14	530	2	2
15	480	3	2
16	600	1	3
17	540	2	3
18	480	3	3
19	580	1	1
20	520	2	1
21	500	3	1
22	530	1	2
23	500	2	2
24	460	3	2

Output Data

One-Way ANOVA (Problem 1)

Variables Sales

By Variables Shelf-space.

Analysis of Variance

Source	*D.F.*	*Sum of Squares*	*Mean Squares*	*F. Ratio*	*F. Prob.*
Between Groups	2	29033.3333	14516.6667	34.3977	.0000
Within Groups	21	8862.5000	422.0238		
Total	23	37895.8333			

Analysis of Variance (Problem 2)

Test of Significance for SALES using UNIQUE sums of squares (Problem 2 output)

Source of Variation	*SS*	*DF*	*MS*	*F*	*Significance Of F*
Within + Residual	7338.89	19	386.26		
Merchandising	1523.61	2	761.81	1.97	0.167
Shelf-space	29033.33	2	14516.67	37.58	0.000
(Model)	30556.94	4	7639.24	19.78	0.000
(Total)	37895.83	23	1647.64		
R-Squared	0.806				
Adjusted R-Squared	0.766				

Test of Significance for SALES using UNIQUE sums of squares (Problem 3 output)

Source of Variation	*SS*	*DF*	*MS*	*F*	*Significance Of F*
Within + Residual	7183.33	15	478.89		
Merchandising	1523.61	2	761.81	1.59	0.236
Shelf-space	28392.06	2	14196.03	29.64	0.000
Merchandising by Shelf-space	155.56	4	38.89	0.08	0.987
(Model)	30712.50	8	3839.06	8.02	0.000
(Total)	37895.83	23	1647.64		
R-Squared	0.810				
Adjusted R-Squared	0.709				

Analysis

Problem 1: If the *F* Probability value in the ANOVA Table is less than .05, we reject the null hypothesis (at the 95 per cent confidence level) that the category of shelf-space has no impact on the sales. From the output table for the one-Way ANOVA, we see that the probability value of *F* is .0000. Therefore, we reject the null hypothesis and conclude that the category of shelf-space (POP, Windows, or Racks) has a significant impact on sales.

Problem 2: Here, there are two null hypotheses to be tested:

1. That the factor (Shelf Space) does not significantly affect sales. To check this, we look at the output of the Table titled Analysis of Variance (Problem 2), and focus on the first of the two tables. We find again that the Significance of *F* value is .000 from the last column for the row labelled Shelf-Space. We conclude that shelf-space does have an impact on sales.
2. The second null hypothesis is that merchandising affects the relationship between shelf-space and sales. If we look at the Merchandising row in the same table of output referred to above, we find that the Significance of *F* is .167 for merchandising. Since this is higher than .05, we conclude at a 95 per cent confidence level that the null hypothesis cannot be rejected, and that merchandising has no effect on the shelf-space to sales relationship.

Problem 3: This time, we treat both MERCHANDISING and SHELF-SPACE as factors in the ANOVA, and test the hypothesis that the two main effects (individually) have no impact on sales. In this case, from the Problem 3 output above, we can see that merchandising has a Significance of *F* value of .236, and shelf-space has a Significance of *F* value of .000. This leads us to the conclusion that when both are treated as factors, **shelf-space** has a significant impact on sales (rejecting the null hypothesis) and **merchandising** has no impact on sales, at the 95 per cent confidence level. Additionally, we are testing the null hypothesis that the interaction of these two factors, **shelf-space** and **merchandising** has no impact on sales. This null hypothesis is accepted (cannot be rejected) at the 95 per cent confidence level, because the row labelled 'Merchandising by Shelf-space' has a significance of *F* value greater than .05. The value is .987 from the last output table.

CASE STUDY 2

ANOVA*

Problem

This problem relates to the taste/quality of food dishes served to twenty-eight customers in the TAJ GROUP Hotel Chain. The customers, who are basically from the higher income group (HIG), were asked to give their opinion about the quality/taste of the four common non-vegetarian dishes served to them. The analysis of the problem gives us the variance between the dependent variable (rating) and the independent variables (non-vegetarian dish).

The following four common dishes served at the above hotel chain's restaurant are coded as follows:

CODE	*DISH NAME*
1.	CHICKEN PLATTER
2.	HONEY CHICKEN
3.	CHICKEN SPINACH
4.	TANDOORI CHICKEN

In this problem we have considered four different non-vegetarian dishes that are being offered by TAJ GROUP Hotel Chain. The Hotel group wants to test, which of the above mentioned non-vegetarian dishes is being preferred by their target customers, that is, the HIG customers. At random these twenty-eight respondents are asked for their preference on the scale of 10 (1 = not liked at all, and 10 = most preferred dish) and these data have been tabulated.

Input

Dish Code is tabulated in the second column of the input data table, and the ratings given by the customer are tabulated in the third column. Here, the dependent variable is the **rating** given by the randomly chosen customer and independent variable is the **dish code**.

Input Data Table for Problem 1

S. No.	*Dish Type*	*Rating*
1	1	6.00
2	1	7.00
3	1	8.00

* Prepared by Biswajit Mishra, Imran Ahmed, Vikrant Sharma, and Sharmistha Kundu

4	1	5.00
5	1	9.00
6	1	8.00
7	1	7.00
8	2	8.00
9	2	8.00
10	2	9.00
11	2	8.00
12	2	7.00
13	2	9.00
14	2	8.00
15	3	7.00
16	3	6.00
17	3	6.00
18	3	5.00
19	3	7.00
20	3	7.00
21	3	5.00
22	4	6.00
23	4	6.00
24	4	7.00
25	4	6.00
26	4	8.00
27	4	7.00
28	4	6.00

Analysis of Variance

Rating by Type of Dish
UNIQUE sums of squares All effects entered simultaneously

Source of Variation	*Sum of Squares*	*DF*	*Mean Square*	*F*	*Sig of F*
Main Effects	15.714	3	5.238	5.641	.005
Dish Type	15.714	3	5.238	5.641	.005
Explained	15.714	3	5.238	5.641	.005
Residual	22.286	24	.929		
Total	38.000	27	1.407		

28 cases were processed.
0 cases (.0 pct) were missing.

Hypothesis

The null hypothesis for this problem can be expressed as

Ho: Dl = D2 = D3 = D4

where Dl, D2, D3, and D4 are the mean ratings for the four types of non-veg. dishes.

Our group tested at 95% confidence level whether any of the above mentioned non-veg dishes is being preferred more or less than the other dishes by the customers.

Analysis of Output

From the output table of one-way ANOVA, the first row is titled as 'source of variation'. Under this title, next comes 'main effect', and then the only independent variable called 'dish type'. In the last column the significance of *F*-test is found to be 0.005. This indicates that at a confidence level of 95 per cent, the *F* -test proves the model is highly significant. In other words the ratings given by the customers for the four non-veg. dishes are significantly different from each other.

The mean rating for the four types of dishes are 7.14, 8.14, 6.14, and 6.57 respectively. This ANOVA on the mean values tells us that the difference shown above is statistically significant at 95% confidence level. So our null hypothesis is rejected and we conclude that in this case, significant differences exist between mean ratings given to the four types of non-veg dishes.

Randomised Block Design This represents a frequently used experimental framework for dealing with multivariable classifications. These designs are typically used when the experimenter desires to eliminate a possible source of uncontrolled variations from the error term in order that the effects due to treatments will not be masked by a larger-than-necessary error term.

Problem

We now focus on the same problem, but with an addition column, that is, the location of TAJ GROUP of Hotel Chains. We assume that the three types of non-veg. dishes were each delivered in all seven locations as described below. Here the locations are coded as shown below.

Input Out of the customers who visited these locations, twenty-eight respondents were picked, one from each location who has tasted a particular type of non-veg. dish. So here in this problem we do have one respondent for every combination of location and the dish.

The seven different locations are coded as follows:

CODE	*LOCATION*
1	CHENNAI
2	BANGALORE
3	MUMBAI
4	NEW DELHI
5	GOA
6	KODAIKANAL
7	JAIPUR

Hypothesis

We assume that the location of hotel in which the type of non-veg. dish is being offered may have an impact on the ratings. We'll test two hypotheses by doing ANOVA with randomised block design. For this purpose we have taken the variable 'rating' as the dependent variable, and 'dish type' as the factor (independent variable), and the 'location' as the block. Here, in this problem, the location where the hotel is situated could influence the rating given to the dish type by the customers. We have tried to remove the effect of location of hotel, by 'blocking' its effect. Because if we don't block on a variable, its effect gets included with the error (residual) term. And this may lead to wrong conclusion about the relationship between the independent variable (dish type) and dependent variable (rating). In this respect a randomised block is more 'powerful' than a simple one-way ANOVA, if the block effect is significantly influencing the relationship.

The two null hypotheses are as follows:

1. The mean rating of the dish type is same for all four dishes.
2. The 'block'(location of hotel) has no effect on mean ratings for the dish type by the respondents.

We want to test these hypotheses at 95% confidence level.

Input data are as follows—

S. No.	*Dish Type*	*Rating*	*Location*
1	1	6.00	1
2	1	7.00	2
3	1	8.00	3
4	1	5.00	4
5	1	9.00	5
6	1	8.00	6
7	1	7.00	7
8	2	8.00	1
9	2	8.00	2
10	2	9.00	3
11	2	8.00	4
12	2	7.00	5
13	2	9.00	6
14	2	8.00	7
15	3	7.00	1
16	3	6.00	2
17	3	6.00	3
18	3	5.00	4

19	3	7.00	5
20	3	7.00	6
21	3	5.00	7
22	4	6.00	1
23	4	6.00	2
24	4	7.00	3
25	4	6.00	4
26	4	8.00	5
27	4	7.00	6
28	4	6.00	7

Analysis of Variance

28 cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
28 non-empty cells.
1 design will be processed.

Analysis of Variance—design-1

Tests of Significance for RATING using UNIQUE sums of squares

Source of Variation	*SS*	*DF*	*MS*	*F*	*Sig*
Residual	11.29	18	.63		
Location	11.00	6	1.83	2.92	.036
Dish Type	15.71	3	5.24	8.35	.001
(Model)	26.71	9	2.97	4.73	.002
(Total)	38.00	27	1.41		

R-Squared = .703
Adjusted R-Squared = .555

ANALYSIS OF OUTPUT

In the output table of randomised block design problem, the last column gives the result of an F-test as 0.036 for *location* and 0.001 for *dish type*. Since the significance of F is less than 0.05, both the hypotheses are rejected. We conclude that the mean ratings given to the four types of 'dishes' are significantly different as well as the 'location' where these non-veg. dishes are being served has an impact on its rating.

Factorial Design This type of design is employed when we have two or more independent variables or factors. The major advantage of this design is that multiple factors can be simultaneously tested. Here, in this factorial design we can test two kinds of effects. One is called the 'main effect' and another is called the 'interaction effect'.

Problem

The fictitious problem that our group has generated is to test, for a brand of ice-cream, the effect of the flavour and price (both independent variables) on sales. We would like to investigate two things:

1. If each of the independent variables (factor) independently affect the sales, which is called the 'main effect' of each factor.
2. If there is a combined effect of flavour of ice-cream and price, which is called the 'interaction effect', on sales.

Input In this problem, this experiment is conducted in simulated environment on 32 randomly selected respondents. Four levels of price, that is, Rs. 100.00, Rs. 110.00, Rs.120.00, and Rs.140.00 per 1kg pack, have been selected. The codes are given below.

Price of 1 kg. Pack	*Code*
Rs. 100.00	1
Rs. 110.00	2
Rs. 130.00	3
Rs. 140.00	4

Similarly four ice-cream flavours namely, strawberry, vanilla, chocolate, and pista are under consideration. These are being coded as follows

Ice-Cream Flavour	*Code*
Strawberry	1
Vanilla	2
Chocolate	3
Pista	4

S. No.	*Sales*	*Flavour*	*Price*
1	1050	1	1
2	1220	2	1
3	1125	3	1
4	1090	4	1
5	1200	1	2
6	1150	2	2

7	1180	3	2
8	1110	4	2
9	1075	1	3
10	1025	2	3
11	1100	3	3
12	1020	4	3
13	1000	1	4
14	1000	2	4
15	930	3	4
16	900	4	4
17	1310	1	1
18	1175	2	1
19	1225	3	1
20	1200	4	1
21	1125	1	2
22	1050	2	2
23	1140	3	2
24	1000	4	2
25	1005	1	3
26	1125	2	3
27	1025	3	3
28	1000	4	3
29	950	1	4
30	1000	2	4
31	940	3	4
32	1000	4	4

Analysis of Variance

Sales by Flavour, Price
Unique sums of squares
All effects entered simultaneously

Source of Variation	*Sum of Squares*	*DF*	*Mean Square*	*F*	*Sig. of F*
Main Effects	212367.187	6	35394.531	7.3	.001
FLAVOUR	14546.094	3	4848.698	1.0	.417
PRICE	197821.094	3	65940.365	13.6	.000
2-Way Interactions	14157.031	9	1573.003	.32	.954
FLAVOUR PRICE	14157.031	9	1573.003	.32	.954
Explained	226524.219	15	15101.615	3.12	.015
Residual	77287.500	16	4830.469		
Total	303811.719	31	9800.378		

32 cases were processed.
0 cases (.0 pct) were missing.

In the input data table, column 2 represents 32 respondents, column 3 is sales, column 4 flavour, and the last column represents price per kg. Price being an independent variable is treated as a categorical variable. We have represented each combination of price and flavour twice in the data set. This replication in design is necessary to reduce the chances of random error affecting the results of our problem. This is similar to the effects of increasing sample size in surveys.

Hypotheses

In this case we have assumed three hypotheses:

1. The mean level of sales remains the same for all four types of flavour (main effect 1).
2. The mean level of sales remains same for all four levels of price (main effect 2).
3. The mean level of sales remains same for all combinations of flavour and price (interaction effect).

We will check if these hypotheses are accepted at 95% confidence level.

Analysis of Output

We found that significance of *F* values are:

FLAVOUR	0.417 (main effect 1)
PRICE	0.000 (main effect 2)
FLAVOUR and PRICE	0.954 (interaction effect)

Therefore only the price effect, one of the two main effects, is statistically significant at 95% confidence level. This means our second hypothesis that the mean level of sales remains same for all four levels of price (main effect 2) is rejected.

Hypotheses one and three are accepted because the significance of F values are 0.417 and 0.954 respectively, which are greater than 0.05 in both the cases.

Thus we conclude that price alone has an impact on sales of ice-creams. Neither the flavour alone nor the combination of flavour and price have any significant impact on the sales of the ice-creams.

The McGraw·Hill Companies

CASE STUDY 3

ANOVA*

Problem Number 1

Statement—To find out if the consumer rating of milk powder differs with respect to the three qualities of milk powder that ABC Co. was supplying to its consumers, by studying 30 consumers.

Quality—Type one coded as 1, Type two coded as 2, and Type three coded as 3.

Rating—The ratings are on a 5-point scale, from 4 to 8.

Null Hypothesis—There is no significant difference in the ratings between the various types of quality of milk powder being supplied by ABC Co.

Alternate Hypothesis—There is a significant difference in the ratings between the various types of quality of milk powder being supplied by ABC Co.

The input data is in Table 1.1 and the output of Problem 1 is in Table 1.2.

Findings—The value of F arrived at through the F-test is 27.961 and its significance is 0.000 which is less than .05 (because we have taken 95% confidence level).

Inference—As 0.000 < .05, the null hypothesis is rejected and we infer that the various types of quality that ABC Co. provides its customers has a direct impact on their rating/purchasing behaviour.

TABLE 1.1 Input data for completely randomised design in a one-way ANOVA using quality as independent variable based on which rating is done

S. No.	*Quality*	*Rating*
1	1	4
2	1	4
3	1	4
4	1	4
5	1	5
6	1	4
7	1	5
8	1	4
9	1	5

* Prepared by Anindya Das, J Padmapriya, and Saurabh Kochar

S. No.	Quality	Rating
10	1	5
11	2	6
12	2	7
13	2	6
14	2	8
15	2	8
16	2	8
17	2	8
18	2	7
19	2	7
20	2	6
21	3	5
22	3	5
23	3	5
24	3	5
25	3	6
26	3	6
27	3	5
28	3	8
29	3	5
30	3	5

TABLE 1.2 Output Completely Randomised Design in a one-way ANOVA (Single Factor)

Source of variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	36.867	2	18.433	27.961	.000
Quality	36.867	2	18.433	27.961	.000
Explained	36.867	2	18.433	27.961	.000
Residual	17.800	27	.659		
Total	54.667	29	1.885		

Problem Number 2

Statement—To find out if the consumer rating differs with respect to the types of quality of milk powder ABC Co. was supplying to its customers, by studying 30 consumers. Further, a block variable

of the kind of packaging in which the milk powder is being supplied was created, to study whether packaging had any impact on the ratings.

Quality—Type I coded as 1, Type II coded as 2, Type III coded as 3.

Rating—The ratings are on a 5-point scale, from 4 to 8.

Packaging Type	*Code*
Loose milk powder	1
Sachet and cardboard	2
Sachet in aluminium foil	3
Small glass jar	4
Small plastic jar	5
Big-sized cardboard pack	6
Big-sized aluminium foil pack	7
Big-sized glass jar	8
Airtight jar	9
Can with flip open/close option	10

Null Hypotheses:

a) There is no significant difference in the ratings between the various types of quality of milk powder being supplied by ABC Co.

b) There is no significant impact on the ratings due to the variety in style of packaging of milk powder being supplied by ABC Co.

Alternate Hypotheses:

a) There is a significant difference in the ratings between the various types of quality of milk powder being supplied by ABC Co.

b) There is a significant impact on the ratings due to the variety in style of packaging of milk powder being supplied by ABC Co.

The input data for the problem is in Table 2.1 and the output is in Table 2.2.

Findings:

a) The value of F for quality arrived at through the F-test is 29.80 and its significance is 0.000 which is less than .05 (because we have taken 95% confidence level).

b) The F-test value of the blocking factor packaging is 1.2 and the significance level is .354 which is greater than .05 (because we have taken 95% confidence level).

Inference:

a) As $0.000 < .05$, the null hypothesis is rejected and we infer that the various types of quality that ABC Co. provides its customers has a direct impact on their rating/purchasing behaviour.

b) As $0.354 > .05$, the null hypothesis is accepted and we infer that the package design does not affect the rating/purchasing behaviour of customers of ABC Co.

TABLE 2.1 Input data for randomised block design with quality as independent factor and packaging as the blocking factor

S. No.	*Quality*	*Rating*	*Packaging*
1	1	4	1
2	1	4	2
3	1	4	3
4	1	4	4
5	1	5	5
6	1	4	6
7	1	5	7
8	1	4	8
9	1	5	9
10	1	5	10
11	2	6	1
12	2	7	2
13	2	6	3
14	2	8	4
15	2	8	5
16	2	8	6
17	2	8	7
18	2	7	8
19	2	7	9
20	2	6	10
21	3	5	1
22	3	5	2
23	3	5	3
24	3	5	4
25	3	6	5
26	3	6	6
27	3	5	7
28	3	8	8
29	3	5	9
30	3	5	10

TABLE 2.2 Output randomised block design (single blocking factor)

Source of variation	*SS*	*DF*	*MS*	*F*	*Sig of F*
Within + Residual	11.13	18	.62		
Packaging	6.67	9	.74	1.20	.354
Quality	36.87	2	18.43	29.80	.000
(Model)	43.53	11	3.96	6.40	.000
(Total)	54.67	29	1.89		
R-Squared = .796					
Adjusted R-Squared = .672					

Problem Number 3

Statement—To find the impact of availability and quality (flavour) on the sales of a brand of cigarettes being sold by XYZ Co.

Availability—Paan shop coded as 1.
Retailer coded as 2.
Wholesaler coded as 3.
Cafeteria coded as 4.

Quality (Flavour)—Type one coded as 1.
Type two coded as 2.
Type three coded as 3.
Type four coded as 4.

Null Hypotheses:

a) There is no significant impact of availability of cigarettes on sales.
b) There is no significant impact of quality of cigarettes on sales.
c) There is no significant impact on sale of cigarettes when both availability and quality interact with each other.

Alternate Hypotheses:

a) There is a significant impact of availability of cigarettes on sales.
b) There is a significant impact of quality (flavour) of cigarettes on sales.
c) There is a significant impact on the sale of cigarettes when both availability and quality interact with each other.

The input data for the problem are in Table 3.1 and the output is in Table 3.2.

Findings:

a) The significance of the *F*-test with respect to the variable availability is 0.00.
b) The significance of the *F*-test with respect to the variable quality is 0.00.
c) The significance of the *F*-test with respect to the interaction of the variables availability and quality is 0.00.

Inference:

a) As 0.00 < 0.05, we reject the null hypothesis and infer that availability has a significant impact on sale of cigarettes for XYZ Co.
b) As 0.00 < 0.05, we reject the null hypothesis and infer that quality has a significant impact on sale of cigarettes for XYZ Co.
c) As 0.00 < 0.05, we reject the null hypothesis and infer that the interaction of the variables availability and quality has a significant impact on sale of cigarettes for XYZ Co.

TABLE 3.1 Input data for factorial design with two factors (quality and availability)

S. No.	*Sales*	*Quality and Flavour*	*Availability*
1	500	1	1
2	440	2	1
3	360	3	1
4	300	4	1
5	600	1	2
6	100	2	2
7	265	3	2
8	575	4	2
9	230	1	3
10	540	2	3
11	115	3	3
12	145	4	3
13	350	1	4
14	310	2	4
15	595	3	4
16	300	4	4
17	510	1	1
18	435	2	1
19	365	3	1
20	305	4	1
21	590	1	2
22	120	2	2
23	285	3	2
24	565	4	2
25	200	1	3
26	505	2	3
27	110	3	3
28	155	4	3
29	320	1	4
30	335	2	4
31	585	3	4
32	325	4	4

The McGraw·Hill Companies

TABLE 3.2 Output factorial design with two or more factors

Tests of Significance for SALES using UNIQUE sums of squares

Source of Variation	*SS*	*DF*	*MS*	*F*	*Sig of F*
Within + Residual	2837.50	16	177.34		
Availability	123852.34	3	41284.11	232.79	.000
Quality	33464.84	3	11154.95	62.90	.000
Availability By Quality	61194.53	9	73466.06	414.26	.000
(Model)	818511.72	15	54567.45	307.69	.000
(Total)	821349.22	31	26495.14		

R-Squared = .997

Adjusted R-Squared = .993

The McGraw·Hill Companies

CHAPTER 10 CORRELATION AND REGRESSION: EXPLAINING ASSOCIATION AND CAUSATION

Learning Objectives

In this chapter, we will

- Introduce the concept of correlation for interval-scaled or metric variables
- Highlight the use of regression in marketing research, particularly for sales forecasting based on marketing mix variables
- Provide a worked example to show a typical application of regression analysis in marketing
- Introduce forward and backward stepwise regression using a software package
- Discuss the use of regression in explaining causation and predicting the value of the dependent variable based on a set of independent variables
- Explain how to judge the significance of the regression model through an *F*-test, how to use R^2 as a measure of explained variation, and the *t*-test as a measure of the significance of individual variables in the regression model

APPLICATION AREAS

Correlation and Regression are generally performed together. The application of correlation analysis is to measure the degree of association between two sets of quantitative data. For example, how are sales of product A correlated with sales of product B? Or, how is the advertising expenditure correlated with other promotional expenditure? There are virtually no limits to applying correlation analysis to any dataset of two or more variables. Whether it makes sense to study a particular variable's correlation with another and how to interpret it, is a different issue. It is the researcher's responsibility to ensure correct use of correlation analysis. Correlation is usually followed by regression analysis in many applications.

The main objective of regression analysis is to explain the variation in one variable (called the dependent variable), based on the variation in one or more other variables (called the independent variables). The applications areas are in 'explaining' variations in sales of a product based on advertising expenses, or number of sales people, or number of sales offices, or on all the above variables. If there is only one dependent variable and one independent variable used to explain the variation in it, then the model is known as a simple regression. If multiple independent variables are used to explain the variation in a dependent variable, it is called a multiple regression model. Even though the form of the regression equation could be either linear or non-linear, we will limit our discussion to linear models. As seen from the preceding discussion, the major application of regression analysis in marketing is in the area of sales forecasting, based on some independent (or explanatory) variables. This does not mean that regression analysis is the only technique used in sales forecasting. There are a variety of quantitative and qualitative methods used in sales forecasting, and regression is only one of the better known (and often used) quantitative techniques.

METHODS

There are basically two approaches to regression:

- A hit-and-trial approach.
- A pre-conceived approach.

1. In the hit and trial approach, we collect data on a large number of independent variables and then try to fit a regression model with a stepwise regression model, entering one variable into the regression equation at a time.

   The general regression model (linear) is of the type

   $$Y = a + b_1x_1 + b_2x_2 + ... + b_nx_n$$

   where y is the dependent variable and $x_1, x_2, x_3...x_n$ are the independent variables expected to be related to y and expected to explain or predict y. $b_1, b_2, b_3...b_n$ are the coefficients of the respective independent variables, which will be determined from the input data.
2. The pre-conceived approach assumes the researcher knows reasonably well which variables explain y and the model is pre-conceived, say, with 3 independent variables x_1, x_2, x_3. Therefore, not too much experimentation is done. The main objective is to find out if the pre-conceived model is good or not. The equation is of the same form as earlier.

Input data on y and each of the x variables is required to do a regression analysis. This data is input into a package to perform the regression analysis. The output consists of the b coefficients for all the independent variables in the model. The output also gives you the results of a t-test for the significance of each variable in the model, and the results of the F-test for the model on the whole.

Assuming the model is statistically significant at the desired confidence level (usually 90 or 95% for typical applications in the marketing area), the coefficient of determination or R^2 of the model is an important part of the output. The R^2 value is the percentage (or proportion) of the total variance in y explained by all the independent variables in the regression equation.

RECOMMENDED USAGE

The hit-and-trial approach may be used for exploratory research. But for serious decision-making, there has to be apriori knowledge of the variables which are likely to affect y, and only such variables should be used in the regression analysis.

It is also recommended that unless the model is itself significant at the desired confidence level (as evidenced by the F-test results printed out for the model), the R^2 value should not be interpreted.

The variables used (both independent and dependent) are assumed to be either interval scaled or ratio scaled. Nominally scaled variables can also be used as independent variables in a regression model, with dummy variable coding. Please refer to either *Marketing Research: Methodological Foundations* by **Churchill** or *Research for Marketing Decisions* by **Green, Tull & Albaum** for further details on the use of dummy variables in regression analysis. Our worked example confines itself to metric interval scaled variables.

If the dependent variable happens to be a nominally scaled one, discriminant analysis should be the technique used instead of regression.

WORKED EXAMPLE

Problem

A manufacturer and marketer of electric motors would like to build a regression model consisting of five or six independent variables, to predict sales. Past data has been collected for 15 sales territories, on sales and six different independent variables. Build a regression model and recommend whether or not it should be used by the company.

We will assume that data are for different territories in which the company operates, and the variables on which data are collected are as follows:

Dependent Variable

Y = sales in Rs. lakh in the territory

Independent Variables

X_1 = market potential in the territory (in Rs. lakh)
X_2 = no. of dealers of the company in the territory
X_3 = no. of salespeople in the territory
X_4 = index of competitor activity in the territory on a 5-point scale
(1 = low, 5 = high level of activity by competitors)
X_5 = no. of service people in the territory
X_6 = no. of existing customers in the territory

Input Data

The data set consisting of 15 observations (from 15 different sales territories), is given in Table 10.1. The dataset is referred to as Regdata 1.

Correlation

First, let us look at the correlations of all the variables with each other. The correlation table (output from the computer for the Pearson Correlation procedure) is shown in Table 10.2. The values in the correlation table are standardised, and range from 0 to 1 (+ ve and – ve). Looking at the last column, we find that except for COMPET (index of competitor activity), all other variables are highly correlated (ranging from .73 to .95) with sales. This means we may have chosen a fairly good set of independent variables (no. of dealers, sales potential, no. of customers, no. of service people, no. of salespeople) to try and correlate with sales. Only the index of competitor activity does not appear to be strongly correlated (correlation coefficient is – .05) with sales. But we must remember that these correlations in Table 10.2 are one-to-one correlations of each variable with the other. So we may still want to do a multiple regression with an independent variable showing low correlation with a dependent variable, because in the presence of other variables, this independent variable may become a good predictor of the dependent variable.

The other point to be noted in the correlation table is whether independent variables are highly correlated with each other. If they are, like in Table 10.2, this may indicate that they are not independent of each other, and we may be able to use only 1 or 2 of them to predict the dependent variables. As we will see later, our regression ends up eliminating some of the independent variables, because all six of them are not required. Some of them, being correlated with other variables, do not add any value to the regression model.

We now move on to the regression analysis of the same data.

Regression

We will first run the regression model of the following form, by entering all the 6 'x' variables in the model

$$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 \qquad \text{...Equation 1}$$

and determine the values of a, b_1, b_2, b_3, b_4, b_5, & b_6.

Regression Output

The results (output) of this regression model are in Table 10.4 in table form. Column 4 of the table, titled 'B' lists all the coefficients for the model. According to this,

$$a \text{ (intercept)} = -3.17298$$
$$b_1 = .22685$$
$$b_2 = .81938$$

$b_3 = 1.09104$

$b_4 = -1.89270$

$b_5 = -0.54925$

$b_6 = 0.06594$

These values can be substituted in the Equation 1 above and we can write the equation (rounding off all coefficients to 2 decimals), as

Sales = – 3.17 + .23 (potential) + .82 (dealers) + 1.09 (salespeople) – 1.89 (competitor activity) – 0.55 (service people) + 0.07 (existing customers)

Before we use this equation, however, we need to look at the statistical significance of the model, and the R^2 value. These are available from Table 10.3, the Analysis of Variance Table, and Table 10.4. From Table 10.3, the analysis of variance table, the last column indicates the *p*-level to be 0.000004. This indicates that the model is statistically significant at a confidence level of (1 – 0.000004)*100 or (0.999996)*100, or 99.9996. The p-level indicates the significance of the *F* value.

The R^2 value is 0.977, from the top of Table 10.4. From Table 10.4, we also note that *t*-tests for significance of individual independent variables indicate that at the significance level of 0.10 (equivalent to a confidence level of 90%), only POTENTL and PEOPLE are statistically significant in the model. The other 4 independent variables are individually not significant.

However, for the time being, we shall use the model as it is, and try to apply it for decision-making. The real use of the regression model would be to try and 'predict' sales in Rs. lakh, given all the independent variable values, or, check the impact of a change in some of them on the sales figure of a territory.

The equation we have obtained means, in effect, that sales will increase in a territory if the potential increases, or if the number of dealers increases, or if level of competitor's activity decreases, if number of service people decreases, and if the number of existing customers increases. The estimated increase in sales for every unit increase or decrease in these variables is given by the coefficients of the respective variables. For instance, if the number of sales people is increased by 1, sales in Rs. lakh, are estimated to increase by 1.09, if all other variables are unchanged. Similarly, if 1 more dealer is added, sales are expected to increase by 0.82 lakh, if other variables are held constant.

There is one coefficient, that of the SERVICE variable, which does not make too much intuitive sense. If we increase the number of service people, sales are estimated to decrease according to the – 0.55 coefficient of the variable 'no. of service people' (SERVICE). But if we look at the individual variable *t*-tests, we find that the coefficient of the variable SERVICE is statistically not significant (*p*-level 0.735204 from Table 10.4). Therefore, it is not to be used in interpreting the regression, as it may lead to wrong conclusions. Strictly speaking, only two variables, potential (POTENTL) and no. of sales people (people) are significant statistically at 90 per cent confidence level since their *p*-level is less than 0.10. One should therefore only look at the relationship of sales with one of these variables, or both these variables.

Predictions

Given the levels of X_1, X_2, X_3, X_4, X_5, and X_6 for a particular territory, we can use the regression model for prediction of sales. Before we do that, we have the option of redoing the regression model so that

the variables not statistically significant are minimised or eliminated. We can follow either the Forward Stepwise Regression method, or the Backward Stepwise Regression method, to try and eliminate the 'insignificant' variables from the full regression model containing all six independent variables.

FORWARD STEPWISE REGRESSION

For example, we could ask the computer for a Forward Stepwise Regression model, in which case the algorithm adds one independent variable at a time, starting with the one which 'explains' most of the variation in sales (y), and adding one more X variable to it, rechecking the model to see that both variables form a good model, then adding a third variable if it still adds to the explanation of Y, and so on. Table 10.5 shows the result of running a forward stepwise regression, which ends up with only 4 out of 6 independent variables remaining in the regression model. The 4 variables in the model are PEOPLE (no. of salespeople) POTENTL (sales potential), DEALERS (no. of dealers) and COMPET (competitive index). Again we notice, that the two significant variables (those with p-value <.10) at 90 % confidence are only PEOPLE and POTENTL (p-levels of .006904 and .002860). But dealers is now at p-levels .108955, very close to significance at 90% confidence level. This could be the equation, instead of the one with 6 independent variables, that we could use. We would be economising on the two variables, which are not required if we decide to use the model from Table 10.5 instead of that from Table 10.4. The F-test for the model in Table 10.5 also indicates it is highly significant (From top of Table 10.5, $F = 105.1296$, $p < .000000$) and R^2 value for the model is 0.9767, which is very close to the 6 independent variable model of Table 10.4. If we decide to use the model from Table 10.5, it would be written as follows

$$\text{Sales} = -3.74 + 1.03\ (\text{PEOPLE}) + .24\ (\text{POTENTL}) + .9\ (\text{DEALERS}) - 1.81\ (\text{COMPET})$$

...Equation 2

BACKWARD STEPWISE REGRESSION

We could, as another alternative, perform a Backward Stepwise Regression, on the same set of 6 independent variables. This procedure starts with all 6 variables in the model, and gradually eliminates those, one after another, which do not explain much of the variation in 'Y', until it ends with an optimal mix of independent variables according to pre-set criteria for the exit of variables. This results in a model with only 2 independent variables POTENTL and PEOPLE remaining in the model. This model is shown in Table 10.6. The R^2 or the model has dropped only slightly, to 0.9599, the F-test for the model is highly significant, and both the independent variables POTENTL and PEOPLE are significant at 90% confidence level (p-levels of .002037 and .000728 from last column, Table 10.6).

If we were to decide to use this model for prediction, we only require data to be collected on the number of salespeople (PEOPLE) and the sales potential (POTENTL), in a given territory. We could form the equation as

$$\text{Sales} = -10.6164 + .2433\ (\text{POTENTL}) + 1.4244\ (\text{PEOPLE})$$

...Equation 3

Thus, if potential in a territory were to be Rs. 50 lakh, and the territory had 6 salespeople, then expected sales, using the above equation would be

$$= -10.6164 + .2433(50) + 1.4244(6)$$

$$= 10.095 \text{ lakh.}$$

Similarly, we could use this model to make predictions regarding sales in any territory for which potential and no. of salespeople were known.

Additional Comments

As we can see from the example discussed, regression analysis is a very simple (particularly on a computer), and useful technique to predict one metric dependent variable based on a set of metric independent variables. Its use, however, gets more complex, for instance, if the independent variables are nominally scaled into two (dichotomous) or more (polytomous) categories. It is also a good idea to define the range of all independent variables, used for constructing the regression model. For prediction of Y values, only those X values which fall within this range (used earlier in the model construction stage) must be used, for the predictions to be effective.

Finally, we have assumed that a linear model is the only option available to us. That is not the only choice. A regression model could be of any non-linear variety, and some of these could be more suitable for particular cases. However, these are outside the scope of the present book. Generally, a look at the plot of Y and X tells us in case of a simple regression model, whether the linear (straight line) approach is best or not. But in a multiple regression, this visual plot may not indicate the best kind of model, as there are many independent variables, and the plot in 2 dimensions is not possible.

In this particular example, we have not used any macroeconomic variables, but in industrial marketing, we may use those types of industry or macroeconomic variables in a regression model. For example, to forecast sales of steel, we may use as independent variables, the growth rate of a country's GDP, the new construction starts, and the growth rate of the automobile industry.

TABLE 10.1 Data file: REGDATA1.STA (15 cases with 7 variables)

STATISTICA : Multiple Regression
Data file : REGDATA1.STA (15 cases with 7 variables)

List of Variables and Variable Labels:

1:	SALES	Sales in a territory (Rs. lakh)
2:	POTENTL	Potential in the territory (Rs. lakh)
3:	DEALERS	No. of dealers
4:	PEOPLE	No. of salespeople
5:	COMPET	Index of competitor activity
6:	SERVICE	No. of service people
7:	CUSTOM	No. of existing customers

missing data deleted casewise

(*Table 10.1 contd*)

(*Table 10.1 contd*)

	1 SALES	2 POTENTL	3 DEALERS	4 PEOPLE	5 COMPET	6 SERVICE	7 CUSTOM
1	5	25	1	6	5	2	20
2	60	150	12	30	4	5	50
3	20	45	5	15	3	2	25
4	11	30	2	10	3	2	20
5	45	75	12	20	2	4	30
6	6	10	3	8	2	3	16
7	15	29	5	18	4	5	30
8	22	43	7	16	3	6	40
9	29	70	4	15	2	5	39
10	3	40	1	6	5	2	5
11	16	40	4	11	4	2	17
12	8	25	2	9	3	3	10
13	18	32	7	14	3	4	31
14	23	73	10	10	4	3	43
15	81	150	15	35	4	7	70

TABLE 10.2 Correlations Table

STAT. MULTIPLE REGRESS.	*Correlations (regdata1.sta)*						
Variable	*POTENT L*	*DEALER S*	*PEOPLE*	*COMPET*	*SERVICE E*	*CUSTOM*	*SALES*
POTENTL	1.00	.84	.88	.14	.61	.83	.94
DEALERS	.84	1.00	.85	–.08	.68	.86	.91
PEOPLE	.88	.85	1.00	–.04	.79	.85	.95
COMPET	.14	–.08	–.04	1.00	–.18	–.01	–.05
SERVICE	.61	.68	.79	–.18	1.00	.82	.73
CUSTOM	.83	.86	.85	–.01	.82	1.00	.88
SALES	.94	.91	.95	–.05	.73	.88	1.00

TABLE 10.3 The ANOVA Table

STAT. MULTIPLE REGRESS.	*Analysis of Variance; Depen.Var: SALES (regdata1.sta)*				
Effect	*Sums of Squares*	*dF*	*Mean Squares*	*F*	*Sig. of F*
Regress	6609.484	6	1101.581	57.13269	.000004
Residual	154.249	8	19.281		
Total	6763.733				

TABLE 10.4 Multiple Regression Results

All independent variables were entered in one block

Dependent Variable:	SALES
Multiple R:	.988531605
Multiple R-Square:	.977194734
Adjusted R-Square:	.960090784
Number of cases:	15
F(6, 8)	= 57.13269 p < .000004
Standard Error of Estimate:	4.391024067
Intercept:	– 3.172982117 Std.Error: 5.813394 t(8) = – .5458 p < .600084

STAT. MULTIPLE REGRESS.	*Regression Summary for Dependent Variable: SALES* $R = .98853160$ $R^2 = .97719473$ *Adjusted* $R^2 = .96009078$ $F(6, 8) = 57.133$ $p < .00000$ *Std. Error of Estimate: 4.3910*					
N = 15	*BETA*	*St. Err. of BETA*	*B*	*St. Err. of B*	*t(8)*	*p-level*
Intercept			– 3.1729	5.813394	– .54581	.600084
POTENTL	.439073	.144411	.22685	.074611	3.04044	.016052
DEALERS	.164315	.126591	.81938	.631266	1.29800	.230457
PEOPLE	.413967	.158646	1.09104	.418122	2.60937	.031161
COMPET	.084871	.060074	– 1.89270	1.339712	– 1.41276	.195427
SERVICE	.040806	.116511	– .54925	1.568233	– .35024	.735204
CUSTOM	.050490	.149302	.06594	.095002	.33817	.743935

TABLE 10.5 STATISTICA : Multiple Regression

Data file: REGDATA1.STA (15 cases with 7 variables)
MULTIPLE REGRESSION RESULTS:
Forward stepwise regression, no. of steps: 4

Dependent variable:	SALES
Multiple R:	.988317862

(*Table 10.5 contd*)

(*Table 10.5 contd*)

Multiple R-square:	.976772197
Adjusted R-square:	.967481076
Number of cases:	15
F (4, 10) = 105.1296	P < .000000
Standard Error of Estimate:	3.963668333

Intercept: – 3.741938802 Std. Error: 4.847682 t(10) = – .7719 P < .458025
No other F to enter exceeds specified limit

STAT. MULTIPLE REGRESS.

Regression Summary for Dependent Variable: Sales
R = .98831786 R^2 = .97677220 Adjusted R^2 = .96748108
F = (4,10) = 105.13 p < .00000 Std. Error of estimate: 3.9637

N = 15	*BETA*	*St. Err. of BETA*	*B*	*St. Err. of B*	*T (10)*	*p-level*
Intercept			–3.74194	4.847683	–.77190	.458025
People	.390134	.115138	1.02822	.303453	3.38841	.006904
Potential	.462686	.117988	.23905	.060959	3.92147	.002860
Dealers	.180700	.102687	.90109	.512065	1.75971	.108955
Compet	–.081195	.053434	–1.81074	–1.191624	–1.51955	.159589

TABLE 10.6 STATISTICA: Multiple Regression

Data file: REGDATA1.STA (15 cases with 7 variables)
MULTIPLE REGRESSION RESULTS:
Backward stepwise regression, no. of steps: 4

Dependent variable:	SALES
Multiple R:	.979756241
Multiple R-square:	.959922293
Adjusted R-square:	.953242675
Number of cases:	15
F(2, 12) = 143.7092	P < .000000
Standard Error of Estimate :	4.752849362

Intercept: – 10.614641069 Std. Error: 2.659532 t(12) = – 3.992 P < .001788
No other F to remove is less than specified limit

STAT. MULTIPLE REGRESS.

Regression Summary for Dependent Variable: SALES
R = .97975624 R^2 = .95992229 Adjusted R^2 = .95324267
F(2, 12) = 143.71 P < .00000 Std. Error estimate: 4.7528

N = 15	*BETA*	*St. Err. of BETA*	*B*	*St. Err. of BETA*	*t(12)*	*p-level*
Intercept			– 10.6164	2.659532	– 3.99183	.001788
POTENTLE	.470825	.120127	.2433	.062065	3.91939	.000728
PEOPLE	.540454	.120127	1.4244	.316602	4.49902	.000728

SUMMARY

Correlation and regression are best applied together to test whether metric variables are associated with each other, and whether the dependent variable can be explained by some independent variables, or predicted from them. In marketing, the dependent variable of interest is usually sales (though not always). The independent variables can be any marketing mix variables which affect sales, such as advertising expenditure, number of salespeople, promotional expenditure, and so on. They may also include macroeconomic variables or sales of other related industries (for example, sales of rubber depending on sales of tyres, cables, etc.).

The researcher may already have an idea of which independent variables may be predictors of the dependent variable, in which case he can use a predetermined model (equation) in regression and test it for significance through an F-test. If he is not sure, the procedures of forward or backward stepwise regression can help him to try out assorted variables as independent variables in the regression equation and let the computer suggest the best ones, that explain the maximum variance in the dependent variable. The correlation coefficients of all the variables also provide a good indicator of which independent variables are highly correlated with the dependent variable. If the independent variables are highly correlated among themselves, it may lead to complications in determining which one is better among two or three. When such cases occur, one or two strongly inter-correlated variables may have to be removed from the regression model.

The regression equation is judged for its usefulness based on:

1. the overall F-test for the model. If this is significant at say, 95 per cent confidence level, it indicates that the model is good overall. This shows up as a p-value of less than .05 on the ANOVA table in the regression output.
2. to decide which variables in the model are good explanatory variables of the dependent, the individual t-test for each variable needs to be looked at. If this value is significant (less than .05), it indicates that the concerned variable is significant in the model.
3. the R^2 value of a model tells you what percentage of the variation in the dependent variable is explained by all the independent variables in the model.

Simple or multiple linear regression is a useful technique for explanation and prediction, if used carefully with an eye on its limitations. It generally should not be used to predict values too far away from the range of values used in building the model (equation). Also, as far as possible, independent variables in the model should not be highly correlated with one another.

The chapter-end case studies further illustrate the use of regression models in explanation and prediction applications in marketing. SPSS commands for getting correlation and regression outputs are listed after the Assignment Questions. (We have used a different package called STATISTICA for the outputs in this chapter).

ASSIGNMENT QUESTIONS

1. In what way does correlation analysis through the Pearson correlation help in the first stage of regression analysis?
2. What are the limitations of correlation analysis and how does regression overcome them?
3. Generally, what are the kinds of variables in marketing that can serve as dependent variables for regression analysis?
4. What are the marketing variables suitable for use as independent variables in a regression analysis?
5. If the output from a regression shows a significance of 0.25 from the ANOVA table, what does it indicate? Suppose the R^2 value is high for the above model. What should the researcher conclude?
6. What do the individual significance of *t*-values indicate for the variables which are in the regression model? Suppose out of 4 independent variables, only 2 have significant *t*-significance at a 95 per cent confidence level. What action should the researcher take?
7. What is the difference between a stepwise regression model and a direct entry method (where all the specified variables are entered into the model)? Which one is better and why?
8. What is the purpose of building a regression model? Explain with the help of an example of a marketing application.
9. Can nominal scale variables be used as independent variables in a regression analysis? If yes, explain how this is done.
10. A fast moving consumer product company's marketing manager thinks there is a strong link between the advertising and promotional expenditure and the sales in the following week. He collects data from his company records on sales, advertising expenditure, and promotional (non-advertising) expenditure for one of the large territories of his company. The data is shown below.

Week No.	*Sales in a week (units)*	*Advertising expenditure in previous week (Rs)*	*Promotional expenditure in previous week (Rs)*
1	120,000	15,000	22,500
2	123,000	25,000	10,000
3	140,000	17,000	17,000
4	115,000	20,000	6,000
5	126,000	15,000	16,000
6	130,000	15,000	18,000
7	115,000	18,000	12,000
8	127,000	10,000	15,000
9	118,000	10,000	10,000
10	121,000	15,000	20,000

11	126,000}	15,000	18,000
12	150,000	25,000	20,000
13	140,000	18,000	17,000
14	135,000	20,000	20,000
15	137,000	17,000	22,000

The marketing manager would like you to perform a regression analysis on the data and advise him on how to use the regression model to predict sales based on advertising and promotional expenditure. What would you tell him?

SPSS COMMANDS FOR CORRELATION AND REGRESSION

Correlation

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the output for a **Correlation** problem similar to that described in the chapter on REGRESSION and CORRELATION in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on CORRELATE, followed by BIVARIATE.
3. on the dialogue box which appears, select all the variables for which correlations are required by clicking on the right arrow to transfer them from the variable list on the left. Then select **Pearson** under the heading Correlation Coefficients, and select **2-tailed** under the heading Tests of Significance.
4. click OK to get the matrix of pairwise Pearson Correlations among all the variables selected, along with the two-tailed significance of each pairwise correlation.

Regression

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the output for a **Regression** problem similar to that described in the chapter on REGRESSION and CORRELATION in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on REGRESSION, followed by LINEAR.
3. in the dialogue box which appears, select a dependent variable by clicking on the arrow leading to the dependent box after highlighting the appropriate variable from the list of variables on the left side.
4. select the independent variables to be included in the regression model in the same way, transferring them from left side to the right side box by clicking on the arrow leading to the box called independent variables or independents.

5. in the same dialogue box, select the METHOD. Choose
   - ENTER as the method if you want all independent variables to be included in the model.
   - STEPWISE if you want to use forward stepwise regression.
   - BACKWARD if you want to use backward stepwise regression.
6. select OPTIONS if you want additional output options, select the ones you want, and click CONTINUE.
7. select PLOTS if you want to see some plots such as residual plots, select those you want, and click CONTINUE.
8. click OK from the main dialogue box to get the REGRESSION output.

Note: You can go back to your data file by clicking on that, change the method at Step 5, and get the regression output using another method of your choice in the same way as described above.

General: All output files can be saved using File Save command. They can be printed using the File Print command.

Input data also can be separately saved, or printed, using the same commands (FILE SAVE, FILE PRINT) while the cursor is on the input data file.

CASE STUDY 1

Correlation and Regression*

Problem

Monsanto inc. a biotech major, specialising in the development of genetically modified seeds wanted to know the degree to which yield of wheat per hectare is influenced by the variety of seeds used for a crop, and a few other variables.

Hence, it deputed an employee in its Bangalore H.O., to conduct a market research to study the effect that six different conditions (independent variables) have on the yield per hectare (dependent variable) for a crop of wheat. The research was conducted by accumulating the data from fifteen major states in India.

The six independent variables are:

X_1 Rainfall (in cms.)

X_2 Soil type (I—low quality, 5—high quality of soil suitable for wheat)

X_3 Quantity of fertilizer (in quintal/sq. kilometer of land)

X_4 Percentage of land being irrigated by state agriculture department.

X_5 Seed quality (l = low, 5 = high level of Genetically modified quality)

X_6 Percentage of automation in cultivation process

The dependent variable is:

Y Yield per hectare (in quintals)

Input Data

The data set for yield of crop in fifteen states is given in tabular form.

ANALYSIS OF OUTPUT

Correlation

The correlation table shows all the pairwise correlations. The values in the correlation table are standardised, and range from 0 to 1 (positive and negative). We observe that except for the RAINFALL

* Prepared by Biswajit Mishra, Imran Ahmed, Vikrant Sharma, and Sharmistha Kundu

column (rainfall in cm), all other variables are highly correlated (ranging from 0.701 to 0.929) with yield. This means Monsanto has chosen a fairly good set of independent variables (soil type, quantity of fertilizer in quintal/sq. kilometer of land, percentage of land being irrigated by state agriculture dept., seed, percentage of automation in cultivation process) to try and correlate with yield. Only the rainfall in cms. doesn't appear to be strongly correlated (correlation coefficient is 0.0668) with yield. These correlations are one-to-one correlations of each variable with the others.

The correlation table shows that independent variables are highly correlated with each other. This indicates that they are not independent of each other and only one or two of them can be used to predict the dependent variable (yield). Regression is helpful in eliminating some of the independent variables as all of them are not required. Some of them, being correlated with other variables, do not add any value to the regression model.

Regression

The regression model of the following form has been used by entering all the six '*X*' variables in the model

$$Y = a + b_1x_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6$$

and the value of $a, b_1, b_2, b_3, b_4, b_5, b_6$.

In the output of regression model the value of '*B*' gives all the coefficients of the model which are as follows:

$$\begin{aligned} X_1 &= 0.081260 \\ X_2 &= -0.464805 \\ X_3 &= -0.386654 \\ X_4 &= 0.192612 \\ X_5 &= 7.661391 \\ X_6 &= 0.276314 \\ a \text{ (constant)} &= 4.467521 \end{aligned}$$

These values can be substituted in the above equation to get the value of *Y* (yield)

Yield = 4.47 + 0.08* (rainfall) – 0.46* (soil type) – 0.39* (quantity of fertilizer)
+ 0.19* (Percentage of land being irrigated) + l66* (seed quality)
+ 0.28* (Percentage of automation).

The *p*-level is observed to be 0.0001, indicating that the model is statistically significant at a confidence level of 99.99.

The R^2 value is 0.94616. We also note that *t*-Test for significance of individual dependent variables indicates that at the significance level of 0.10 (confidence level of 90%), only SEED TYPE is statistically significant in the model. The other five independent variables are individually not significant at 90% confidence limit.

The equation we have obtained, mentioned above, means that yield of wheat will increase in a state if rainfall increases, or if soil quality decreases, or if quantity of fertilizer per sq. km decreases, or if percentage of land being irrigated by state irrigation department increases, or if seed quality increases or if the degree of automation in cultivation increases. The estimated increase in yield of wheat for every unit increase or decrease in these variable is given by the respective variables. That means, if the rainfall in cm is increased by one, yield in quintal per hectare, is estimated to increase by 0.08, if all other variables are unchanged. Similarly, if automation in cultivation increases one percent; yield is expected to increase by 0.28 quintal per hectare, other variables remaining constant.

There is one coefficient, that of the SOIL TYPE variable, which doesn't make too much intuitive sense. If we increase the quality of soil, yields are estimated to decrease by 0.46 quintal per hectare. But if we look at the individual variable *t*-tests, we find that coefficient of the independent variable SOIL TYPE is statistically not significant (significance level 0.8058). Therefore it is not to be used when interpreting regression, as it may lead to wrong conclusion. So we conclude that only one independent variable, SEED TYPE is significant statistically at 90% confidence level since sig *T* is less than 0.10. Therefore one should look at the relationship of yield with this independent variable.

Given the levels of X_1, X_2, X_3, X_4, X_5 and X_6 for a particular state, Monsanto inc. can use the regression model for prediction of yield of wheat per hectare of land.

Forward Stepwise Regression

In forward stepwise regression the algorithm adds one independent variable at a time, starting with the one which explains most of the variation in yield (*Y*), and adding one more *X* variable to it, rechecking the model to see that both variables form a good model, then adding a third variable if it still adds to the explanation of *Y*, and so on. In our output of forward stepwise regression, the regression ends up with 2 out of 6 independent variables remaining in the regression model. The two variables in the model (Refer to the FINAL OUTPUT TABLES for forward stepwise method) are AREAIRRG (percentage of land being irrigated by state agriculture department) and SEED TYPE (seed quality). We notice that only one significant variable (with sig $T < 0.10$) at 90% confidence is SEED TYPE (sig *T* 0.0122). But AREAIRRG is now at sig *T* 0.1012, very close to significance at 90% confidence level. The *F*-test for the model also indicates it is highly significant, $F = 69.66873$ at sig $F = .0000$ and R^2 value for the model is 0.92071, which is very close to the six independent variable model (0.94616). If we decide to use this model the equation would be

Yield = 7.77 + 6.11* (seed type) + 0.31* (areairrg)

Backward Stepwise Regression

Backward stepwise regression was performed on the same set of six independent variables. This regression model starts with six variables and gradually eliminates one after another those which do not explain much of variation in *Y*, until it ends with an optimal mix of independent variables according to preset criteria for the exit of variables, resulting in a model with only two independent variables, that is, SEED TYPE and AREAIRRG remaining, which is same as the outcome of forward stepwise regression. So the equation of yield will remain same.

Yield = 7.77 + 6.11* (seed type) + 0. 31 *(areairrg)

Input Data Table

S. No	Statecod	Yield	Rainfall	Soiltype	Frtilizer	Areairrg	Seedtype	Automate
1	1	40	23	3	8	78	2	45
2	2	30	12	1	3	47	1	26
3	3	42	28	3	9	82	2	39
4	4	35	15	2	5	51	2	25
5	5	44	30	3	4	80	3	35
6	6	48	31	4	7	86	3	40
7	7	55	37	5	10	90	4	42
8	8	34	15	2	5	50	2	25
9	9	28	64	1	4	42	1	21
10	10	60	44	5	9	92	5	52
11	11	35	16	2	7	49	2	25
12	12	48	30	1	4	82	2	38
13	13	29	65	1	3	41	1	20
14	14	46	30	3	8	80	3	36
15	15	58	42	4	8	89	4	49

Correlation Coefficients

	Rainfall	Soiltype	Fertilizer	Areairrg	Seedtype	Automa	Yield
Rainfall	1.0000 (15) P = .	.0240 (15) P = .932	– .0674 (15) P = .811	– .0086 (15) P = .976	– .1190 (15) P = .673	– .0009 (15) P = .998	.0668 (15) P = .813
Soiltype	.0240 (15) P = .932	1.0000 (15) P = .	.8468 (15) P = .000	.8103 (15) P = .000	.8959 (15) P = .000	.8177 (15) P = .000	.8483 (15) P = .000
Fertilizer	– .0674 (15) P = .811	.8468 (15) P = .000	1.0000 (15) P = .	.7059 (15) P = .003	.7580 (15) P = .001	.7360 (15) P = .002	.7010 (15) P = .004
Areairrg	– .0086 (15) P = .976	.8103 (15) P = .000	.7059 (15) P = .003	1.0000 (15) P = .	.8151 (15) P = .000	.9407 (15) P = .000	.9291 (15) P = .000
Seedtype	– .1190 (15) P = .673	.8959 (15) P = .000	.7580 (15) P = .001	.8151 (15) P = .000	1.0000 (15) P = .	.8032 (15) P = .000	.8961 (15) P = .000
Automate	– .0009 (15) P = .998	.8177 (15) P = .000	.7360 (15) P = .002	.9407 (15) P = .000	.8032 (15) P = .000	1.0000 (15) P = .	.9155 (15) P = .000

Yield	.0668 (15) P = .813	.8483 (15) P = .000	.7010 (15) P = .004	.9291 (15) P = .000	.8961 (15) P = .000	.9155 (15) P = .000	1.0000 (15) P = .

(Coefficient/(Cases)/2-tailed Significance)
" . " is printed if a coefficient cannot be computed

Multiple Regression Output with all Variables (Table 1)

Listwise Deletion of Missing Data
Equation Number 1 Dependent Variable. Yield
Block Number 1. Method: Enter
RAINFALL SOILTYPE FRTILZER AREAIRRG SEEDTYPE AUTOMATE

Variable (s) Entered on Step Number

1. AUTOMATE
2. RAINFALL
3. FRTILZER
4. SEEDTYPE
5. SOILTYPE
6. AREAIRRG

Multiple R	.97271
R Square	.94616
Adjusted R Square	.90579
Standard Error	3.19378

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	6	1434.13142	239.02190
Residual	8	81.60191	10.20024
$F = 23.43297$		Signif $F = .0001$	

Variable in the Equation

Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
RAINFALL	.081260	.056255	.127400	1.444	.1866
SOILTYPE	– .464805	1.829003	– .062417	– .254	.8058
FRTILZER	– .386654	.697963	– .088235	– .554	.5947
AREAIRRG	.192612	.134588	.363586	1.431	.1903
SEEDTYPE	7.661391	3.077901	.518164	2.489	.0376
AUTOMATE	.276314	.257878	.273381	1.071	.3152
(Constant)	4.467521	4.826069		.926	.3817

End Block Number 1 All requested variables entered.

FORWARD STEPWISE REGRESSION

Multiple Regression Output Using Forward Stepwise Method (Stage 1)

Listwise Deletion of Missing Data

Equation Number 1 Dependent Varable. YIELD

Block Number 1. Method: Forward Criterion Pin .0500

RAINFALL SOILTYPE FRTILZER AREAIRRG SEEDTYPE AUTOMATE

Variable (s) Entered on Step Number

1. AREAIRRG

Multiple R	.92915
R Square	.86332
Adjusted R Square	.85280
Standard Error	3.99204

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	1	1308.55994	1308.55994
Residual	13	207.17339	15.93641
$F = 82.11131$		Signif $F = .0000$	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
Areairrg	.492224	.054320	.929149	9.062	.0000
(Constant)	8.038648	3.901205		2.061	.0599

Variables Not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
RAINFALL	.074783	.202270	.999926	.715	.4880
SOILTYPE	.277839	.440402	.343417	1.699	.1150
FRTILZER	.089817	.172068	.501636	.605	.5564
SEEDTYPE	.413532	.647975	.335590	2.947	.0122
AUTOMATE	.360267	.330662	.115141	1.214	.2482

Multiple Regression (Forward Stepwise Method, Final Output)

Equation Number 1 Dependent Variable. YIELD

Variable(s) Entered on Step Number

2. SEEDTYPE

Multiple R	.95953
R Square	.92071
Adjusted R Square	.90749
Standard Error	3.16474

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	2	1395.54630	697.77315
Residual	12	120.18703	10.01559

$F = 69.66873$ Signif $F = .0000$

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
AREAIRRG	.313655	.074336	.592073	4.219	.1012
SEEDTYPE	6.114345	2.074735	.413532	2.947	.0122
(Constant)	7.771172	3.094059		2.512	.0273

Variables not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
RAINFALL	.125828	.438392	.323035	1.618	.1340
SOILTYPE	– .010883	– .016318	.174201	– .054	.9578
FRTILZER	– .075714	– .170538	.269118	– .574	.5775
AUTOMATE	.237487	.281212	.105146	.972	.3520

BACKWARD STEPWISE REGRESSION

Multiple Regression Output Using Backward Stepwise Regression (Stage 1)

Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable. YIELD

Block Number 1. Method : Enter

Variable(s) Entered on Step Number

1. AUTOMATE
2. RAINFALL
3. FRTILZER

4. SEEDTYPE
5. SOILTYPE
6. AREAIRRG

Multiple R	.97271
R Square	.94616
Adjusted R Square	.90579
Standard Error	3.19378

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	6	1434.13142	239.02190
Residual	8	81.60191	10.20024
$F = 23.43297$		Signif $F = .0001$	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
RAINFALL	.081260	.056255	.127400	1.444	.1866
SOILTYPE	– .464805	1.829003	– .062417	– .254	.8058
FRTILZER	– .386654	.697963	– .088235	– .554	.5947
AREAIRRG	.192612	.134588	.363586	1.431	.1903
SEEDTYPE	7.661391	3.077901	.518164	2.489	.0376
AUTOMATE	.276314	.257878	.273381	1.071	.3152
(Constant)	4.467521	4.826069		.926	.3817

End Block Number 1 All requested variables entered.

Multiple Regression (Backward Stepwise, Stage 2)

Equation Number 1 Dependent Variable. YIELD

Block Number 2. Method: Backward Criterion POUT .1000

RAINFALL SOILTYPE FRTILZER AREAIRRG SEEDTYPE AUTOMAT

Variable(s) Removed on Step Number

7. SOILTYPE

Multiple R	.97249
R Square	.94573
Adjusted R Square	.91558
Standard Error	3.02326

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	5	1433.47267	286.69453
Residual	9	82.26067	9.14007
$F = 31.36676$		Signif $F = .0000$	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
RAINFALL	.076789	.050580	.120391	1.518	.1633
FRTILZER	– .483457	.553626	– .110326	– .873	.4052
AREAIRRG	.191126	.127282	.360780	1.502	.1674
SEEDTYPE	7.165460	2.253081	.484623	3.180	.0112
AUTOMATE	.271053	.243321	.268176	1.114	.2942
(Constant)	5.287892	3.395927		1.557	.1539

Variables not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
SOILTYPE	– .062417	– .089488	.103377	– .254	.8058

Multiple Regression (Backward Stepwise, Stage 3)

Equation Number 1 Dependent Variable. YIELD

Variable(s) Removed on Step Number

8. FRTILZER

Multiple R	.97012
R Square	.94113
Adjusted R Square	.91758
Standard Error	2.98715

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	4	1426.50266	356.62567
Residual	10	89.23067	8.92307
$F = 39.96671$		Signif $F = .0000$	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig. T*
RAINFALL	.077479	.049970	.121473	1.551	.1521
AREAIRRG	.199047	.125442	.375732	1.587	.1436
SEEDTYPE	6.366498	2.034364	.430586	3.129	.0107
AUTOMATE	.218638	.232986	.216318	.938	.3701
(Constant)	5.148615	3.351669		1.536	.1555

Variables Not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
SOILTYPE	– .136675	– .224529	.104993	– .691	.5069
FRTILZER	– .110326	– .279486	.104048	– .873	.4052

Multiple Regression (Backward Stepwise, Stage 4)

Equation Number 1 Dependent Variable. YIELD

Variable(s) Removed on Step Number

9. AUTOMATE

Multiple R	.96744
R Square	.93595
Adjusted R Square	.91848
Standard Error	2.97090

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	3	1418.64476	472.88159
Residual	11	97.08858	8.82623

$F = 53.57682$ Signif $F = .0000$

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig. T*
RAINFALL	.080257	.049611	.125828	1.618	.1340
AREAIRRG	.296087	.070623	.558911	4.193	.1115
SEEDTYPE	6.735492	1.985142	.455543	3.393	.0060
(Constant)	5.125424	3.333342		1.538	.1524

Variables Not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
SOILTYPE	– .094875	– .152635	.156385	– .488	.6358
FRTILZER	– .075609	– .189479	.261005	– .610	.5553
AUTOMATE	.216318	.284491	.104993	.938	.3701

Multiple Regression (Backward Stepwise, Final Stage)

Equation Number 1 Dependent Variable. YIELD

Variable(s) Removed on Step Number

10. RAINFALL

Multiple R	.95953
R Square	.92071
Adjusted R Square	.90749
Standard Error	3.16474

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	2	1395.54630	697.77315
Residual	12	120.18703	10.01559
$F = 69.66873$		Signif $F = .0000$	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
AREAIRRG	.313655	.074336	.592073	4.219	.1012
SEEDTYPE	6.114345	2.074735	.413532	2.947	.0122
(Constant)	7.771172	3.094059		2.512	.0273

Variables Not in the Equation					
Variable	*Beta In*	*Partial*	*MinToler*	*T*	*Sig T*
RAINFALL	.125828	.438392	.323035	1.618	.1340
SOILTYPE	– .010883	– .016318	.174201	– .054	.9578
FRTILZER	– .075714	– .170538	.269118	– .574	.5775
AUTOMATE	.237487	.281212	.105146	.972	.3520

End Block Number 2 POUT = .100 Limits reached.

CASE STUDY 2

Correlation and Regression*

Problem

The owner of Pizza Corner, Bangalore would like to build a regression model consisting of six factors, to predict the sales of pizzas. Data for the past fifteen months on sales and six different factors were collected for the purpose.

Methodology

- The data is self-generated.
- There are 15 observations.
- The correlation between dependent variable and independent variables is checked using the correlation table generated using STATISTICA.
- The three regression models are created using SPSS.

Input

The variables for which the data have been collected are as follows:

Dependent Variable:

Y = Monthly sales (in Rs. 000's)

Independent Variables:

X_1 = Number of delivery boys
X_2 = Cost of advertisements in Rs. 000's
X_3 = Number of outlets
X_4 = Varieties of pizzas
X_5 = Competitors activities index
X_6 = Number of existing customers ('000s)

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala

The dataset consisting of fifteen observations is given in Table 1.1.

TABLE 1.1

S. No.	SALES	BOYS	ADCOST	OUTLETS	VARIETIES	COMPETITOR	CUSTOMER
1	81	15	20	35	17	4	70
2	23	10	12	10	13	4	43
3	18	7	11	14	14	3	31
4	8	2	6	9	13	3	10
5	16	4	10	11	12	4	17
6	4	1	5	6	12	5	8
7	29	4	14	15	15	2	39
8	22	7	12	16	16	3	40
9	15	5	10	18	15	4	30
10	6	3	5	8	13	2	16
11	45	13	17	20	14	2	30
12	11	2	9	10	12	3	20
13	20	5	12	15	12	3	25
14	60	12	18	30	15	4	50
15	5	1	5	6	12	5	20

Output (Correlation Table)

TABLE 1.2

	SALES	BOYS	ADCOST	OUTLETS	VARIETIES	COMPETITOR	CUSTOMER
SALES	1	0.9	0.93	0.95	0.72	– 0.04	0.88
BOYS	0.9	1	0.9	0.85	0.67	– 0.1	0.84
ADCOST	0.94	0.9	1	0.9	0.7	– 0.19	0.87
OUTLETS	0.95	0.85	0.9	1	0.79	– 0.04	0.86
VARIETIES	0.72	0.68	0.7	0.79	1	– 0.18	0.82
COMPETITOR	– 0.04	– 0.1	– 0.19	– 0.04	– 0.18	1	0.01
CUSTOMERS	0.88	0.84	0.87	0.86	0.82	0.01	1

Table 1.2 shows the output for correlation table. The values in the correlation table are standardised and range from 0–1. First column of Table 1.2 shows that except for competition (index of competitor activity), all the other variables are highly correlated (ranging from .73 to .95) with sales. This means

that we may have chosen a fairly good set of independent variables (number of delivery boys, cost of advertisements, number of outlets, varieties of pizzas, number of existing customers) to try and correlate with the sales. It is only the index of competitor activity that is negatively correlated with sales. The correlation table can also be used to find out the correlation between the independent variables. If they are like in Table 1.2, then it can be inferred that all the factors are not independent of each other. Thus, only one or two of them can be used to predict the dependent variable.

Regression Analysis

A regression analysis is done to explain the variation in one variable (dependent variable), based on variation in one or more other variables (independent variables). In case there is only one independent variable to explain the variation in one dependent variable, it is known as simple regression. If there are multiple independent variables to explain the variation in a single dependent variable, it is known as a multiple regression model.

The linear equation commonly used for a regression analysis is $Y = a + bX_1 + cX_2 + dX_3 + ...$

Where, Y is the dependent variable and X_1, X_2, X_3,..... are the independent variables, and b, c, d are the coefficients of the respective independent variables.

From Table 1.3, the above equation can be written as follows:

Sales = 6.3719 + 0.918 (no. of boys) + 0.699 (adcost) + 1.620 (no. of outlets) – 1.977 (varieties) + 0.066 (competitors activity) + 0.241 (no. of customers)

From the above equation it can be inferred that, if the number of delivery boys is increased by 1, sales in 000's are estimated to increase by 0.918, assuming all the other variables to be constant. Similarly the influence on the sales for every unit increase or decrease in the given 6 factors can be explained by their coefficients. The negative value for varieties shows that as the varieties of pizza increases the sale decreases. The other variables shows direct proportionality.

The measure of strength of association in the regression analysis is given by the coefficient of determination denoted by R^2. This coefficient varies between 0 and 1 and represents the proportion of total variation in the dependent variable that is accounted for by the variation in the factors. From Table 1.3, the R^2 value is 0.95536 which shows that 95% of the variation in sales can be explained by the 6 factors or independent variables.

We also have the t-test value for the significance of individual independent variables to indicate the significance level at 90% confidence level. From the above Table 1.3 we can see that only the number of outlets is statistically significant with a value of 0.0306 which is less than 0.1 (significant). The other five independent variables are individually not significant.

From the Analysis of Variance Sub-Table in Table 1.3, the significance of F is 0.0001. This indicates that the model is statistically significant at a confidence level of 99.999%.

Stepwise Multiple Linear Regression

When the problem involves six independent variables, all the variables may not be equally important. It may be that some combinations can effectively explain the variation better than others. To resolve

this uncertainty, one must evaluate various combinations of independent variables used in different regression equations. The number of variables to be included in the analysis should be decided upon. Stepwise multiple regression analysis is done for the above purpose.

In this analysis the six factors are evaluated separately to determine which one results in the largest R^2, that is, the one which explains most of the variation in the dependent variable, say X_3. Next, X_3 is evaluated in combination with each of the remaining factors to determine which one of the latter results in the largest increase in the R^2, say X_5. Then, X_5 becomes the second factor for the regression equation.

Stepwise regression is of two types. They are:

- Forward stepwise regression
- Backward stepwise regression

Forward Stepwise Regression

It is a model, in which the algorithm adds one factor at a time, starting with the one which explains most of the variation in sales and adding any one of the given six factors to it, rechecking the model to see that both the variables form a good model, then adding a third variable to see that it could still add to the explanation of sales and so on.

Table 1.4 shows the result of running a forward stepwise regression, which ends up with only two out of six factors in the regression model. The 2 variables in the model are the number of delivery boys and the number of outlets. Again we noticed that 'Boys' which was not significant in Table 1.3 is significant in this model. This explains the principle of multicollinearity, that is, there is some dependency between the independent variables taken.

Based on the model from Table 1.4, the equation can be written as follows:

$$\text{Sales} = -11.817 + 1.639\ (\text{Boys}) + 1.753\ (\text{Outlets})$$

The F-test for the model is highly significant and the R^2 value for the model is 0.94, which is very close to the 6 factor model as shown in Table 1.3 (0.95).

Backward Stepwise Regression

This is another method of stepwise regression. This procedure is started with all the six variables in the model, and gradually eliminates those, one after another, which do not explain much of the variation in sales, until it ends with an optimal mix of factors according to pre-set criteria for the exit of variables, results in a model with only two factors, boys and outlets, remaining in the model. This model is shown in Table 1.5. The R^2 for the model is again 0.94 and also the F-test for the model is highly significant and both the factors, that is, boys and outlets are significant at 90% confidence level.

Based on this model the equation is as follows:

$$\text{Sales} = -11.817 + 1.639\ (\text{Boys}) + 1.753\ (\text{Outlets}).$$

Output

TABLE 1.3

Multiple R	.97640
R Square	.95336
Adjusted R Square	.91838
Standard Error	6.26036

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*	*F*	*Signif F*
Regression	6	6408.86352	1068.14392	27.25409	.0001
Residual	8	313.53648	39.19206		

Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
ADCOST	.699223	1.303175	.151823	.537	.6062
COMPET	.066647	2.211029	.002998	.030	.9767
CUSTOMER	.241571	.298900	.182220	.808	.4423
BOYS	.918788	.909790	.189158	1.010	.3421
OUTLETS	1.620447	.618360	.616727	2.621	.0306
VARIETY	– 1.977971	2.309882	– .147403	– .856	.4167
(Constant)	6.371946	32.586445		.196	.8498

Forward Stepwise Regression

TABLE 1.4

Multiple R	.96962
R Square	.94017
Adjusted R Square	.93020
Standard Error	5.78925

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*	*F*	*Signif F*
Regression	2	6320.21515	3160.10758	94.28821	.0000
Residual	12	402.18485	33.51540		

Variable	B	SE B	Beta	T	Sig T
BOYS	1.639945	.641491	.337628	2.556	.0252
OUTLETS	1.753464	.347010	.667351	5.053	.0003
(Constant)	– 11.817157	3.171856		– 3.726	.0029

Backward Stepwise Regression

TABLE 1.5

Multiple R	.96962
R Square	.94017
Adjusted R Square	.93020
Standard Error	5.78925

Analysis of Variance

	DF	Sum of Squares	Mean Square	F	Signif F
Regression	2	6320.21515	3160.10758	94.28821	.0000
Residual	12	402.18485	33.51540		

Variable	B	SE B	Beta	T	Sig T
BOYS	1.639945	.641491	.337628	2.556	.0252
OUTLETS	1.753464	.347010	.667351	5.053	.0003
(Constant)	– 11.817157	3.171856		– 3.726	.0029

CASE STUDY 3

Correlation and Regression*

Problem

An organisation would like to build a regression model consisting of four independent variables to predict the compensation (dependent variable) of its employees. Past data has been collected for 15 different employees, and four independent variables. Build a regression model and recommend its proper usage.

The data is as follows:

Dependent variables

Y = compensation in rupees.

Independent variables

1. Experience (in years)
2. Education (in years, after 10th standard)
3. Number of employees supervised
4. Number of projects handled

The dataset consisting 15 observations, is given in Table 1.1

Correlation

The correlation table is shown in Table 1.2. The values in the correlation table are standardised and range from zero to one, positive and negative. Looking at the last column, we can say that all the variables are highly correlated to compensation ranging from 0.73 to 0.95 except experience whose correlation is 0.6049. This means that the independent variables have been chosen in a fairly good manner. This correlation shown in Table 1.2 is a one-to-one correlation of each variable with the other, so we still have to do multiple regression with an independent variable showing low correlation with a dependent variable because in the presence of other variables this independent variable may become a good predictor of the dependent variable.

The other point to be noted in the correlation table is whether independent variables are highly correlated with each other. In certain cases, they are highly correlated. This may indicate that they are not independent of each other and we may be able to use only 2 or 3 of them to predict the dependent variables.

* Prepared by Anupama B., Anil P., Jagan James, and Sujith S.

Regression

We will first run the regression model of the following form by entering all the four X variables in the model.

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

and determine the values of the above.

Output values are:

$$a = 456.84$$
$$b_1 = 110$$
$$b_2 = -93.43$$
$$b_3 = 44.21$$
$$b_4 = 77.51$$

These values can be substituted in the above equation and the equation can be written as follows

Sales = 456.84 + 110 (education) – 93.43 (experience) + 44.21 (no. of supervisors) + 77.51 (no. of projects)

Now we need to look the statistical significance of the model and the R^2 value. These are available from Table 1.3, the analysis of variance table and Table 1.4, the regression output. From Table 1.3, the last column indicates the p-level to be 0.0001. This indicates that the model is statistically significant at a confidence level of (1 – 0.0001) 100 or .999 × 100 which is equal to 99.9.

The R^2 value is 0.88764 from the top of Table 1.4. From the same figure we can also note that t-test for the significance of independent variables indicate that at a significance level of 95% only experience, education, and number of projects are statistically significant in the model. The number of employees supervised is not significant.

There is one negative coefficient, that of experience, which can be interpreted to mean that if we increase the employee's experience, the compensation will decrease (according to – 93.43 coefficient of the variable).

We have another independent variable, that is the no. of people supervised, that shows a 't' value as .4459, that is, this value is statistically not significant.

Therefore education, experience, and number of projects are significant, and should be used for the interpretation. Therefore one should look at these to determine the compensation.

We can use the forward stepwise regression method or the backward stepwise regression method to try and eliminate the insignificant variable from the full regression model containing all the four independent variables.

Forward Stepwise Regression

Table 1.5 shows the result of forward stepwise regression, which ends up with only three out of four independent variables of the regression model. The three variables are projects (no. of projects handled), experience, and education. We notice that these variables are statistically significant at 95% confidence level.

F-Test of the model is also highly significant and R^2 value is 0.88057. We need not take the no. of people supervised in case we decide to use this model. If we decide to use this model it would be written as follows.

$$Y = 462.79 + 114.344 \text{ (education)} - 82.08 \text{ (experience)} + 89.18 \text{ (projects)}.$$

Backward Stepwise Regression

Table 1.6 shows the output for the backward stepwise regression. The results show that only education, experience, and projects remain in the model. R^2 value is 0.88057. The F-test for the model is highly significant and the independent variables are statistically significant at 95% confidence level. P-levels = 0.0323, 0.0456, 0.0031.

If we were to decide to use this model for prediction, we only require the data to be collected on the above 3 independent variables. We could form the equation as

$$Y = 462.79 + 114.344 \text{ (education)} - 82.08 \text{ (experience)} + 89.18 \text{ (projects)}.$$

Input Data

TABLE 1.1

	Compensa	*Experien*	*Educatio*	*Nosuper*	*Projects*
1	1500.00	2.00	5.00	4.00	10.00
2	1650.00	3.00	6.00	5.00	10.00
3	1750.00	3.00	3.00	5.00	12.00
4	1400.00	2.00	3.00	3.00	9.00
5	2000.00	4.00	4.00	6.00	15.00
6	2200.00	5.00	6.00	6.00	14.00
7	2100.00	1.00	5.00	4.00	12.00
8	2750.00	5.00	8.00	7.00	15.00
9	2900.00	8.00	9.00	8.00	25.00
10	1100.00	3.00	3.00	2.00	7.00
11	1000.00	4.00	2.00	1.00	5.00
12	1350.00	6.00	4.00	4.00	12.00
13	1550.00	4.00	6.00	4.00	11.00
14	1375.00	8.00	4.00	8.00	13.00
15	1400.00	4.00	3.00	5.00	10.00

Output

TABLE 1.2 Correlation Table

Statistical Multiple Regression	*Correlations*				
Variables	*Compensation*	*Education*	*Experience*	*No. of employees supervised*	*No. of projects*
Compensation	1	0.8517	0.2953	0.7106	0.8582
Education	0.8517	1.00	0.3824	0.6504	0.7712
Experience	0.2953	0.3824	1	0.6554	0.6049
No. of employees supervised	0.7106	0.6504	0.6554	1	0.8239
No. of projects	0.8582	0.7712	0.6049	0.8239	1

Output

TABLE 1.3 ANOVA Table

Statistical Multiple Regression	*Analysis of Variance (Dependent Variable, Compensation)*				
Effect	*Sum of squares*	*DF*	*Mean Square*	*F*	*Sig. of F*
Regression	3861033.354	4	965258.33	19.75	0.0001
Residual	488716.64	10	48871.66		

Output

TABLE 1.4

Equation Number 1 Dependent Variable.. COMPENSA
Block Number 1. Method: Enter EDUCATIO EXPERIEN NOSUPER PROJECTS

Variable (s) Entered on Step Number

1. PROJECTS
2. EXPERIEN
3. EDUCATIO
4. NOSUPER

Multiple R	.94215
R Square	.88764
Adjusted R Square	.84270
Standard Error	221.06937

(*Table 1.4 contd*)

(*Table 1.4 contd*)

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
EDUCATIO	110.269955	47.788179	.391871	2.307	.0437
EXPERIEN	−93.432248	39.717225	−.340391	−2.352	.0405
NOSUPER	44.210909	55.720773	.159198	.793	.4459
PROJECTS	77.515223	28.216230	.630741	2.747	.0206
(Constant)	456.847130	173.948594		2.626	.0253

End Block Number 1 All requested variables entered.

TABLE 1.5 Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable.. COMPENSA
Block Number 1. Method: Forward Criterion PIN .0500
EDUCATIO EXPERIEN NOSUPER PROJECTS

Variable (s) Entered on Step Number

1. PROJECTS

Multiple R	.85820
R Square	.73651
Adjusted R Square	.71624
Standard Error	296.92484

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	1	3203613.28125	3203613.28125
Residual	13	1146136.71875	88164.36298
F = 36.33683		Signif *F* = .0000	

Multiple R	.93839
R Square	.88057
Adjusted R Square	.84800
Standard Error	217.31490

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	3	3830266.59489	1276755.53163
Residual	11	519483.40511	47225.76410
F = 27.03515		Signif *F* = .0000	

(*Table 1.5 contd*)

(*Table 1.5 contd*)

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
EDUCATIO	114.344398	46.704574	.406350	2.448	.0323
EXPERIEN	– 82.081200	36.421949	– .299037	– 2.254	.0456
PROJECTS	89.187246	23.669086	.725716	3.768	.0031
(Constant)	462.791851	170.835697		2.709	.0203

Variables Not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
NOSUPER	.159198	.243363	.213141	.793	.4459

End Block Number 1 PIN = .050 Limits reached.

TABLE 1.6

Equation Number 1 Dependent Variable.. COMPENSA
Block Number 2. Method: Backward Criterion POUT .1000
EDUCATIO EXPERIEN NOSUPER PROJECTS

Variable (s) Removed on Step Number

5. NOSUPER

Multiple R	.93839
R Square	.88057
Adjusted R Square	.84800
Standard Error	217.31490

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	3	3830266.59489	1276755.53163
Residual	11	519483.40511	47225.76410
F = 27.03515		Signif *F* = .0000	

Variables in the Equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
EDUCATIO	114.344398	46.704574	.406350	2.448	.0323
EXPERIEN	– 82.081200	36.421949	– .299037	– 2.254	.0456
PROJECTS	89.187246	23.669086	.725716	3.768	.0031
(Constant)	462.791851	170.835697		2.709	.0203

Variables not in the Equation					
Variable	*Beta In*	*Partial*	*Min Toler*	*T*	*Sig T*
NOSUPER	.159198	.243363	.213141	.793	.4459

End Block Number 2 POUT = .100 Limits reached.

The McGraw·Hill Companies

CHAPTER 11 DISCRIMINANT ANALYSIS FOR CLASSIFICATION AND PREDICTION

Learning Objectives

In this chapter, we will

- Introduce Discriminant Analysis as a useful technique for assigning objects to groups
- Discuss its data requirements and applications in selection, classification, and prediction
- Explain how to judge the usefulness of a discriminant function
- Illustrate a decision-making application for a credit card issuer
- Introduce and explain the terms Classification Matrix, Group Centroids, Decision Rule for Classifying Objects, and Relative Strength of Explanatory Variables as used in discriminant analysis

APPLICATION AREAS

The major application area for this technique is where we want to be able to distinguish between two or three sets of objects or people, based on the knowledge of some of their characteristics. For example, the selection process for a job, the admission process of an educational programme in a college, or dividing a group of people into potential buyers and non-buyers. Discriminant analysis can be, and is in fact, used by credit rating agencies to rate individuals, to classify them into good lending risks or bad lending risks. The detailed example later in this chapter shows how it is done.

To summarise, we can use linear discriminant analysis when we have to classify objects into two or more groups based on the knowledge of some variables (characteristics) related to them. Typically, these groups would be users/non-users, potentially successful salesman/potentially unsuccessful salesman, high risk/low risk consumer, or on similar lines.

The McGraw·Hill Companies

Methods

Discriminant analysis is very similar to the multiple regression technique, explained in Chapter 10. The form of the equation in a two-variable discriminant analysis is:

$$Y = a + k_1x_1 + k_2x_2$$

which is similar to a regression equation. It is called the discriminant function. Also, like in a regression analysis, Y is the dependent variable and x_1 and x_2 are independent variables. k_1 and k_2 are the coefficients of the independent variables, and a is a constant. Please note that Y in this case is a categorical variable (unlike in regression analysis, where it is continuous). x_1 and x_2 are however, continuous (metric) variables. k_1 and k_2 are determined by appropriate algorithms in the computer package used, but the underlying objective is that these two coefficients should maximise the separation between the two groups of the Y variable.

k_1 and k_2 are also called the unstandardised discriminant function coefficients.

VARIABLES AND DATA

As mentioned above, Y is a classification into 2 or more groups and therefore, a 'grouping' variable, in the terminology of discriminant analysis. That is, groups are formed on the basis of existing data, and coded as 1 and 2 or similar to dummy variable coding. The independent variables are continuous scale variables, and used as predictors of the group to which the objects will belong. Therefore, to be able to use discriminant analysis, we need to have some data on Y and the x variables from experience and/or past records.

PREDICTING THE GROUP MEMBERSHIP FOR A NEW DATA POINT

We then build a model which is a linear equation of the form shown earlier, and use the coefficients of the equation to calculate the Y discriminant score—for any new data points that we want to classify into one of the groups. A decision rule is formulated for this process—to determine the cutoff score, which is usually the midpoint of the mean discriminant scores of the two groups.

ACCURACY OF CLASSIFICATION

Then, the classification of the existing data points is done using the equation, and the accuracy of the model is determined. This output is given by the classification matrix (also called the confusion matrix), which tells us what percentage of the existing data points is correctly classified by this model. This percentage is somewhat analogous to the R^2 in regression analysis (percentage of variation in dependent variable explained by the model). Of course, the actual predictive accuracy of the discriminant model may be less than the figure obtained by applying it to the data points on which it was based.

Stepwise/Fixed Model

Just as in regression, we have the option of entering one variable at a time into the discriminant equation, or entering all variables which we plan to use. Depending on the correlations between the independent variables, and the objective of the study (exploratory or predictive/confirmatory), the choice is left to the student.

Relative Importance of Independent Variables

Suppose we have two independent variables, x_1 and x_2. How do we know which one is more important in discriminating between groups? The coefficients of x_1 and x_2 are the ones which provide the answer, but not the raw (unstandardised) coefficients. To overcome the problem of different measurement units, we must obtain standardised discriminant coefficients. These are available from the computer output. The difference between unstandardised and standardised discriminant functions will be clear after looking at the worked out example, and the outputs in Tables 11.4 and 11.5. However, one indicator of whether we are looking at standardised coefficients or unstandardised coefficients is to look at the equation and see if it has the constant '*a*' or not. The standardised function will not have the '*a*' constant, but only the coefficients for x_1 and x_2. The raw or unstandardised function will have the constant '*a*' in addition to coefficients for x_1 and x_2.

The higher the standardised discriminant coefficient of a variable, the higher its discriminating power.

Apriori Probability of Classification into Groups

The discriminant analysis algorithm requires us to assign an *apriori* (before analysis) probability of a given case belonging to one of the groups. There are two ways of doing this.

1. We can assign an equal probability of assignment to all groups. Thus, in a 2 group discriminant analysis, we can assign 0.5 as the probability of a case being assigned to any group.
2. We can formulate any other rule for the assignment of probabilities. For example, the probabilities could be proportional to the group size in the sample data. If two-thirds of the sample is in one group, the a priori probability of a case being in that group would be 0.66 (two-thirds).

WORKED EXAMPLE

We will turn now to a complete worked example which will clarify many of the concepts explained above. We will begin with the problem statement and input data.

Problem

Suppose State Bank of Bhubaneswar wants to start credit card division. They want to use discriminant analysis and set up a system to screen applicants and classify them as either 'low risk' or 'high risk'

(risk of default on credit card bill payments), based on information collected from their applications for a credit card.

Suppose SBB has managed to get from SBI, its sister bank, some data on SBI's credit card holders who turned out to be 'low risk' (no default) and 'high risk' (defaulting on payments) customers. These data on 18 customers are given in Table 11.1

We will perform a discriminant analysis and advise SBB on how to set up its system to screen potential good customers (low risk) from bad customers (high risk). In particular, we will build a discriminant function (model) and find out

1. the percentage of customers that it is able to classify correctly.
2. statistical significance of the discriminant function.
3. which variables (age, income, or years of marriage) are relatively better in discriminating between 'low' and 'high' risk applicants.
4. how to classify a new credit card applicant into one of the two groups—'low risk' or 'high risk', by building a decision rule and a cut off score.

Input Data

Data is given in Table 11.1.

The code for low risk customers is 1 and the code for high risk customers is 2 in the RISKLOHI column in the table.

INTERPRETATION OF COMPUTER OUTPUT

We will now find answers to all the four questions we have raised earlier.

1. How good is the Model? How many of the 18 data points does it classify correctly?

   To answer this question, we look at the computer output labelled Table 11.3. This is a part of the discriminant analysis output from any computer package such as SPSS, SYSTAT, STATISTICA, SAS, and so on (there could be minor variations in the exact numbers obtained, and major variations could occur if options chosen by the student are different. For example, if a priori probabilities chosen for the classification into the two groups are equal, as we have assumed while generating this output, then we will very likely see similar numbers in our output).

   This output (Table 11.3) is called the classification matrix (also known as the confusion matrix), and it indicates that the discriminant function we have obtained is able to classify 94.44 per cent of the 18 objects correctly. This figure is in the 'per cent correct' column of the classification matrix. More specifically, it also says that out of 10 cases predicted to be in group 1, 9 were observed to be in group 1 and 1 in group 2, (from column G-1). Similarly, from the column G-2, we understand that out of 8 cases predicted to be in group 2, all 8 were found to be in group 2. Thus, on the whole, only 1 case out of 18 was misclassified by the discriminant model, thus giving us a classification (or prediction) accuracy level of (18-1)/18, or 94.444 per cent.

As mentioned earlier, this level of accuracy may not hold for all future classification of new cases. But it is still a pointer towards the model being a good one, assuming the input data was relevant and scientifically collected. There are ways of checking the validity of the model, but these will be discussed separately.

2. Statistical Significance–How significant, statistically speaking, is the discriminant function?

   This question is answered by looking at the Wilks' Lambda and the probability value for the *F*-test given in Table 11.3. The value of Wilks' Lambda is 0.318. This value is between 0 and 1, and a low value (closer to 0) indicates better discriminating power of the model. Thus, 0.318 is an indicator of the model being good. The probability value of the *F*-test indicates that the discrimination between the two groups is highly significant. This is because $p < .00089$, which indicates that the *F*-test would be significant at a confidence level of upto $(1 - .00089) \times 100$ or (.99911) 100 or 99.91 percent.

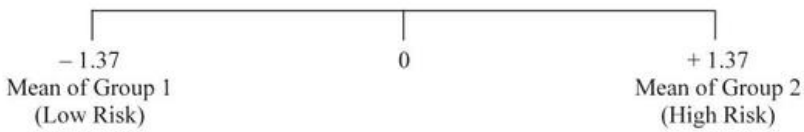
3. We have 3 independent (or predictor) variables—age, income, and no. of years married for. Which of these is a better predictor of a person being a low credit risk or high credit risk?

   To answer this question, we look at the standardised coefficients in the output. These are given in Table 11.5. This output shows that age is the best predictor, with the coefficient of – 0.92, followed by income, with a coefficient of – 0.77, years of marriage is the last, with a coefficient of – 0.15. Please recall that the absolute value of the standardised coefficient of each variable indicates its relative importance.

4. How do we classify a new credit card applicant into either a 'high risk' or 'low risk' category, and make a decision on accepting or refusing him a credit card?

   This is the most important question to be answered. Please remember why we started out with the discriminant analysis in this problem. State Bank of Bhubaneswar wished to have a decision model for screening credit card applicants.

   The way to do this is to use the outputs in Table 11.4 (raw or unstandardised coefficients in the discriminant function) and Table 11.6 (means of canonical variables). Table 11.6, the means of canonical variables, gives us the new means for the transformed group centroids. Thus, the new mean for group 1 (low risk) is – 1.37793, and the new mean for group 2 (high risk) is + 1.37792. This means that the midpoint of these two is 0. This is clear when we plot the two means on a straight line, and locate their midpoint, as shown below



This also gives us a decision rule for classifying any new case. If the discriminant score of an applicant falls to the right of the midpoint, we classify him as 'high risk', and if the discriminant score of an applicant falls to the left of the midpoint, we classify him as 'low risk'. In this case, the midpoint is 0. Therefore, any positive (greater than 0) value of the discriminant score will lead to classification as 'high risk', and any negative (less than 0) value of the discriminant score will lead to classification as 'low risk'. But how do we compute the discriminant scores of an applicant?

We use the applicant's age, income, and years of marriage (from his application) and plug these into the unstandardised discriminant function. This gives us his discriminant score.

Remember, the unstandardised (or raw) discriminant function is, from Table 11.4,

$$Y = 10.0036 - \text{age } (.24560) - \text{income } (.00008) - \text{yrs. married } (.08465)$$

Where Y would give us the discriminant score of any person whose age, income, and yrs. married were known.

Let us take an example of a credit card application to SBB of a person who is aged 40, has an income of Rs. 25,000 per month, and has been married for 15 years. Plugging these values into the discriminant function or model above, we find his discriminant score Y to be

$$\begin{aligned} & 10.0036 - 40\ (.24560) - 25000\ (.00008) \\ & - 15\ (.08465), \text{ which is} \\ = \ & 10.0036 - 9.824 - 2 - 1.26975 \\ = \ & -3.09015 \end{aligned}$$

According to our decision rule, any discriminant score to the left of the midpoint of 0 leads to a classification in the 'low risk' group. Therefore, we should give this person a credit card, as he is a 'low risk' customer. The same process is to be followed for any new applicant. If his discriminant score is to the right of the midpoint of 0, he should be denied a credit card, as he is a 'high risk' customer.

We have completed answering the four questions raised by State Bank of Bhubaneswar.

Additional Comments

One way to derive the new means of the groups 1 and 2 is to take the original means (from Table 11.2, which lists the means for all independent variables after classifying them into group 1 and group 2) and plug them into the unstandardised discriminant function (Table 11.4). This will result in the new group means shown in Table 11.6.

Similarly, the discriminant score for each of the original 18 cases can be computed, and a manual check of their classification can be made. Since this is not too important to be able to use and apply discriminant analysis, it is not explicitly referred to in this book. The computer can also generate an output of these scores, which is shown in Table 11.7. The scores in the last column are the discriminant scores for the 18 cases, used to classify them into the two groups as per the same decision rule we have used to classify a new applicant. These results are summarised in the classification matrix (Table 11.3), as already explained.

Discriminant analysis is a useful procedure to use in marketing applications such as sales force recruitment, discriminating between buyers/non-buyers, heavy users/light users, and any other natural groupings of customers. It provides an unbiased decision model, which is as good as the data on which it is based.

TABLE 11.1

S. No.	RISKLOHI	AGE	INCOME	YRSMARID
1	1	35	40000	8
2	1	33	45000	6
3	1	29	36000	5
4	2	22	32000	0
5	2	26	30000	1
6	1	28	35000	6
7	2	30	31000	7
8	2	23	27000	2
9	1	32	48000	6
10	2	24	12000	4
11	2	26	15000	3
12	1	38	25000	7
13	1	40	20000	5
14	2	32	18000	4
15	1	36	24000	3
16	2	31	17000	5
17	2	28	14000	3
18	1	33	18000	6

TABLE 11.2

STAT. DISCRIM. ANALYSIS	Means (discrbkl.sta)			
RISKLOHI	AGE	INCOME	YRSMARID	Valid N
G_1	33.77778	32333.33	5.777778	9
G_2	26.88889	21777.78	3.222222	9
All Grps	30.33333	27055.55	4.500000	18

TABLE 11.3 VARIABLES

2:	AGE	–	age in years
3:	INCOME	–	monthly income in Rs.
4:	YRSMARID	–	no. of years married for
1:	RISKLOHI	–	risk of default (Low = 1, Hi = 2)

Grouping variable: RISKLOHI 2 Groups

Codes: 1 2

Missing data are casewise deleted

Discriminant Function Analysis Results

Number of variables in the model: 3

Wilks' Lambda: .3188764 approx. F(3, 14) = 9.968056 p < .00089

STAT. DISCRIM. ANALYSIS	*Classification Matrix (discrbkl.sta)* *Rows: Observed classifications* *Columns: Predicted classifications*		
Group	*Percent Correct*	*G_1 (Predicted) P = .50000*	*G_2 (Predicted) P = .50000*
G1 (Observed)	100.0000	9	0
G2 (Observed)	88.8889	1	8
Total	94.4444	10	8

TABLE 11.4

STAT. DISCRIM. ANALYSIS	*Raw Coefficients (discrbkl.sta) for Canonical Variables*
Variable	*Root 1*
AGE	– .24560
INCOME	_ .00008
YRSMARID	_ .08465
Constant	10.00335
Eigenval	2.13601
Cum.Prop	1.00000

TABLE 11.5

STAT. DISCRIM. ANALYSIS	Standardized Coefficients (discrbkl.sta) for Canonical Variables
Variable	*Root 1*
AGE	_ .923955
INCOME	_ .774780
YRSMARID	_ .151298
Eigenval	2.136012
Cum.Prop	1.000000

TABLE 11.6

STAT. DISCRIM. ANALYSIS	Means of Canonical Variables (discrbkl.sta)
Group	*Root 1*
G_1:1	– 1.37793
G_2:2	1.37792

TABLE 11.7

STAT. DISCRIM. ANALYSIS	Unstandardized Canonical Scores (discrbkl.sta)	
Case	*Group*	*Root 1*
1	G_1:1	– 2.48184
2	G_1:1	– 2.22283
3	G_1:1	– .43309
4	G_2:2	2.03057
5	G_2:2	1.12412
6	G_1:1	– .19184
7	G_2:2	– .44650
8	G_2:2	2.01716
9	G_1:1	– 2.21813

(*Table 11.7 contd*)

(*Table 11.7 contd*)

10	G_2:2	2.80674
11	G_2:2	2.15930
12	G_1:1	– 1.92951
13	G_1:1	– 1.84991
14	G_2:2	.36014
15	G_1:1	– 1.01940
16	G_2:2	.60139
17	G_2:2	1.74839
18	G_1:1	– .05476

SUMMARY

Discriminant Analysis is somewhat similar to regression analysis. There is a dependent variable, and there are some independent variables used to predict the dependent variable in both techniques. But in discriminant analysis, the dependent variable is categorical, not metric. It is used to classify people or objects into two or more groups based on some knowledge of their characteristics. For example, selection of salespeople, or MBA students, based on some known scores in selection tests, could be an application. Or, whether a particular consumer is likely to be brand-loyal or a brand-switcher can be predicted based on some demographic or psychographic characteristics of the consumers.

The use of discriminant analysis requires data on all the independent variables and the categories to which each object belongs (for example, if it is a salespeople selection application, whether each salesperson was selected or rejected for a given set of scores in selection tests). Then, a discriminant model (equation) is built, and tested for its usefulness based on

1. its significance. A low value of Wilk's Lambda indicates high significance. The F-test should show a p-value of less than .05.
2. which variables are relatively better at predicting the dependent? This is judged by looking at the standardised coefficients of the independent variables. The larger the absolute value of the standardised coefficient, the better the predictive or explanatory power of the variable.
3. the number of data points from the original dataset the model classifies correctly. This is given on the classification matrix or can be computed by adding the diagonal numbers on it and dividing by the total number of cases. This yields a percentage. The closer it is to 100, the better is the model, but with the caution that its classification accuracy will probably reduce when applied to fresh data.

This model can be used for predicting where a new object should be classified, if we know the independent variable values for the new object. We substitute the 'x' values (independent variables or known characteristics) of the new object in the unstandardised (raw) discriminant function, and

calculate the value (called the unstandardised canonical score). We then check with our decision rule for classifying new objects using the means of canonical variables (canonical variables are transformed variables), also known as group centroids. Based on the decision rule, we classify the object into one group or the other. This is the predicted group membership for the new object.

The chapter-end case studies provide further illustrations of various applications using discriminant analysis in marketing. SPSS commands for discriminant analysis are given after the Assignment Question.

ASSIGNMENT QUESTIONS

1. What are the application areas for discriminant analysis in marketing?
2. What is the recommended scale of the dependent and independent variables in discriminant analysis?
3. Discuss the differences and similarities between linear regression, linear discriminant analysis, and ANOVA.
4. What is the unstandardised discriminant function used for?
5. How can we tell which of the independent variables is better at predicting group membership in discriminant analysis? Mention specifically the output used for this purpose, and explain how to use it to answer the question above.
6. What is a classification matrix? Suppose a classification matrix shows a 90 per cent accuracy level of classification, can we expect the same level of accuracy in classifying a future set of data points? Explain.
7. What are means of canonical variables (or group centroids)? How are they useful in determining the cutoff point for classifying a new data point in 2 group discriminant analysis? Explain the concept of a 'decision rule' based on this cutoff.
8. The Yamaguchi School of Business selects students for its MBA Programme every year through a written test, group discussion, and interview. It then tracks the performance by means of Grade Point Average (GPA) during the 2 year program. The school has past data for 30 students admitted in an earlier year, who either were successful students (GPA over 3.00 on a 4-point scale is defined as successful), or unsuccessful students (GPA below 3.00). The data are listed below:

S. No.	*Score on Written Test*	*Score on Group Discussion*	*Score on Interview*	*Successful/ Unsuccessful*
1	200	10	15	U
2	270	15	20	S
3	300	30	30	S
4	250	35	30	S
5	260	45	35	S

6	220	40	45	U
7	210	30	40	S
8	200	10	25	U
9	240	15	30	S
10	230	25	25	U
11	200	30	35	U
12	300	35	40	S
13	280	20	20	U
14	290	30	25	U
15	275	40	40	U
16	263	25	30	S
17	285	30	25	S
18	291	35	35	S
19	300	25	30	S
20	205	15	20	U
21	220	25	40	U
22	230	40	25	U
23	240	40	15	U
24	270	45	25	S
25	290	35	30	S
26	280	30	35	U
27	250	30	28	U
28	255	15	23	U
29	260	20	40	S
30	260	25	15	U

Develop a model using linear discriminant analysis for the Yamaguchi School of Business. The model should be able to predict whether a student will be successful or unsuccessful, based on his written test score, group discussion score, and interview score. Explain the decision rule to be used for classifying a student as potentially successful or unsuccessful.

What is the classification accuracy of the model? Which of the three scores is the best predictor of a student's future success?

9. Think of how a discriminant analysis model can be built by a sales manager to decide which of his salespeople he should give a promotion to. What variables can be used as predictors of a good future performance?

10. Consider how the technique of discriminant analysis can be used to predict whether a person whose demographics are known would be a buyer or a non-buyer of a given brand or product. Try and suggest the demographic variables which could be predictors for a given product, say, a washing machine.

SPSS COMMANDS FOR DISCRIMINANT ANALYSIS

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the output for a **Discriminant Analysis** problem similar to that described in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on CLASSIFY, followed by DISCRIMINANT.
3. on the dialogue box which appears, select the GROUPING VARIABLE (dependent categorical variable in discriminant analysis) by clicking on the right arrow to transfer it from the variable list on the left to the grouping variable box on the right.
4. define the range of values of the grouping variable by clicking on DEFINE RANGE just below the grouping variable box. Fill in the minimum and maximum values (the codes used in your problem, say 1 and 2) of the variable in the box which appears. Then click CONTINUE.
5. select all the independent variables for discriminant analysis from the variable list by clicking on the arrow which transfers them to the INDEPENDENTS box on the right.
6. just below the INDEPENDENTS box select "Enter independents together" if you want all selected independent variables (that are in the box) in the discriminant model. (Here you have an option to use STEPWISE discriminant analysis by selecting 'Use Stepwise Method' instead of 'enter independents together').
7. click on STATISTICS on the lower part of the main dialogue box. This opens up a smaller dialogue box. On this, under the title FUNCTION COEFFICIENTS, choose UNSTANDARDISED to obtain the unstandardised coefficients of the discriminant function. These are used to classify a new object in discriminant analysis, as explained in the chapter. Click on CONTINUE to return to the main dialogue box.
8. click on CLASSIFY on the lower part of the main dialogue box. Select SUMMARY TABLE under the heading DISPLAY in the smaller dialogue box that appears. This gives you the classification table (also called the confusion matrix) that judges the accuracy of the discriminant model when applied to the input data points. Click on CONTINUE to return to the main dialogue box.
9. click OK to get the discriminant analysis output.

CASE STUDY 1

Discriminant Analysis*

Problem

A retail outlet wants to know the consumer behavioral pattern of the purchase of products in two categories—national brands and local brands respectively, which would help it to place orders depending on demand and requirements of the customer. This retail outlet uses data from a retail outlet in another location to arrive at a decision about customers visiting at their end.

This retail outlet wants to use discriminant analysis to screen the responsiveness of customers towards national brand and local brand categories and find out the following:

- The percentage of customers that it is able to classify correctly.
- Statistical significance of the discriminant function.
- Which variable (annual income and household size) are relatively better in discriminating between consumers for national and local brand.
- Classification of new customers into one of the two groups namely—national and local brand acceptors.

Input Data

S. No	*Brand*	*Annual Income*	*Household Size*
1	1	16.8	3
2	1	21.4	2
3	2	17.3	4
4	2	15.4	3
5	1	17.3	4
6	1	18.4	1
7	2	14.3	4
8	2	14.5	5
9	1	23.2	2
10	1	21.1	5
11	2	17.4	2

* Prepared by Bhaskar, M., KB Srinivas, G. Nagaraj, and RC Rajiv

12	2	16.7	6
13	1	14.5	4
14	1	18.9	1
15	2	13.9	7
16	2	12.4	7
17	1	17.8	2
18	1	19.3	1
19	2	15.3	6
20	2	13.3	4

Codes for brand: 1—For National Brand, 2—For Local Brand

Output from Statistica

TABLE 1

STAT DISCRIM ANALYSIS	*Classification Matrix (discrbkl.sta)* *Rows: Observed Classifications* *Columns: Predicted Classifications*		
Group	*Percent Correct*	*G_1 (predicted)* *P = 0.50* *Annual Income*	*G_2 (predicted)* *P = 0.5* *Household Size*
G 1 (observed)	90.00	9	1
G 2 (observed)	80.00	2	8
Total	85.00	11	9

TABLE 2

STAT DISCRIM ANALYSIS	*Raw Coefficients (new.sta) for Canonical Variables*
Variable	*Root 1*
ANNUALIN	– .335427
HOUSESZ	.313361
Constant	4.545077
Eigenval	1.113402
Cum.Prop	1.000000

TABLE 3

Wilks' Lambda: .4731707 $p < 0.0017$

STAT DISCRIM ANALYSIS	*Standardized Coefficients (new.sta) for Canonical Variables*
Variable	*Root 1*
ANNUALIN	– 0.721519
HOUSESZ	.490487
Eigenval	1.113402
Cum.Prop	1.000000

TABLE 4

STAT DISCRIM ANALYSIS	*Means of Canonical Variables*
Group	*Root 1*
G_1	– 1.00103
G_2	1.00103

ANALYSIS

1. Strength of the model in classifying correctly the exact number of data points.

   From the classification matrix as represented by Table 1, it can be inferred that the discriminant function as obtained for the above mentioned problem, is able to classify 85.0% of the 20 objects correctly.

   This figure is shown in the 'per cent correct' column of the classification matrix. It also shows that out of 11 cases predicted to be in group 1, 9 were observed to be group 1, and 2 in group 2. Similarly from column G_2 we understand that out of 9 cases predicted to be in group 2, 8 were in group 2, and 1 in group 1. Thus on the whole only 3 cases out of 20 were misclassified by the discriminant model, thus giving us a classification (or prediction) accuracy level of (20-3/20), or 85.0%.

2. Statistical significance of the discriminant functions.

   This can be answered by looking at the Wilks' lambda and the probability value for the *F*-test as given in Table 3. The value of Wilks' lambda ranges between 0 and 1 with a lower value indicating better discriminating power of the model. In our case, the magnitude of Wilks' lambda stands at 0.4731707 which is reasonably good, being close to 0 and less than 0.5.

   The probability value of *F*-test indicates that discrimination between two groups is highly significant because $P < 0.0017$ indicates that the *F*-test would be significant at a confidence level of up to (1-0.0017)*100 or 0.9983*100 or 99.83%.

3. We have two independent variables, annual income and household size, with one being a better predictor on purchase of national/local brand. From Table 3 it can be inferred that the best predictor is annual income (variable) with a higher coefficient of 0.721519.
4. Classification of a new customer into purchase behaviour for national/local brand: This can be judged from Table 4 where the means of group 1 (national brand) and group 2 (local brand) are plotted on a scale ranging from – 1.00103 to + 1.00103. If the discriminant score of a customer falls to the left of the mid point (zero) we classify him as one with a preference for a national brand and the one on the right signifies a preference for the local brand.

   The unstandardised (or raw) **discriminant function** from Table 2 is:

   $\rightarrow$ ***Y* = 4.545077 – 0.335427 (annual income) + 0.313361 (household size)**

   If we have the customer's annual income and household size data respectively then the above discriminant function gives us the score that helps the retail outlet to classify the customer.

CASE STUDY 2

Discriminant Analysis*

The Problem

The manager of the Raymond's showroom in Bangalore proposes to his head office that as the sales have touched the target, they should promote further sales by showing some special gratitude to their loyal customers by providing then with 'First Citizen Club Card'. With this card the customer will get points on every purchase that they make and after a certain number of points they will be benefited with 'Gift Hampers' and 'Free Purchases' based on the total number of points accrued.

To facilitate the head office, the branch manager provides a discriminant analysis and sets up a system to screen the customers and classify them as either 'loyal' or 'unloyal' based on the following three independent variables:

1. Freq : frequency of purchase in a year
2. Avgpurc : average purchase by customer in a year
3. Years : no. of years since the customer has been purchasing from the Raymond showroom

Now we perform the discriminant analysis and advise Raymonds Bangalore division on how to set up its system to screen its potential 'loyal' customer from its 'unloyal' customers. We shall build a discriminant function and find out:

1. the percentage of customers that it is able to classify correctly.
2. statistical significance of the discriminant function.
3. which of the independent variable are relatively better in discriminating between 'loyal' and 'unloyal' customers.
4. how to classify a new customer into one of those two groups—loyal/unloyal.

Input Data

The dependent variable is loyalty of customer.

Loyal = 1

Disloyal = 2

* Prepared by Aditi Sood, Mukesh Kumar Sinha, Vikram Rathi, and Vinita Nair

S. No.	Loyalty	Freq	Avgpurc	Years
1	1	15	24765	3
2	1	17	18654	4
3	1	29	20320	1
4	2	25	41230	7
5	2	29	31462	5
6	1	41	7232	6
7	1	14	45352	4
8	2	27	45320	5
9	2	32	51500	5
10	2	29	45782	7
11	2	40	59990	9
12	1	13	8920	3
13	2	33	23250	5
14	1	3	35000	6
15	1	18	14235	2
16	1	21	25550	3
17	2	39	33330	7
18	2	31	31654	4

Independent variables are

Freq : frequency of purchase in a year

Avgpurc : average purchase by customer in a year

Years : no. of years since the customer has been purchasing from the Raymond showroom

Interpretation of Computer Output

Means (muk.sta)

	FREQ	AVGPURC	YEARS	Valid N
G_1:1	19	22225.33398	3.5555556	9
G_2:2	31.66667	40390.89063	6	9
All Grps	25.33333	31308.11133	4.7777777	18

Standard Deviations (muk.sta)

	FREQ	AVGPURC	YEARS	Valid N
G_1:1	10.75872	12235.19629	1.6666666	9
G_2:2	5.074446	11518.06348	1.5811388	9
All Grps	10.44312	14840.06836	2.0162737	18

1. HOW GOOD IS THIS MODEL?

From the output table *'Classification Matrix'*, we can observe that the discriminant function, obtained, is able to classify 94.44% of the 18 customers correctly. Specifically it says that out of 8 cases predicted to be in group 1, all 8 were observed to be in group 1. Similarly out of 10 cases predicted to be in group 2, 9 were found to be in group 2 and 1 in group 1. Thus, on the whole, only 1 case out of 18 was misclassified by the discriminant model, giving us a classification (or prediction) accuracy level of (18-2)/18, or 94.44%.

The above level of accuracy may not hold for all future classification of new customers. But it is still a pointer towards the model being a good one, assuming the input data was relevant and scientifically collected.

Classification Matrix
(muk.sta)
Rows: Observed classifications
Columns: Predicted classifications

	Percent	*G_1:1*	*G_2:2*
	Correct	*p = .50000*	*p = .50000*
G_1:1	88.88889	8	1
G_2:2	100	0	9
Total	94.44444	8	10

2. STATISTICAL SIGNIFICANCE

From the table of *'Chi-Square Tests with Successive Roots Removed'*, the Wilks' Lambda was found to be 0.3372.

Since this value is closer to 0 it indicates better discriminating power of the model. The probability value for the Chi-square test indicates that the discrimination between the two groups is highly significant. This is because *p*-value is 0.00127, which signifies that the Chi-square test is significant at a confidence level of upto (1– 0.00127)* 100 or 99.873%.

Chi-Square Tests with Successive Roots Removed
(muk.sta)

	Eigen-value	*Canonicl R*	*Wilks' Lambda*	*Chi-Sqr.*	*df*	*p-level*
0	1.965108	0.814090967	0.3372559	15.76024	3	0.001272

3. PREDICTOR

We have three independent variables—frequency of purchase in a year, average purchase by customer in a year, number of years since the customer has been purchasing from the Raymond show room.

Now we have to find out which of these is a better predictor of a 'Loyal' or 'Disloyal' customer. From the table for *'Standardised coefficients for canonical variables'*, it is observed that frequency of visiting the showroom is the best predictor with a coefficient of – 0.777, followed by average purchase in rupees by customer per year, with a coefficient of – 0.739 and number of years since the customer has been frequenting the showroom with a coefficient of – 0.227. Here one of the most vital points is that the absolute value of the standardised coefficient of each independent variable indicates its relative importance.

Standardized Coefficients (muk.sta) for Canonical Variables		
	Root 1	
FREQ	– 0.77738	
AVGPURC	– 0.73963	
YEARS	– 0.22719	
Eigenval	1.965108	
Cum.Prop	1	

4. CLASSIFICATION OF NEW CUSTOMER

Since the Raymonds Bangalore division is planning to provide a 'First Citizen Club Card' and 'Gift Hampers', this is the most important question to be answered.

From the output table of *'Raw Coefficients for Canonical Variables' and 'Means of Canonical Variables'* the mean of canonical variable gives us the new means for the transformed group centroids. Thus the new mean for group 1 (Loyal) is 1.32165, and the new mean for group 2 (Disloyal) is – 1.32165. This means that the midpoint of the two is 0. This also gives us a decision rule for classifying any new customer case. If the discriminant score of a customer is negative, we classify him as 'Disloyal' and if the discriminant score of a customer is positive, we classify him as 'Loyal'.

Raw Coefficients (muk.sta) for Canonical Variables

	Root 1
FREQ	– 0.09242
AVGPURC	– 6.2E-05
YEARS	– 0.13986
Constant	4.958395
Eigenval	1.965108
Cum.Prop	1

Means of Canonical Variables (muk.sta)	
	Root 1
G_1:1	1.321651
G_2:2	–1.32165

In order to compute the discriminant score of any customer the unstandardised discriminant function has to be used.

$$Y = 4.95839 - \text{FREQ } (0.09242) - \text{AVGPURC } (0.00006) - \text{YEARS } (0.139856)$$

Unstandardized Canonical Scores (muk.sta)		
	Group	*Root 1*
1	G_1:1	1.610945225
2	G_1:1	1.666642189
3	G_1:1	0.873456001
4	G_2:2	– 0.89760113
5	G_2:2	– 0.3795357
6	G_1:1	– 0.12018206
7	G_1:1	0.282015294
8	G_2:2	– 1.0573225
9	G_2:2	– 1.90411854
10	G_2:2	– 1.55063677
11	G_2:2	– 3.73139715
12	G_1:1	2.782102346
13	G_2:2	– 0.23804143
14	G_1:1	1.663321257
15	G_1:1	2.129007578
16	G_1:1	1.007554531
17	G_2:2	– 1.69973803
18	G_2:2	– 0.43647245

Where Y would give Raymond, the discriminant score of any customer whose frequency of purchase in a year, average purchase in a year, and number of years for which the customer has been purchasing from the Raymond showroom were known.

According to the Raymond decision rule, any positive discriminant score leads us to classify the customer in the 'loyal' group. Therefore Raymond should provide that loyal customer with a 'First Citizen Club Card' and 'Gift Hampers'.

CASE STUDY 3

Discriminant Analysis*

Problem

Procter and Gamble (Bangalore Divn), an FMCG Major, is planning to arm all its TSI (Territorial Sales Inspector) agents with a palmtop system. This would enhance their performance as they could book more orders faster and accurately which would help it to compete with its main rival. (HLL has already distributed the systems to all its TSI).

The Bangalore division has been achieving their sales target for the last two years. Hence the head of Bangalore Divn. has decided to recommend to the H.O. at Mumbai to provide all its TSI with the latest palmtops.

P & G Mumbai has decided to use discriminant analysis to screen TSIs of Bangalore Divn. on the basis of efficiency and Potential to service dealers in their respective territories. The classification will be 'potentially successful TSI' and 'potentially unsuccessful TSI', based on the information supplied by the Bangalore Divn. In total cases of 24 TSIs are being considered in terms of the above discriminant function in order to find out:

1. the percentage of TSI that the discriminant function is able to classify correctly.
2. statistical significance of the discriminant function.
3. which independent variables are relatively better in discriminating between 'successful' and 'unsuccessful' TSI.
4. how to classify a future TSI into one of the two groups—and 'potentially unsuccessful TSI'.

Input Data

The dependent variable is a categorical variable having two categories with codes as follows:

Potentially Unsuccessful TSI = 1

Potentially Successful TSI = 2

Independent variables are experience in terms of number of years, sales in rupees per year, and number of dealers covered by each TSI.

* Prepared by Biswajit Mishra, Imran Ahmed, Sharmistha Kundu, and Vikrant Sharma

Input Data for Discriminant Analysis

S. NO.	POTENTIAL	YEAREXP	SALES_YR	DLR_CVR
1	1	5	74500	32
2	2	4	185000	21
3	1	2	115000	29
4	2	7	145600	19
5	1	4	89500	37
6	1	5	79000	31
7	2	4	135000	23
8	1	3	115500	29
9	2	5	152000	25
10	2	7	175000	19
11	1	2	98000	17
12	2	5	130000	27
13	2	8	189500	21
14	2	5	150000	20
15	1	3	112000	39
16	1	2	94500	30
17	2	3	164500	29
18	1	4	78500	34
19	2	2	125500	17
20	1	5	135000	29
21	1	8	85000	36
22	2	5	105500	21
23	2	8	148500	27
24	1	2	94500	24

Analysis of Output

Means

	YEAREXP	*SALES_YR*	*DLR_CVR*	*Valid N*
G_1:1	3.75	97583.33594	30.58333397	12
G_2:2	5.25	150508.3281	22.41666603	12
All Grps	4.5	124045.8359	26.5	24

Standard Deviations				
	YEAREXP	SALES_YR	DLR_CVR	Valid N
G_1:1	1.815338731	18383.33398	5.9613657	12
G_2:2	1.912875056	24951.78906	3.776923656	12
All Grps	1.978141427	34497.80078	6.420077324	24

1. Model

Classification Matrix Rows: Observed Classifications Columns: Predicted Classifications			
	Percent	G_1:1	G_2:2
	Correct	p = .50000	p = .50000
G_1:1	91.66666412	11	1
G_2:2	100	0	12
Total	95.83333588	11	13

From the output table 'Classification Matrix', we can observe that the discriminant function obtained, is able to classify 95.83% of the 24 TSI correctly. Specifically it says that out of 11 cases predicted to be in group 1, all 11 were observed to be in group 1. Similarly out of 13 cases predicted to be in group 2, 12 were found to be in group 2 and 1 in group 1. Thus on the whole only one case out of 24 was misclassified by the discriminant model, thus giving us a classification (or prediction) accuracy level of (24-1)/24, or 95.83% .

The above level of accuracy may not hold for all future classification of new TSIs. But it is still a pointer towards the model being a good one, assuming the input data were relevant and scientifically collected.

1. Statistical Significance

Chi-square Tests with Successive Roots Removed

Wilks' Lambda	Chi-Sqr.	df	p-level
0.28583199	25.67319679	3	1.12178E-05

From the table of **'Chi-Square Tests with Successive Roots Removed'**, the Wilks' Lambda was found to be 0.285. Since this value is closer to 0 it indicates **better discriminating power of the model**. The probability value for the Chi-square test indicates that **the discrimination between the two groups is highly significant.** This is because *p*-value is 0.00001, which signifies that Chi-square test would be significant at a confidence level of upto (1-0.00001)*100 or 99.999%.

2. Better Predictor

We have three independent variables namely, YEARS OF EXPERIENCE, SALES IN RUPEES PER YEAR, DEALERS COVERED. Now we have to find out which of these is a better predictor of a TSI being 'Potentially unsuccessful' and 'Potentially successful'.

Standardized Coefficients for Canonical Variables	
	Root 1
YEAREXP	0.384273112
SALES_YR	0.7218467
DLR_CVR	– 0.595397234
Eigenval	2.49855876
Cum.Prop	1

From the output table for 'Standardised coefficients for canonical variables', it was observed that sales in Rs. per year by a TSI is the best predictor with a coefficient of 0.721, followed by no. of dealers covered by TSI, with a coefficient of – 0.595. Experience in terms of years is the last one, with a coefficient of 0.384. Here we note that the absolute value of the standardised coefficient of each independent variable indicates its relative importance.

3. Classification of New TSI

Since P & G Bangalore division is planning to provide palmtops to its TSI, this is the most important question to be answered.

Raw Coefficients (bisda.sta) for Canonical Variables	
	Root 1
YEAREXP	0.206072778
SALES_YR	3.29384E-05
DLR_CVR	– 0.119314693
Constant	– 1.851359606
Eigenval	2.49855876
Cum.Prop	1

Means of Canonical Variables	
	Root 1
G_1:1	– 1.513388753
G_2:2	1.513388753

From the output table of **'Raw Coefficients for Canonical Variables'** and **'Means of Canonical Variables'**, the mean of canonical variables gives us the new means for the transformed group centroids. Thus the new mean for group 1 (Potentially unsuccessful TSI) is – 1.513388, and the new means for group 2 (Potentially successful TSI) is 1.513388. The means that the midpoint of the two is 0. This also gives us a decision rule for classifying any new TSI case. If the discriminant score of a TSI falls to the left of the midpoint (negative), we classify him as 'Potentially unsuccessful TSI' and if the discriminant score of a TSI falls to the right of the midpoint (positive), we classify him 'Potentially successful TSI'.

Unstandardised Canonical Scores		
Sl. No.	*Group*	*Root 1*
1	G_1:1	– 2.185155153
2	G_2:2	2.560926914
3	G_1:1	– 1.111424088
4	G_2:2	2.120001793
5	G_1:1	– 2.4937253
6	G_1:1	– 1.91761744
7	G_2:2	0.675377965
8	G_1:1	– 0.88888222
9	G_2:2	1.202774167
10	G_2:2	3.088390827
11	G_1:1	– 0.239600435
12	G_2:2	0.239499435
13	G_2:2	3.533441067
14	G_2:2	1.733470559
15	G_1:1	– 2.197313547
16	G_1:1	– 1.905975938
17	G_2:2	0.725099444
18	G_1:1	– 2.498103619
19	G_2:2	0.666205525
20	G_1:1	0.165562496
21	G_1:1	– 1.698342085
22	G_2:2	0.148397073
23	G_2:2	1.467078567
24	G_1:1	– 1.190087795

In order to compute the discriminant score of any TSI the unstandardised discriminant function has to be used.

$$Y = -1.85135 + \text{YEAREXP}\ (0.20607) + \text{SALES} - \text{YR}\ (0.00003) - \text{DLR_CVR}\ (0.11931)$$

Where Y would give P&G the discriminant score of any TSI whose experience, sales per year, and number of dealers covered by that TSI were known. According to P&G's decision rule, any positive discriminant score leads them to classify the TSI into the 'Potentially successful TSI' group. Therefore P&G should arm that TSI with a palmtop as he would be a potentially successful TSI. Similarly any negative discriminant score leads them to classify the TSI into the 'Potentially unsuccessful TSI' group, therefore P&G should not provide that TSI with a palmtop as he would be a potentially unsuccessful TSI.

The McGraw·Hill Companies

CHAPTER 12 LOGISTIC REGRESSION FOR CLASSIFICATION AND PREDICTION

Learning Objectives

In this chapter, we will

- Introduce the concept of logistic regression
- Compare it with discriminant analysis
- Explain its working through an example
- Discuss pros and cons of its usage in the real world

APPLICATION AREAS

Like Discriminant Analysis, Logistic Regression is used to distinguish between two or more groups. While either technique can be used for such applications, Logistic Regression is generally preferred when there are only two categories of the dependent variable, while Discriminant Analysis is the method of choice when there are more than two. Typical application areas are cases where one wishes to predict the likelihood of an entity belonging to one group or another, such as in response to a marketing effort (likelihood of purchase/non-purchase), creditworthiness (high/low risk of default), insurance (high/low risk of accident claim), medicine (high/low risk of some ailment like heart attack), sports (likelihood of victory/loss).

METHODS

Unlike Multiple Linear Regression or Linear Discriminant Analysis, Logistic Regression fits an S-shaped curve to the data. To visualize this graphically, consider a simple case with only one independent variable, as in Figure 12.1 below:

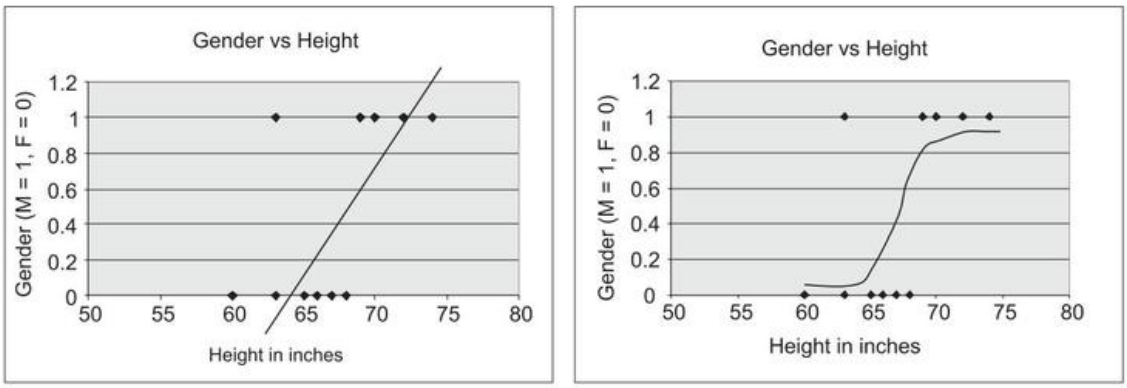


FIGURE 12.1 A Linear model vs Logistic Regression (S-curve on the right)

This curved relationship ensures two things—first, that the predicted values are always between 0 and 1 and secondly, that the predicted values correspond to the probability of Y being 1, or Male, in the above example. (Note that the two values of the dependent variable Y are coded as 0 and 1).

To achieve this, a regression is first performed with a transformed value of Y, called the *Logit* function. The equation (shown below for two independent variables) is:

$$Logit(\mathrm{Y}) = \ln(odds) = a + k_1x_1 + k_2x_2 \tag{12.1}$$

Where *odds* refers to the odds of Y being equal to 1. To understand the difference between odds and probabilities, consider the following example:

When a coin is tossed, the probability of Heads showing up is 0.5, but the odds of belonging to the group "Heads" are 1.0. Odds are defined as the probability of belonging to one group divided by the probability of belonging to the other.

$$\text{odds} = \mathrm{p}/(1-\mathrm{p}) \tag{12.2}$$

For the coin toss example, odds = 0.5/0.5 = 1.

The above can be rewritten in terms of probability p as:

$$\mathrm{p} = \text{odds}/(1+\text{odds}) \tag{12.3}$$

Substitute the values for a coin toss to convince yourself that this makes sense.

Now going back to equation 12.1, we see that the right hand side is a linear function like regression that does not guarantee values between 0 and 1. We can take the exponent of each side of the equation to get

$$\mathrm{e}^{\ln(odds)} = odds = \mathrm{e}^{(a + k1x1 + k2x2)} \tag{12.4}$$

Dividing both sides of Equation (12.4) by (1+*odds*) gives us

$$odds\ /\ (1+odds) = \mathrm{e}^{(a + k1x1 + k2x2)} / (1 + \mathrm{e}^{(a + k1x1 + k2x2)}) \tag{12.5}$$

From equations (12.3) and (12.5), we see that

$$\mathrm{p} = [\ \mathrm{e}^{(a + k1x1 + k2x2)}] / [(1 + \mathrm{e}^{(a + k1x1 + k2x2)})] \tag{12.6}$$

This equation yields p, the probability of belonging to a group (Y = 1) rather than the log of the odds of the same. SPSS will automatically compute the coefficients $k1$ and $k2$ for the regression shown in Equation (12.1). We can then compute the probabilities of belonging to group 1 by using the above transformation. It should be fairly easy to see that the right hand side of Equation (12.6) can only yield values that are between 0 and 1.

The Algorithm

While Linear models use the *Ordinary Least Squares* (OLS) estimation of coefficients, Logistic regression uses the *Maximum Likelihood Estimation* (MLE) technique. In other words, it tries to estimate the odds that the dependent variable values can be predicted using the independent variable values. This is done by starting out with a random set of coefficients and then iteratively improving them based on improvements to the *log likelihood* measure. After a few iterations, the process stops when further improvement is insignificant, based on some predetermined criteria.

Logistic Regression *Versus* Linear Discriminant Analysis

In terms of predicting power, there is a debate over which technique performs better and there is no clear winner. As stated before, the general view is that Logistic Regression is preferred for binomial dependent variables, while discriminant is better when there are more than 2 values of the dependent.

However, there is a practical advantage to Logistic Regression. An organization such as a credit card lender typically has several models of this type (with binary dependents) to predict various aspects of a customers' behaviour. For instance, each month it may evaluate all its customers to see how likely they are to become delinquent in the near future. They may have another model to predict the likelihood of a customer leaving them to switch to a competitor, and yet another to predict the ones most likely to purchase additional financial products. If each of these models is built using Logistic regression, then all the different scores that a customer receives on these models can be interpreted the same way. For instance, a score of 0.75 on any of the models means the same thing—it implies a 75% chance that the customer will belong to the group with value 1 (as opposed to 0) for that model. From a manager's point of view, this is much easier to deal with. With discriminant analysis, two different models that both yield the same score can mean different things. A score of .75 may be a "good" score on one model and a "bad" one on another.

Mathematically too, Logistic Regression is less encumbered by the assumptions of Discriminant Analysis. The independent variables in Logistic Regression may be anything from Nominal to Ratio scaled and there are no distribution assumptions.

NUMERICAL EXAMPLE WITH SPSS

In order to see how Logistic Regression works and to compare it with Discriminant Analysis, consider the case study described in the Discriminant Analysis chapter (Case Study 2) on *Customer Loyalty* at a Raymond's showroom. The data are shown below again. The only change is that Loyalty has been coded as 0 (Loyal) and 1 (disloyal) rather than 1 and 2.

TABLE 12.1 Input Data

Freq	*AvgPurc*	*Years*	*Loyalty*
15	24765	3	0
17	18654	4	0
29	20320	1	0
25	41230	7	1
29	31462	5	1
41	7232	6	0
14	45352	4	0
27	45320	5	1
32	51500	5	1
29	45782	7	1
40	59990	9	1
13	8920	3	0
33	23250	5	1
3	35000	6	0
18	14235	2	0
21	25550	3	0
39	33330	7	1
31	31654	4	1

Independent Variables are

Freq : Frequency of purchase in a year

Avgpurc : Average purchase by customer in a year

Years : Number of years the customer has been purchasing from Raymond

Interpretation of Output

Note that Logistic Regression was able to classify all 100% of the cases correctly.

TABLE 12.2 Classification Table

			Predicted LOYALTY		Percentage Correct
	Observed		**0**	**1**	
Step 1	LOYALTY	**0**	9	0	100.0
		1	0	9	100.0
	Overall Percentage				100.0

The cut value is .500

Statistical Significance

Statistical significance can be tested using the *-2 log likelihood* measure, which essentially measures how likely it is that one can predict the observed values of Y based on the model. Since likelihood varies between 0 and 1, the log value varies from negative infinity to 0. The negative log, therefore goes from 0 to infinity. A prediction likelihood of 1, or a perfect model, has a corresponding *-2 log likelihood* value of 0. This is the case in this example, as shown below.

TABLE 12.3 Model Summary

Step	*-2 Loglikelihood*	*Cox & Snell R Square*	*Nagelkerke R Square*
1	.000	.750	1.000

Another statistical test that is used is the *Hosmer and Lemeshow goodness-of-fit test* (Chi-Square test), which tests how well the model fits. Lack of significance on this test is a good thing – in this case, with a significance value of 1.000, the test is completely insignificant, meaning that there is no significant difference between the model predictions and observed values.

TABLE 12.4 Hosmer and Lemeshow Test

Step	*Chi-square*	*df*	*Sig.*
1	.000	5	1.000

Predictors

The variables in the equation and their beta coefficients are shown in the table below.

TABLE 12.5 Variables in the Equation

		B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
Step 1	FREQ	9.478	1203.310	.000	1	.994	13075.018
	AVGPURC	.006	.786	.000	1	.994	1.006
	YEARS	–5.733	2257.418	.000	1	.998	.003
	Constant	–416.973	51162.247	.000	1	.993	.000

Variable(s) entered on Step 1: FREQ, AVGPURC, YEARS.

As in a regression, we can compute the score for any observation. Consider the first observation in our data, with values of 15, 24765 and 3 for each of the three independent variables, respectively. The score for this person is

$$-416.973 + 9.478\ (15) + 0.006\ (24765) - 5.733\ (3) = -133.68$$

This score indicates the log of the odds of being disloyal (dependent value of 1), as shown earlier in Equation (12.1). To convert this into a probability (p) of being disloyal, we use the transformation shown in Equation (12.6):

$$p = e^{-133.68}/[1+ e^{-133.68}] = 0$$

Since the probability of disloyalty is 0, this person will be classified by the model as loyal (forecasted value of dependent is 0). Note that in the Table 12.5, the Wald's test indicates that the variables are not significant. For very small samples such as this one, the Wald's test is not meaningful and should be ignored.

Classification of New Customer

As shown above for an existing customer, the values of the independent variables are used to compute a score, which is then transformed to get a probability of disloyalty. If this probability is greater than 0.5, the customer will be classified as disloyal (1). If less than 0.5, then he/she will be classified as loyal (0).

SUMMARY

Logistic Regression proved to be a better predictor than discriminant analysis in the example shown in this chapter. While this is no guarantee that it will always outperform discriminant analysis, it is a method that can be applied to the same class of problems with fewer restrictive assumptions about the data. For binary classification problems, it is generally considered a superior technique.

This technique also has the advantage of providing us with the probability of an observation belonging to one of the two classes defined by the binary dependent variable. Thus, upon scoring an entity, the score (suitably transformed) will range between 0 and 1, and will directly reflect the probability of it belonging to group 1. Thus score interpretation is an easy task. The tradeoff is the extra computation involved in the transformation (a minor one given the computing power available).

The coefficients in the model are computed using the maximum likelihood estimation (MLE) and the significance of the model is tested using the—*2 log likelihood* measure or the *Hosmer and Lemeshow Chi-Square test*.

ASSIGNMENT QUESTIONS

1. Discuss the types of situations where Logistic Regression is the appropriate technique to use.
2. Compare and contrast Logistic Regression with Linear Discriminant Analysis.
3. Think of a binary dependent variable in any industry of your choice. Collect data on the dependent variable and about 5 to 10 variables that may be useful in predicting it. Develop a Logistic Regression Model to predict the dependent variable.

SPSS COMMANDS FOR LOGISTIC REGRESSION

1. Click on **Analyze, Regression, Binary Logistic**.
2. Select the dependent variable, independent variables (covariates) and method (The **Enter** Method is preferred, where you choose to enter all the independent variables in the model. For small sample sizes, the stepwise procedure will tend to throw out most variables, since the statistical tests like Wald's will find them insignificant.
3. Click on **SAVE** and select Probabilities and Group Membership, then choose Continue.
4. Click on **OPTIONS** and select the statistics (like the Hosmer Lemeshow test) and plots you wish to see.
5. Click **OK** to get the output.

CASE STUDY 1

Logistic Regression

Problem

We wish to predict the likelihood of India winning a One Day International match against the West Indies. Data is collected from 18 recent matches on the following variables.

1. **DravidScr** is Dravid's score in the match. Since Dravid's batting is seen as instrumental to India's chances, we wish to see if his score has in fact had an impact on India's victories.
2. Does batting first help or hinder? The variable **BatFirst** is coded 1 if India bats first, 0 otherwise.
3. Taking early wickets helps, so say the experts. **Wicket15** shows the number of West Indies wickets to fall in the first 15 overs of their batting.
4. Finally, **IndiaWin** is the dependent variable, a 1 indicating a victory for India, 0 a loss.

Input Data

The following data has been collected (note that this data is fictitious, for illustration purposes only).

TABLE 1 Raw Data

Obs	*DravidScr*	*BatFirst*	*Wicket15*	*IndiaWin*
1	15	0	3	1
2	17	0	4	1
3	29	0	1	0
4	25	1	0	0
5	29	0	2	0
6	41	1	2	1
7	14	0	3	0
8	27	1	1	0
9	32	1	0	0
10	75	1	2	0
11	88	1	0	1
12	105	0	0	1
13	123	1	1	1
14	86	1	1	0
15	90	1	2	0
16	65	0	3	1
17	89	0	2	1
18	77	0	4	1

Output from SPSS

TABLE 2 Hosmer and Lemeshow Test

Step	*Chi-square*	*df*	*Sig.*
1	8.555	7	.286

The Hosmer and Lemeshow test shows lack of significance, indicating that the model predictions are not significantly different from observed values. In other words, the model fits well.

TABLE 3 Classification Table

		Predicted Did India Win? (1 = Yes)		**Percentage Correct**
Observed		0	1	
Did India Win? (1 = Yes)	0	8	1	88.9
	1	3	6	66.7
Overall Percentage				77.8
The cut value is .500				

The classification table above shows that the overall correct classification rate of the model is 77.8%. The model predicts India's losses better than it predicts the wins.

TABLE 4 Variables in the Equation

	B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
DRAVDSCR	.043	.023	3.388	1	.066	1.044
BATFIRST	–1.854	1.548	1.434	1	.231	.157
WICKET15	.661	.640	1.066	1	.302	1.936
Constant	–2.708	2.180	1.543	1	.214	.067

Variable(s) entered on Step 1: DRAVDSCR, BATFIRST, WICKET15.

Note that the Wald's test shows that all the variables in the model are insignificant. This is not reliable for small sample sizes and should be ignored. A good test of a variable is how the model classifies with or without it present.

Interpretation of Individual Coefficients

Consider the coefficient 0.661 for the variable Wicket15. Since the equation predicts log of odds, taking the exponent ($e^{0.661}$) gives us 1.936, which represents the impact of Wickets in the first 15 overs

on the odds of India winning. In other words, if we assume for simplicity that 1 West Indies wicket in the first 15 overs gives India odds of 1 to 1 of winning (prob. of 0.5), then 1 additional wicket will almost double the odds to 1.936 to 1 (probability of 0.66, approximately).

The prediction equation is given by:

Log (odds India Win) = –2.708 + 0.043(DravidScr) – 1.854(BatFirst) + .661(Wickets15)

For example, consider the first observation in the raw data, where Dravid's score is 15, India batted second, and took 3 West Indies wickets in the first 15 overs. The equation yields –2.708 + 0.043(15) – 1.854 (0) + 0.661(3) = – 0.08.

Thus, Odds of India Win = $e^{-0.08} = 0.923116$

Probability of India Win = $e^{-0.08} / (1+ e^{-0.08}) = 0.48$

Based on this probability, SPSS predicts a loss for India (using a probability cutoff of 0.5 for a predicted win). In this instance, the prediction does not match the observed value of an India win.

SPSS prints out the predicted probabilities and the predicted classification (win/loss), as shown in Table 5:

TABLE 5 Predicted Probabilities and Classification

DravidScr	*BatFirst*	*Wicket15*	*IndiaWin*	*Pred_Prob*	*Pred_Class*
15	0	3	1	.47883	0
17	0	4	1	.65959	1
29	0	1	0	.30836	0
25	1	0	0	.02950	0
29	0	2	0	.46329	0
41	1	2	1	.18417	0
14	0	3	0	.46818	0
27	1	1	0	.06024	0
32	1	0	0	.03938	0
75	1	2	0	.49119	0
88	1	0	1	.30980	0
105	0	0	1	.85564	1
123	1	1	1	.79502	1
86	1	1	0	.44378	0
90	1	2	0	.64698	1
65	0	3	1	.88616	1
89	0	2	1	.91813	1
77	0	4	1	.96179	1

CASE STUDY 2

Logistic Regression

A pharmaceutical firm that developed a particular drug for women wants to understand the characteristics that cause some of them to have an adverse reaction to a particular drug. They collect data on 15 women who had such a reaction and 15 who did not. The variables measured are:

1. Systolic Blood Pressure
2. Cholesterol Level
3. Age of the person
4. Whether or not the woman was pregnant (1 = yes).

The dependent variable indicates if there was an adverse reaction (1 = yes).

TABLE 1 Raw Data

BP	*Cholesterol*	*Age*	*Pregnant*	*DrugReaction*
100	150	20	0	0
120	160	16	0	0
110	150	18	0	0
100	175	25	0	0
95	250	36	0	0
110	200	56	0	0
120	180	59	0	0
150	175	45	0	0
160	185	40	0	0
125	195	20	1	0
135	190	18	1	0
165	200	25	1	0
145	175	30	1	0
120	180	28	1	0
100	180	21	1	0
100	160	19	1	1
95	250	18	1	1
120	200	30	1	1
125	240	29	1	1
130	172	30	1	1

(*Table 1 contd*)

(*Table 1 contd*)

120	130	35	1	1
120	140	38	1	1
125	160	32	1	1
115	185	40	1	1
150	195	65	0	1
130	175	72	0	1
170	200	56	0	1
145	210	58	0	1
180	200	81	0	1
140	190	73	0	1

SPSS Output

TABLE 2 Model Summary

Step	*-2 Log likelihood*	*Cox & Snell R Square*	*Nagelkerke R Square*
1	21.841(a)	.482	.643

Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

TABLE 3 Hosmer and Lemeshow Test

Step	*Chi-square*	*df*	*Sig.*
1	4.412	8	.818

The lack of significance of the Chi-Squared test indicates that the model is a good fit.

TABLE 4 Classification Table

Observed			*Predicted*		
			DrugReaction		
			0	*1*	*Percentage Correct*
Step 1	DrugReaction	0	11	4	73.3
		1	2	13	86.7
	Overall Percentage				80.0

The cut value is .500.

The classification table shows that the model makes a correct prediction 80% of the time overall. Of the 15 women with no reaction, the model correctly identified 11 of them as not likely to have one. Similarly, of the 15 who did have a reaction, the model correctly identifies 13 as likely to have one.

TABLE 5 Variables in the Equation

		B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
Step 1(a)	**BP**	–.018	.027	.463	1	.496	.982
	Cholesterol	.027	.025	1.182	1	.277	1.027
	Age	.265	.114	5.404	1	.020	1.304
	Pregnant	8.501	3.884	4.790	1	.029	4918.147
	Constant	–17.874	10.158	3.096	1	.078	.000

Variable(s) entered on Step 1: BP, Cholesterol, Age, Pregnant.

Since BP and Cholesterol show up as not significant, one can try to run the regression again without those variables to see how it impacts the prediction accuracy. Since the sample size is low, one cannot assume that they are insignificant. Wald's test is best suited to large sample sizes.

The prediction equation is:

Log (odds of a reaction to drug) = –17.874 – 0.018 (BP) + 0.027 (Cholesterol) + 0.265 (Age) + 8.501 (Pregnant).

As with any regression, the positive coefficients indicate a positive relationship with the dependent variable.

TABLE 6 Predicted Probabilities and Classification

BP	*Cholesterol*	*Age*	*Pregnant*	*Drug Reaction*	*Pred_Prob*	*Pred_Class*
100	150	20	0	0	.00003	0
120	160	16	0	0	.00001	0
110	150	18	0	0	.00002	0
100	175	25	0	0	.00023	0
95	250	36	0	0	.03352	0
110	200	56	0	0	.58319	1
120	180	59	0	0	.60219	1
150	175	45	0	0	.01829	0
160	185	40	0	0	.00535	0
125	195	20	1	0	.24475	0
135	190	18	1	0	.12197	0

(*Table 6 contd*)

(*Table 6 contd*)

165	200	25	1	0	.40238	0
145	175	30	1	0	.65193	1
120	180	28	1	0	.66520	1
100	180	21	1	0	.30860	0
100	160	19	1	1	.13323	0
95	250	18	1	1	.58936	1
120	200	30	1	1	.85228	1
125	240	29	1	1	.92175	1
130	172	30	1	1	.69443	1
120	130	35	1	1	.76972	1
120	140	38	1	1	.90642	1
125	160	32	1	1	.75435	1
115	185	40	1	1	.98365	1
150	195	65	0	1	.86545	1
130	175	72	0	1	.97205	1
170	200	56	0	1	.31892	0
145	210	58	0	1	.62148	1
180	200	81	0	1	.99665	1
140	190	73	0	1	.98260	1

The table above shows the predicted probabilities of an adverse reaction, and the classification of each into group 0 or 1 on the basis of that probability, using 0.5 as the cut-off score.

CASE STUDY 3

Logistic Regression

A market researcher wants to understand the characteristics of people that determine which one of two brands of red wine they prefer. Data is collected on 30 people (15 of whom prefer Brand A and 15 who prefer Brand B), with age, gender and income as the three independent variables.

Choice: 0 = Brand A, 1= Brand B

Age in years

Gender: 0 = Female, 1 = Male

Income: Thousands of Rupees per month.

Age	*Gender*	*Income*	*Choice*
21	0	3	0
22	0	8	0
29	0	7	0
25	0	5	0
29	0	7	0
41	0	9	0
35	0	12	0
42	0	12	0
35	0	15	0
30	0	6	0
25	1	8	0
26	1	9	0
23	1	11	0
26	1	11	0
25	1	12	0
65	1	13	1
89	1	12	1
77	1	11	1
50	1	15	1
40	1	16	1
38	1	10	1
35	1	12	1
28	1	13	1

29	1	15	1
40	0	20	1
44	0	22	1
56	0	25	1
22	1	9	1
30	1	7	1
32	1	18	1

SPSS Output

TABLE 1 Model Summary

Step	*-2 Log likelihood*	*Cox & Snell R Square*	*Nagelkerke R Square*
1	12.661	.619	.825

TABLE 2 Hosmer and Lemeshow Test

Step	*Chi-square*	*df*	*Sig.*
1	2.837	8	.944

The high p-value (0.944) on the Hosmer-Lemeshow test suggests that it is extremely likely that there is no significant difference between the model predictions and observed values, indicating a good model.

TABLE 3 Classification Table

		Predicted Product Choice (0=Wine A,	1 = Wine B)	Percentage Correct
Observed		0	1	
Product Choice	0	15	0	100.0
(0=Wine A, 1 = Wine B)	1	1	14	93.3
Overall Percentage				96.7

The cut value is .500.

The classification table shows that the model does an excellent job of correctly classifying people into two groups—those that prefer Brand A and those that prefer Brand B. It correctly classifies 29 of the 30 cases overall, based on the age, income and gender.

TABLE 4 Variables in the Equation

	B	*S.E.*	*Wald*	*df*	*Sig.*	*Exp(B)*
AGE	.443	.288	2.356	1	.125	1.557
GENDER	8.783	4.785	3.370	1	.066	6524.416
INCOME	.398	.248	2.564	1	.109	1.488
Constant	–24.706	12.916	3.659	1	.056	.000

Variable(s) entered on Step 1: AGE, GENDER, INCOME.

The prediction equation is

Log (odds of Choice 1) = –24.706 + 0.443 (Age) + 8.783 (Gender) + 0.398 (Income)

Each of the coefficients is positive, indicating that older people, males, and those with higher incomes will tend to prefer Brand B.

TABLE 5 Predicted Probabilities and Classification

Age	*Gender*	*Income*	*Choice*	*Pred_Prob*	*Pred_Class*
21	0	3	0	.00000	0
22	0	8	0	.00001	0
29	0	7	0	.00011	0
25	0	5	0	.00001	0
29	0	7	0	.00011	0
41	0	9	0	.04867	0
35	0	12	0	.01169	0
42	0	12	0	.20793	0
35	0	15	0	.03752	0
30	0	6	0	.00012	0
25	1	8	0	.15814	0
26	1	9	0	.30327	0
23	1	11	0	.20340	0
26	1	11	0	.49081	0
25	1	12	0	.47950	0
65	1	13	1	1.00000	1
89	1	12	1	1.00000	1
77	1	11	1	1.00000	1

(*Table 5.1 contd*)

(*Table 5.1 contd*)

50	1	15	1	.99999	1
40	1	16	1	.99971	1
38	1	10	1	.99246	1
35	1	12	1	.98721	1
28	1	13	1	.83807	1
29	1	15	1	.94694	1
40	0	20	1	.72253	1
44	0	22	1	.97135	1
56	0	25	1	.99996	1
22	1	9	1	.06894	0
30	1	7	1	.53605	1
32	1	18	1	.99552	1

As an illustration of the computation of predicted probability, consider the last observation, with values 32, 1 and 18 for the independent variables age, gender and income respectively.

$$\text{Log (odds)} = -24.706 + 0.443\ (\text{Age}) + 8.783\ (\text{Gender}) + 0.398\ (\text{Income})$$

$$= -24.706 + 0.443\ (32) + 8.783\ (1) + 0.398\ (18) = 5.417$$

Hence $p = e^{5.417} / (1 + e^{5.417}) = 0.9955$, as shown in the table above. This probability is greater than 0.5, so the person is classified into group 1 (shown in last column), which corresponds to Brand B. This predicted classification agrees with the observed value of *Choice*.

CHAPTER 13 FACTOR ANALYSIS FOR DATA REDUCTION

Learning Objectives

In this chapter, we will

- Introduce Factor Analysis as a technique for data reduction or for combining variables
- Differentiate between factors and original variables from which they are obtained
- Explain the concept of Factor Extraction and Factor Rotation
- Demonstrate the above two concepts with an application
- Introduce the use of Eigen values to determine the number of factors to be extracted
- Explain the use of factor loadings to identify and interpret factors
- Discuss how results of factor analysis can be used by marketers

APPLICATION AREAS

Factor analysis is a very useful method of reducing data complexity by reducing the number of variables being studied. It is a common experience, for example, to find a marketing decision maker wondering what exactly makes a consumer buy his product. The possible purchasing criteria could range from just one or two to fifteen or twenty, and often, the marketing manager is shooting in the dark, trying to figure out what *really* drives buyer behaviour. In other words, what are the underlying significant drivers of his behaviour? Factor analysis is a good way of resolving this confusion and identifying *latent* or underlying factors from an array of seemingly important variables. In a more general way, factors analysis is a set of techniques which, by analysing correlations between *variables*, reduces their number into fewer *factors* which *explain* much of the original data, more economically. Even though a subjective interpretation can result from a factor analysis output, the procedure often provides an insight into relevant psychographic variables, and results in economical use of data

collection efforts. The subjective element of factor analysis could be reduced by splitting the sample randomly into two, and extracting factors separately from both parts. If similar factors result, the analysis could be assumed as reliable or stable.

Other creative uses of factor analysis could be to do it separately for 2 groups such as users and non-users of a brand, and check what differences exist in the factors extracted. This would be an indirect way of finding out differences in buying criteria for the same product category among the two groups. Of course, if the objective is to classify a person by predicting whether he will be a buyer or non-buyer, the reader is advised to consider discriminant analysis instead of factor analysis.

Methods

There are two stages in factor analysis. **Stage 1** can be called the Factor Extraction process, where our objective is to identify how many factors will be extracted from the data. The most popular method for this is called Principal Components Analysis. There is also a rule-of-thumb based on the computation of an Eigen value, to determine how many factors to extract.

The concept of Eigen value translates approximately to the 'variance explained' concept of regression analysis. The higher the Eigen value of a factor, the higher is the amount of variance explained by the factor. What we are attempting to do is to extract the least number of factors possible which will maximise the explained variance. Before extraction, it is assumed that each of the original variables has an Eigen value = 1. Therefore, it stands to reason that we would expect any single factor, which is a linear combination of some of the original variables, to exceed the value of 1. Indeed, that is exactly what the rule-of-thumb for factor extraction says. Theoretically, we can have as many factors as there are original variables. But since the objective is to reduce the *variables* to a fewer number of *factors*, we usually retain those with an Eigen value of 1 or more (in other words, a factor must explain at least as much of the variance if not more, than a single original variable).

Stage 2 is called *Rotation* of principal components. This is actually optional, but highly recommended. After the number of extracted factors is decided upon in Stage 1, the next task of the researcher is to interpret and name the factors. This is done by the process of identifying which factors are associated with which of the original variables. The factor matrix is used for this purpose. The original factor matrix is unrotated, and is a part of the output from Stage 1. The rotated factor matrix comes about in Stage 2, where we request that the computer package perform a rotation and give us a rotated factor matrix. The factor matrix (whether unrotated or rotated) gives us the loading of each variable on each of the extracted factors. This is similar to a correlation matrix, with 'loadings' having values between 0 and 1. Values close to 1 represent high loadings and those close to 0, low loadings. The objective is to find variables which have a high loading on one factor, but low loadings on other factors. If factor 1 is loaded highly by variables 3, 8, and 10, for example, it is assumed that factor 1 is a linear combination of these three variables, and it is given a suitable name, representing the essence of the original variables, of which it is a combination. This process is somewhat subjective, but in skilful hands, results in a very useful interpretation, as we will illustrate.

There are two popular methods of orthogonal rotation, the varimax and the quartimax. Any of these two is adequate for the average user of factor analysis in marketing research.

Recommended Usage

The author recommends using factor analysis as described above, to reduce data variables into a smaller set of *factors*. The analysis could be started by observing through a correlation matrix, if correlations exist between at least some of the original variables. Unless some of the variables are correlated with each other, factor analysis is not recommended. It may result in forced extraction of non-existing factors. Formal tests like Bartlett's test of sphericity can also be used to ensure that there are some significant correlations among the variables in the input data.

Stage 1 should be performed using Principal Components Analysis followed by varimax (or quartimax) rotation in Stage 2. Both the unrotated and rotated factor matrix should be looked at for interpretation of factors. Unless the researcher is doing only an exploratory factor analysis, the thumb rule of Eigen value being equal to 1 or more should be used to determine the number of factors to be extracted. The procedure will be illustrated through the worked example which follows.

WORKED EXAMPLE

The complete problem is given in Appendix 1 at the end of this chapter. In this example, we assume that a two-wheeler manufacturer is interested in determining which variables his potential customers think about when they consider his product. Let us assume that twenty two-wheeler owners were surveyed by this manufacturer (or by a marketing research company on his behalf). They were asked to indicate on a seven point scale (1 = completely agree, 7 = completely disagree), their agreement or disagreement with a set of ten statements relating to their perceptions and some attributes of the two-wheelers. The objective of doing Factor Analysis is to find underlying 'factors' which would be fewer than 10 in number, but would be linear combinations of some of the original 10 variables.

The ten statements were as follows;

1. I use a two-wheeler because it is affordable.
2. It gives me a sense of freedom to own a two-wheeler.
3. Low maintenance cost makes a two-wheeler very economical in the long run.
4. A two-wheeler is essentially a man's vehicle.
5. I feel very powerful when I am on my two-wheeler.
6. Some of my friends who don't have their own vehicle are jealous of me.
7. I feel good whenever I see the ad for my two-wheeler on TV, in a magazine or on a hoarding.
8. My vehicle gives me a comfortable ride.
9. I think two-wheelers are a safe way to travel.
10. Three people should be legally allowed to travel on a two-wheeler.

Input Data

The answers given by the twenty respondents on a scale of 1 to 7 are in the data matrix labelled Table 13.1. These are the same as the numbers given in Appendix 1.

The McGraw·Hill Companies

Interpretation of Computer Output

The output of factor analysis is obtained by requesting Principal Components Analysis and specifying a rotation. In Tables 13.2, 13.3 and 13.4, we have the output of the factor analysis for this problem—the unrotated factor matrix (Table 13.2), the final statistics (Table 13.3) comprising the communality for all 10 variables, and the Eigen values of all factors having Eigen values of 1 or more than 1. We assume that we are extracting factors of Eigen values 1 or more, and have specified this in the request to the computer package while doing factor analysis.

The first step in interpreting the output is to look at the factors extracted, their Eigen values and the cumulative percentage of variance (Table 13.3). We see from the Cum. Pct. column that the three factors extracted together account for 80.3 per cent of the total variance (information contained in the original ten variables). This is a pretty good bargain, because we are able to economise on the number of variables (from 10 we have reduced them to 3 underlying factors), while we lost only about 20 per cent of the information content (80 per cent is retained by the 3 factors extracted out of the 10 original variables).

Now the second task which follows is that of interpreting what these 3 extracted factors represent. This we can accomplish by looking at Tables 13.4 and 13.2, the rotated and unrotated factor matrices.

Looking at Table 13.4, the rotated factor matrix, we notice that variable nos. 4, 5, 6 and 7 have loadings of 0.96986, 0.96455, 0.94544 and 0.97214 on factor 1 (we look down the factor 1 column in Table 13.4, and look for high loadings close to 1.00). This suggests that factor 1 is a combination of these four original variables. Table 13.2 also suggests a similar grouping. Therefore, there is no problem interpreting factor 1 as a combination of 'a man's vehicle' (statement in variable 4), 'feeling of power' (variable 5), 'others are jealous of me' (variable 6), and 'feel good when I see my two-wheeler ads'.

At this point, the researcher's task is to find a suitable phrase which captures the essence of the original variables which continue to form the underlying concept or 'factor'. In this case, factor 1 could be named 'male ego', or 'machismo', or 'pride of ownership' or something similar to these. With the same mathematical output, interpretations of different researchers may differ.

Now we will attempt to interpret factor 2. We look in Table 13.4, down the column for factor 2, and find that variables 8 and 9 have high loadings of 0.85203 and 0.87772, respectively. This indicates that factor 2 is a combination of these two variables.

But if we look at Table 13.2, the unrotated factor matrix, a slightly different picture emerges. Here, variable 3 also has a high loading on factor 2, along with variables 8 and 9. It is left to the researcher which interpretation he wants to use, as there are no hard and fast rules. Assuming we decide to use all three variables, the related statements are 'low maintenance', 'comfort' and 'safety' (from statements 3, 8, and 9). We may combine these variables into a factor called 'utility' or 'functional features' or any other similar word or phrase which captures the essence of these three statements/variables.

For interpreting factor 3, we look at the column labelled factor 3 in Table 13.4 and find that variables 1 and 10 are loaded high on factor 3. According to the unrotated factor matrix of Table 13.2, only variable 10 loads high on factor 3. Supposing we stick to Table 13.4, then the combination of 'affordability' and 'cost-saving by 3 people legally riding on a 2-wheeler' give the impression that factor 3 could be 'economy' or 'low cost'.

We have now completed the interpretation of the 3 factors with Eigen values of 1 or more. We will now look at some additional issues which may be of importance in using factor analysis.

ADDITIONAL COMMENTS

While interpreting the factor matrix to decide which variables load highly on each factor, we looked at only the concerned factor column. For example, we looked at column factor of Table 13.4, while interpreting factor 1. What we must also guard against is the possibility that a variable may load highly on more than one factors. Strictly speaking, a variable should load close to 1.00 on one and only one factor, and load close to 0 on the other factors. If this is not the case, it indicates that either the sample of respondents has more than one opinion about the variable, or that the question/variable may not be clear in its phrasing.

The other issue important in the practical use of factor analysis is the answer to the question 'what should be considered a high loading and what is not a high loading?' Here, unfortunately, there is no clear-cut guideline, and many a time, we must look at relative values in the factor matrix. Sometimes, 0.7 may be treated as a high value, while sometimes 0.9 could be the cutoff for high values. The researcher has to bring his experience with similar problems to solve this problem.

Communality is the other issue. The proportion of variance in any one of the original variables which is captured by the extracted factors is known as communality. For example, Table 13.3 tells us that after 3 factors were extracted and retained, the communality is 0.72243 for variable 1, 0.45214 for variable 2 and so on (all values are in column labelled communality in Table 13.3). This means that 0.72243 or 72.24 per cent of the variance (information content) of variable 1 is being captured by our 3 extracted factors together. Variable 2 exhibits a low communality value of 0.45214. This implies that only 45.214 per cent of the variance in variable 2 is captured by our extracted factors. This may also partially explain why variable 2 is not appearing in our final interpretation of the factors (in the earlier section). It is possible that variable 2 is an independent variable which is not combining well with any other variable, and therefore should be further investigated separately. 'Freedom' could be a different concept in the minds of our target audience.

As a final comment, it is again the author's recommendation that we use the rotated factor matrix (rather than unrotated factor matrix) for interpreting factors, particularly when we use the principal components method for extraction of factors in stage 1.

Appendix 1

Assignment: Factor Analysis

Twenty two-wheeler users were surveyed about their perceptions and image attributes of the vehicles they owned. Ten questions were asked to each of them, all answered on a scale of 1 to 7 (1 = completely agree, 7 = completely disagree).

1. I use a two-wheeler because it is affordable.
2. It gives me a sense of freedom to own a two-wheeler.
3. Low maintenance cost makes a two-wheeler very economical in the long-run.
4. A two-wheeler is essentially a man's vehicle.
5. I feel very powerful when I am on my two-wheeler.
6. Some of my friends who don't have their own vehicle are jealous of me.
7. I feel good whenever I see the ad for my two-wheeler on TV, in a magazine or on a hoarding.
8. My vehicle gives me a comfortable ride.
9. I think two-wheelers are a safe way to travel.
10. Three people should be legally allowed to travel on a two-wheeler.

S. No.	*QUESTION NO.*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	1	4	1	6	5	6	5	2	3	2
2	2	3	2	4	3	3	3	5	5	2
3	2	2	2	1	2	1	1	7	6	2
4	5	1	4	2	2	2	2	3	2	3
5	1	2	2	5	4	4	4	1	1	2
6	3	2	3	3	3	3	3	6	5	3
7	2	2	5	1	2	1	2	4	4	5
8	4	4	3	4	4	5	3	2	3	3
9	2	3	2	6	5	6	5	1	4	1
10	1	4	2	2	1	2	1	4	4	1
11	1	5	1	3	2	3	2	2	2	1
12	1	6	1	1	1	1	1	1	2	2
13	3	1	4	4	4	3	3	6	5	3

14	2	2	2	2	2	2	2	1	3	2
15	2	5	1	3	2	3	2	2	1	6
16	5	6	3	2	1	3	2	5	5	4
17	1	4	2	2	1	2	1	1	1	3
18	2	3	1	1	2	2	2	3	2	2
19	3	3	2	3	4	3	4	3	3	3
20	4	3	2	7	6	6	6	2	3	6

Factor Analyze the above 20 × 10 matrix and determine if there are any 'factors' underlying the ten variables, which can explain the variance in the responses. Name these factors.

TABLE 13.1

S.No.	*Var 1*	*Var 2*	*Var 3*	*Var 4*	*Var 5*	*Var 6*	*Var 7*	*Var 8*	*Var 9*	*Var 10*
1	1.00	4.00	1.00	6.00	5.00	6.00	5.00	2.00	3.00	2.00
2	2.00	3.00	2.00	4.00	3.00	3.00	3.00	5.00	5.00	2.00
3	2.00	2.00	2.00	1.00	2.00	1.00	1.00	7.00	6.00	2.00
4	5.00	1.00	4.00	2.00	2.00	2.00	2.00	3.00	2.00	3.00
5	1.00	2.00	2.00	5.00	4.00	4.00	4.00	1.00	1.00	2.00
6	3.00	2.00	3.00	3.00	3.00	3.00	3.00	6.00	5.00	3.00
7	2.00	2.00	5.00	1.00	2.00	1.00	2.00	4.00	4.00	5.00
8	4.00	4.00	3.00	4.00	4.00	5.00	3.00	2.00	3.00	3.00
9	2.00	3.00	2.00	6.00	5.00	6.00	5.00	1.00	4.00	1.00
10	1.00	4.00	2.00	2.00	1.00	2.00	1.00	4.00	4.00	1.00
11	1.00	5.00	1.00	3.00	2.00	3.00	2.00	2.00	2.00	1.00
12	1.00	6.00	1.00	1.00	1.00	1.00	1.00	1.00	2.00	2.00
13	3.00	1.00	4.00	4.00	4.00	3.00	3.00	6.00	5.00	3.00
14	2.00	2.00	2.00	2.00	2.00	2.00	2.00	1.00	3.00	2.00
15	2.00	5.00	1.00	3.00	2.00	3.00	2.00	2.00	1.00	6.00
16	5.00	6.00	3.00	2.00	1.00	3.00	2.00	5.00	5.00	4.00
17	1.00	4.00	2.00	2.00	1.00	2.00	1.00	1.00	1.00	3.00
18	2.00	3.00	1.00	1.00	2.00	2.00	2.00	3.00	2.00	2.00
19	3.00	3.00	2.00	3.00	4.00	3.00	4.00	3.00	3.00	3.00
20	4.00	3.00	2.00	7.00	6.00	6.00	6.00	2.00	3.00	6.00

TABLE 13.2 Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*
VAR00001	.17581	.66967	.49301
VAR00002	– .13577	– .60774	.25369
VAR00003	– .10651	.81955	.21827
VAR00004	.96647	– .03627	– .09745
VAR00005	.95098	.16594	– .13593
VAR00006	.95184	– .08442	– .02522
VAR00007	.97128	.09591	– .04636
VAR00008	– .32171	.77498	– .03757
VAR00009	– .06890	.73502	– .48213
VAR00010	.16143	.31862	-.81356

TABLE 13.3 Final Statistics

Variable	*Communality*	*	*Factor*	*Eigenvalue*	*Pact of Var*	*Cum Pct*
VAR00001	.72243	*	1	3.88282	38.8	38.8
VAR00002	.45214	*	2	2.77701	27.8	66.6
VAR00003	.73056	*	3	1.37475	13.7	80.3
VAR00004	.94488	*				
VAR00005	.95038	*				
VAR00006	.91376	*				
VAR00007	.95474	*				
VAR00008	.79869	*				
VAR00009	.77745	*				
VAR00010	.78946	*				

TABLE 13.4 Rotated Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*
VAR00001	.13402	.34749	.76402
VAR00002	– .18143	– .64300	– .07596
VAR00003	– .10944	.62985	.56742
VAR00004	.96986	– .06383	– .01338
VAR00005	.96455	.13362	.04660
VAR00006	.94544	– .13868	.02600
VAR00007	.97214	.02862	.09411
VAR00008	– .26169	.85203	.06517
VAR00009	.00891	.87772	– .08347
VAR00010	.07209	– .10990	.87874

SUMMARY

The techniques of factor analysis provide a fascinating way of reducing the number of variables in a research problem to a smaller and more manageable number by combining related ones into factors. This relieves the researcher from the confusion arising through overlapping measures of the same underlying variables. Also, the cost of further research may be reduced by focussing efforts on fewer variables for study.

Factor analysis has many alternative algorithms that can be used to extract factors out of a set of variables. The method used in this chapter is known as **Principal Components Analysis.** The primary decision in stage 1 of factor analysis is to decide how many factors to extract from the data. The simple rule of thumb normally used says that all factors with an **Eigen value** of 1 or more should be extracted. This is also the default on many computer packages.

In stage 2, we look at either the Unrotated or Rotated factor matrix, to assign variables to factors and to interpret factors. The recommended method in this chapter uses the **Rotated factor matrix**. This matrix should be viewed columnwise. For each column (factor), the variables which have a high (close to 1) loading should be identified and a combined meaning for the factor found. This leads to a phrase which is the name given to the factor.

These factors can be used to design communication or marketing strategies. Factor analysis also provides an estimate of the variance explained by each factor, which can be used as a measure of its relative importance.

The chapter-end cases provide further illustrations of the technique. SPSS commands for Factor Analysis are given after the Assignment Questions.

ASSIGNMENT QUESTIONS

1. Explain how factor analysis is useful to a marketing manager.
2. Are there any independent and dependent variables in factor analysis? Explain your answer.
3. Explain the concept of an Eigen value, and how it is used as a criterion to decide on the number of factors to extract.
4. What is the maximum number of factors that can be extracted from a set of 'K' variables?
5. If all the original variables in the data are uncorrelated with each other, are you likely to get a good, interpretable factor analysis output? Explain.
6. 'A factor is a linear combination of some of the original variables'. Is the statement correct, or wrong?
7. How do you use the loadings in the factor matrix to decide on which variables are associated with which factor? Explain.
8. What do you do after you locate the variables associated with a factor?
9. How can you use the results of a factor analysis?

10. Is it better, in general, to use the rotated factor matrix, or the unrotated factor matrix for interpreting the factors? Why?
11. Formulate a marketing problem where factor analysis could be useful. Solve it using any computer package available to you.

SPSS COMMANDS FOR FACTOR ANALYSIS

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the output for a **Factor Analysis** problem similar to that described in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on DATA REDUCTION, followed by FACTOR.
3. on the dialogue box which appears, select all the variables required for the factor analysis by clicking on the right arrow to transfer them from the variable list on the left to the variables box on the right.
4. click on EXTRACTION in the lower part of the dialogue box.
   i. Select "Principal Components" as the Method.
   ii. Under DISPLAY, select "Unrotated Factor Solution".
   iii. Under EXTRACT, select "Eigen values over 1".
   iv. Under ANALYZE, choose "Correlation Matrix".
   v. Click CONTINUE.
5. click on ROTATION in the lower part of the main dialogue box. Select VARIMAX from the options under METHOD. Click CONTINUE.
6. click OK to get the FACTOR ANALYSIS output, including the unrotated factor matrix, the rotated factor matrix using varimax rotation and the extracted factors along with Eigen values and cumulative variance. Communality figures would also be a part of the output.

Note: It is possible to use other methods such as Generalised Least Squares to get the factor analysis output instead of Principal Components. It is also possible to use other rotation methods instead of varimax.

CASE STUDY 1

Factor Analysis*

The Problem

Private domestic airline operators raised their airfare in December 2000. Indian Airlines, the government-owned carrier, was the only domestic airline which did not follow suit. Information available (source—*Economic Times*) showed that people still prefer to fly Jet Airways. The reasons for the above preference had to be ascertained, as according to Indian Airlines officials domestic flyers are price-conscious customers.

This study consisted of 20 respondents who had recently flown with Jet Airways. They were asked to indicate on a seven point scale (1 = completely agree, 7 = completely disagree), their agreement or disagreement with the set of 10 statements relating to their perceptions and attributes of the airline.

The 10 statements were as follows:

1. They (Jet Airways) are always on time.
2. The seats are very comfortable.
3. I love the food they provide.
4. Their air-hostesses are very beautiful.
5. My boss/friend flies with the same airline.
6. The airlines have younger aircrafts.
7. I get the advantage of a frequent flyer program.
8. It (the flight timing) suits my schedule.
9. My mom feels safe when I fly Jet.
10. Flying Jet compliments my lifestyle and social standing in the society.

Methodology

We tabulated data from the above study and analysed it using the SPSS package.

Input data are in Table 1.

* Prepared by Aditi Sood, Mukesh Sinha, Vikram Rathi, and Vinita Nair

TABLE 1 Input Data

S.No.	Ontime	Comfort	Food	Hostess	Boss	Airyoung	Freqfly	Schedule	Mom	Lifestyl
1	1	2	2	3	1	1.00	2	2.00	1	2
2	2	1	2	2	5	2.00	2	1.00	2	2
3	1	3	1	4	6	2.00	3	2.00	5	3
4	3	4	2	2	4	3.00	2	4.00	1	3
5	4	2	4	3	2	4.00	3	2.00	3	3
6	5	3	4	2	6	5.00	2	2.00	1	2
7	2	1	2	1	5	2.00	1	1.00	4	1
8	1	2	1	4	2	2.00	4	2.00	2	4
9	1	1	1	2	3	1.00	2	2.00	5	2
10	2	5	1	3	1	1.00	2	5.00	3	2
11	2	4	2	6	3	2.00	5	5.00	2	5
12	2	1	2	4	5	2.00	4	1.00	1	4
13	3	2	3	7	6	3.00	6	2.00	2	6
14	2	2	2	5	2	1.00	5	2.00	3	5
15	2	1	1	2	2	2.00	2	1.00	1	2
16	1	1	1	3	4	1.00	3	1.00	2	3
17	4	1	4	2	6	4.00	2	1.00	5	2
18	1	2	2	2	2	2.00	2	1.00	7	2
19	2	3	2	1	2	2.00	2	3.00	7	3
20	2	1	2	3	1	2.00	3	1.00	5	3

Factor Analysis

TABLE 2 Initial Statistics

Analysis number 1 Listwise deletion of cases with missing values

Extraction 1 for analysis 1, Principal Components Analysis (PC)

Variable	Communality	*	Factor	Eigenvalue	Pct of Var	Cum Pct
AIRYOUNG	1.00000	*	1	3.17748	31.8	31.8
BOSS	1.00000	*	2	3.04991	30.5	62.3
COMFORT	1.00000	*	3	1.84465	18.4	80.7
FOOD	1.00000	*	4	.90728	9.1	89.8
FREQFLY	1.00000	*	5	.66616	6.7	96.5

(*Table 2 contd*)

(*Table 2 contd*)

HOSTESS	1.00000	*	6	.13481	1.3	97.8
LIFESTYL	1.00000	*	7	.11183	1.1	98.9
MOM	1.00000	*	8	.05619	.6	99.5
ONTIME	1.00000	*	9	.04697	.5	100.0
SCHEDULE	1.00000	*	10	.00470	.0	100.0

PC extracted 3 factors.

Factor Matrix

	Factor 1	*Factor 2*	*Factor 3*
AIRYOUNG	.95311	– .01970	.11627
BOSS	.60655	.03874	– .30474
COMFORT	– .04379	.40603	.87620
FOOD	.91327	.02804	.01859
FREQFLY	– .00961	.92653	– .33551
HOSTESS	– .05238	.94071	– .23625
LIFESTYL	– .00669	.93326	– .22845
MOM	– .18286	– .39749	.03273
ONTIME	.93655	.05809	.22847
SCHEDULE	– .15542	.48397	.83448

Factor Analysis

TABLE 3 Final Statistics

Variable	*Communality*	*	*Factor*	*Eigenvalue*	*Pct of Var*	*Cum Pct*
AIRYOUNG	.92233	*	1	3.17748	31.8	31.8
BOSS	.46227	*	2	3.04991	30.5	62.3
COMFORT	.93450	*	3	1.84465	18.4	80.7
FOOD	.83519	*				
FREQFLY	.97111	*				
HOSTESS	.94349	*				
LIFESTYL	.92321	*				
MOM	.19251	*				
ONTIME	.93271	*				
SCHEDULE	.95474	*				

Skipping Rotation 1 for extraction 1 in analysis 1

TABLE 4 Rotated Factor Matrix

Variable	*Factor 1*	*Factor 2*	*Factor 3*
AIRYOUNG	.95933	– .03985	.02061
BOSS	.57807	.14912	– .32537
COMFORT	.03743	.09013	.96175
FOOD	.91167	.03695	– .05175
FREQFLY	– .02820	.98503	– .00529
HOSTESS	– .06199	.96456	.09631
LIFESTYL	– .01589	.95577	.09722
MOM	– .18386	– .38898	– .08606
ONTIME	.95351	– .00419	.15331
SCHEDULE	– .07652	.17546	.95817

INTERPRETATION OF OUTPUT

The output of factor analysis is obtained by requesting principal component analysis and specifying the rotation. There are two stages in factor analysis. Stage one being the factor extraction process, wherein the objective is to identify how many factors are to be extracted from the data. The most popular method for this is called principal component analysis. There is also a rule-of-thumb based on the computation of an Eigen value, to determine how many factors to extract. The higher the Eigen value of a factor, the higher is the amount of variance explained by the factor.

As evident from Table 3 (looking at the Cumulative Pct column), we find that the three factors extracted together account for 80.7% of the total variance (information contained in the original 10 variables). Hence we have reduced the number of variables from 10 to 3 underlying factors.

Looking at Table 4 we see that the variables; airyoung, food, and ontime have loadings of 0.959, 0.911, and 0.953 on **factor 1**. This suggests that factor 1 is a combination of these three variables. Table 2 also suggests the same grouping. Therefore this factor can be interpreted as 'Customer Service'.

Now for **factor 2** (in Table 4), we see that freqfly, hostess, and lifestyle have a high loading of 0.985, 0.964, and 0.955 respectively, indicating that factor 2 is a combination of these variables. Also Table 2 projects the same picture. These variables can be clubbed into a single factor called 'Flyer Incentive' factor.

As for **factor 3** (as seen in Table 4), it is evident that comfort and schedule have the highest loading of 0.961 and 0.958 and this is also reflected in the unrotated factor matrix. This factor consisting of the above two variables can be termed as 'Convenience'.

CASE STUDY 2

Factor Analysis*

Problem

Alpha Car manufacturers want to release a small car into the market. For this purpose they have surveyed twenty respondents about their perceptions and image attributes of small cars. Ten questions were asked to each of them, all answered on a scale of 1 to 7 (1 = completely agree, 7 = completely disagree).

1. The price of the car should be reasonable.
2. Fuel mileage of the car should be at least 17 km/hr.
3. A small car should be easy to maintain and to be serviced.
4. Seating should be comfortable for four adults.
5. A small car should have adequate leg space and headroom.
6. Brakes are the most critical part of a small car.
7. Collapsible steering column should be standard equipment in all the new cars.
8. Power steering is a must.
9. Interior accessories of a small car should be attractive.
10. Adequate space for luggage.

Methodology

- Data is self-generated.
- A sample of 20 respondents is taken.
- Factor analysis technique is used to determine the variables, which are to be considered.
- Quartimax rotation is performed.

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala

Input Data

TABLE 1.1

S.No.	VAR 00001	VAR 00002	VAR 00003	VAR 00004	VAR 00005	VAR 00006	VAR 00007	VAR 00008	VAR 00009	VAR 00010
1	2.00	1.00	2.00	4.00	4.00	1.00	1.00	4.00	2.00	1.00
2	6.00	5.00	6.00	1.00	4.00	2.00	1.00	3.00	2.00	5.00
3	1.00	2.00	1.00	4.00	4.00	2.00	5.00	2.00	5.00	2.00
4	4.00	4.00	5.00	2.00	3.00	4.00	3.00	4.00	3.00	3.00
5	3.00	3.00	3.00	6.00	5.00	3.00	3.00	2.00	3.00	3.00
6	5.00	4.00	4.00	1.00	1.00	1.00	2.00	2.00	2.00	4.00
7	2.00	2.00	2.00	3.00	2.00	5.00	3.00	1.00	4.00	2.00
8	1.00	2.00	1.00	7.00	6.00	2.00	2.00	4.00	3.00	1.00
9	4.00	3.00	3.00	5.00	5.00	2.00	2.00	3.00	2.00	3.00
10	6.00	5.00	6.00	2.00	3.00	1.00	2.00	4.00	1.00	5.00
11	7.00	6.00	6.00	2.00	3.00	4.00	6.00	3.00	2.00	6.00
12	3.00	4.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	4.00
13	1.00	2.00	2.00	3.00	2.00	2.00	2.00	3.00	1.00	2.00
14	2.00	1.00	2.00	1.00	1.00	1.00	3.00	4.00	2.00	1.00
15	2.00	1.00	2.00	1.00	1.00	5.00	3.00	6.00	3.00	1.00
16	2.00	2.00	2.00	1.00	3.00	2.00	2.00	2.00	2.00	2.00
17	4.00	4.00	3.00	6.00	5.00	3.00	3.00	1.00	4.00	3.00
18	1.00	1.00	1.00	1.00	2.00	1.00	2.00	6.00	1.00	1.00
19	3.00	2.00	3.00	2.00	2.00	1.00	1.00	5.00	1.00	2.00
20	3.00	2.00	3.00	2.00	1.00	2.00	6.00	5.00	1.00	2.00

The first step in interpreting the output is to look at the factors extracted, their Eigen values and the cumulative percentage of variance. We see from the cum.pct column of Table1.3 that the three factors extracted together account for 82.2% of the total variance (information contained in the ten original variables). This is pretty good bargain, because we are able to economise on the number of variables (from 10 we have reduced them to 3 underlying factors), while we lost only 17.8% of the information content (82.2% is retained by the 3 factors extracted out of the 10 original variables).

The second column of Table 1.3 shows the communality values. Communality can be defined as the proportion of variance in any one of the original variables, which is captured by the extracted factors. From Table 1.3, it can be seen that after three factors were extracted and retained, the communality is 0.94134 for variable 1, 0.98862 for variable 2, and so on. This means that 94.134% of the variance (information content of variable is being captured by the three extracted factors together. Whereas, variable 8 exhibits a low communality value of 0.56933. This implies that only 56.93% of the variance in variable 8 is captured by the extracted factors. This may also partially explain as to why variable 8 is

not considered in the final interpretation of the factors. It is possible that variable 8 is an independent variable, which may not combine well with another variable, and therefore should be further investigated separately.

Now, the second task that follows is that of interpreting what these three extracted factors represent. This we can understand by using Tables 1.4 and 1.2, the rotated and unrotated factor matrices.

Looking at Table 1.4, the rotated factor matrix, we notice that variables 1, 2, 3, and 10 have loadings of 0.96863, 0.96268, 0.95977, and 0.97469 on factor 1 respectively. This suggests that factor 1 is a combination of these four variables. Table 1.2 also suggests a similar grouping. Thus factor 1 can be interpreted as a combination of 'the price should be reasonable' (VAR 1), 'fuel mileage should be at least 17 kms/hr' (VAR 2), 'easy maintenance and servicing' (VAR 3), and 'adequate space to keep the luggage' (VAR 10). At this point, our task is to find a suitable phrase, which captures the essence of the original variables, which combine to form the underlying concept or 'Factor'. In this case factor 1 can be named as **'Economy'**.

Now we will attempt to interpret factor 2. From Table 1.4, we find that the variables 4 and 5 have high loadings of 0.88911 and 0.90771 on factor 2. This indicates that factor 2 is combination of these two variables. Similarly columns 2 and 3 of Table 1.2 also show the same grouping. Thereby, factor 2 can be interpreted as a combination of 'seating should be comfortable for four adults' (VAR 4) and 'a small car should have adequate leg space and headroom' (VAR 5). Hence factor 2 can be named as **'Spaciousness'.**

To interpret factor 3, we use last column of Table 1.4 and find that variables 6 and 7 are loaded high on factor 3. Table 1.2 also shows the similar grouping. Thereby factor 3 can be interpreted as a combination of 'Brakes form the most critical part' (VAR 6) and 'Collapsible steering column should be a standard equipment in all new cars' (VAR 7). Hence factor 3 can be named as **'Safety'**.

We have now completed the interpretation of the three factors with the help of factor loadings.

Output Data

TABLE 1.2 Factor Matrix

Variable	*Factor 1*	*Factor 2*	*Factor 3*
VAR00001	.95786	– .14241	– .05964
VAR00002	.97525	.11117	– .07180
VAR00003	.94272	– .22469	– .03990
VAR00004	– .13963	.82080	– .36332
VAR00005	.12891	.74127	– .55233
VAR00006	.18169	.42622	.69991
VAR00007	.19120	.19341	.73669
VAR00008	– .32590	– .67984	.03056
VAR00009	– .08923	.83612	.33253
VAR00010	.97882	– .00256	– .04409

TABLE 1.3 Final Statistics

Variable	*Communality*	*	*Factor*	*Eigen value*	*Pct of Var.*	*Cum Pct*
VAR00001	.94134	*	1	3.93527	39.4	39.4
VAR00002	.96862	*	2	2.68668	26.9	66.2
VAR00003	.94080	*	3	1.59342	15.9	82.2
VAR00004	.82521	*				
VAR00005	.87118	*				
VAR00006	.70454	*				
VAR00007	.61668	*				
VAR00008	.56933	*				
VAR00009	.81764	*				
VAR00010	.96004	*				

TABLE 1.4 Rotated Factor Matrix

Variable	*Factor 1*	*Factor 2*	*Factor 3*
VAR00001	.96863	– .05558	.00218
VAR00002	.96268	.18144	.09460
VAR00003	.95977	– .13942	– .01459
VAR00004	– .18533	.88911	– .01846
VAR00005	.10501	.90771	– .19028
VAR00006	.07851	.11249	.82809
VAR00007	.10700	– .11424	.76953
VAR00008	– .26082	.64915	– .28269
VAR00009	– .19737	.62187	.62607
VAR00010	.97469	.06665	.07471

CASE STUDY 3

Factor Analysis*

Problem

A cigarette manufacturer is interested in determining which variables his potential customers think about when they consider his product. To get an answer to the above, a set of 10 statements were prepared by the manufacturer with the help of a market research agency. The statements are listed below. The sample size was 20, which comprised male smokers only.

The respondents were asked the following questions and were requested to rate them on a 7-point scale (1 = completely disagree and 7 = completely agree). The answers given by the respondents were related to their perceptions and some attributes of the cigarettes.

- It makes me feel very manly.
- I feel women get attracted to me when I smoke.
- When I am in a group, it makes me feel superior with a cigarette in my hand, if others in the group are non-smokers.
- I support the ban of cigarette advertisements.
- I prefer lighter cigarettes, which cause less harm to the body.
- Tension is relieved /relaxation when I smoke.
- I endorse the view that all cigarette boxes should carry the statutory warning—"Cigarette smoking is injurious to health".
- I tend to smoke more when tension prevails in my mind.
- I smoke because I like the smell of tobacco.
- A cigarette enhances my overall personality.

Input Data

TABLE 1.1

S.No.	Var. 1	Var. 2	Var. 3	Var. 4	Var. 5	Var. 6	Var. 7	Var. 8	Var. 9	Var. 10
1	1	1	1	7	3	3	3	7	7	2
2	5	4	4	6	4	7	4	6	7	5
3	2	3	2	6	5	5	1	7	6	1
4	3	1	2	6	4	5	5	6	7	2

(*Table 1.1 contd*)

* Prepared by Harakuni Rajeev, Jitender Singh, Siddharth Agarwala, and Vidya Thekke Cherupilli

(*Table 1.1 contd*)

5	4	3	3	5	1	5	2	7	6	4
6	6	6	5	7	4	1	1	2	3	7
7	7	1	1	7	6	7	5	7	7	4
8	1	1	1	7	1	7	7	7	7	1
9	4	5	6	7	4	6	2	6	7	3
10	5	3	1	6	4	3	5	6	5	4
11	1	2	1	6	2	2	3	3	6	2
12	2	1	1	6	2	3	3	4	3	3
13	2	1	1	7	1	2	1	1	7	1
14	7	6	5	7	5	1	2	7	3	7
15	2	5	4	5	2	7	2	6	7	2
16	5	1	1	7	5	1	7	6	7	2
17	2	1	1	7	3	4	5	7	7	3
18	7	7	6	5	5	2	4	6	4	5
19	2	4	3	7	6	6	4	7	6	3
20	3	1	5	7	4	3	7	7	5	4

Factor Matrix

TABLE 1.2

Variable	*Factor 1*	*Factor 2*	*Factor 3*
VAR00001	.77112	.27848	– .18824
VAR00002	.84576	– .01080	.41042
VAR00003	.77488	.14957	.33485
VAR00004	– .15129	.20147	– .71795
VAR00005	.45747	.65975	– .25427
VAR00006	– .38258	.48225	.64023
VAR00007	– .35218	.61831	– .38608
VAR00008	– .09983	.83597	.22909
VAR00009	– .73658	.29934	.25642
VAR00010	.88771	.12770	.14985

Final Statistics

TABLE 1.3

Variable	*Communality*	*	*Factor*	*Eigen value*	*Pct of Var.*	*Cum Pct*
VAR00001	.70760	*	1	3.75348	37.5	37.5
VAR00002	.88387	*	2	1.99553	20.0	57.5
VAR00003	.73493	*	3	1.59574	16.0	73.4
VAR00004	.57892	*				
VAR00005	.70920	*				
VAR00006	.78883	*				
VAR00007	.65540	*				
VAR00008	.76129	*				
VAR00009	.69790	*				
VAR00010	.82680	*				

Rotated Factor Matrix

TABLE 1.4

Variable	*Factor 1*	*Factor 2*	*Factor 3*
VAR00001	.82814	– .04949	.13906
VAR00002	.77907	.00171	– .52623
VAR00003	.76469	.11194	– .37100
VAR00004	– .04434	– .16675	.74105
VAR00005	.65000	.30528	.43989
VAR00006	– .25009	.81709	– .24216
VAR00007	– .12542	.39503	.69543
VAR00008	.14853	.83076	.22153
VAR00009	– .62228	.55290	.07052
VAR00010	.89087	– .18173	.01107

Interpretation of the Computer Output

The output of the factor analysis is obtained by requesting the Principal Component Analysis (PCA). We get the output in Tables 1.2, 1.3, and 1. 4. Table 1.3 (the final statistics) comprises the communality for all 10 variables and the Eigen values of all factors which have Eigen values of 1 or more than 1 (in this case, we have assumed only extracting factors having Eigen values 1 or more).

The first step in interpreting the output is to look at the factors extracted. Table 1.3 states that there are 3 factors, which have an Eigen value of 1 or more than 1. The last column in the table (cum. pct.) shows that the 3 factors extracted together account for 73.4% of the total variance. This is a good deal because with only 3 factors (reducing them from 10) we have lost only about 27% of the information content, while 73% is retained by the 3 factors extracted out of the 10 original variables.

We now shift our attention to interpret what these 3 extracted factors represent. This can be done with the help of Table 1.2 (factor matrix) and Table 1.4 (rotated factor matrix).

Looking at Table 1.4, the rotated factor matrix we notice that variable no.'s 1, 2, 3, and 10, have loadings of .8281, .7790, .7646, and .8908 respectively. This suggests that factor 1 is a combination of these 4 original variables, which are:

- I feel manly when I smoke
- I feel women get attracted to me when I smoke.
- When I am in a group, it makes me feel superior with a cigarette in my hand, if others in the group are non-smokers.
- A cigarette enhances my overall personality, respectively.

At this point the researcher's task is to find a suitable phrase, which captures the essence of the original variable, which continues to form the underlying concept, or 'factor'. In this case factor 1 could be named as **Male Glorification** or something similar.

We now attempt to interpret factor 2. The 3rd column of Table 1.4 states that variables 6 and 8 have high loadings of .81709 and .83076 respectively. This indicates that factor 2 is a combination of these 2 variables which are:

- Tension is relieved /relaxation when I smoke and
- I tend to smoke more when tension prevails in my mind, respectively.

However, if we look at Table 1.2 at the unrotated factor matrix, we observe that variable 5 (I prefer lighter cigarettes which cause less harm to the body) and variable 7 (I endorse the view that all cigarette boxes should carry the statutory warning "Cigarette smoking is injurious to health)" Have a high loading on factor 2 along with variables 6 and 8. Now the choice is left to us (researchers) to decide which combination should be chosen. We choose variable 6 and 8 in the rotated factor matrix. This is because both of them show a high loading of .81709 and, 83706 respectively and more importantly the 2 can be clubbed under the heading **Smoking pattern and effect under Tension**.

For interpreting factor 3 we find that variables 4 and 7 have a high loading in Table 1.4. In this case the unrotated factor matrix (Table 1.2) is not considered, as there is no conclusive relationship between any of the variables. We therefore stick to Table 1.4 and combine the following 2 variables under the heading **Sweet Poison**. The variables are:

- I support the ban on cigarette advertisements and
- I endorse the view that all cigarette boxes should carry the statutory warning—"Cigarette smoking is injurious to health".

CHAPTER 14 CLUSTER ANALYSIS FOR MARKET SEGMENTATION

Learning Objectives

In this chapter, we will

- Introduce Cluster Analysis as a technique for combining similar objects, ideally used in segmentation applications
- Explain how to use hierarchical and non-hierarchical clustering methods
- Provide a worked example for learning to interpret the agglomeration schedule, and a dendrogram to determine the number of clusters in a dataset
- Introduce the use of *k*-means clustering method after the number of clusters has been determined
- Provide an interpretation of Final Cluster Centres to name the clusters
- Explain how these techniques are useful in marketing

APPLICATION AREAS

Cluster Analysis is a multivariate procedure ideally suited to segmentation applications in marketing research. This is because a cluster, by definition, is a group of similar objects. And segmentation involves identifying groups of target customers who are similar in buying habits, demographic characteristics, or psychographics. Of course, it is not necessary that a cluster should be only of people (customers or potential customers). There could be clusters of brands for instance, which are similar to each other, and different from other clusters.

The McGraw·Hill Companies

METHODS

The basic methods of clustering used in computer packages are of two types:

1. Hierarchical clustering or Linkage methods.
2. Non-hierarchical clustering or Nodal methods.

The first type includes methods such as single linkage, complete linkage, and average linkage. Here, we do not have to specify in advance how many clusters are to be extracted. A range of solutions is provided by the computer, from a 1-cluster solution to an n-cluster solution, where n is the number of objects being studied. The second type includes the K-means approach, where you specify in advance how many clusters are required from the data. The specified number of nodes and points closest to them are used to form initial clusters and through an iterative rearrangement, the final K clusters are determined by the package.

In each case, a distance measure has to be calculated for distances between objects (or nodes) being clustered. A choice of distance (also called similarity or proximity, or opposite of distance) measures is offered on computer programmes.

One of the most commonly used measures is Euclidean distance.

Data/Scales of Variables

Generally, interval-scaled variables are ideally suited for cluster analysis. Continuous or ratio-scaled variables can also be used but the instances of such use are rarer. The section 'Additional Comments' at the end of this chapter gives tips on using nominal or ordinal data in cluster analysis. Standardisation may be necessary if the units of measurement of different variables are widely different from each other.

RECOMMENDED USAGE

The author's recommendation is to approach a typical clustering problem in two stages.

1. Find the number of clusters in the data by running a hierarchical clustering programme on the variables (if necessary after standardisation). To do this, the programme output called the agglomeration schedule could be used.

   The agglomeration schedule shows all possible solutions from 1 cluster to (n-1) clusters, where n is the number of respondents or objects (cases). Going up from the bottom of the agglomeration schedule we look at the column called 'coefficient' to decide on the number of clusters. In this column, starting at the bottom, we calculate the difference in the value of the coefficient in neighbouring rows.

   If the maximum value of this difference occurs between say, the *third* and the *fourth* row (from the bottom) coefficients, it indicates that there might be *three* clusters (the lower row in this case). The dendrogram and the icicle plot (vertical or horizontal) can also be requested and will give essentially the same information as the agglomeration schedule, in a graphical form.

2. Once the number of clusters has been identified, a *k*-means clustering option can be run on the data. The number of clusters identified above (say, 3) is specified and the output obtained. The initial cluster centres and final cluster centres for each variable are a part of this output. *The final cluster centres* are the best solution, and are used to interpret the average values of each variable for a cluster, and thereby describe the clusters.

The above recommended procedure (both stage 1 and 2) is illustrated with the help of a worked example below.

WORKED EXAMPLE

Problem: A major Indian FMCG company wants to map the profile of its target audience in terms of lifestyle, attitudes, and perceptions. The company's managers prepare, with the help of their marketing research team, a set of 15 statements, which they feel measure many of the variables of interest. These 15 statements are given below. The respondent had to agree or disagree (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree) with each statement.

1. I prefer to use email rather than write a letter.
2. I feel that quality products are always priced high.
3. I think twice before I buy anything.
4. Television is a major source of entertainment.
5. A car is a necessity rather than a luxury.
6. I prefer fast food and ready-to-use products.
7. People are more health-conscious today.
8. Entry of foreign companies has increased the efficiency of Indian companies.
9. Women are active participants in purchase decisions.
10. I believe politicians can play a positive role.
11. I enjoy watching movies.
12. If I get a chance, I would like to settle abroad.
13. I always buy branded products.
14. I frequently go out on weekends.
15. I prefer to pay by credit card rather than in cash.

Input Data

For the purpose of this illustration, we will assume that 20 respondents answered the questionnaire above (In a real life situation, the sample size could vary, and would be higher). The input data matrix of 20 respondents × 15 variables is shown in Table 14.1.

Output and Its Interpretation

As per the author's recommended usage, the output is derived by first doing a hierarchical cluster analysis to find the number of clusters that exist in the data. These outputs are in Table 11.2 and

Figs 14.1 and 14.2 (agglomeration schedule, vertical icicle plot, and dendrogram using average linkage, respectively).

The second stage is a *K*-means (quick cluster) output with a predetermined number of clusters to be specified. In this case, the output is for 4 clusters. We will look at both stage 1 and stage 2 outputs to understand the interpretation of both stages.

Stage 1

A look at Table 14.2, the agglomeration schedule, can help us to identify large differences in the coefficient (4th column). The agglomeration schedule from top to bottom (stage 1 to 19) indicates the sequence in which cases get combined with others (or one cluster combines with another), until all 20 cases are combined together in one cluster at the last stage (stage 19). Therefore, stage 19 represents a 1-cluster solution, stage 18 represents a 2-cluster solution, stage 17 represents a 3-cluster solution, and so on, going up from the last row to the first row. We have to identify how many clusters are there in the data. We use the difference between rows in a measure called coefficient (also known as fusion coefficient) in column 4 to identify the number of clusters in the data. We must look at this figure from the last row upwards, because we would like to have lowest possible number of clusters, for reasons of economy and ease of interpretation. We see that there is a difference of (58.15 – 51.79) in the coefficients between the 1-cluster solution (stage 19) and the 2-cluster solution (stage 18). This is a difference of 6.36. The next difference is of (51.79–47.00) which is equal to 4.79 (between stage 18, the 2-cluster solution and stage 17, the 3-cluster solution). The next one after that is (47 – 46.34), only 0.66, between stage 17 and stage 16. After this, there is again a large difference between the 4-cluster and 5-cluster solutions, of (46.34 – 41.660) or 4.68. Thereafter, the differences are smaller between subsequent rows of coefficients. A large difference in the coefficient values between any two rows indicates a solution pertaining to the number of clusters which the lower row represents.

Ignoring the first difference of 6.36 which would indicate only 1 cluster in the data, we look at the next largest differences. 4.79 is the difference between row 2 from the bottom and row 3 from the bottom, indicating a 2-cluster solution. But almost the same is the difference between stage 16 and 15, indicating a 4-cluster solution. At this point, it is the judgement of the researcher, which should decide whether to go for a 2-cluster or a 4-cluster solution. Just for illustration, let us choose the 4-cluster solution.

We can also look at the icicle plot (Fig. 14.1, for the graphically minded reader) or the dendrogram (Fig. 14.2) for information as to which cases link up in what sequence to form clusters. The numbers in columns 2 and 3 of the agglomeration schedule also give the same information. The dendrogram, additionally, provides a rescaled distance measure between various cluster combines at various stages (similar to the coefficient in the agglomeration schedule but depicted in a rescaled distance measure). The case membership of each cluster is also much clearer in the dendrogram. For instance, for a 4-cluster solution, from the dendrogram, cluster 1 would consist of cases no. 4, 5, 3, 19, 20, 11, and 10. Cluster 2 would consist of cases no. 1, 2, 14, 15, and 9. Cluster 3 would consist of cases no. 13, 16, 7 and 17. Cluster 4 consists of cases no. 6, 18, 8, and 12. Another way to do this would be to decide on a cutoff distance, on the scale at the top of the dendrogram and then look at the cases which are in independent clusters at that distance. In this case, a cutoff of 19 would yield the 4 clusters we have just mentioned above.

Stage 2

Why do we need to do a *K*-means (quick clustering) at all? Why not stop at stage 1? This is because a *K*-means procedure generally gives more stable clusters, since it is an interactive procedure compared with the single-pass hierarchical methods. But it needs a pre-specified number of starting points, to get an initial position. Therefore, it is best used in combination with stage 1.

In the given problem, Tables 14.3, 14.4, 14.5 and 14.6 indicate the outputs of *K*-means clustering for a 4-cluster solution. These outputs give us the initial cluster centres, the case listing of cluster membership (i.e., which case belongs to which of the clusters), the final cluster centres (the solution), and an ANOVA table.

The final cluster centres describe the mean value of each variable for each of the 4 clusters. For example, cluster 1 is described by the mean values of variable 1 = 2.2, variable 2 = 2.2, variable 3 = 3.8, variable 4 = 3.2, and so on. Similarly, cluster 3 is described by variable 1 = 1.75, variable 2 = 2.0, variable 3 = 2.25, and variable 4 = 3.0, and so on.

We now go back to the original variables (in this case the 15 statements in our questionnaire), and interpret the clusters in terms of the 15 variables. For example, cluster 3 consists of people who are on the email rather than writing conventional letters (variable 1 value = 1.75 which is equivalent to 'agree' on the scale of 1 to 5). Similarly, they are also people who tend to think twice before buying anything (variable 3 value 2.25) in other words, careful spenders. They also agree (variable 2 value = 2.00) that quality products are always priced high—that is, they have a positive correlation in their minds about a product's quality and price.

On these same variables, cluster 2 shows people who prefer conventional mail to email (variable 1 value = 3.5 or close to 'disagree'), people who do not necessarily associate high price with good quality (variable 2 value = 3.67), and tend to be neutral about care in spending (variable 3 value = 2.67). In this way, when we compare final cluster centre values on each of the 15 variables, for 1 cluster at a time, a complete picture of the clusters emerges.

In this case, we will briefly describe each of the 4 clusters as follows:

Cluster 1

People belonging to this cluster are email users, they feel quality comes at a price. They are not careful spenders and do not think a car is a necessity. They do not like television, fast food or ready-to-use products. They are not sure whether people are more health-conscious now, they disagree that women are active purchasing decision makers and they think foreign companies have somewhat increased the efficiency of Indian companies. They feel politicians can play an active role. They do not enjoy watching movies or going out on weekends, they tend to buy branded products and like to pay cash. They might consider settling abroad.

It is thus a cluster exhibiting many traditional values, except that they have adapted to email use. They are also beginning to loosen their purse strings, and are probably in transition in some other factors like acceptance of women as decision makers and the advent of credit cards.

Cluster 2

People belonging to this cluster are regular mail writers, bargain hunters or aggressive buyers. They are not too particular about thinking before spending, not great valuers of TV and they believe the car is a luxury. They are not too fond of fast food and convenience products, do not think people are very health-conscious, they feel foreign companies have done us good and think women are active purchasing decision makers. They do not believe in politicians, do not like movies, do not want to settle abroad, do not stress on branded products, do not go out on weekends, but do prefer credit cards for payments.

It is a group which likes to use credit, spends more freely, believes in woman power, believe in economics rather than politics, and feel quality products can be cheap. Also, they seem to have a patriotic streak, as they do not want to settle abroad.

Cluster 3

People belonging to this cluster are email users, and feel quality is measured by price. They think twice before buying, they are indifferent to TV and a car is a luxury to them. They are not too fond of fast food, agree that people are health-conscious, do not think foreign companies have made us efficient, do not believe in woman power, detest politicians and enjoy watching movies. They are willing to settle abroad, always buy branded products, go out on weekends and slightly prefer cash to credit cards.

This group is not a free spending one, but health-conscious, more patriarchical, more brand loyal to branded products, but outgoing compared to other groups, even willing to go abroad to settle.

Cluster 4

People belonging to this cluster are not too particular about email, measure quality by price, and are free spending. They enjoy watching TV, think a car is necessary, are not fond of fast food, think people are health-conscious, do not think foreign companies have made us efficient, and believe in woman power. They are somewhat positive about politicians, not movie watchers, do not want to settle abroad, indifferent to branding, moderately outgoing and moderately in favour of credit cards rather than cash.

This group is optimistic, free spending and a good target for TV advertising, particularly consumer durables and entertainment. But they are not necessarily influenced by brands. They may want value for money, but if they see value, they may spend a lot.

In summary, the cluster analysis of this sample of respondents tells us a lot about the possible segments which exist in the target population.

ANOVA

The ANOVA table (Table 14.6) tells us which of the 15 variables is significantly different across the 4 clusters. The last column indicates that variables 2, 7, 8, 10, 11, 12, 13 are significant at the 0.10 level (equivalent to 90% confidence level) as they have prob. values less than 0.10. The other variables are

not statistically significant, as they all have prob. values greater than 0.10. But there is divided opinion about the utility of statistical testing for cluster analysis. Most established writers seen to feel that these tests (ANOVA or other tests) are not valid. Therefore, it is left to the researcher's judgement whether he would like to use these in determining which variables are significant. If the tests were used, then the interpretation of clusters and differences across clusters should be only on the basis of those variables which are (statistically) significantly different across clusters at 0.10 or 0.05 or some other level.

ADDITIONAL COMMENTS ON CLUSTER ANALYSIS

Objects

We have looked at an example of classifying people with interval scaled data. It is possible to classify objects such as brands, products, cities, and so on with cluster analysis. For example, which brands are clustered together in terms of consumer perceptions for a positioning exercise, or which cities are clustered together in terms of income, education, and age profile of its residents. Once the reader is aware of the basics of cluster analysis, he can begin to use it creatively. For example, a cluster analysis can be done on some of the measured variables, and then other variables can be checked to see if they also exhibit differences across clusters. In the worked out example discussed earlier, only psychographics or behavioural variables were used to get the 4 clusters. We could then see if they belonged to different places, had different education levels, or whether one gender figured predominantly in any one of the clusters.

Scale

The method of cluster analysis is ideally suited to similar interval scaled variables, because Euclidean distance is a commonly used distance measure used in the clustering process. But nominal and ordinal level data can be used after standardisation if appropriate. This may also necessitate the use of other measures of distance, more appropriate with the scales of variables being used. But this should be done with care. In general, it is a good idea to standardise the variables before clustering, if the units of measurement are radically different.

Statistical Tests

As mentioned briefly earlier, some statistical tests for cluster analysis are available. But their validity being questionable, caution is recommended in using either ANOVA or any other tests.

A general caution about cluster analysis itself is that it tends to produce different results with different methods and some methods are quite vulnerable to errors in data. So, the stability of the clusters can be checked through splitting the sample and repeating the cluster analysis.

TABLE 14.1

S. No.	Var 00001	Var 00002	Var 00003	Var 00004	Var 00005	Var 00006	Var 00007	Var 00008	Var 00009	Var 00010	Var 00011	Var 00012	Var 00013	Var 00014	Var 00015
1.	1.00	3.00	5.00	4.00	3.00	5.00	3.00	2.00	3.00	2.00	4.00	1.00	1.00	1.00	5.00
2.	2.00	3.00	2.00	3.00	4.00	4.00	3.00	2.00	2.00	2.00	4.00	2.00	2.00	2.00	4.00
3.	3.00	2.00	3.00	4.00	3.00	5.00	3.00	3.00	4.00	2.00	4.00	3.00	4.00	4.00	3.00
4.	3.00	2.00	4.00	2.00	2.00	4.00	3.00	4.00	5.00	4.00	5.00	4.00	5.00	5.00	5.00
5.	2.00	2.00	4.00	2.00	2.00	5.00	3.00	3.00	4.00	4.00	5.00	5.00	3.00	3.00	4.00
6.	2.00	4.00	3.00	3.00	5.00	4.00	4.00	2.00	3.00	4.00	5.00	4.00	3.00	3.00	3.00
7.	1.00	1.00	2.00	4.00	4.00	1.00	2.00	4.00	2.00	5.00	4.00	3.00	3.00	3.00	1.00
8.	4.00	5.00	1.00	4.00	5.00	4.00	5.00	1.00	1.00	5.00	3.00	3.00	5.00	5.00	2.00
9.	2.00	1.00	5.00	3.00	4.00	4.00	2.00	1.00	2.00	1.00	2.00	2.00	4.00	4.00	3.00
10.	5.00	2.00	4.00	3.00	2.00	5.00	1.00	5.00	3.00	2.00	5.00	1.00	2.00	2.00	1.00
11.	4.00	3.00	3.00	2.00	1.00	2.00	1.00	5.00	2.00	2.00	4.00	5.00	1.00	1.00	2.00
12.	3.00	4.00	4.00	4.00	3.00	2.00	5.00	1.00	5.00	3.00	2.00	4.00	4.00	4.00	3.00
13.	4.00	3.00	2.00	2.00	3.00	3.00	4.00	2.00	2.00	3.00	4.00	3.00	5.00	5.00	4.00
14.	1.00	2.00	2.00	4.00	2.00	5.00	1.00	3.00	5.00	4.00	3.00	2.00	2.00	2.00	5.00
15.	2.00	3.00	4.00	1.00	5.00	4.00	2.00	4.00	4.00	5.00	2.00	1.00	1.00	1.00	4.00
16.	3.00	2.00	1.00	3.00	4.00	3.00	2.00	3.00	2.00	5.00	1.00	2.00	5.00	5.00	3.00
17.	5.00	1.00	1.00	5.00	1.00	2.00	4.00	2.00	2.00	4.00	4.00	3.00	3.00	3.00	2.00
18.	3.00	5.00	5.00	3.00	5.00	5.00	5.00	1.00	2.00	3.00	4.00	4.00	2.00	2.00	1.00
19.	3.00	2.00	4.00	2.00	4.00	4.00	1.00	4.00	1.00	3.00	4.00	5.00	3.00	3.00	2.00
20.	1.00	3.00	3.00	2.00	2.00	5.00	2.00	5.00	1.00	3.00	4.00	4.00	3.00	3.00	3.00

TABLE 14.2 Agglomeration Schedule

	Clusters	*Combined*	*Stage Cluster 1st Appears*			*Next*
Stage	*Cluster 1*	*Cluster 2*	*Coefficient*	*Cluster 1*	*Cluster 2*	*Stage*
1	4	5	14.000000	0	0	5
2	19	20	16.000000	0	0	7
3	6	18	17.000000	0	0	9
4	1	2	17.000000	0	0	8
5	3	4	20.000000	0	1	11
6	13	16	25.000000	0	0	13
7	11	19	28.000000	0	2	11
8	1	14	28.500000	0	0	10
9	6	8	32.500000	0	0	12
10	1	15	34.666668	0	0	14
11	3	11	36.444443	0	7	15
12	6	12	36.666668	0	0	19
13	7	13	39.500000	0	6	17
14	1	9	41.000000	10	0	16
15	3	10	41.666668	11	0	16
16	1	3	46.342857	14	15	18
17	7	17	47.000000	13	0	18
18	1	7	51.791668	16	17	19
19	1	6	58.156250	18	12	0

Vertical Icicle Pot Using Average Linkage (Between Groups)

TABLE 14.3 Initial Cluster Centres

Cluster	*VAR00001*	*VAR00002*	*VAR00003*	*VAR00004*
1	2.0000	1.0000	5.0000	3.0000
2	4.0000	5.0000	1.0000	4.0000
3	1.0000	1.0000	2.0000	4.0000
4	4.0000	3.0000	3.0000	2.0000

Cluster	*VAR00005*	*VAR00006*	*VAR00007*	*VAR00008*
1	4.0000	4.0000	2.0000	1.0000
2	5.0000	4.0000	5.0000	1.0000
3	4.0000	1.0000	2.0000	4.0000
4	1.0000	2.0000	1.0000	5.0000

Cluster	VAR00009	VAR00010	VAR00011	VAR00012
1	2.0000	1.0000	2.0000	2.0000
2	1.0000	5.0000	3.0000	3.0000
3	2.0000	5.0000	4.0000	3.0000
4	2.0000	2.0000	4.0000	5.0000

Cluster	VAR00013	VAR00014	VAR00015
1	3.0000	4.0000	3.0000
2	5.0000	5.0000	2.0000
3	1.0000	3.0000	1.0000
4	4.0000	1.0000	2.0000

Convergence achieved due to no or small distance change. The maximum distance by which any centre has changed is .0000. Current iteration is 2.
Minimum distance between initial centres is 7.5498.

TABLE 14.4 Case Listing of Cluster Membership

Case ID	Cluster	Distance
1	1	3.583
2	1	2.871
3	1	2.417
4	1	4.609
5	4	3.990
6	2	2.925
7	3	4.257
8	2	3.902
9	1	3.826
10	4	4.486
11	4	3.704
12	2	3.986
13	2	3.771
14	3	4.016
15	3	4.016
16	3	3.623
17	2	5.088
18	2	4.346
19	4	2.706
20	4	3.053

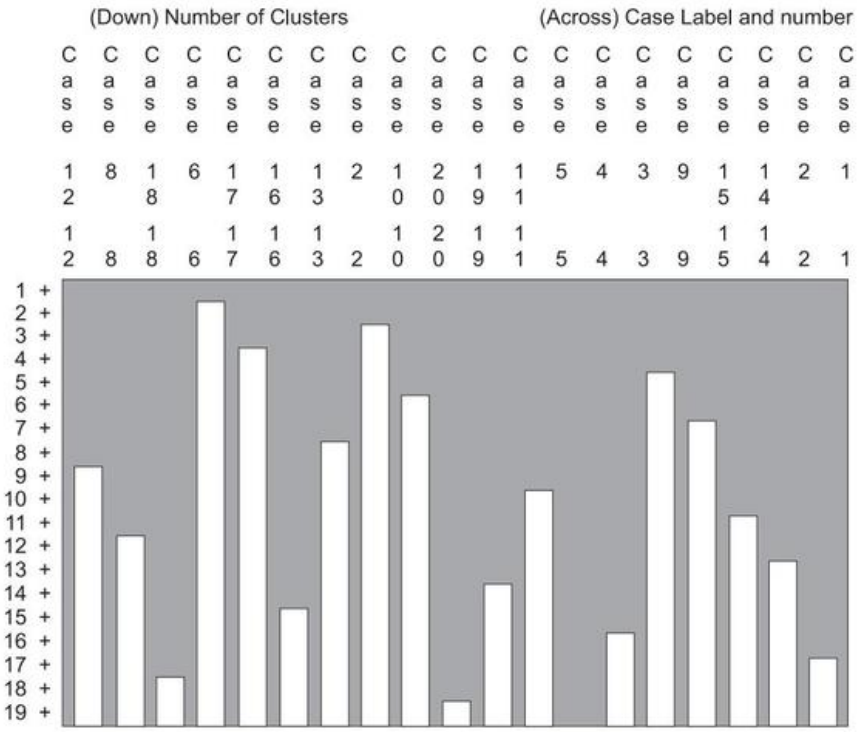


FIGURE 14.1 Vertical Icicle Plot Using Average Linkage (Between Groups)

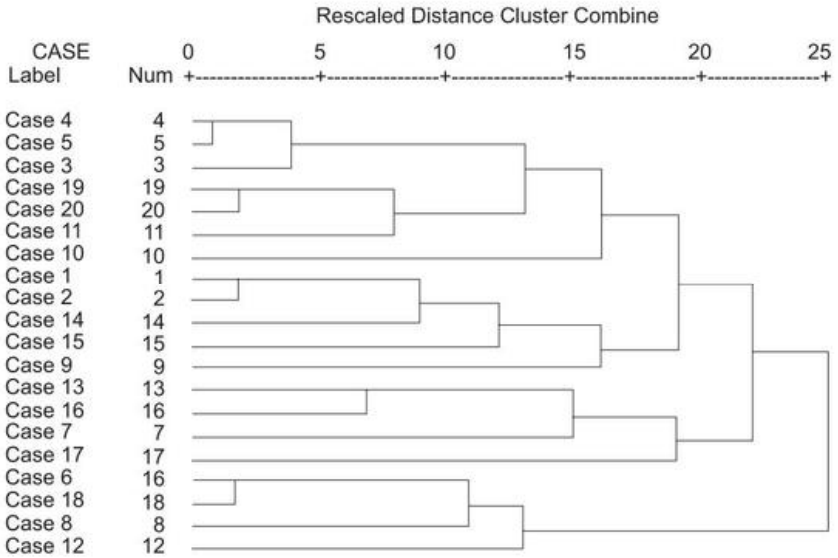


FIGURE 14.2 Dendrogram Using Average Linkage (Between Groups)

TABLE 14.5 Final Cluster Centres

Cluster	*VAR00001*	*VAR00002*	*VAR00003*	*VAR00004*
1	2.2000	2.2000	3.8000	3.2000
2	3.5000	3.6667	2.6667	3.5000
3	1.7500	2.0000	2.2500	3.0000
4	3.0000	2.4000	3.6000	2.2000

Cluster	*VAR00005*	*VAR00006*	*VAR00007*	*VAR00008*
1	3.2000	4.4000	2.8000	2.4000
2	3.6667	3.3333	4.5000	1.5000
3	3.7500	3.2500	1.7500	3.5000
4	2.2000	4.2000	1.6000	4.4000

Cluster	*VAR00009*	*VAR00010*	*VAR00011*	*VAR00012*
1	3.2000	2.2000	3.8000	2.4000
2	2.5000	3.6667	3.6667	3.5000
3	3.2500	4.7500	2.5000	2.0000
4	2.2000	2.8000	4.4000	4.0000

Cluster	*VAR00013*	*VAR00014*	*VAR00015*
1	2.4000	3.2000	4.0000
2	4.1667	3.6667	2.5000
3	1.2500	2.7500	3.2500
4	3.0000	2.4000	2.4000

Distances between Final Cluster Centres.

Cluster	*1*	*2*	*3*	*4*
1	.0000			
2	4.4417	.0000		
3	4.1310	5.7530	.0000	
4	4.0050	5.3612	4.9744	.0000

TABLE 14.6 Analysis of Variance

Variable	*Cluster MS*	*DF*	*Error MS*	*DF*	*F*	*Prob*
VAR00001	3.0500	3	1.315	16.0	2.3183	.114
VAR00002	3.0722	3	1.083	16.0	2.8359	.071
VAR00003	2.5722	3	1.630	16.0	1.5778	.234
VAR00004	1.6333	3	.943	16.0	1.7307	.201
VAR00005	2.5056	3	1.605	16.0	1.5609	.238
VAR00006	1.7056	3	1.505	16.0	1.1331	.365
VAR00007	9.6500	3	.390	16.0	24.7040	.000
VAR00008	8.5500	3	.681	16.0	12.5505	.000
VAR00009	1.3000	3	1.865	16.0	.6968	.567
VAR00010	5.556	3	.730	16.0	7.5397	.002
VAR00011	2.7389	3	1.020	16.0	2.6830	.082
VAR00012	4.0833	3	1.293	16.0	3.1562	.054
VAR00013	7.2556	3	.799	16.0	9.0813	.001
VAR00014	1.6222	3	1.880	16.0	.8628	.480
VAR00015	2.8500	3	1.465	16.0	1.9446	.163

Number of Cases in each Cluster.

Cluster	*Unweighted cases*	*Weighted cases*
1	5.0	5.0
2	6.0	6.0
3	4.0	4.0
4	5.0	5.0
Valid cases	20.0	20.0

SUMMARY

Clustering methods are many. The basic idea of cluster analysis is to group similar objects together. Some measure of similarity (or distance) is used to do this. Two basic types of clustering methods are hierarchical and non-hierarchical. This chapter demonstrates the use of both used sequentially. We first start with a hierarchical method, and try to identify the number of clusters in the data.

This can be done by looking at one of the following:

1. the agglomeration schedule, or the order in which variables combine with each other. If we use this, we can find from the bottom two rows going up, the maximum difference between the coefficients at each stage. The last row indicates one cluster, the row before that indicates a

2-cluster solution, and so on. Wherever the maximum difference between coefficients occurs, the lower row indicates the number of clusters.

2. One other way is to look at the dendrogram, and define a cutoff point for the rescaled distance given at the top. Based on this, the groupings of variables looking from the left indicate the number of clusters in the data. Generally, the cutoff point is subjective, but is based on the distance between two successive cluster formation points, as indicated by the small vertical lines.

Once the number of clusters is identified, we can use a *k*-means clustering algorithm (a non-hierarchical method). We have to specify that we want a particular number of clusters, for this to work. Once we specify this, and ask for a *k*-means clustering solution, we get the initial cluster centers, and the final cluster centres. The final cluster centres contain the mean values for each variable in each cluster. By combining the information on all variables for each cluster, we interpret the cluster and give it a descriptive name.

Cluster analysis can produce different results when different methods are used, or when different distance measures are used in the same method. To check for stable clusters, one way is to split the sample into two parts and see if similar clusters emerge from both subsamples. The cases at the end of the chapter provide further illustrations of using cluster analysis for marketing applications. It is also possible to cluster brands or other objects instead of people.

ASSIGNMENT QUESTIONS

1. What are the application areas for cluster analysis in marketing?
2. What is the major difference between hierarchical clustering methods and the *k*-means clustering method?
3. What are the two stages recommended in the chapter, for using cluster analysis?
4. Suppose we do a cluster analysis of brands by measuring some of their attributes, what would the result be? Would it be comparable in interpretation to the perceptual map/positioning map from multidimensional scaling? (You will have to read the chapter on Multidimensional Scaling to answer this question).
5. How do we decide on the number of clusters in the solution from the agglomeration schedule?
6. Describe how to interpret the final cluster centres for all the variables in the data.
7. How do we arrive at names for clusters?
8. How useful are satisfied tests in cluster analysis?
9. Will the clusters differ depending on the distance measure used? The clustering method used?
10. Is standardisation of the variables used in cluster analysis necessary? If yes, under what conditions?

SPSS COMMANDS FOR CLUSTER ANALYSIS

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the first stage output for a **Cluster Analysis** problem similar to that described in the Cluster Analysis chapter of the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on CLASSIFY followed by **HIERARCHICAL CLUSTER**.
3. on the dialogue box which appears, select all the variables which are required to be used in cluster analysis by clicking on the right arrow to transfer them from the variable list on the left. Then, under the small section called "Cluster", select CASES because you usually would be clustering cases (rows of data which normally are the respondents or objects to be grouped into clusters). In another small box called "DISPLAY", select (tick indicates selection) "Statistics" and "Plots".
4. click on METHOD. In the dialogue box which will open up, Choose "Between Groups Linkage" as the CLUSTERING METHOD. In the box titled "Measure" choose "Squared Euclidean Distance".
5. click CONTINUE to return to the main dialogue box.
6. click STATISTICS on the main dialogue box. Choose "Agglomeration Schedule" so that it will appear in the final output. Click CONTINUE.
7. click PLOTS on the main dialogue box. Choose DENDROGRAM. Then on the box called ICICLE, choose "All Clusters" and "Vertical". This will get you all the required plots on the output. Click CONTINUE to return to main dialogue box.
8. click OK on the main dialogue box to get the output of the hierarchical cluster analysis.

Stage 2

After the number of clusters has been identified using hierarchical cluster analysis, you can proceed to the second stage recommended in the chapter on cluster analysis. This stage is called K-MEANS CLUSTERING. This generally provides a stable solution, and is used if you know how many clusters you want. Some packages call it QUICK CLUSTERING too.

1. Click CLASSIFY, followed by K-MEANS CLUSTER.
2. Fill in the desired number of clusters you have identified from stage 1 above.
3. Click OPTIONS on the main dialogue box. Select "Initial Cluster Centers", "ANOVA Table" and "Cluster Information for each case" in the box labeled STATISTICS. Then click CONTINUE to return to the main dialogue box.
4. Click OK on the main dialogue box to get the output which contains the final cluster Centres from the *K*-means clustering method described in the text.

Note: Different methods will yield different results in cluster analysis at the first stage. Some subjectivity is also involved in decisions regarding how many clusters the data should be grouped into. See the text for further discussion of these issues.

CASE STUDY 1

Cluster Analysis*

Problem

Godrej India Ltd. wants to map the profile of its target customer in terms of lifestyle, attitude, and perceptions. Godrej's marketing managers prepare a set of fifteen statements, which according to the market research team will measure many of the variables of interest. The respondent had to agree or disagree with each statement on a scale of 1 to 5.

1 = Completely Agree
2 = Agree
3 = Neither Agree nor Disagree.
4 = Disagree.
5 = Completely disagree.

The following fifteen statements were prepared by the Godrej marketing team:

1. I feel foreign made products are always superior in quality.
2. I prefer to pay by credit card as a matter of convenience.
3. A computer is a necessity rather than luxury.
4. The liberalisation of the Indian Economy has increased the efficiency of Indian companies.
5. I prefer old Hindi songs than the latest ones.
6. I feel vegetarian food is more nutritious than non-vegetarian food.
7. I enjoy surfing on the net.
8. Television has become an integral part of Indian urban life.
9. Women's education is an important aspect for the overall development of the country.
10. A movie is a major source of entertainment.
11. People are more conscious about quality of product.
12. I believe the economic status of India will improve.
13. I prefer readymade wear to tailored clothes.
14. I prefer to take my food outside every weekend.
15. Basic Computer Education should be included in primary education level.

* Prepared by Biswajit Mishra, Imran Ahmed, Vikrant Sharma, and Sharmistha Kundu

Input Data

Godrej India Limited has done this market research with twenty respondents who answered the above questionnaire. The input data matrix (20 × 15) is shown in Table 1.1.

Analysis of Output

The output is derived by first doing a hierarchical cluster analysis to find the number of clusters that exist in the data. The second stage is a *K*-means (quick cluster) output with a predetermined number of clusters to be specified. In our case, the output is for 4 clusters. Now we will interpret both the stage 1 and stage 2 outputs.

Stage 1

From the first output table (Table 1.2) titled 'Agglomeration Schedule (Squared Euclidean measure)', we can try and identify how many clusters are present in the data. The agglomeration schedule from top to bottom (stage 1 to stage 19) indicates the sequence in which cases get combined with others, until all 20 cases are combined together in one cluster at the last stage (stage 19). Therefore stage 19 represents a 1-cluster solution, stage 18 represents a 2-cluster solution, stage 17 a 3-cluster solution and so on, going off from the last row to the first row. Now we have to identify how many clusters are in the data. We used the difference between rows in a measure called coefficient (fusion coefficient) to identify the number of clusters in the data. Since we would like to have the lowest possible numbers of clusters for reasons of economy and ease of interpretation, we will look at this coefficient figure from the last row upward. There is a difference of 8.169998. (i.e.; 44.419998-36.250000) in the coefficients between the cluster 1 solution (stage 19) and 2-cluster solution (stage 18). The next difference is 3.450001 between the 2-cluster solution and the 3-cluster solution. Between stage 17 & stage 16 a difference of only 0.049999 is observed. After this there is again a large difference of 4.75 between the 4-cluster (stage 16) and 5-cluster solution (stage 15). Thereafter the differences are smaller between subsequent rows of coefficients. Here, a large difference in the coefficient values between any two rows indicates a solution pertaining to the number of clusters which the lower row represents.

Ignoring the first difference of 8.169998, which would indicate only 1 cluster in the data, we look at the next largest difference, that in, 3.450001 between row 2 from the bottom and row 3 from the bottom, indicating a 2-cluster solution. But again, a difference of 4.75 was observed between the 4-cluster (stage 16) and 5-cluster solution (stage 15), indicating a 4-cluster solution. Here, since the higher difference indicated 4 clusters, we preferred to choose the 4-cluster solution.

We can also look into the tabulation of 'Vertical Icicle Plot using Average Linkage (between groups)' or 'Dendrogram using Average Linkage (between groups)' for the information as to which cases link up in what sequence to form clusters. The numbers in column 2 and 3 of the table 'Agglomeration Schedule' also give the same information. The dendrogram, additionally provides a rescaled distance measure between the various cluster combines at various stages. The case membership of each cluster is also much clearer in the dendrogram. For a 4-cluster solution, from the

TABLE 1.1 Input Data

S. No	var 0001	var 0002	var 0003	var 0004	var 0005	var 0006	var 0007	var 0008	var 0009	var 00010	var 00011	var 00012	var 00013	var 00014	var 00015
1	1.00	2.00	3.00	2.00	4.00	1.00	1.00	3.00	2.00	1.00	1.00	1.00	1.00	2.00	3.00
2	2.00	3.00	3.00	2.00	4.00	3.00	2.00	2.00	2.00	4.00	3.00	3.00	3.00	1.00	4.00
3	4.00	4.00	3.00	3.00	3.00	3.00	3.00	5.00	2.00	5.00	4.00	2.00	3.00	1.00	3.00
4	3.00	2.00	2.00	4.00	2.00	3.00	1.00	2.00	3.00	4.00	3.00	2.00	2.00	4.00	2.00
5	1.00	2.00	2.00	3.00	1.00	2.00	2.00	5.00	2.00	3.00	2.00	1.00	2.00	3.00	2.00
6	3.00	2.00	3.00	3.00	1.00	1.00	3.00	2.00	2.00	1.00	1.00	2.00	2.00	3.00	2.00
7	4.00	4.00	3.00	2.00	4.00	5.00	1.00	2.00	5.00	3.00	5.00	3.00	2.00	3.00	4.00
8	2.00	4.00	3.00	2.00	3.00	4.00	2.00	2.00	2.00	2.00	5.00	3.00	1.00	3.00	5.00
9	2.00	4.00	5.00	2.00	4.00	3.00	3.00	2.00	3.00	2.00	4.00	1.00	2.00	3.00	4.00
10	1.00	2.00	3.00	1.00	2.00	2.00	4.00	1.00	2.00	1.00	1.00	2.00	3.00	1.00	2.00
11	3.00	3.00	2.00	1.00	2.00	1.00	3.00	1.00	1.00	3.00	4.00	3.00	1.00	2.00	1.00
12	3.00	2.00	3.00	5.00	4.00	2.00	1.00	3.00	4.00	2.00	1.00	1.00	2.00	2.00	1.00
13	2.00	2.00	2.00	1.00	1.00	3.00	2.00	3.00	4.00	2.00	1.00	3.00	2.00	3.00	3.00
14	2.00	4.00	1.00	2.00	1.00	4.00	2.00	4.00	4.00	2.00	5.00	3.00	2.00	2.00	2.00
15	4.00	4.00	1.00	3.00	5.00	5.00	1.00	5.00	4.00	2.00	5.00	2.00	2.00	2.00	5.00
16	1.00	1.00	5.00	4.00	4.00	3.00	2.00	4.00	3.00	3.00	4.00	3.00	2.00	2.00	4.00
17	2.00	3.00	4.00	4.00	3.00	3.00	3.00	3.00	3.00	2.00	4.00	4.00	1.00	1.00	1.00
18	3.00	5.00	1.00	3.00	2.00	4.00	2.00	3.00	3.00	2.00	4.00	4.00	3.00	3.00	5.00
19	1.00	2.00	2.00	2.00	3.00	2.00	1.00	3.00	2.00	1.00	3.00	3.00	1.00	2.00	3.00
20	3.00	2.00	2.00	1.00	3.00	2.00	2.00	2.00	2.00	3.00	2.00	1.00	1.00	2.00	2.00

dendrogram, cluster 1 would consist of cases no. 1, 19, 5, 13, 11, 20, 6, and 10. Cluster 2 consists of 4 and 12. Cluster 3 would consist of 14, 18, 7, and 15 and similarly cluster 4 consists of 8, 9, 2, 16, 17, and 3. Another way of doing this would be to decide on a cutoff distance, on the scale at the top of the dendrogram and then looking at the cases, which form independent clusters at that distance. In our case, a cutoff of 15 has yielded the 4-cluster solution.

Data Information

20 unweighted cases accepted.

0 cases rejected because of missing value.

Squared Euclidean measure used.

Hierarchical Cluster Analysis

TABLE 1.2 Agglomeration Schedule using Average Linkage (Between Groups)

Stage	*Clusters Combined*		*Coefficient*	*Stage Cluster 1st Appears*		*Next*
	Cluster 1	*Cluster 2*		*Cluster 1*	*Cluster 2*	*Stage*
1	1	19	11.000000	0	0	11
2	11	20	15.000000	0	0	12
3	8	9	15.000000	0	0	7
4	6	10	17.000000	0	0	12
5	14	18	19.000000	0	0	15
6	7	15	20.000000	0	0	15
7	2	8	20.500000	0	3	14
8	5	13	21.000000	0	0	11
9	16	17	22.000000	0	0	14
10	4	12	23.000000	0	0	16
11	1	5	23.500000	1	8	13
12	6	11	24.000000	4	2	13
13	1	6	26.375000	11	12	16
14	2	16	28.000000	7	9	17
15	7	14	28.000000	6	5	18
16	1	4	32.750000	13	10	19
17	2	3	32.799999	14	0	18
18	2	7	36.250000	17	15	19
19	1	2	44.419998	16	18	0

Hierarchical Cluster Analysis

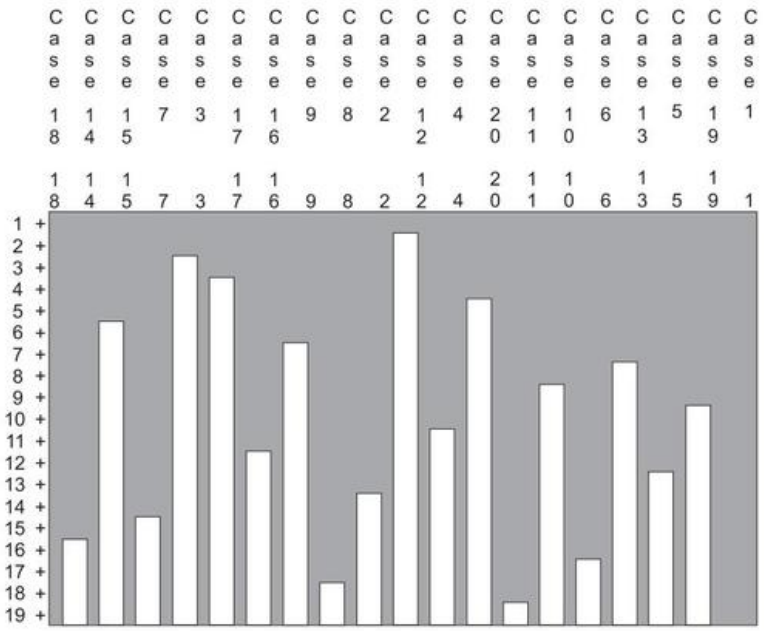


FIGURE 1 Vertical Icicle Plot Using Average Linkage (Between Groups)
(Down) Number of Clusters (Across) Case Label and Number

Hierarchical Cluster Analysis

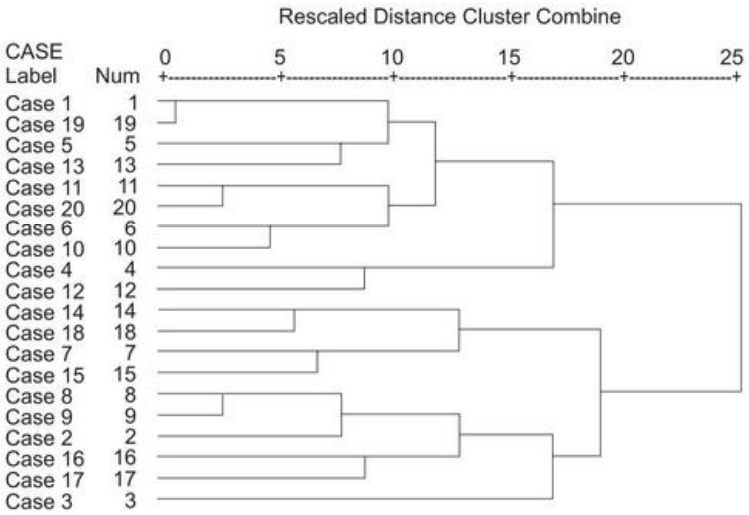


FIGURE 2 Dendrogram using Average Linkage (Between Groups)

Quick Cluster

TABLE 1.3 Initial Cluster Centres

Cluster	VAR00001	VAR00002	VAR00003	VAR00004
1	3.0000	5.0000	1.0000	3.0000
2	3.0000	3.0000	2.0000	1.0000
3	1.0000	1.0000	5.0000	4.0000
4	3.0000	2.0000	3.0000	5.0000
Cluster	VAR00005	VAR00006	VAR00007	VAR00008
1	2.0000	4.0000	2.0000	3.0000
2	2.0000	1.0000	3.0000	1.0000
3	4.0000	3.0000	2.0000	4.0000
4	4.0000	2.0000	1.0000	3.0000
Cluster	VAR00009	VAR00010	VAR00011	VAR00012
1	3.0000	2.0000	4.0000	4.0000
2	1.0000	3.0000	4.0000	3.0000
3	3.0000	3.0000	4.0000	3.0000
4	4.0000	2.0000	1.0000	1.0000
Cluster	VAR00013	VAR00014	VAR00015	
1	3.0000	3.0000	5.0000	
2	1.0000	2.0000	1.0000	
3	2.0000	2.0000	4.0000	
4	2.0000	2.0000	1.0000	

Convergence achieved due to no or small distance change.
The maximum distance by which any centre has changed is .0000
Current iteration is 3

Minimum distance between initial centres is 6.0828

Iteration	*Change in Cluster Centres*			
	1	2	3	4
1	2.7811	3.2125	2.8174	3.0414
2	.7796	.6438	.0000	.0000
3	.0000	.0000	.0000	.0000

Quick Cluster

TABLE 1.4 Case Listing of Cluster Membership

Case ID	*Cluster*	*Distance*
1	4	3.428
2	3	2.905
3	1	4.333

(*Table 1.4 contd*)

(*Table 1.4 contd*)

Case ID	*Cluster*	*Distance*
4	4	3.279
5	4	3.041
6	2	2.833
7	1	3.180
8	1	3.180
9	3	3.152
10	2	3.219
11	2	3.563
12	4	3.041
13	2	3.219
14	1	3.480
15	1	3.528
16	3	2.817
17	3	3.307
18	1	2.789
19	2	3.060
20	2	2.455

TABLE 1.5 Final Cluster Centres

Cluster	VAR00001	VAR00002	VAR00003	VAR00004
1	3.1667	4.1667	2.0000	2.5000
2	2.1667	2.1667	2.3333	1.5000
3	1.7500	2.7500	4.2500	3.0000
4	2.0000	2.0000	2.5000	3.5000
Cluster	VAR00005	VAR00006	VAR00007	VAR00008
1	3.0000	4.1667	1.8333	3.5000
2	2.0000	1.8333	2.5000	2.0000
3	3.7500	3.0000	2.5000	2.7500
4	2.7500	2.0000	1.2500	3.2500
Cluster	VAR00009	VAR00010	VAR00011	VAR00012
1	3.3333	2.6667	4.6667	2.8333
2	2.1667	1.8333	2.0000	2.3333
3	2.7500	2.7500	3.7500	2.7500
4	2.7500	2.5000	1.7500	1.2500
Cluster	VAR00013	VAR00014	VAR00015	
1	2.1667	2.3333	4.0000	
2	1.6667	2.1667	2.1667	
3	2.0000	1.7500	3.2500	
4	1.7500	2.7500	2.0000	

Distances between Final Cluster Centres.

Cluster	*1*	*2*	*3*	*4*
1	.0000			
2	5.3203	.0000		
3	3.7924	4.1592	.0000	
4	5.3098	3.1929	4.0850	.0000

Stage 2

The *K*-means procedure (also called Quick Cluster) gives more stable clusters, since it is an iterative procedure compared with the single pass hierarchical methods. But it needs a pre-specified number of starting points to get an initial position. Therefore it is best used in combination with stage one.

The output tables number 1.3 to 1.6 titled 'Initial Cluster Centres', 'Case Listing Of Cluster Membership', 'Final Cluster Centres' and 'Analysis Of Variance' indicate the outputs of *K*-means clustering for a 4-cluster solution. The final cluster centres describe the mean value of each variable for each of the 4 clusters, for instance, cluster 1 is described by the mean values of variable 1 = 3.1667, variable 2 = 4.1667, variable 3 = 2.0000, variable 4 = 2.5000, and so on. Similarly cluster 3 is described by variable 1 = 1.7500, variable 2 = 2.7500, variable 3 = 4.2500, variable 4 = 3.0000, and so on. (from Table 1.5).

We will now interpret the clusters in terms of 15 original variables (15 statements). For example cluster 2 consists of people preferring foreign made products because of their superior quality (variable 1 value = 2.1667 which is equivalent to 'agree' on the scale of 1 to 5). Similarly, there are also people who agree to pay through credit card for the reason of convenience (variable 2 value = 2.1667) and people tend to think that computers are a necessity rather than a luxury (variable 3 value = 2.3333) which is equivalent to 'agree'.

On these same variables, cluster 4 shows that quality-conscious people prefer to buy foreign made products (variable 1 value = 2.0000 or 'agree'), people prefer to pay their bill through credit card (variable 2 value = 2.0000), and tend to show the similar response for computer as a necessity (variable 3 value = 2.5000, close to 'agree'). In this way, when we compare final cluster centre values on each of the fifteen variables, for 1 cluster at a time, a complete picture of the cluster emerges. Now we will briefly describe each of the four clusters as follows:

Cluster 1

People belonging to this group are neutral to foreign products, don't prefer to pay by credit card, agree that a computer is a necessity rather than luxury, agree that liberalization of the Indian Economy has increased the efficiency of Indian companies, are indifferent to old Hindi songs and also to the latest ones, do not feel vegetarian food is more nutritious than non-vegetarian, are exuberant about surfing the net, are neutral about whether television has become an integral part of Indian urban life, are indifferent about women's education bringing about a turnaround in the overall development of the country, are undecided whether a movie is a major source of entertainment or not, disagree that people are more conscious about quality, are indifferent about the improvement of the economic status of

India, prefer readymade wear, prefer to dine outside every weekend, and are averse to the idea that basic computer education be included at primary level.

This group holds a mixed bag kind of opinion on diverse subjects. They do not seem to hold any view (undecided or indifferent) on a number of topics as well.

Cluster 2

People belonging to this cluster are foreign product users; prefer to pay by credit card, agree that a computer is a necessity rather than luxury, are optimistic that liberalisation of the Indian Economy has increased the efficiency of Indian companies, subscribe to old Hindi songs; feel vegetarian food is more nutritious than non-vegetarian, are exuberant about surfing the net, agree with the opinion that television has become an integral part of Indian urban life, are sure that women's education brings about a turnaround in the overall development of the country, agree that a movie is a major source of entertainment, subscribe to the school of thought that people are more conscious about quality, are optimistic about the improvement of the economic status of India, prefer readymade wear, like to wine and dine outside every weekend, and strongly support the idea that basic computer education be included at the primary level.

This group is optimistic because they believe in quality, are optimistic about improvement of country's economic status, and favour introduction to computer education at a primary level.

Cluster 3

People belonging to this group are foreign product users, neutral in paying by credit card, disagree that a computer is a necessity rather than luxury, are neutral to the idea that liberalisation of the Indian Economy has increased the efficiency of Indian companies, do not subscribe to old Hindi songs, are indifferent about the idea that vegetarian food is more nutritious than non-vegetarian, are exuberant about surfing the net, are neutral to the opinion that television has become an integral part of Indian urban life, are not quite sure that women's education brings about a turnaround in the overall development of the country, are indifferent that a movie is a major source of entertainment, do not subscribe to the school of thought that people are more conscious about quality, are not confident about the improvement of the economic status of India, prefer readymade wear, like to wine and dine outside every weekend, and are not firm whether basic computer education be included at the primary level.

Cluster 4

People belonging to this group are foreign product users, prefer to pay by credit card, agree that computer is a necessity rather than luxury, are not quite sure that liberalisation of the Indian Economy has increased the efficiency of Indian companies, do not differentiate between old Hindi songs and latest ones, feel vegetarian food is more nutritious than non-vegetarian, are exuberant about surfing the net, are indifferent to the opinion that television has become an integral part of Indian urban life; are not quite sure that women's education brings about a turnaround in the overall development of the country, agree that a movie is a major source of entertainment, subscribe to the school of thought that people are

more conscious about quality, are optimistic about the improvement of the economic status of India, prefer readymade wear, do not have any special liking towards wining and dining outside every weekend, and strictly support the idea that basic computer education be included at the primary level.

TABLE 1.6 Analysis of Variance

Variable	*Cluster MS*	*DF*	*Error MS*	*DF*	*F*	*Prob*
VAR00001	2.0444	3	.901	16.0	2.2690	.120
VAR00002	5.3778	3	.401	16.0	13.4095	.000
VAR00003	4.4889	3	.692	16.0	6.4802	.004
VAR00004	3.6667	3	.875	16.0	4.1905	.023
VAR00005	2.5667	3	1.343	16.0	1.9101	.169
VAR00006	6.5111	3	.479	16.0	13.5884	.000
VAR00007	1.6222	3	.630	16.0	2.5741	.090
VAR00008	2.5167	3	1.312	16.0	1.9175	.167
VAR00009	1.3611	3	.979	16.0	1.3901	.282
VAR00010	.9611	3	1.244	16.0	.7721	.526
VAR00011	10.3222	3	.802	16.0	12.8693	.000
VAR00012	2.2944	3	.729	16.0	3.1467	.054
VAR00013	.2944	3	.557	16.0	.5283	.669
VAR00014	.6944	3	.729	16.0	.9524	.439
VAR00015	4.7389	3	1.224	16.0	3.8718	.029

Quick Cluster

TABLE 1.7 Number of Cases in each Cluster

Cluster	*Unweighted cases*	*Weighted cases*
1	6.0	6.0
2	6.0	6.0
3	4.0	4.0
4	4.0	4.0
Missing	0	
Valid cases	20.0	20.0

ANOVA

The ANOVA table (Table 1.6) tells us which of the fifteen variables is significantly different across the 4 clusters. Since there is a divided opinion about the utility of statistical testing in cluster analysis, researcher's judgement is given priority in order to determine which variables are statistically significant. Based on the assumption that ANOVA is a valid test, the interpretation of clusters and differences across clusters were made on the basis of those variables, which were significantly different across clusters at 0.10 level. The last column in ANOVA table indicates that variables 2, 3, 4, 6, 7, 11, 12, & 15 are significant at 0.10 level (equivalent to 90% confidence level) as they have probability values less than 0.10. The other variables having prob. values more than 0.10 are not statistically significant.

CASE STUDY 2

Cluster Analysis*

Problem

A consumer durable goods company wants to know various features and services the consumers perceive when purchasing through Catalogs. The purpose is to determine how the various features of catalog marketing go together in terms of its convenience features, risk-reducing features, and so on.

10 statements are made in order to measure the variables of perception. The 10 statements are given below. The respondents have to agree or disagree (1 = strongly agree, 2 = agree, 3 = neither agree nor disagree, 4 = disagree, 5 = strongly disagree) with each statement.

1. The company should provide toll-free numbers.
2. The reputation of the company should be good.
3. They should have discount schemes based on quantity.
4. The company should provide guarantee for the product.
5. The company should give a trial period.
6. The sales catalog should be attractive and eye-catchy.
7. The company should make on-time delivery.
8. Number of years in business is an important factor.
9. Advertisements play an important role in decision-making.
10. It should have the facility to return the goods, if not satisfied.

Input Data

TABLE 1.1 Input Data Matrix

VAR00001	*VAR00002*	*VAR00003*	*VAR00004*	*VAR00005*	*VAR00006*	*VAR00007*	*VAR00008*	*VAR00009*	*VAR00010*
1.00	3.00	5.00	1.00	3.00	5.00	3.00	2.00	3.00	2.00
2.00	3.00	2.00	3.00	4.00	4.00	3.00	2.00	2.00	2.00
3.00	2.00	3.00	4.00	3.00	5.00	3.00	3.00	4.00	2.00
3.00	4.00	4.00	2.00	2.00	4.00	3.00	4.00	5.00	4.00
2.00	2.00	4.00	2.00	2.00	5.00	3.00	3.00	4.00	4.00

(*Table 1.1 contd*)

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala

(*Table 1.1 contd*)

VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006	VAR00007	VAR00008	VAR00009	VAR00010
2.00	4.00	3.00	3.00	5.00	4.00	4.00	2.00	3.00	4.00
1.00	3.00	2.00	4.00	4.00	1.00	2.00	4.00	2.00	5.00
4.00	5.00	1.00	1.00	5.00	4.00	5.00	1.00	1.00	5.00
2.00	1.00	5.00	3.00	4.00	4.00	2.00	1.00	2.00	1.00
5.00	2.00	4.00	3.00	2.00	5.00	1.00	5.00	3.00	2.00
4.00	3.00	3.00	2.00	1.00	2.00	1.00	5.00	2.00	2.00
3.00	4.00	4.00	4.00	3.00	2.00	5.00	1.00	5.00	3.00
4.00	3.00	2.00	2.00	3.00	3.00	4.00	2.00	2.00	3.00
1.00	3.00	2.00	4.00	2.00	5.00	1.00	3.00	5.00	4.00
2.00	2.00	4.00	1.00	5.00	4.00	2.00	4.00	4.00	5.00
3.00	3.00	1.00	3.00	4.00	3.00	2.00	3.00	2.00	5.00
5.00	2.00	1.00	5.00	1.00	2.00	4.00	2.00	2.00	4.00
3.00	1.00	5.00	3.00	5.00	5.00	5.00	1.00	2.00	3.00
3.00	5.00	4.00	2.00	4.00	4.00	1.00	4.00	1.00	3.00
1.00	2.00	3.00	2.00	2.00	5.00	2.00	5.00	1.00	3.00

ANALYSIS OF OUTPUT

Stage 1

The output is derived by first doing a hierarchical cluster analysis to find the number of clusters that exist in the data. The output for this purpose is shown in Table 1.2, Figs. 1.1 and 1.2. Table 1.2, the agglomeration schedule, helps us to identify the large differences in the coefficient, that is, 4th column. This column from top to bottom (1 to 19) indicates the sequence in which the cases get combined with others, until all the 20 cases are combined together in one cluster at the last stage (Stage 19). Therefore, Stage 19 represents a 1-cluster solution, Stage 18 represents a 2-cluster solution, and so on, going up from the last row to the first row. To identify the number of clusters in the data, the difference between rows in a measure called coefficient is used. One should look at this figure from the last row upwards, because we want to identify the lowest possible number of clusters, for reasons of economy and ease of interpretation. We see that there is a difference of (49-41.29) in the coefficients between the 1-cluster solution and the 2-cluster solution. This is a difference of 7.71. The next difference of (41.29-34), which is equal to 7.29 (between Stage 18, the 2-cluster solution and Stage 17, the 3-cluster solution). The next one after that is (34-33.83), only 0.17, between Stage 17 and Stage 16). Again after this, there is a large difference between 14th and 15th-cluster solution, of (30.45-25.87) or 4.58. Thereafter the differences are smaller between subsequent rows of coefficients. Large difference in the coefficient values between any two rows indicates a solution pertaining to the number of clusters, which the lower row represents.

Ignoring the first difference of 7.71, which would indicate only one cluster in the data, we look at the next largest difference i.e., 7.29, between row 2 and row 3 from the bottom, indicating a 2-cluster solution.

We can also look at the dendrogram for information as to which cases link up in what sequence to form clusters. The numbers in column 2 and 3 of the agglomeration schedule also give the same information. The dendrogram additionally provides a rescaled distance measure between various cluster combines at various stages. The case membership of each cluster is also much clearer in the dendrogram.

Stage 2

K-means cluster analysis generally gives more stable clusters, since it is an interactive procedure compared with the single pass hierarchical methods. This method needs a pre-specified number of starting points, to get an initial position. Therefore it is best used in combination with Stage 1.

In the given problem, Tables 1.3, 1.4, 1.5, 1.6, 1.7 and 1.8 indicate the outputs of *k*-mean clustering for a 2-cluster solution. These outputs give us the initial cluster centres, case listing of cluster membership (i.e., which case belongs to which of the clusters), the final cluster centres (the solution), and an ANOVA table.

The final cluster centres describe the mean value of each variable for each of the 2 clusters. For example, cluster 1 is described by the mean values of variable 1 = 2.45, variable 2 = 2.63, variable 3 = 3.72, and so on. Similarly, cluster 2 is described by variable 1 = 3, variable 2 = 3.11, variable 3 = 2.33, and so on.

Now, we can interpret the clusters in terms of the 10 variables in the questionnaire. For example, cluster 1 consists of people who give more importance to delivery time (variable 7 has a mean value = 2, which is equivalent to agree on the scale of 1 to 5). Similarly, there are also people, who do not give importance to the number of years the company is in business (variable 8 = 3.54).

On these same variables, cluster 2 shows people who do not give importance to the delivery times in the mode of their purchase (variable 7 value = 3.77) and people who give much importance to the number of years the company is in business (variable 8 = 2). In this way, when we compare final cluster centre values on each of the 10 variables, for one cluster at a time, a complete picture of the cluster emerges.

A brief description of each cluster and its characteristics follows.

Cluster 1

- insist on toll-free numbers for better communication
- weightage to company reputation
- do not want discounts based on quantity
- guarantees are to be provided
- need trial period

- sales catalog need not be attractive and eye-catching
- much importance is given to delivery times
- importance is not given to the number of years the company is in business
- not able to decide whether advertisement plays an important role in decision-making
- want return facility if not satisfied with the product

It is thus a cluster giving a lot of importance to convenience and safe mode of purchase.

Cluster 2

- do not insist on toll-free numbers for better communication
- due weightage is not given to company reputation
- want discounts based on quantity
- guarantees are not needed
- do not need trial period
- sales catalog need not be attractive and eye-catchy
- importance is not given to delivery times
- importance is given to the number of years the company is in business
- they feel that advertisements plays an important role in decision-making
- don't want return facility, if not satisfied with the product

This group gives importance to the promotional activities and the performance of the company. For this they are ready to sacrifice their convenience.

In summary, cluster analysis of this sample of respondents tells us about the possible segments, which exist in the target population.

ANOVA

The ANOVA table (Table 1.7) tells us which of the 10 variables is significantly different across the 2 clusters. Last column indicates that variable 3, 5, 6, 7, and 8 are significant at the 0.1 level (equivalent to 90% confidence level) as they have prob. values less than 0.1. The other variables are not statistically significant, as they all have prob. values greater than 0.1.

Output

Squared Euclidean measure used.

TABLE 1.2 Agglomeration Schedule Using Average Linkage (Between Groups)

Stage	Clusters Combined		Coefficient	Stage Cluster 1st Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	5	8.000000	0	0	5
2	2	13	9.000000	0	0	4
3	7	16	11.000000	0	0	10
4	2	6	11.500000	2	0	10
5	3	4	14.000000	0	1	8
6	10	11	15.000000	0	0	13
7	1	9	15.000000	0	0	9
8	3	14	16.666666	5	0	12
9	1	18	19.500000	7	0	16
10	2	7	20.166666	4	3	14
11	19	20	21.000000	0	0	13
12	3	15	22.000000	8	0	14
13	10	19	23.500000	6	11	15
14	2	3	25.879999	10	12	15
15	2	10	30.450001	14	13	16
16	1	2	33.833332	9	15	18
17	12	17	34.000000	0	0	18
18	1	12	41.294117	16	17	19
19	1	8	49.000000	18	0	0

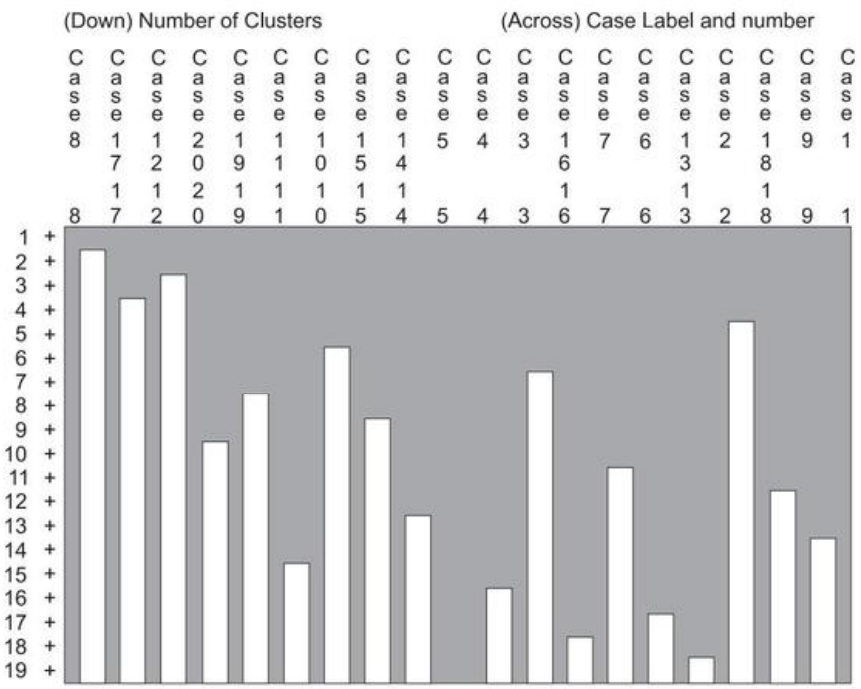


FIGURE 1.1 Vertical Icicle Plot using Average Linkage (Between Groups)

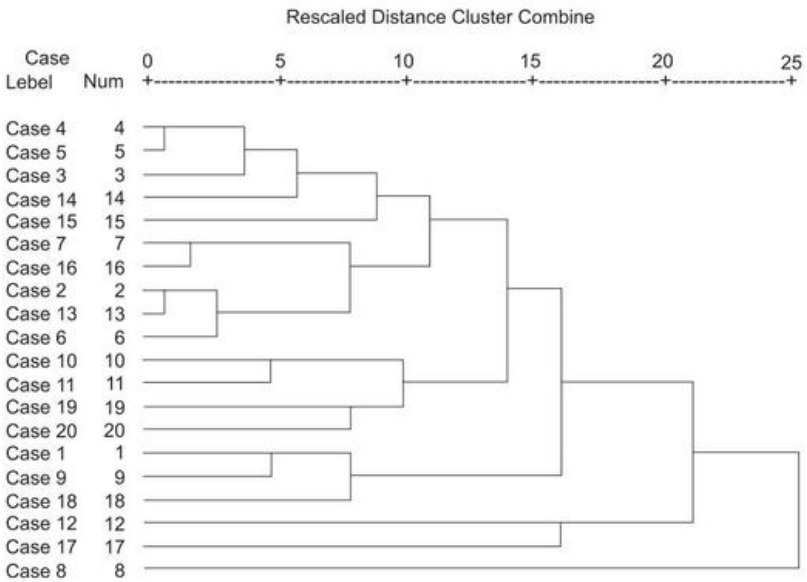


FIGURE 1.2 Dendrogram using Average Linkage (Between Groups)

TABLE 1.3 Initial Cluster Centers

Cluster	*VAR00001*	*VAR00002*	*VAR00003*	*VAR00004*
1	5.0000	2.0000	4.0000	3.0000
2	4.0000	5.0000	1.0000	1.0000
Cluster	VAR00005	VAR00006	VAR00007	VAR00008
1	2.0000	5.0000	1.0000	5.0000
2	5.0000	4.0000	5.0000	1.0000
Cluster	VAR00009	VAR00010		
1	3.0000	2.0000		
2	1.0000	5.0000		

TABLE 1.4 Case Listing of Cluster membership.

Case ID	*Cluster*	*Distance*
1	1	3.230
2	2	4.419
3	1	2.709
4	1	3.011
5	1	2.251
6	2	2.253
7	2	4.264
8	2	4.342

(*Table 1.4 contd*)

(*Table 1.4 contd*)

Case ID	*Cluster*	*Distance*
9	1	4.240
10	1	3.500
11	1	4.109
12	2	4.051
13	2	1.934
14	1	3.870
15	1	3.640
16	2	2.762
17	2	4.443
18	2	4.456
19	1	3.665
20	1	3.258

TABLE 1.5 Final Cluster Centres

Cluster	VAR00001	VAR00002	VAR00003	VAR00004
1	2.4545	2.6364	3.7273	2.3636
2	3.0000	3.1111	2.3333	3.1111
Cluster	VAR00005	VAR00006	VAR00007	VAR00008
1	2.7273	4.3636	2.0000	3.5455
2	3.7778	3.1111	3.7778	2.0000
Cluster	VAR00009	VAR00010		
1	3.0909	2.9091		
2	2.3333	3.7778		

Table 1.6 Distances between Final Cluster Centres.

Cluster	1	2
1	.0000	
2	3.5461	.0000

Table 1.7 Analysis of Variance

Variable	*Cluster MS*	*DF*	*Error MS*	*DF*	*F*	*Prob*
VAR00001	1.4727	1	1.596	18.0	.9228	.349
VAR00002	1.1157	1	1.301	18.0	.8569	.367
VAR00003	9.6182	1	1.343	18.0	7.1594	.015
VAR00004	2.7657	1	1.190	18.0	2.3225	.145
VAR00005	5.4626	1	1.541	18.0	3.5449	.076
VAR00006	7.7657	1	1.190	18.0	6.5214	.020
VAR00007	15.6444	1	1.086	18.0	14.4000	.001
VAR00008	11.8227	1	1.373	18.0	8.6063	.009
VAR00009	2.8409	1	1.717	18.0	1.6544	.215
VAR00010	3.7354	1	1.359	18.0	2.7483	.115

TABLE 1.8 Number of Cases in each Cluster

Cluster	unweighted cases	weighted cases
1	11.0	11.0
2	9.0	9.0
Missing	1	
Valid cases	20.0	20.0

CASE STUDY 3

Cluster Analysis*

Methodology

Our group formulated a problem, and developed data for a sample size of 20. The data was entered into the computer and analysed using the SPSS package. We used both the average linkage hierarchical clustering method and the *K*-means clustering technique.

The Problem

A large housing finance corporation wants to identify and cluster the basic characteristics, lifestyles, and mindsets of persons who would be prone to taking housing loans. We formulated a set of 17 statements, which we feel measure many of the relevant variables. These are as follows;

Var1 (SALARY)	My salary suffices for me to indulge in luxuries (over and above necessities).
Var 2 (HOUSE)	Owning a house is a necessity.
Var 3 (CAR)	Owning a car is very important to me.
Var 4 (QLLTY)	I do not mind paying a high price for quality.
Var 5 (SPOUSE)	In my opinion, my wife/husband plays an important role in all major decisions.
Var 6 (SAVINGS)	I feel everyone should put away money for a rainy day.
Var 7 (STOCKS)	I feel investing in the stock market is highly risky.
Var 8 (CHILDREN)	I believe that every parent should invest for his or her children's future.
Var 9 (FLY.FRNDS)	I feel one could borrow money from family/friends in times of financial need.
Var 10 (INSTLMNT)	I believe I will be able to meet monthly financial obligations arising out of a loan without much trouble.
Var 11 (PRESENT)	I believe that one should enjoy the present without worrying too much about the future.
Var 12 (ADVT)	I think advertisements play an important role in influencing my purchase decisions.
Var 13 (COMPUTER)	I feel owning a computer in today's world is very important.

* Prepared by Aditi Sood, Mukesh Kumar Sinha, Vikram Rathi, and Vinita Nair

Var 14 (BORROWING) I believe that borrowing money even in times of need is not a good idea.

Var 15 (FINDAILY) I believe that Indian financial dailies are a good source of information and help me in my financial decision making.

Var 16 (TV) Television is a major source of information and entertainment for me.

Var 17 (DECFF) Friends/family are an important factor for my major decisions.

The respondents were to agree or disagree with each of these statement based on the following scale:

1 = strongly agree
2 = agree
3 = neither agree nor disagree
4 = disagree
5 = strongly disagree

The input data table follows as Table 1.1

TABLE 1.1

S.No.	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17
1	5	5	5	2	1	4	3	4	1	5	4	5	4	1	5	1	4
2	2	3	3	3	2	2	1	2	2	2	2	4	1	3	2	5	3
3	3	2	2	4	4	4	2	4	5	3	1	2	3	4	3	4	2
4	5	4	4	1	1	5	4	5	2	5	4	1	4	2	5	3	5
5	1	2	2	2	4	1	2	1	4	1	2	3	2	3	1	2	5
6	3	2	2	4	3	4	3	4	4	3	2	5	3	4	3	5	1
7	4	4	4	4	5	2	1	2	2	4	5	4	3	1	4	4	2
8	2	1	1	1	1	1	3	1	5	2	1	2	1	4	2	1	3
9	1	2	2	3	4	2	3	2	4	1	3	1	2	5	1	2	3
10	4	5	5	2	3	5	4	5	1	4	2	4	3	2	4	5	5
11	3	2	2	5	2	1	2	1	5	3	1	5	4	4	3	2	4
12	5	4	4	5	1	4	3	4	1	5	4	2	4	2	5	5	2
13	3	3	3	2	5	2	1	2	3	3	4	3	2	3	3	4	4
14	4	5	5	4	4	5	4	5	2	4	5	1	4	1	4	1	3
15	4	4	4	5	2	3	2	3	2	4	3	4	3	2	4	3	2
16	2	1	1	4	3	5	1	5	5	2	1	5	2	4	2	5	1
17	5	4	4	3	2	4	1	3	3	5	4	2	4	2	5	2	5
18	1	2	2	1	4	1	5	5	4	1	1	2	2	3	1	4	3
19	4	5	5	2	3	3	2	4	1	4	3	3	4	1	4	2	4
20	3	1	1	2	5	1	5	1	5	3	5	4	2	5	3	3	4

THE OUTPUT INTERPRETATION

First, the data was input in the SPSS Package and hierarchical clustering was carried out. This was done in order to find out the number of clusters that exists in the data. The basic output of this stage is the agglomeration schedule, vertical icicle plot, and dendrogram using average linkages (Table 1.2, Figs 1 and 2).

The fourth column in the agglomeration schedule is the coefficient column, which helps us to identify the large differences in the coefficients. The schedule from stages 1 to 19 indicates the sequence in which cases get combined with others. At stage 19 all cases are combined together and thus we obtain a 1-cluster solution. Similarly stage 18 represents 2-cluster solution, stage 17 a 3-cluster solution, and so till the first row. To identify how many clusters are present in the data we use the fusion coefficient. We find the difference between the coefficients of 1-cluster solution and a 2-cluster solution, the next difference between a 2-cluster solution and a 3-cluster solution, and so on.

Looking at the agglomeration schedule obtained we find the following differences,

(87.166664 – 59.142847)	= 28.023817	1-cluster solution
(59.142847 – 51.916668)	= 7.226179	2-cluster solution
(51.142857 – 48.333332)	= 3.583336	3-cluster solution
(48.333332 – 43.25)	= 5.083332	4-cluster solution
(43.25 – 35.33332)	= 7.91668	5-cluster solution

We ignore the first difference of 28.023817 as it would indicate only one cluster in the data. The next largest difference is 7.91668, which indicates a 5-cluster solution. Therefore in this case we choose a 5-cluster solution.

NON-HIERARCHICAL CLUSTERING (K-MEANS TECHNIQUE)

Once the number of clusters has been identified using the hierarchical clustering a *K*-Means clustering is run on the data. The number of clusters identified above, in our case 5 clusters, is specified and the output obtained. The *K*-Means cluster (also called Quick Cluster) is run because it gives us more stable clusters. The basic output of this technique is a case listing of cluster memberships, final cluster centres, and an ANOVA Table. The final cluster centres are the best solution. These outputs are in Tables 1.3 to 1.8.

The final cluster centres (Table 1.5) describe the mean values of each variable for each of the five-clusters. For example, cluster one is described by mean values of ADVT = 4, BORROWING = 1, CAR = 4, and CHILDREN = 2. Cluster two is described by mean values of ADVT = 4 and so on.

In order to interpret each of the clusters, we must refer back to the 17 statements originally formulated. For example, cluster one consists of people who claim that their purchase decisions are not highly influenced by advertisements (ADVT = 4), believe that one should avoid borrowing even when in financial need (BORROWING = 1), disagree that owning a car is a necessity (CAR = 4), and believe that parents should put away some money for their children's futures. Similarly, we may interpret each of the other variable.

The five clusters identified are briefly discussed here:

Cluster 1

The persons belonging to this cluster, believe that they earn a modest salary which enables them to take care of their necessities but does not leave much room for luxuries. They do not agree that owning a house is a necessity. They also believe that it is not a necessity to own a car. They are not very quality conscious consumers and would rather go for cost effectiveness than for high priced quality products. These individuals also do not feel that their spouse must play an important role in all their major decisions. These people do believe in saving for any future contingencies. They are generally risk averse and believe that stocks are a highly risky option. They believe that it is every parents responsibility to save money for their children's futures. They believe that in times of financial crisis it is acceptable to borrow from family and friends. They feel that payment of monthly instalments would prove to be a burden considering their current financial position. These individuals are very 'future-conscious' and disagree that one should live for the present with scant regard for the future. Persons belonging to this cluster claim that advertisements do not play an important role in influencing their purchase decisions. Also, they believe that owning a computer even in today's 'techno-savvy' world is very important. These persons would rather avoid any borrowings even in time of financial need. They also do not rely on Indian financial dailies as major source of information and decision-making. For people in this category, television is not a major source of entertainment. Also, these individuals do rely very heavily on family and friends for making any major decisions.

In summary, people belonging to this cluster have a modest salary and do not really believe in indulging themselves. We can observe in them some traditional values such as saving for posterity, and still relying on family and friends for borrowings and risk averseness. This cluster may be termed as 'price-conscious' rather than 'quality-conscious'.

Cluster 2

These are individuals whose salary does not permit them to indulge in luxury but believe that owning a house and car is a necessity. They are quality-conscious buyers. However, they don't consider their spouse's opinion in major decisions. They believe in saving money for contingencies and do not see the stock market as a risky investment option. The people in this cluster believe that every parent should save for their children. However they don't perceive family and friends as a major source of borrowing. Their financial position allows them to meet monthly financial obligations relatively comfortably and they do not believe in living life for the present. Advertisements do not play a major role in purchase decision and for them, it is important to own a computer in today's world. They opine that borrowing in times of financial need is acceptable. Financial dailies are an important source of information and decision-making, just as television is an important source of entertainment. Family and friends don't play a decisive role in their decision-making.

In short, these individuals belong to a mediocre salary class, and are very independent decision-makers. They may be termed as progressive in their outlook. However, some traditional values still persist. They are also open to taking calculated risks, and are very aware of the financial world. Owning a house is very important to this cluster of people.

Cluster 3

People belonging to this cluster believe that their salary permits indulgences in luxuries and also feel that owning a house and a car is very necessary. However, they are not quality-conscious buyers, and price takes precedence in their purchase decisions. They do not consider their spouses opinion as important in making major decisions and also do not believe in saving for a rainy day and in relation to this they also believe in living life to the fullest without worrying too much about the future. They are risk-averse and see the stock market as a risky investment option. They do not see it as a responsibility on the part of the parents to save for their children's futures. They do not believe in borrowing from family and friends in times of financial need and due to their current financial position believe that paying monthly instalments would not be a strain on their resources. Advertisements are not a very important influence just as television is not a major source of entertainment. Owning a computer is important for these individuals, and they believe it is not unacceptable to borrow in times of financial need. Indian financial dailies are an important source of information and decision-making and similarly, the opinions of family and friends are not considered important in making major decisions.

In short, this cluster of individuals are spenders, believing in living for the present and consequently, they do not indulge in too many savings. They are risk-averse. In spite of their comfortable financial position, they feel that meeting monthly financial obligations would be a burden. They are information oriented, relying on financial dailies for a lot of their decisions. They believe that it is not acceptable to borrow from friends and family. For them, owning a house of their own is a necessity.

Cluster 4

The salary of individuals belonging to this cluster does not permit many indulgences, and consequently, they do not see owning a house and a car as very necessary. However, they are quality-conscious buyers and believe that better quality warrants a higher price. Interestingly, this cluster considers their spouse's opinion as very important in major decisions and do not really believe in saving for contingencies. However, in contradiction to this, they do believe that one must have due concern for the future. They do not consider stocks as a risky option and disagree that every parent should save money for their children. They feel that borrowing from family and friends in time of need is acceptable, and believe that their current financial position entails that it would be a burden to meet monthly financial obligations. Advertisements play a very major role in their purchase decisions and television is a major source of entertainment for them. They, however, do not consider it very important to own a computer in today's world, and consider borrowing even in times of need a bad idea. They do not consider Indian financial dailies as a major source of information and decision-making and do not rely on the opinions of friends and family in making major decisions.

In summary, this cluster of individuals belongs to a modest income category. They are quality-conscious and risk-takers. They are not very progressive in terms of information and decision-making and neither do friends and family play a major role in their decisions. They feel that monthly financial obligations are a burden, and do not consider owning a house a necessity.

Cluster 5

This cluster of individuals have a comfortable income level permitting them certain indulgences, and they consider owning a car and a house a necessity. They are quality-conscious buyers and do not consider their spouse's opinion as very important in decision-making. They strongly believe in saving for future contingencies. However, at the same time they believe in living for the present. They do not view the stock exchange as a risky investment option. Thus, they may be looked upon as calculated risk-takers. They strongly believe in saving for their children's futures and do not see borrowing from friends and family as a good option. Their comfortable income level permits them to meet any monthly financial obligations without much trouble. Advertisements are a major influence in their purchase decisions, just as television is a major source of entertainment. They believe that owning a computer is important and rely on financial dailies for a lot of their information and financial decision-making. They do not depend too much on friends and family to make any major decisions.

In short, this cluster may be viewed as a group of calculated risk-takers, who weigh risk and return carefully before making any decisions. They are also information oriented relying on financial dailies and probably the Internet for their information. They do not consider borrowing from friends and family as a very good option, and can meet financial obligations comfortably. They are conscious about quality and are influenced by advertisements in their purchase decisions. Owning a house is a necessity for the individuals belonging to this cluster.

ANOVA

The ANOVA (Table 1.7) is one of the outputs of the *K*-means clustering and tells us which of the 17 variables are significantly different across the clusters.

Testing at confidence levels of 90% (0.10 level of significance), we find that the following variables are significant:

- BORROWING: I believe that borrowing money even in times of financial need is not a good idea.
- CHILDREN: I believe that every parent should invest for his or her children's future.
- CAR: Owning a car is very important to me.
- COMPUTER: I feel owning a computer in today's world is very important.
- DECFF: Friends/family are an important factor for my major decisions.
- FINDAILY: I feel that the Indian financial dailies are a very good source of information and help me in my financial decisions.
- FL.FRNDS: I believe that one could borrow from family and friends in times of financial need.
- HOUSE: Owning a house is a necessity for me.
- INSTLMNT: I believe I will be able to meet monthly financial obligations arising out of a loan without much trouble.
- PRESENT: I believe that one should enjoy the present without worrying too much about the future.

- SALARY: My salary suffices for me to indulge in luxuries (over and above necessities).
- SAVINGS: I feel everybody should put away money for a rainy day.
- STOCK: I believe investment in the stock market is a highly risky option.

As the significance values of all the above variables are below 0.10, these are significant at 90% level of confidence.

Hierarchical Cluster Analysis

TABLE 1.2 Agglomeration Schedule using Average Linkage (Between Groups)

Stage	*Clusters Combined*		*Coefficient*	*Stage Cluster 1st Appears*		*Next Stage*
	Cluster 1	*Cluster 2*		*Cluster 1*	*Cluster 2*	
1	6	16	14.000000	0	0	5
2	1	19	16.000000	0	0	9
3	5	9	17.000000	0	0	8
4	12	15	18.000000	0	0	14
5	3	6	18.000000	0	1	16
6	4	17	19.000000	0	0	9
7	7	13	21.000000	0	0	12
8	5	18	21.500000	3	0	10
9	1	4	25.500000	2	6	11
10	5	8	28.333334	8	0	15
11	1	14	32.250000	9	0	13
12	2	7	32.500000	0	7	17
13	1	10	35.000000	11	0	14
14	1	12	35.833332	13	4	19
15	5	20	43.250000	10	0	18
16	3	11	48.333332	5	0	17
17	2	3	51.916668	12	16	18
18	2	5	59.142857	17	15	19
19	1	2	87.166664	14	18	0

Hierarchical Cluster Analysis

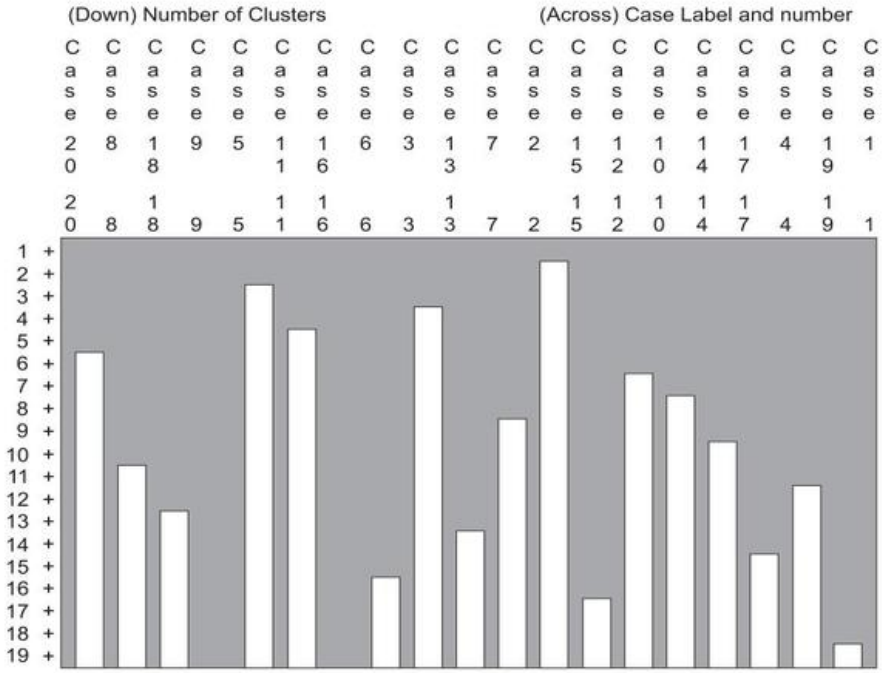


FIGURE 1 Vertical Icicle Plot using Average Linkage (Between Groups)

Hierarchical Cluster Analysis

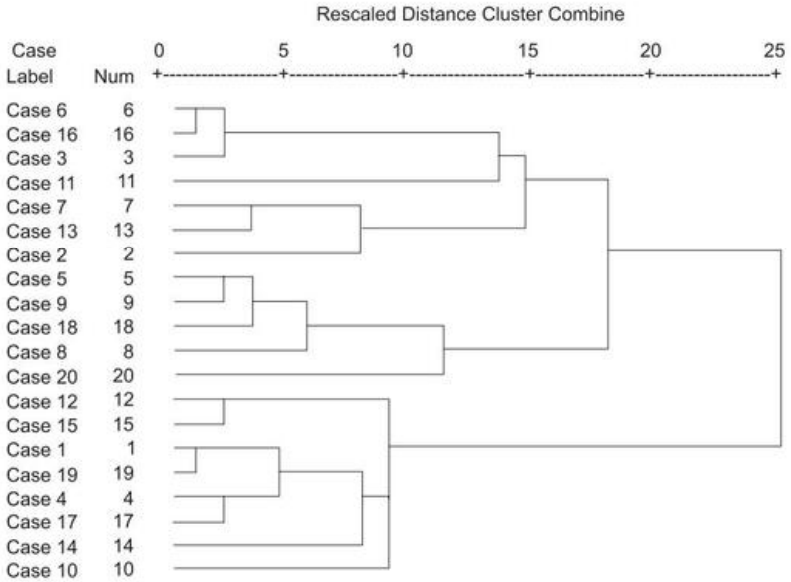


FIGURE 2 Dendrogram using Average Linkage (Between Groups)

Quick Cluster

TABLE 1.3 Initial Cluster Centres

Cluster	ADVT	BORROWIN	CAR	CHILDRE
1	4.0000	1.0000	4.0000	2.0000
2	4.0000	5.0000	1.0000	1.0000
3	5.0000	4.0000	1.0000	5.0000
4	1.0000	2.0000	4.0000	5.0000
5	2.0000	4.0000	1.0000	1.0000
Cluster	COMPUTER	DECFF	FINDAILY	FL.FRNDS
1	3.0000	2.0000	4.0000	2.0000
2	2.0000	4.0000	3.0000	5.0000
3	2.0000	1.0000	2.0000	5.0000
4	4.0000	5.0000	5.0000	2.0000
5	1.0000	3.0000	2.0000	5.0000
Cluster	HOUSE	INSTALMN	PRESENT	QLTY
1	4.0000	4.0000	5.0000	4.0000
2	1.0000	3.0000	5.0000	2.0000
3	1.0000	2.0000	1.0000	4.0000
4	4.0000	5.0000	4.0000	1.0000
5	1.0000	2.0000	1.0000	1.0000
Cluster	SALARY	SAVINGS	SPOUSE	STOCKS
1	4.0000	2.0000	5.0000	1.0000
2	3.0000	1.0000	5.0000	5.0000
3	2.0000	5.0000	3.0000	1.0000
4	5.0000	5.0000	1.0000	4.0000
5	2.0000	1.0000	1.0000	3.0000
Cluster	TV			
1	4.0000			
2	3.0000			
3	5.0000			
4	3.0000			
5	1.0000			

Convergence achieved due to no or small distance change.
The maximum distance by which any centre has changed is .0000
Current iteration is 3
Minimum distance between initial centres is 7.1414

Quick Cluster

TABLE 1.4 Case Listing of Cluster membership

Case ID	*Cluster*	*Distance*
1	4	3.534
2	1	3.666
3	3	2.539
4	4	3.390
5	5	2.561
6	3	2.028
7	1	3.152
8	5	3.487
9	5	3.219
10	4	4.131
11	5	4.976
12	4	4.250
13	1	2.990
14	4	4.131
15	1	3.307
16	3	2.472
17	4	3.104
18	5	3.655
19	4	2.962
20	2	.000

TABLE 1.5 Final Cluster Centres

Cluster	ADVT	BORROWIN	CAR	CHILDRE
1	3.7500	2.2500	3.5000	2.2500
2	4.0000	5.0000	1.0000	1.0000
3	4.0000	4.0000	1.6667	4.3333
4	2.5714	1.5714	4.5714	4.2857
5	2.6000	3.8000	1.8000	1.2000
Cluster	COMPUTER	DECFF	FINDAILY	FL.FRNDS
1	2.2500	2.7500	3.2500	2.2500
2	2.0000	4.0000	3.0000	5.0000
3	2.6667	1.3333	2.6667	4.6667
4	3.8571	4.0000	4.5714	1.5714
5	2.2000	3.6000	1.6000	4.4000
Cluster	HOUSE	INSTLMN	PRESENT	QLTY
1	3.5000	3.2500	3.5000	3.5000
2	1.0000	3.0000	5.0000	2.0000
3	1.6667	2.6667	1.3333	4.0000
4	4.5714	4.5714	3.7143	2.7143
5	1.8000	1.6000	1.6000	1.4000

(*Table 1.5 contd*)

(*Table 1.5 contd*)

Cluster	SALARY	SAVINGS	SPOUSE	STOCKS
1	3.2500	2.2500	3.5000	1.2500
2	3.0000	1.0000	5.0000	5.0000
3	2.6667	4.3333	3.3333	2.0000
4	4.5714	4.2857	2.1429	3.0000
5	1.6000	1.2000	3.0000	3.0000
Cluster	TV			
1	4.0000			
2	3.0000			
3	4.6667			
4	2.7143			
5	2.2000			

TABLE 1.6 Distances between Final Cluster Centres.

Cluster	*1*	*2*	*3*	*4*	5
1	.0000				
2	7.3782	.0000			
3	5.7921	8.0069	.0000		
4	5.4313	9.8954	8.2075	.0000	
5	6.0496	5.6178	6.3512	9.0739	.0000

Quick Cluster

TABLE 1.7 Analysis of Variance.

Variable	*Cluster MS*	*DF*	*Error MS*	*DF*	*F*	*Prob*
ADVT	2.0339	4	1.977	15.0	1.0285	.425
BORROWIN	6.4839	4	.484	15.0	13.3886	.000
CAR	8.6923	4	.278	15.0	31.1852	.000
CHILDRE	9.8262	4	.376	15.0	26.1092	.000
COMPUTER	2.8690	4	.604	15.0	4.7428	.011
DECFF	4.2833	4	.974	15.0	4.3957	.015
FINDAILY	6.7173	4	.555	15.0	12.0945	.000
FL. FRNDS	9.6548	4	.422	15.0	22.8751	.000
HOUSE	8.6923	4	.278	15.0	31.1852	.000
INSTALMN	6.7173	4	.555	15.0	12.0945	.000
PRESENT	6.5637	4	.953	15.0	6.8873	.002
QLTY	1.8304	4	1.841	15.0	.9937	.441
SALARY	6.7173	4	.555	15.0	12.0945	.000

(*Table 1.7 contd*)

(*Table 1.7 contd*)

Variable	*Cluster MS*	*DF*	*Error MS*	*DF*	*F*	*Prob*
SAVINGS	9.8262	4	.376	15.0	26.1092	.000
SPOUSE	2.6065	4	1.768	15.0	1.4741	.259
STOCKS	4.0125	4	1.116	15.0	3.5933	.030
TV	3.9137	4	1.659	15.0	2.3581	.100

TABLE 1.8 Number of Cases in each Cluster.

Cluster	*unweighted cases*	*weighted cases*
1	4.0	4.0
2	1.0	1.0
3	3.0	3.0
4	7.0	7.0
5	5.0	5.0
Missing	0	
Valid cases	20.0	20.0

The McGraw·Hill Companies

CHAPTER 15 MULTIDIMENSIONAL SCALING FOR BRAND POSITIONING

Learning Objectives

In this chapter, we will

- Introduce two approaches to multidimensional scaling—attribute based and similarity/dissimilarity based
- Explain the similarity/dissimilarity based approach with a detailed example, including the format for data collection
- Discuss the use of multidimensional scaling in perceptual mapping for positioning applications
- Explain how to determine the number of dimensions in the data using a measure of stress or lack-of-fit
- Illustrate how to name the dimensions with some prior knowledge of attributes

APPLICATION AREAS

The most common and useful marketing application of multidimensional scaling is in product positioning or brand positioning. Positioning is essentially concerned with mapping a consumer's mind and placing all the competing brands of a product category in appropriate slots or 'positions' on it. One obvious way to do that is to ask customers what they think of competing brands on say, six attributes with a rating scale of 5 to 10 points. This would result in ratings for all brands on all attributes, which could be taken two attributes at a time, and plotted on a graph.

For example, a product category of shampoos could be identified as having 5 attributes important to the consumer—price, lather, fragrance, consistency, and favorable effects on hair. If these were to be rated on a 7-point scale for say, six leading brands of shampoo *A*, *B*, *C*, *D*, *E*, and *F*, then we could pick

up any two attributes and plot the six brands on a map according to the consumer ratings. For example, if we plotted rating on 'price' versus rating on 'favorable effect on hair', we may find that all the six brands are positioned in different places based on consumer ratings or perceptions. This is called a perceptual map of consumer perception about competing brands in a product category. This is the type of map useful for deliberate positioning of a new brand, based on 'gaps' in the current map, or for finding out the current position of an existing brand on the map. If the desired position of an existing brand owned by a company is different from the one perceived by consumers, an option is to 'reposition' the brand.

Even though technically the above method is quite straightforward, it may not capture the consumer's mind accurately. If we assume that the consumer simultaneously thinks of several product dimensions or attributes rather than one attribute at a time, the above method is only an approximation of that process, because it forces consumers to rate brands on only one attribute at a time. Multidimensional scaling, on the other hand, captures the complex interactions between attributes and brands in a particular way, and then 'derives' attributes or dimensions which explain the 'positions' given by consumers to various brands.

Methods

There are two basic methods used in multidimensional scaling:

(a) Attribute based approach

(b) Similarity/Dissimilarity based approach

The attribute-based approach is similar to what we have described in the previous section, except that these input data are then further analysed using either factor analysis or discriminant analysis. The first approach is covered in detail in Chapter 17, and requires knowledge of the discriminant analysis procedure to understand it. The second approach is very easy to understand intuitively, and quite useful in gaining a good understanding of consumer psyche, so we will discuss only this (similarity and dissimilarity based) approach in this chapter.

In the similarity/dissimilarity-based approach, we need some kind of a distance measure between the brands being rated. The distance measure being input, could be a simple ranking of distances between a brand and all other brands by a customer. One way to do this is to provide a customer (respondent) with cards, each containing a pair of brands written on it, and asking him to write down a number indicating the difference between the two brands on any numerical scale which can represent distance. This can be repeated for all pairs of brands being included in the research. No attributes are specified by which the customer is asked to decide on the difference.

It may be assumed that a consumer would tend to include parameters such as price, quality of product, after-sales-service, delivery time, his satisfaction with promised benefits, and perceptions about corporate image and so on, but he would not specify these. He would only indicate distance (or dissimilarity) in some numerical value.

This distance measure or dissimilarity measure can be compiled into a matrix of the type shown in Table 15.1.

RECOMMENDED USAGE

The author's recommendation is to use as input the dissimilarity (or similarity) measure without specifying any particular attributes, and derive both the number of dimensions as well as the interpretation for the derived dimensions by plotting the stimulus coordinates. For the input, pairwise rating or ranking of distance can be done by each respondent, and the average ratings can be used for a preliminary or 'exploratory' multidimensional scaling exercise. The **number of dimensions** is decided upon by a measure known as 'stress' which is given by the computer output for a solution in each dimension. The objective should be to get the solution with an acceptably low value of stress (indicating a good fit, because **stress is a measure of lack-of-fit**), in the least possible number of dimensions. Once the number of dimensions are decided on, the solution for that dimensionality should be looked at, and the objects (brands) plotted on a map if possible, for visual clarity. Then comes the difficult part, of naming the dimensions. This has to be done keeping in mind the attributes of the brands, their target segments by age, price, quality, or attempted positioning through brand communication, and so on. The worked example from the TV industry in India will clarify what we mean.

WORKED EXAMPLE

Problem

Let us take the example of eight brands of TV available in the Indian market, and use multidimensional scaling to determine how they are perceived by Indian consumers, and try to plot a positioning map of the eight brands. We will also attempt to find out how many dimensions the consumers seem to be using, when they think of TV brands.

Input Data

Table 15.1 shows input data collected from a sample of respondents each of whom was asked to rate the dissimilarity between all pairs of TV brands on a numerical scale (essentially, an ordinal measure). There are actually two options at this point. Each consumer's distance matrix can be separately analysed through multidimensional scaling, or the ratings can be averaged across respondents and a single distance matrix can be constructed. The assumption here is that ratings have been aggregated, assuming some homogeneity among the respondents. For the sake of simplicity, integers are used as distance measures in our example, even though it is not necessary for them to be integers.

In this symmetrical 8×8 matrix, the columns and rows both denote the brands, and the body of the matrix is a measure of the perceived dissimilarity between brands. For example, row 3 column 1 represents a distance (or dissimilarity) of 6.00 between brand 3 and brand 1 of TVs, and so on. In other words, both column numbers and row number represents brands of TV.

We have said the data are measured ordinally. This is to demonstrate that multidimensional scaling can work even with ordinal data. Interval scaled data can also be used if available.

Interpretation of Computer Output

In Tables 15.2(a), 15.2(b), 15.3(a), 15.3(b), 15.4(a) and 15.4(b) we have the outputs of a multidimensional scaling program using one of the standard packages. Table 15.2(a) and Table 15.2(b) contain the 3-dimensional solution. Table 15.3(a) and Table 15.3(b) contain the 2-dimensional solution. Table 15.4(a) and Table 15.4(b) contain the 1-dimensional solution.

Our first task is to determine the number of dimensions in which we feel the best solution exists. The decision is based on the stress value for various solutions in different dimensions. From Table 15.2(a), Table 15.3(a), and Table 15.4(a), we see the following values of stress.

3-dimensional solution	: 0.05230
2-dimensional solution	: 0.24015
1-dimensional solution	: 0.43159

Clearly, the 1-dimensional solution is not a good one. Remember, the stress value indicates lack of fit, so it should be as close to zero as possible. The 2-dimensional solution is better, but the 3-dimensional solution looks the best, as the stress value is a low 0.05.

Since we have only 8 brands of TV being compared, it is not possible to get a solution in 4 or more dimensions in this particular case. But if we were comparing 12–15 brands, for example, it would be possible to get a solution in a higher number of dimensions. But even there, ease of interpretation may require a tradeoff between stress value and number of dimensions. We may choose a slightly lower number of dimensions, in the interest of ease of interpretation or for parsimony (economy), instead of looking only at the lowest stress value, in exceptional cases.

3-Dimensional Solution

Assuming that we have decided to use the 3-dimensional solution for interpretation in this case, the next task now would be to **name** the dimensions. For doing so, our previous knowledge of the brands may become important. For example, let us assume that the eight brands of TV were as follows:

1. Aiwa
2. Videocon
3. LG
4. Samsung
5. Sony
6. Onida
7. Thomson
8. BPL

If these had been the eight brands, then we must look at the qualities of various attributes offered by these brands either through our judgment or knowledge of the market or through a survey of consumers, or a combination of these methods. This process of interpretation tends to be subjective, regardless of the methods used. For example, we could look at the output of multidimensional scaling, and the scores for the eight brands on the 3 dimensions, and decide on the following names for the 3 dimensions:

Dimension 1 : Value for money
Dimension 2 : After sales service
Dimension 3 : Current brand image

We could then look at the scores on the three dimensions and conclude that some brands like BPL, and Videocon, currently enjoy a good brand image, but brands like Aiwa, Onida and Thomson are leading in 'Value for money' perceptions. Also, Videocon and Thomson may be perceived as having the best after-sales service.

If we were to choose the 2-dimensional solution instead of the 3-dimensional one, it could be plotted on a graph and would be visually easier to interpret. The plot of the 2-dimensional solution is shown in Fig. 15.1 and the brands can be seen to form distinct clusters based on their perceived similarity.

For example, brands 1 and 6 are perceived to be similar, whereas brand 5 is a standalone brand. So is brand 3, to some extent. Here again, knowledge of the brand names and their attributes or qualities would be used to name the two dimensions. Again, dimension 1 could be value for money. Dimension 2 could be after-sales service. But notice that we are losing some information on the third dimension which we had called brand image in the 3-dimensional solution. The loss of information may turn out to be critical in some cases.

ADDITIONAL COMMENTS

MDS can be performed even with a sample size of 1. It can be used to get a composite picture of a segment's perception, by combining the responses of any one segment, and repeating the MDS for each of the major segments. It can also be done across segments (a single MDS) by aggregating responses for the entire sample. If we have a significant marketing decision hinging on the results, both approaches 2 and 3 (segmentwise and across segments) can be used and if there are significant differences, it should be seen whether the positioning needs to be different for different segments. That may indeed be the case, in these days of diversity of consumer preferences. It would be tempting to do one MDS for each respondent, but the analysis would remain meaningless unless there are sufficient numbers of each consumer type which means determining the segments after the MDS. This is a possibility, but would involve a lot of work in the analysis stage. It is best left to the judgment of the researcher which approach he would like to follow.

TABLE 15.1

	Var 1	*Var 2*	*Var 3*	*Var 4*	*Var 5*	*Var 6*	*Var 7*	*Var 8*
Var 1	.00	3.00	6.00	8.00	1.00	2.00	7.00	8.00
Var 2	3.00	.00	4.00	6.00	4.00	5.00	2.00	5.00
Var 3	6.00	4.00	.00	3.00	2.00	4.00	6.00	1.00
Var 4	8.00	6.00	3.00	.00	3.00	5.00	4.00	7.00
Var 5	1.00	4.00	2.00	3.00	.00	2.00	8.00	5.00
Var 6	2.00	5.00	4.00	5.00	2.00	.00	3.00	6.00
Var 7	7.00	2.00	6.00	4.00	8.00	3.00	.00	5.00
Var 8	8.00	5.00	1.00	7.00	5.00	6.00	5.00	.00

TABLE 15.2(a) Iteration history for the 3-dimensional solution (in squared distances) Young's S-stress formula 1 is used

Iteration	*S-stress*	*Improvement*
1	.16119	
2	.10360	.05759
3	.09295	.01065
4	.08850	.00445
5	.08589	.00261
6	.08417	.00172
7	.08301	.00116
8	.08222	.00078
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities)
- in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix

Stress = .05230 RSQ = .96043

Configuration derived in 3 dimensions

TABLE 15.2(b) Stimulus Coordinates

Stimulus Number	*Stimulus Name*	*Dimension*		
		1	*2*	*3*
1	VAR00001	1.9512	.2028	.0664
2	VAR00002	– .1995	1.3140	.7743
3	VAR00003	– .6043	– 1.3429	.4680
4	VAR00004	– .9038	– .2969	– 1.8497
5	VAR00005	.8931	– 1.0092	– .0350
6	VAR00006	1.1045	.1529	– .7070
7	VAR00007	– 1.1031	1.6088	– .1289
8	VAR00008	– 1.1381	– .6295	1.4121

TABLE 15.3(a) Iteration history for the 2-dimensional solution (in squared distances) Young's S-stress formula 1 is used.

Iteration	*S-stress*	*Improvement*
1	.37358	
2	.29359	.07999
3	.28026	.01332
4	.27591	.00436
5	.27400	.00191
6	.27284	.00116
7	.27205	.00079
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress value are Kruskal's stress formula 1.

For matrix

Stress = .24015 RSQ – .58148

Configuration Coordinates
Stimulus Coordinates

TABLE 15.3(b)

Stimulus Number	*Stimulus Name*	*Dimension*	
		1	*2*
1	VAR00001	1.6156	.4756
2	VAR00002	– .2760	1.3796
3	VAR00003	– .2540	– 1.0558
4	VAR00004	– 1.2851	– .7799
5	VAR00005	.9602	– .9335
6	VAR00006	1.1044	.0666
7	VAR00007	– .5683	1.5124
8	VAR00008	– 1.2968	– .6620

TABLE 15.4(a) Iteration history for the 1 dimensional solution (in squared distances) Young's S-stress formula 1 is used

Iteration	*S-stress*	*Improvement*
0	.53492	
1	.50736	
2	.49658	.01078
3	.49595	.00063
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress value are Kruskal's stress formula 1.

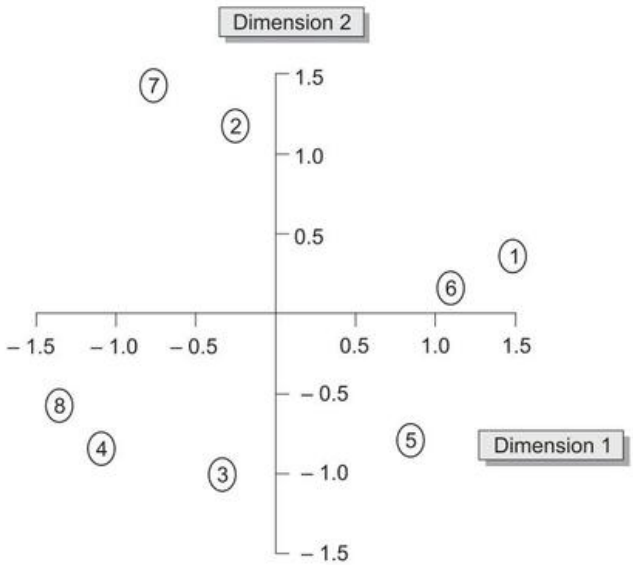
For matrix

Stress = .43159 RSQ = .35255

Configuration derived in 1 dimension
Stimulus Coordinates

TABLE 15.4(b)

Stimulus Number	*Stimulus Name*	*Dimension 1*
1	VAR00001	– 1.6474
2	VAR00002	– .4073
3	VAR00003	– .0704
4	VAR00004	1.2044
5	VAR00005	– 1.0409
6	VAR00006	– .2644
7	VAR00007	1.2424
8	VAR00008	.9837



BRANDS :

1	= AIWA	5	= SONY
2	= VIDEOCON	6	= ONIDA
3	= LG	7	= THOMSON
4	= SAMSUNG	8	= BPL

FIGURE 15.1 MULTIDIMENSIONAL SCALING 2-DIMENSIONAL OUTPUT : 8 COLOUR TVS

SUMMARY

The technique of multidimensional scaling has received a lot of attention in marketing research because it is a visually and intuitively appealing way to obtain perceptual maps of brands in a given product category. These maps can be used to determine the current positioning of brands and to reposition them if there is a need to do so.

Attribute based multidimensional scaling is similar to attribute based cluster analysis or even factor analysis. What is discussed in this chapter is the simpler (and intuitively more appealing) similarities (or dissimilarities) based approach.

Usually, a pairwise comparison of a set of brands forms the input data, where the respondent has to indicate the distance between each pair of brands, without any criterion or attribute being specified. Or, the respondent can fill up a table containing all the pairs of brands, and mention his dissimilarity ratings on any numerical scale. Ordinal data is adequate for multidimensional scaling of this type.

The data are analysed by the computer and the output provides a range of solutions in a range of dimensions specified (usually from 1 to 4 or 5 for a limited size of dataset). The output contains the solution for each dimension specified, and the researcher has to select one based on low stress value.

Once a particular number of dimensions is chosen, the researcher looks at the possible interpretations for the dimensions, and names the dimensions. At this stage, the process is subjective, and a good judgement will produce a useful interpretation.

Similar brands can be seen close to each other on a map if the solution chosen is a 2-dimensional one.

Multidimensional scaling provides an unbiased view of consumer perceptions regarding similarities between brands or objects. Sometimes this can provide startling new insights for marketing managers and may lead to strategic changes in brand positioning or communications.

ASSIGNMENT QUESTIONS

1. What is the major marketing application of multidimensional scaling?
2. Are there any conceptual or application-based similarities between multidimensional scaling and cluster analysis? Discuss.
3. What kind of input data is required for multidimensional scaling?
4. What is the minimum sample size required to perform multidimensional scaling?
5. What are the implications of doing multidimensional scaling on (1) a segmentwise sample, and (2) across segments on the full sample by averaging data values?
6. How do we decide on the number of dimensions that exist in the data?
7. On what basis can we name the dimensions, after we have zeroed in on a particular dimensionality for our solution?
8. A toothpaste manufacturer wants to know the current status of positions occupied by all the brands in the market, so that he can decide where to position his new brand, and whether to reposition his existing brand.

With this objective in mind, he conducts a study and gets from a sample of consumers in a target segment, the following distance matrix through a pairwise comparison method.

	Close Up	*Pepsodent*	*Colgate Dental Cream*	*Aquafresh*	*Colgate Gel*	*Colgate Total*
Close Up	0	1.5	5.7	4.8	2.2	6.5
Pepsodent	1.5	0	2.0	3.3	6.0	5.2
CDC	5.7	2.0	0	1.9	4.8	3.3
Aquafresh	4.8	3.3	1.9	0	5.0	3.8
Colgate Gel	2.2	6.0	4.8	5.0	0	6.1
Colgate Total	6.5	5.2	3.3	3.8	6.1	0

Is it possible to do a multidimensional scaling based positioning map of consumer perceptions for toothpastes? How many dimensions does it yield? What are these dimensions?

SPSS COMMANDS FOR MULTIDIMENSIONAL SCALING

After the input data has been typed along with variable labels and value labels in an SPSS file, to get the output for a **Multidimensional Scaling** problem similar to that described in the text,

1. click on ANALYZE at the SPSS menu bar (in older versions of SPSS, click on STATISTICS instead of ANALYZE).
2. click on SCALE followed by MULTIDIMENSIONAL SCALING.
3. select all the variables from the list on the left and transfer them to the right side by clicking on the arrow. If you have used pairwise distance data as mentioned in the text, select "Data are distances" under the box labeled "Distances".
4. click MODEL and type in the minimum and maximum dimensions you want in the solution. Generally, minimum can be either 1 or 2, and maximum can be about 3 or 4 for sample sizes that are moderate, like in student projects. As the sample size increases, the dimensionality can increase.

   Specify "Level of Measurement" as "Ordinal" or "Interval" depending on your scale of measurement. For pairwise comparison data, it should be "Ordinal".

   Click CONTINUE to return to the main dialogue box.
5. click OPTIONS and select "Individual Subject Plots". Then click CONTINUE to return to main dialogue box.
6. click OK to get the MDS output including plots.

CASE STUDY 1

Multidimensional Scaling*

Problem

A set of eight brands of detergent cakes available in the Indian market is taken and multidimensional scaling is used to determine how the Indian consumers perceive them. Also, it is used to find out how many dimensions the consumers seem to be considering, when they think of these brands.

The eight brands of detergents cake taken are:

1. Rin
2. Nirma
3. Ariel
4. OK
5. Eta
6. Wheel
7. Trilo
8. Super Bar 501

Input Data

TABLE 1.1

	VAR001	VAR002	VAR003	VAR004	VAR005	VAR006	VAR007	VAR008
var001	0.00	2.00	3.00	5.00	4.00	2.00	1.00	8.00
var002	2.00	0.00	5.00	7.00	6.00	4.00	1.00	3.00
var003	3.00	5.00	0.00	2.00	5.00	1.00	8.00	7.00
var004	5.00	7.00	2.00	0.00	3.00	4.00	7.00	8.00
var005	4.00	6.00	5.00	3.00	0.00	7.00	3.00	2.00
var006	2.00	4.00	1.00	4.00	7.00	0.00	5.00	6.00
var007	1.00	1.00	8.00	7.00	3.00	5.00	0.00	2.00
var008	8.00	3.00	7.00	8.00	2.00	6.00	2.00	0.00

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala

In this symmetrical 8 × 8 matrix, the columns and rows (var001 to var008) both denote the brands, and the body of the matrix is the measure of the perceived dissimilarity between brands. For example, row 4 of column 1 represents a distance of 5 between brand 4 (OK) and brand 1 (Rin) and so on. In other words both column numbers and row numbers represents the brands of the detergent cakes.

The two options that we have for collecting the above data are:

- Each consumer's distance matrix can be separately analysed through multidimensional scaling, or
- The ratings can be averaged across respondents and the single distance matrix can be constructed.

Here, we assume that the ratings have been aggregated, considering some homogeneity among the respondents.

Interpretation of Output

Tables 1.2(a) and 1.2(b) contain the 3-dimensional solution.

Tables 1.3(a) and 1.3(b) contain the 2-dimensional solution and

Tables 1.4(a) and 1.4(b) contain the 1-dimensional solution.

Firstly, we have to determine the number of dimensions in which we feel the best solution exists. The decision is based on the stress value for various solutions in different solution.

From the Tables 1.2(a), 1.3(a), and 1.4(a), the following values of stress can be obtained

3-dimensional solution: 0.05975

2-dimensional solution: 0.08609

1-dimensional solution: 0.36852

Clearly, the 1-dimensional solution is not a good one. Stress value indicates lack-of-fit, thereby it should be as close to zero as possible. The 2-dimensional solution looks better. But the 3-dimensional solution is the best because its stress value is the least.

Since we have only 8 brands to be compared, it is not possible to get a solution in more than 3-dimensions in this particular case. Ease of interpretation demands a trade-off between stress value and number of dimensions.

3-Dimensional Solution

Based on the stress value, it is desirable to use the 3-dimensional solution for interpretation in this case. For this purpose we should first name the dimensions.

The dimensions are usually the attributes offered by these brands either through our judgement or knowledge of the market or through a survey of the consumers, or a combination of these methods. This process of interpretation tends to be subjective regardless of the methods used.

We have named the three dimensions for this problem as follows

Dimension 1 : Value for money dimension

Dimension 2 : Dirt removal dimension

Dimension 3 : Fabric care

After having a look at the score on the three dimensions, we can conclude the brands like Super Bar 501 and Wheel are leading in Fabric care perception. Brands like OK and Eta are the best in Dirt removal perception and brands like Nirma, Trilo, and Super Bar 501 may be leading in Value for Money perception.

2-Dimensional Solution

If we were to choose the 2-dimensional solution instead of the 3-dimensional one, it could be plotted on a graph and would be easier for interpretation. This is shown in Fig 1.1 and the brands can be seen to form distant cluster based on their perceived similarity.

From the graph, brands 5 and 8, 3 and 4, 1 and 6, 2 and 7 are perceived to be similar. But we have to realise that we are losing some information on the third dimension which we called 'fabric care' in the 3-dimensional solution. The loss of information may turn out to be critical in some cases. Thus, it is better to analyse the data from a 3-dimensional solution instead of a 2-dimensional one, if the stress values indicates a 3-dimensional solution.

Output

TABLE 1.2(a) Iteration history for the 3-dimensional solution (in squared distances) Young's S-stress formula 1 is used.

Iteration	*S-stress*	*Improvement*
1	.13166	
2	.10232	.02934
3	.09078	.01154
4	.08724	.00354
5	.08631	.00093
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix

Stress = .05975 RSQ = .98047

TABLE 1.2(b) Configuration derived in 3-Dimensions Stimulus Coordinates

Stimulus Number	*Stimulus Name*	*Dimension*		
		1	*2*	*3*
1	VAR00001	– .1860	– .8867	– 1.1559
2	VAR00002	1.0175	– 1.1665	.3782
3	VAR00003	– 1.6674	– .1959	.4654
4	VAR00004	– 1.4639	1.3148	– .0511
5	VAR00005	.4844	1.5134	– .7565
6	VAR00006	– 1.0923	– .9168	.8062
7	VAR00007	1.3535	– .4514	– .7064
8	VAR00008	1.5543	.7892	1.0199

TABLE 1.3(a) Iteration History for the 2-Dimensional Solution (in squared distances) Young's S-stress formula 1 is used.

Iteration	*S-stress*	*Improvement*
1	.18037	
2	.15401	.02636
3	.14929	.00472
4	.14871	.00058
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix

Stress = .08609 RSQ = .96647

TABLE 1.3(b) Configuration derived in 2-Dimensions Stimulus Coordinates

Stimulus Number	*Stimulus Name*	*Dimension*	
		1	*2*
1	VAR00001	– .1114	– 1.0329
2	VAR00002	.7469	– 1.1393
3	VAR00003	– 1.4314	.1206
4	VAR00004	– 1.2552	1.0474
5	VAR00005	.5300	1.4018
6	VAR00006	– 1.0878	– .8479
7	VAR00007	1.2175	– .3499
8	VAR00008	1.3913	.8001

Output

TABLE 1.4(a) Iteration history for the 1-Dimensional solution (in squared distances) Young's S-stress formula 1 is used.

Iteration	*S-stress*	*Improvement*
1	.44803	
2	.44000	.00804
3	.44198	– .00198

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix

Stress = .36852 RSQ = .64662

TABLE 1.4(b) Configuration derived in 1-Dimensions Stimulus Coordinates

Stimulus Number	*Stimulus Name*	*Dimension 1*
1	VAR00001	.5342
2	VAR00002	– .9214
3	VAR00003	1.2041
4	VAR00004	1.3024
5	VAR00005	– .5174
6	VAR00006	.7732
7	VAR00007	– 1.0582
8	VAR00008	– 1.3168

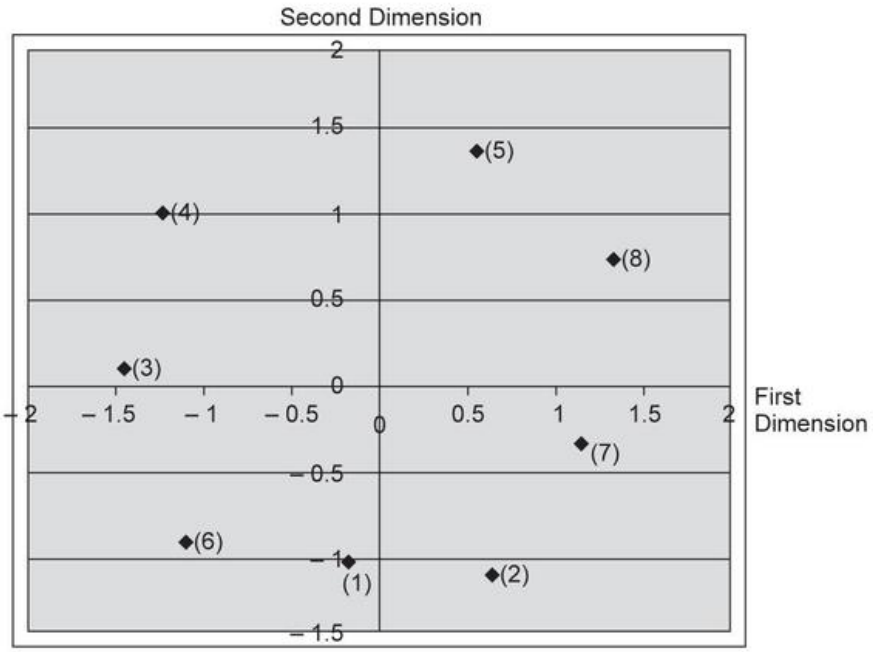


FIGURE 1.1 Multidimensional Scaling
2-Dimensional Output: 8 Brands of Detergent Cakes

Brands:

1. Rin
2. Nirma
3. Ariel
4. OK
5. Eta
6. Wheel
7. Trilo
8. Super Bar 501

CASE STUDY 2

Multidimensional Scaling*

Problem

Let us take the example of eight brands of newspapers circulating in the print media and use multidimensional scaling to determine the consumer perception, and try to plot a positioning map of these eight brands. An attempt is also made to find out the number of dimensions the consumers (populace) seem to be using immediately when they think of these newspaper brands.

Input Data

TABLE 1.1

	VAR1	*VAR2*	*VAR3*	*VAR4*	*VAR5*	*VAR6*	*VAR7*	*VAR8*
var1	0	3	4	7	1	5	1	8
var2	3	0	2	4	7	7	8	6
var3	4	2	0	1	3	6	7	3
var4	7	4	1	0	2	4	7	7
var5	1	7	3	2	0	2	4	5
var6	5	7	6	4	2	0	3	6
var7	1	8	7	7	4	3	0	2
var8	8	6	3	7	5	6	2	0

Input Data

The table above shows the input data collected from a sample of respondents each of whom was asked to rate the dissimilarity between all pairs of newspaper brands on a numerical scale (essentially, an ordinal measure). At this point two options arise. Each consumers distance matrix can be separately analysed through MDS, or the ratings can be averaged across respondents and a single distance matrix can be constructed. The assumption made here is that the ratings have been aggregated, assuming some homogeneity among the respondents.

For the sake of simplicity, integers have been used as distance measures in our framed example.

* Prepared by Bhaskar Majumdar, K B Srinivas, Nagaraj Gokale, and Rajiv Roy Chaudhuri

The symmetrical 8×8 matrix shows columns and rows denoting the brands, with the body of matrix containing a measure of the perceived dissimilarity between the brands.

Analysis

The decision on the best solution is determined by the number of dimensions that fits the same. It is based on the stress value for various solutions in different dimensions. In this case the results are as follows.

3-dimensional solution : 0.09039 (from Tables 1.1(a) and 1.1(b))
2-dimensional solution : 0.20997 (Tables 1.2(a) and 1.2(b))
1-dimensional solution : 0.42502 (Tables 1.3(a) and 1.3(b))

The results show that the 3-dimensional solution is the best one as the stress is close to zero. The 2-dimensional solution follows next, while the 1-dimensional solution having the highest stress value can not be taken for analysis.

3-dimensional Solution

The dimensions that we have taken for consideration on the eight brands of newspaper are as under:

Dimension 1 : Entertainment/leisure
Dimension 2 : Politics
Dimension 3 : Regional coverage

The 2-dimensional solution is plotted on the graph (Fig. 1.1) with the *X*-axis representing the Entertainment/leisure dimension while the *Y*-axis represents the political dimension respectively considering all the eight brands of newspapers.

It is seen from the graph that the brands *Hindustan Times* and *Times of India* enjoy good position in the dimension pertaining to Entertainment/leisure.

Brands like *Statesman*, *Indian Express*, and *Telegraph* enjoy good position in the market on the political front or dimension.

It is observed that there is a similarity between the brands *Hindustan Times* and *Times of India* considering both the dimensions. Brands like *Telegraph* and *Indian Express* are found to have a similarity. The brand *The Hindu* has been observed to be a standalone brand.

TABLE 1.1(a) Iteration History for the 3-Dimensional Solution (in Squared Distances) Young's S-stress Formula 1 is Used.

Iteration	*S-stress*	*Improvement*
1	.17513	
2	.14763	.02750
3	.13638	.01125
4	.13200	.00439
5	.13103	.00096
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances
  RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's Stress formula 1.

For matrix

Stress = .09039 RSQ = .95469

Configuration derived in 3 dimensions

TABLE 1.1(b) Stimulus Coordinates (3-dimensional solution)

Stimulus Number	*Stimulus Name*	*Dimension*		
		1	*2*	*3*
1	VAR00001	.6520	1.3540	.9128
2	VAR00002	– 1.4415	1.2010	– .3740
3	VAR00003	– 1.3894	.2482	– .3071
4	VAR00004	– 1.2919	– 1.0666	.5063
5	VAR00005	.4485	– .5762	1.1585
6	VAR00006	.7146	– 1.5062	.4564
7	VAR00007	1.6938	.4142	– .4534
8	VAR00008	.6139	– .0684	– 1.8995

TABLE 1.2(a) Iteration History for the 2-Dimensional Solution (in Squared Distances)
Young's S-stress formula 1 is Used.

Iteration	*S-stress*	*Improvement*
1	.34690	
2	.30015	.04676
3	.28979	.01035
4	.28479	.00501
5	.28146	.00332
6	.27936	.00211
7	.27803	.00132
8	.27744	.00059
Iterations stopped because		
S-stress improvement is less than .001000		

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix

Stress = .20997 RSQ = .80725

TABLE 1.2(b) Stimulus Coordinates (2-dimensional Solution)

Stimulus Number	*Stimulus Name*	*Dimension*	
		1	*2*
1	VAR00001	– .1735	– 1.4487
2	VAR00002	1.6737	– .4389
3	VAR00003	.9264	.6538
4	VAR00004	1.1472	.0258
5	VAR00005	– .6590	.3820
6	VAR00006	– .8459	– 1.0233
7	VAR00007	– 1.6313	.0970
8	VAR00008	– .4375	1.7522

TABLE 1.3(a) Iteration History for the 1-Dimensional Solution (in Squared Distances)
Young's S-stress formula 1 is used.

Iteration	*S-stress*	*Improvement*
0	.56542	
1	.54741	
2	.53711	.01030
3	.54144	.00433

- Stress and squared correlation (RSQ) in distances RSQ values are the proportion of variance of the scaled data (disparities) in the partition (row, matrix, or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress fromula 1.

For matrix

Stress = .42502 RSQ = .54243

TABLE 1.3(b) Stimulus Coordinates (1-Dimensional Solution)

Stimulus Number	*Stimulus Name*	*Dimension 1*
1	VAR00001	1.1318
2	VAR00002	– 1.3469
3	VAR00003	– .9812
4	VAR00004	– 1.3057
5	VAR00005	.5582
6	VAR00006	.5938
7	VAR00007	1.2504
8	VAR00008	.0995

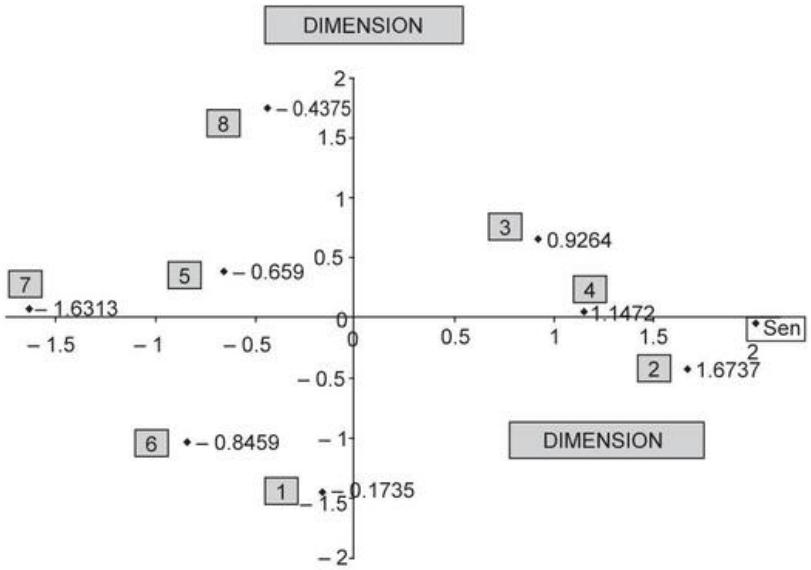


FIGURE 1.1 Multi Dimensional Scaling 2-Dimensional Output: 8 Newspapers

BRANDS:

1 = INDIAN EXPRESS
2 = TELEGRAPH
3 = STATESMAN
4 = DECCAN CHRONICLE
5 = DECCAN HERALD
6 = TIMES OF INDIA
7 = HINDUSTAN TIMES
8 = THE HINDU

CASE STUDY 3

Multidimensional Scaling*

Problem

We have taken the example of eight brands of toothpaste available in the Indian market, and use multidimensional scaling to determine how they are perceived by Indian consumers, and try to plot a positioning map of the eight brands. We will also attempt to find out how many dimensions the consumers seem to be using, when they think of toothpaste brands.

Input Data

TABLE 1.1

	Var00001	*Var00002*	*Var00003*	*Var00004*	*Var00005*	*Var00006*	*Var00007*	*Var00008*
Var00001	0.00	1.00	2.00	6.00	5.00	5.00	2.00	1.00
Var00002	2.00	0.00	5.00	4.00	4.00	6.00	4.00	3.00
Var00003	4.00	2.00	0.00	3.00	2.00	2.00	6.00	5.00
Var00004	6.00	4.00	3.00	0.00	3.00	3.00	8.00	7.00
Var00005	3.00	3.00	4.00	1.00	0.00	4.00	1.00	2.00
Var00006	5.00	5.00	6.00	5.00	1.00	0.00	3.00	4.00
Var00007	7.00	7.00	1.00	7.00	8.00	1.00	0.00	6.00
Var00008	1.00	8.00	7.00	2.00	6.00	7.00	5.00	0.00

Table 1.1 shows input data collected from a sample of respondents each of whom was asked to rate the dissimilarity between all pairs of toothpaste brands on a numerical scale (essentially an ordinal measure). The ratings obtained from each consumer's distance matrix have been aggregated, assuming some homogeneity among the respondents.

In this symmetrical 8×8 matrix, the columns and rows both denote the brands, and the body of the matrix is a measure of the perceived dissimilarity between brands. For example, row 3 column 1 represents a distance or dissimilarity of 4 between brand 3 and brand 1 of toothpastes and so on. In other words, both the column numbers and row numbers represent brands of toothpastes.

* Prepared by Harakuni Rajeev, Jitender Singh, Siddharth Agarwala, and Vidya Thekke Cherupilli

Output

TABLE 1.2(a) Iteration History for the 3-Dimensional Solution (in squared distances) Young's S-stress Formula 1 is Used

Iteration	*S-stress*	*Improvement*
1	0.33988	
2	0.29517	0.04471
3	0.28948	0.00569
4	0.28744	0.00204
5	0.28618	0.00126
6	0.28512	0.00105
7	0.28436	0.00076
Iterations stopped because		
S-stress improvement is less than .001		

- Stress and squared correlation (RSQ) in distance RSQ values are the proportion of variance of the scaled data (disparities).
- In the partition (row, matrix or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix:

Stress = 0.17287 RSQ = 0.83263

TABLE 1.2(b) Stimulus Coordinates (3-Dimensional Solution)

Stimulus Number	*Stimulus Name*	*Dimension*		
		1	*2*	*3*
1	Var00001	1.6998	– 0.2421	– 0.4185
2	Var00002	0.8403	– 1.0802	0.9735
3	Var00003	0.0569	0.7188	1.5506
4	Var00004	– 1.1509	– 1.3265	0.3901
5	Var00005	– 1.3633	– 0.3915	– 0.2848
6	Var00006	– 1.0400	1.0353	– 1.0043
7	Var00007	– 0.1164	1.7500	0.2535
8	Var00008	1.036	– 0.4637	– 1.4603

TABLE 1.3(a) Iteration History for the 2-Dimensional Solution (in Squared Distances)
Young's S-stress formula 1 is used

Iteration	*S-stress*	*Improvement*
1	0.47844	
2	0.40748	0.07097
3	0.40136	0.00612
4	0.39981	0.00155
5	0.39976	0.00005
	Iterations stopped because	
	S-stress improvement is less than .001	

- Stress and squared correlation (RSQ) in distance RSQ values are the proportion of variance of the scaled data (disparities).
- In the partition (row, matrix or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix:
Stress = 0.30203 RSQ = 0.63852

TABLE 1.3(b) Stimulus Coordinates (2-Dimensional Solution)

Stimulus Number	*Stimulus Name*	*Dimension*	
		1	*2*
1	Var00001	– 1.5303	0.2218
2	Var00002	– 0.7548	1.2020
3	Var00003	1.0302	0.7452
4	Var00004	1.0279	0.9942
5	Var00005	0.9870	– 0.1844
6	Var00006	0.5589	– 1.4296
7	Var00007	0.2086	– 1.4789
8	Var00008	– 1.5275	– 0.0704

TABLE 1.4(a) Iteration History for the 1-Dimensional Solution (in Squared Distances)

	Young's S-stress formula 1 is used	
Iteration	*S-stress*	*Improvement*
0	0.63976	
1	0.61186	
2	0.58324	0.02861
3	0.58343	– 0.00019
Iterations stopped because		
S-stress improvement is less than .001		

- Stress and squared correlation (RSQ) in distance RSQ values are the proportion of variance of the scaled data (disparities).
- In the partition (row, matrix or entire data) which is accounted for by their corresponding distances.
- Stress values are Kruskal's stress formula 1.

For matrix:

Stress = 0.52914 RSQ = 0.38298

TABLE 1.4(b) Stimulus Coordinates (1-Dimensional Solution)

Stimulus Number	*Stimulus Name*	*Dimension 1*
1	Var00001	1.2698
2	Var00002	1.3594
3	Var00003	– 0.7331
4	Var00004	– 0.8530
5	Var00005	– 0.3969
6	Var00006	– 0.9953
7	Var00007	– 0.8416
8	Var00008	1.1908

Interpretation of Computer Output

In Tables 1.2(a) and (b), 1.3(a) and (b), 1.4(a) and (b), we have the output of a multidimensional scaling programme using one of the standard packages. Tables 1.2(a) and 1.2(b) contain the 3-dimensional solution, Tables 1.3(a) and 1.3(b) contain the 2-dimensional solution and Tables 1.4(a) and 1.4(b) contain the 1-dimensional solution.

Our first task is to determine the number of dimensions in which we feel the best solution exists. The decision is based on the stress value for various solutions in different dimensions. From Tables 1.2(a), 1.3(a) and 1.4(a), we see the following stress values.

3-dimensional solution : .17287
2-dimensional solution : .30203
1-dimensional solution : .52914

Cloearly, the 1-dimensional solution is not a good one. Since, the stress value indicates lack-of-fit, it should be as close to 0 as possible. The 2-dimensional solution is better, but the 3-dimensional looks the best, as the stress value is a low 17.

3-Dimensional Solution

We assume that we use the 3-dimensional solution for interpretation in this case, our next task is to name a dimension. For doing so our previous knowledge of the brands, which as stated below may become important.

1. Colgate
2. Cibaca
3. Close-up
4. Pepsodent
5. Babool
6. Aquafresh
7. Meswak
8. Neem

For these 8 brands, we must look at the qualities of the various attributes offered by these brands either through our judgement or knowledge of the market or through the survey of the consumers or a combination of these methods. This process of interpretation tends to be subjective, regardless of the method used. For instance, we could look at the output of multidimensional scaling and the scores for the 8 brands on the 3 dimensions and decide on the following names for the 3 dimensions:

Dimension 1 : Whitening of teeth
Dimension 2 : Stopping bad breath
Dimension 3 : Fighting tooth decay

We could then look at the scores on the 3 dimensions and conclude that that some brands like Cibaca and Close-up are perceived by the customers as being highly effective in ‘fighting tooth-decay’. Aquafresh and Meswak are leading in ‘stopping bad breath’ perceptions. Colgate and Neem maybe perceived as having the most ‘whitening of teeth’ property.

CHAPTER 16 CONJOINT ANALYSIS FOR PRODUCT DESIGN

Learning Objectives

In this chapter, we will

- Introduce Conjoint Analysis as a technique for analysing consumer utility levels for specific product attributes
- Elaborate on the concept of a tradeoff between various attribute values in consumer decision-making
- Emphasise the practical use of conjoint analysis through a worked out example
- Explain how conjoint can convert rank order data into interval scaled utility measures
- Explain how to perform conjoint analysis by recoding the input data as dummy variables through Effects Coding and running it as a regression model

APPLICATION AREAS

One of the eternal questions in a marketing manager's mind has been "how does the customer value the various tangible and intangible benefits/features offered by my product?" Or, in other words, what is the 'Utility' he sees in my price level, in my after-sales service level, my product features, say speed, accuracy, or any other? Conjoint analysis is a multivariate technique that captures the exact levels of utility that an individual customer puts on various attributes of the product offering. It enables a direct comparison between say, the utility of a price level of Rs. 400 versus Rs. 500, a delivery period of 1 week versus 2 weeks, or an after-sales response of 24 hours versus 48 hours. Once we know utility levels for every attribute (and at every level), we can combine these to find the best combination of attributes that gives him the highest utility, the second best combination that gives the second highest utility, and so on. If this is done across a sample of customers say, segmentwise, it can also be used to predict market-share, and the response of customers to changes in the competitive strategy through changes in the marketing elements.

METHODS

Essentially, conjoint analysis is an attempt to convert ordinal scale rankings given by respondents into an interval scale 'value' or 'utility' scale. The method is quite straight-forward in the data collection stage. The researcher determines a set of attributes and their levels, say 3 attributes, each at 2 levels, which he feels are critical decision-making variables for his consumers. Now, all possible combinations of these levels are listed out. For example, in a readymade shirt, price could be one factor, at levels Rs. 300 or Rs. 350, stores could be exclusive or non-exclusive, and design could be checks or solid colours. We would then take all the possible combinations as follows:

1. Rs. 300 Exclusive store — Checks
2. Rs. 300 Exclusive store — Solid colours
3. Rs. 350 Exclusive store — Checks
4. Rs. 350 Exclusive store — Solid colours
5. Rs. 300 Non-exclusive store — Checks
6. Rs. 300 Non-exclusive store — Solid colours
7. Rs. 350 Non-exclusive store — Checks
8. Rs. 350 Non-exclusive store — Solid colours

These eight combination can be presented to the respondent of our survey, and he is asked to rank the combination he prefers from rank 1 to rank 8. This forms the input data for conjoint analysis. The objective, as stated earlier, is to convert these rankings into utilities, so we know how this respondent's utility varies with any change in the level of any of the attributes. In other words, the output of conjoint analysis will generate utility levels for combinations given above. For example, the computer output may generate a utility table that looks like this:

Rs. 300	— Utility 5
Rs. 350	— Utility 1
Checks	— Utility 10
Solid solours	— Utility 6
Exclusive stores	— Utility 4
Non-exclusive stores	— Utility 2

Thus, this table indicates that relatively, checks have the highest utility of 10, and solid colours 6. Price at the given price points has lower utility, but still, Rs. 300 has a much higher utility than Rs. 350. Relatively, exclusivity or otherwise of the stores has less utility. But exclusive stores have 4, and non-exclusive stores have 2. The best combined utility can also be calculated for the original eight combinations. For example, for this consumer, the best utility combination would be a price of Rs. 300, checks and exclusive stores—5 + 10 + 4 = 19 points. The second best would be 5 + 10 + 2 = 17 points. (Rs. 300, checks, non-exclusive stores). The third, fourth, fifth….eighth best combination and their utilities can similarly be found.

RECOMMENDED USAGE

The usage of conjoint analysis can be at three levels:

1 Individual consumer.
2 Segment level.
3 Across segments.

For industrial marketing usage, individual level usage is recommended. This is because the industrial marketing consumers are usually smaller in number, and larger in importance individually, as compared with consumer goods. Each significant consumer may be a segment in itself.

In the case of consumer goods or services, it is advisable to do the exercise segmentwise, without losing the benefit of getting opinions of individuals. Since sample sizes would be large, running conjoint analyses for every consumer individually could mean running 2,000 plus analyses, and it would involve a lot of time to analyse and interpret the data. In any case, even if the analysis were done individually, to make sense of the findings, it would have to be grouped post-facto into meaningful segments. The author's recommendation is therefore to group the consumers into segments, and use average values to do the conjoint analysis in the case of consumer goods and services if the sample sizes are high.

Number of Attributes and Levels

To avoid creating masses of data, the researcher has to be careful in selecting both the number of attributes and the number of levels of each. Only those attributes and levels must be used, which are feasible offerings from the manufacturer's/marketer's view point.

Number of Combinations

Another point of interest is that the number of combinations being offered for ranking by respondents should not be too high. For example, beyond about 25 or 30 combinations, respondent fatigue would probably induce inaccurate or disinterested responses, affecting the validity of the procedure. In such cases, a partial list of combinations can be chosen. An orthogonal design employing a subset of the full list of attribute combinations can be generated by many of the statistical packages. This pruned list can be offered for ranking by respondents for the input data.

WORKED EXAMPLE

Let us take the example of an industrial product—a CNC machine tool which is used to perform a variety of manufacturing operations—to illustrate the application of conjoint analysis. We first identify the attributes of the product which are important to customers, and then the levels for each attribute that we are willing to design and offer to a customer. Thus, this will be an application of conjoint analysis for product design of an industrial machine tool. Let us assume that three attributes of a CNC machine tool are important:

1 Setup time in minutes. This is the time it takes to prepare or setup the machine for operations.
2 Delivery period in days. This is the time the manufacturer needs to deliver after the customer has placed an order.
3 Number of different tools the machine can accommodate. This is a measure of machine flexibility in performing different operations.

These are the three attributes. The levels of these attributes are:

1 Setup time—3 minutes, 6 minutes, 9 minutes, 12 minutes (4 levels)
2 Delivery period—18 days, 22 days, 28 days (3 levels)
3 Number of tools—4, 8 or 10 (3 levels)

To recapitulate, the objective of conjoint analysis is to determinate from a potential customer of this machine tool, how important each attribute is to him. Also, we want to know how much utility he derives from a given combination of these levels of attributes. Since we have 4, 3 and 3 levels of the three attributes, we get a total of $4 \times 3 \times 3 = 36$ different combinations of attribute levels.

Ranking

The next stage of the input process is to collect from a respondent his ranking for all the 36 combinations of attribute levels. This table would look like Table 16.1.

RUNNING CONJOINT AS A REGRESSION MODEL

For those who do not have a conjoint analysis module on their statistical package, it is quite easy to convert the conjoint analysis input into an equivalent regression model and run it as a regression. The coding of the attribute levels for this purpose is known as 'Effects coding' and Table 16.2 shows our machine tool example coded for a regression run. In this input data matrix (Table 16.2), which is similar to coding of dummy variables, the four levels of Setup time are coded as shown in the following table.

Setup time in minutes	*Var 1*	*Var 2*	*Var 3*
S3	1	0	0
S6	0	1	0
S9	0	0	1
S12	– 1	– 1	– 1

Thus, 3 Variables, Var 1, Var 2, Var 3 are used to indicate 4 levels of setup time, as per the coding scheme above.

Similarly, the coding scheme for the 3 levels of the attribute Delivery period is as shown below:

Delivery period in days	*Var 4*	*Var 5*
D18	1	0
D22	0	1
D28	– 1	– 1

Finally, the coding scheme for Number of tools is as shown below-

Number of tools	*Var 6*	*Var 7*
T4	1	0
T8	0	1
T10	– 1	– 1

Thus, seven variables var 1 to var 7 are used to represent the 4 levels of Setup time (S3, S6, S9 and S12), 3 levels of Delivery period (D18, D22 and D28), and 3 levels of Number of tools (T4, T8 and T10). All the 7 variables are independent variables in the regression run. Var 8 is the rating of each combination given by a respondent, and forms the **dependent** variable for the regression run.

If the conjoint analysis is run as a regression model, the **rating** (which is a reverse of ranking) is used as a dependent variable. All combinations from the first to the thirty-sixth were ranked by the respondent. Rank 1 can be considered as highest rating and given a rating of 36. Rank 2 can be given a rating of 35, and so on. Strictly speaking, this is not an interval scale rating, and should have only ordinal interpretation.

Output and its Interpretation

If run as a regression model, the output is shown in Table 16.3. Variables 1 to 7 are treated as independent variables. Now, the column titled 'B' (the regression coefficients column) provides the part utility of each level of attributes.

For example, Setup time of S3 (3 minutes) is represented by variable 1 as per our coding scheme. Its utility is equal to 5.5 (looking under column B of Table 16.3, for variable 1). Similarly, the utility for variable 2 representing S6 (Setup time of 6 minutes) is 4.16 and for variable 3 representing S9, it is – 1.05. The utility for the fourth level of Setup time (S12), is not in the table, but is derived from the property of this coding, that all the utilities for a given attribute should sum to 0. Thus, utility for S12 should be equal to – (5.5 + 4.16 – 1.05), or – 8.61.

Similarly, for Delivery Period, the utilities of D18 and D22 are given by the numbers 3.33 and 1.25, against var 4 and var 5 in Table 16.3. But the utility for D28 is derived from the same property, that the sum of the utilities for Delivery period should sum to zero. Therefore D28 has a utility of – (3.33 + 1.25) or – 4.58.

Finally, for Number of tools, T4 has a utility of – 10.33 (variable 6 in Table 16.3) and T8 has a utility of 1.58 (variable 7 in Table 16.3). T10 has a derived utility of – (– 10.33 + 1.58) or + 8.75.

Now, we have the utilities for all the levels of all attributes, and we can put them into a table, as follows (rounded off to 1 decimal points).

Utilities Table for Conjoint Analysis

Attribute	*Level*	*Part Utility*	*Range of Utilities*
1. Setup time in minutes	S3	5.5	(Max. – Min.)
	S6	4.2	= 5.5 – (– 8.6)
	S9	– 1.1	= **14.1**
	S12	– 8.6	for Setup time

2. Delivery period in days

D18	3.3	(Max. – Min.)
D22	1.3	= 3.3 – (– 4.6)
D28	– 4.6	= **7.9**
		for Delivery period.

3. Number of tools (flexibility)

T4	– 10.3	(Max. – Min.)
T8	1.6	= 8.7 – (– 10.3)
T10	8.7	= **19.0**
		for No. of tools

Now, with the part utilities of every level of every attribute available to us, we can come to several conclusions. First, we can conclude that machine flexibility is the most important attribute for this customer. There are two indicators for this. One, the range of utility values is highest (19.0) for number of tools (flexibility). Two, the highest individual value of utility for any level of any attribute is 8.7, for T10 (number of tools = 10). Both these figures indicate that number of tools is the most important attribute at given levels of attributes. The Setup time seems to be the second most important attribute, as its range of utilities is 14.1, as shown in the above table. The last attribute in relative importance is the Delivery period, with a utility range of 7.9.

Combination Utilities

We can also pick up one attribute level from each attribute and combine their part utilities to calculate the total utility of the combination. For example, S3, D18 and T4 have a combined utility of 5.5 + 3.3 – 10.3 = – 1.5. Similarly, S3, D22 and T4 have a combined utility of 5.5 + 1.3 – 10.3 = – 3.5.

If we want the best combination, we pick the highest utilities from each attribute, and add them. S3 + D18 + T10 in this case is the most preferred combination with a combined utility of 5.5 + 3.3 + 8.7 = 17.5. The next best combination is S6 + D18 + T10, with a combined utility of 4.2 + 3.3 + 8.7, or 16.2.

Individual Attributes

We can also check what difference in utility a change of one level in one attribute makes. For example, S3 to S6 (Setup time change from 3 to 6 minutes) induces only a 1.3 units drop in utility, but it gets progressively more at the next stage—S6 to S9 has a difference in utility of 5.3.

Similarly, increase in Delivery period from 18 to 22 days costs 2.0 units (3.3 – 1.3) of utility drop, whereas 22 to 28 days causes 5.9 units of drop in utility (1.3 – (– 4.6)).

Finally, Number of tools causes a drastic change in utility of 11.9 units from T8 to T4, and a significant drop in utility by 7.1 units from T10 to T8.

ADDITIONAL COMMENTS

We have seen an example of conjoint analysis for a single respondent in an industrial marketing situation. But the same process is useful in any consumer product/service situation when designing or

redesigning the product offering. As we have seen, service aspects of a product can also be incorporated into the conjoint analysis.

As we saw earlier, any number of attributes and levels of these attributes can be tested, subject only to respondent fatigue. If the number of combinations is larger than about 25–30, it is advisable to use fractional factorial designs, using a subset of the total combinations. The conjoint analysis module of the computer package would explain how to do this. For example, SPSS has a feature called Orthoplan in its conjoint analysis module which helps the researcher to generate a subset of all the possible combinations of attribute levels. This can generate a sample of the total combinations for any specified sample size, with minimum loss of information.

The input data matrix of Table 16.1 can be directly input into a conjoint analysis program if available in the package being used. If not, the approach we have used above can be used, with effects coding, to run the conjoint analysis using a regression model. The results are equivalent, and will be as useful. If using the CONJOINT module of SPSS, you have to create two data files instead of one. The Ratings column is entered into a separate data file in a single row, and the Setup Time, Delivery Period and No. of Tools data are entered in a separate data file. This procedure is elaborated in the SPSS commands section at the end of this chapter.

TABLE 16.1 Input Data for Conjoint Analysis

Machine Tool Problem

S. No	*Setup Time (Minutes) 3, 6, 9, 12*	*Delivery Period (Days) 18, 22, 28*	*No. of Tools 4, 8, 10*	*Rating 36 to 1*
1	3	18	10	36
2	6	18	10	35
3	3	22	10	34
4	6	22	10	33
5	9	18	10	32
6	9	22	10	31
7	3	18	8	30
8	6	18	8	29
9	3	28	10	28
10	6	28	10	27
11	3	22	8	26
12	6	22	8	25
13	3	28	8	24
14	9	28	10	23

(Table 16.1 contd)

(Table 16.1 contd)

S. No	Setup Time (Minutes) 3, 6, 9, 12	Delivery Period (Days) 18, 22, 28	No. of Tools 4, 8, 10	Rating 36 to 1
15	6	28	8	22
16	9	18	8	21
17	9	22	8	20
18	12	18	10	19
19	12	22	10	18
20	3	18	8	17
21	6	18	4	16
22	3	22	4	15
23	12	18	8	14
24	12	22	8	13
25	9	22	4	12
26	12	28	10	11
27	9	28	8	10
28	9	18	4	9
29	9	22	4	8
30	12	28	8	7
31	3	28	4	6
32	6	28	4	5
33	12	18	4	4
34	9	28	4	3
35	12	22	4	2
36	12	28	4	1

TABLE 16.2 Conjoint Problem Input Data Coded for Regression

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8
1	1.00	.00	.00	1.00	.00	– 1.00	– 1.00	36.00
2	.00	1.00	.00	1.00	.00	– 1.00	– 1.00	35.00
3	1.00	.00	.00	.00	1.00	– 1.00	– 1.00	34.00
4	.00	1.00	.00	.00	1.00	– 1.00	– 1.00	33.00

(Table 16.2 contd)

(Table 16.2 contd)

	Var 1	*Var 2*	*Var 3*	*Var 4*	*Var 5*	*Var 6*	*Var 7*	*Var 8*
5	.00	.00	1.00	1.00	.00	– 1.00	– 1.00	32.00
6	.00	.00	1.00	.00	1.00	– 1.00	– 1.00	31.00
7	1.00	.00	.00	1.00	.00	.00	1.00	30.00
8	.00	1.00	.00	1.00	.00	.00	1.00	29.00
9	1.00	.00	.00	– 1.00	– 1.00	– 1.00	– 1.00	28.00
10	.00	1.00	.00	– 1.00	– 1.00	– 1.00	– 1.00	27.00
11	1.00	.00	.00	.00	1.00	.00	1.00	26.00
12	.00	1.00	.00	.00	1.00	.00	1.00	25.00
13	1.00	.00	.00	– 1.00	– 1.00	.00	1.00	24.00
14	.00	.00	1.00	– 1.00	– 1.00	– 1.00	– 1.00	23.00
15	.00	1.00	.00	– 1.00	– 1.00	.00	1.00	22.00
16	.00	.00	1.00	1.00	.00	.00	1.00	21.00
17	.00	.00	1.00	.00	1.00	.00	1.00	20.00
18	– 1.00	– 1.00	– 1.00	1.00	.00	– 1.00	– 1.00	19.00
19	– 1.00	– 1.00	– 1.00	.00	1.00	– 1.00	– 1.00	18.00
20	1.00	.00	.00	1.00	.00	1.00	.00	17.00
21	.00	1.00	.00	1.00	.00	1.00	.00	16.00
22	1.00	.00	.00	.00	1.00	1.00	.00	15.00
23	– 1.00	– 1.00	– 1.00	1.00	.00	.00	1.00	14.00
24	– 1.00	– 1.00	– 1.00	.00	1.00	.00	1.00	13.00
25	.00	1.00	.00	.00	1.00	1.00	.00	12.00
26	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	11.00
27	.00	.00	1.00	– 1.00	– 1.00	.00	1.00	10.00
28	.00	.00	1.00	1.00	.00	1.00	.00	9.00
29	.00	.00	1.00	.00	1.00	1.00	.00	8.00
30	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	.00	1.00	7.00
31	1.00	.00	.00	– 1.00	– 1.00	1.00	.00	6.00
32	.00	1.00	.00	– 1.00	– 1.00	1.00	.00	5.00
33	– 1.00	– 1.00	– 1.00	1.00	.00	1.00	.00	4.00
34	.00	.00	1.00	– 1.00	– 1.00	1.00	.00	3.00
35	– 1.00	– 1.00	– 1.00	.00	1.00	1.00	.00	2.00
36	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	1.00	.00	1.00

TABLE 16.3 Multiple Regression Output for the Conjoint Problem

Equation Number 1 Dependent Variable.. VAR00008
Method: Enter
Variable(s) Entered on Step Number

1. VAR00007
2. VAR00005
3. VAR00003
4. VAR00006
5. VAR00004
6. VAR00002
7. VAR00001

Multiple R	.98119
R Square	.96273
Adjusted R Square	.95342
Standard Error	2.27390

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	7	3740.22222	534.31746
Residual	28	144.77778	5.17063
F = 103.33691	Signif *F* = .0000		

Variables in the Equation

Variable	*B*	*SE B*	*Beta*	*T*	*Sig. T*
VAR00001	5.500000	.656419	.374372	8.379	.0000
VAR00002	4.166667	.656419	.283615	6.348	.0000
VAR00003	– 1.055556	.656419	– .071849	– 1.608	.1190
VAR00004	3.333333	.535964	.261992	6.219	.0000
VAR00005	1.250000	.535964	.098247	2.332	.0271
VAR00006	– 10.333333	.535964	– .812177	– 19.280	.0000
VAR00007	1.583333	.535964	.124446	2.954	.0063
(Constant)	18.500000	.378984		48.815	.0000

All requested variables entered.

SUMMARY

Conjoint analysis is ideally suited for product (or service) design problems. This is because the technique is able to put numerical values on the mysteries of the consumer's mind. It tries to map his decision-making process and the tradeoffs he makes (which he may be unable to articulate if directly questioned) while choosing a particular product offering. The result of a conjoint analysis is a set of

utility values for every product variation and attribute level on offer. This is achieved by first listing out (at the input stage) all possible combinations of product or service attributes that the marketer is willing to offer. Let us assume there are a total of 64 different combinations. We then ask the respondent to rank all 64 (or a part thereof according to a planned selection of a subset) in order of his preference. These rankings form the input data.

The output can be obtained in two ways. One, by directly using the CONJOINT program module available on a computer package such as SPSS. Two, by coding the variables in a form known as Effects coding explained in the chapter, to convert the problem into an equivalent regression problem. The output of the equivalent regression problem yields coefficients which are the 'utilities' of the conjoint problem. These utilities can be analysed to map the individual customer's utility levels for levels of individual attributes and his preference levels for different combinations of attributes (product offerings).

Conjoint analysis can be done for a sample size of 1, or across a segment, or across a sample. If done for individual customers (as in industrial marketing applications), respondents with similar preference values (utilities) can be added together to get an estimate of the size of various segments, and used for market share predictions.

This technique is widely used by all types of industries today. The case studies at the end of the chapter further illustrate applications in various industries. SPSS Commands for Conjoint Analysis are given after the Assignment Questions.

ASSIGNMENT QUESTIONS

1. A package tour operator in Goa wants to find out from his customers their utility levels for different features and combinations of features of a package tour. In particular, there are three features he chooses for his research:
   (a) *Choice of hotel:* the three choices are beach-front cottages, motels, and city hotels.
   (b) *Mode of transport:* luxury coaches, private luxury cabs, and private ordinary cabs are the three choices of transport he offers.
   (c) *Guides:* he offers the services of a male guide, or a female guide or no guide as the three options.

   The operator wants to administer the questionnaire to one or two representative individual tourists from each of the following countries—Italy, Germany, and France, from where bulk of his customers come.

   You have to help him with the questionnaire design. How would you do it?
2. Assume you have collected from one German tourist the rankings for all the 27 different combinations of variables (attributes and levels), mentioned in Question 1 above. Put the (assumed) data into a table form for input into a conjoint analysis package.
3. Perform the conjoint analysis for the above problem, either by using the conjoint analysis package available to you, or by recoding the data for analysis using regression analysis.
4. Interpret the output of the above analysis and explain the following.

(a) Which of the three attributes is the most important to this customer—the choice of hotel, mode of transport, or type of guide?
(b) What is the best combination for the tourist, and its utility for her?
(c) What is the second best combination for this tourist, and its utility?
(d) What is the worst combination, and its utility level?

5. Comment on whether it is a good idea to perform conjoint analysis by combining data (a) segmentwise or (b) across segments in a survey.
6. A major Indian Multi Utility Vehicle (MUV) manufacturer wants to launch a new vehicle in the market. He wants to know the price level, seating capacity, service level, and warranty period that will strike a right balance with his customers. He is prepared to offer the following:
   (a) Price: Rs. 5 lakh, Rs. 6 lakh
   (b) Seating capacity: 8 or 10
   (c) Warranty period: 1 or 2 years
   (d) Service level: every 500 km or every 800 km on any major highway.

Some potential customers of MUVs are contacted, and a rating from them collected for all the possible combinations of the 4 attributes and their options. The data from one customer who is considered a representative of the target market are shown in Table 1 below.

Perform a conjoint analysis on the data and advise the MUV manufacturer on how to interpret it.

TABLE 1

S. No.	*Price in Rs. Lakh*	*Seating Capacity*	*Warranty Period (Yrs)*	*Service (500/800 km)*	*Rating*
1	5	8	1	500	13
2	5	8	2	500	14
3	5	10	1	500	15
4	5	10	2	500	16
5	5	8	1	800	5
6	5	8	2	800	6
7	5	10	1	800	7
8	5	10	2	800	8
9	6	8	1	500	9
10	6	8	2	500	10
11	6	10	1	500	11
12	6	10	2	500	12
13	6	8	1	800	1
14	6	8	2	800	2
15	6	10	1	800	3
16	6	10	2	800	4

SPSS COMMANDS FOR CONJOINT ANALYSIS

There are two approaches to doing conjoint analysis using SPSS.

Approach 1 (Using Effects coding and Regression commands of SPSS)

This approach is described in the Conjoint Analysis chapter in the book, and is very simple to use, particularly for small sample sizes or a sample of size 1.

1. Make a list of all possible combinations (also called a **full profile**) of attribute levels. Say, there are 36 combinations. These are coded as independent variables, using Effects coding, as described in the chapter.
2. Get rankings or ratings of all these combinations (also called product profiles) from a respondent (say, RANKS 1 to 36, or RATINGS from 36 to 1, obtained by reversing the ranks).
3. Do a regression by entering all the independent variables in the regression model using the ENTER command in SPSS, and using RATINGS as the dependent variable (the SPSS command structure is already described in the earlier section on SPSS commands for regression).
4. The regression coefficients obtained in the regression output are the part utilities of each attribute level, except one attribute level for each variable, which is determined by using the property that all the attribute levels for a variable should sum to 0. Please look at the chapter on Conjoint Analysis for details.

Approach 2 (Using the SPSS procedure called Conjoint)

This is a slightly complex procedure, but sometimes necessary when summary utilities are needed to be calculated for a large sample. It also requires that the **Conjoint** module of SPSS be available on your computer. This procedure utilises the capability of SPSS to generate orthogonal fractions (sub-samples) of the entire attribute-level combination set, if you have too many of these. There is a procedure called ORTHOPLAN in SPSS which can be used to generate a fractional factorial design. A brief simplified description follows. You may refer to the SPSS manual for further details.

1. Create a data file in SPSS (we assume it will be named DATAFILE1) containing either the full profile data (all the attribute level combinations) or a subset of the data called an ORTHOGONAL FRACTIONAL DESIGN. (To create the orthogonal fractional design, you have to use a command called ORTHOPLAN in SPSS. We will assume you are using either a full profile or the fractional design, stored in a file called DATAFILE1).
2. Get the RATINGS or RANKINGS from each respondent for all the combinations of attributes contained in your file called DATAFILE1. Enter these RATINGS or RANKINGS in a separate data file (to be called DATAFILE2) in SPSS, using ONE ROW for each respondent's ratings or rankings. If there are 36 combinations in DATAFILE1, there will be 36 columns in DATAFILE2. The number of rows will be equal to the number of respondents in your sample.
3. After creating DATAFILE1 and DATAFILE 2, open a third file called a SYNTAX file using the FILE, OPEN command followed by SYNTAX.
4. Type in the following-CONJOINT PLAN = 'DATAFILE1.sav'/DATA 'DATAFILE2.sav'/ SCORES = SCORE 1 TO SCORE 36/FACTOR VAR1 (DISCRETE) VAR2 (DISCRETE) VAR3 (DISCRETE) /PLOT ALL.

Here, the variables VAR1, VAR2, etc., are the independent variables appearing in DATAFILE1 of your conjoint problem. For the example given in the text, these are SETUP, DELIVERY and NUMTOOLS. DISCRETE is the type of variable, to be specified for each variable. SCORES is the term used for Ratings. There are 36 SCORES because in our example there were 36 Ratings.The word RANK should be used in the syntax instead of SCORES if Rankings are contained in DATAFILE2.

5. Click RUN from the menu of the SYNTAX file you created. Click ALL in the menu which appears. If the syntax is correct, the output for CONJOINT will appear. On the output file, look for SUMMARY UTILITIES for graphs containing the part utilities for each attribute level. Look for **Importance Summary** to get the relative importance of the attributes.

   For any variations, consult the SPSS manual.For instance, in SPSS 10.1.3, to create an orthogonal sub-sample of the combinations available, click on DATA, then click on ORTHOGONAL DESIGN, and GENERATE, and then fill up the details of the variables and their values. Then, under OPTIONS, specify the minimum number of cases to be generated, click CONTINUE. Then specify a destination file to save the orthogonal design, by clicking on FILE. Then click OK to generate the combinations (orthogonal design data).

CASE STUDY 1

Conjoint Analysis*

Problem

XYZ paint company identified the attributes which are important to their customers and also classified each of the attributes into their levels. Based on this, they want to use the technique of conjoint analysis to determine from a potential customer's point of view, how important each attribute is to him.They also want to know how much utility the customer derives from a given combination of these levels of attributes. It also helps to understand the feasible offerings from the marketer's point of view. The three important attributes identified for the paint are:

1. Life—this is the number of years the paint coat lasts.
2. Price—the price of one litre of paint.
3. Colour—the colour of paint

The levels of the above mentioned attributes are as follows:

- Life—3 years, 4 years, 5 years
- Price—Rs. 50 per litre, Rs. 60 per litre, Rs. 70 per litre
- Colour—Green, Blue, Cream

Input Data

After the attributes and their levels are decided, the next stage is to collect from the respondent, the ranking for all 27 combinations of attribute levels. This can be seen from Table 1.1.

TABLE 1.1 Input Data for Conjoint Analysis

S. No.	Life (in years)	Price (Rs/Litre)	Colour	Rating (27 to 1)
1	5	50	Green	27
2	4	50	Green	26
3	5	50	Cream	25
4	5	50	Blue	24
5	5	60	Green	23
6	4	60	Green	22

(Table 1.1 contd)

* Prepared by Divya Sharma, Manoj Jacob, Shilpy Tiwari, and Swapna Gurijala

(Table 1.1 contd)

S. No.	Life (in years)	Price (Rs/Litre)	Colour	Rating (27 to 1)
7	5	70	Green	21
8	5	60	Blue	20
9	5	60	Cream	19
10	4	50	Blue	18
11	4	50	Cream	17
12	5	70	Blue	16
13	3	50	Green	15
14	5	70	Cream	14
15	3	50	Blue	13
16	4	60	Blue	12
17	4	60	Cream	11
18	3	50	Cream	10
19	4	70	Green	9
20	3	60	Green	8
21	4	70	Blue	7
22	3	60	Blue	6
23	4	70	Cream	5
24	3	60	Cream	4
25	3	70	Green	3
26	3	70	Blue	2
27	3	70	Cream	1

TABLE 1.2 Shows different codes assumed for various levels of attributes for a regression run. The coding of the attribute levels for this purpose is known as 'effects coding'. In this table, which is similar to the coding of dummy variables, the three levels of life are coded as follows:

Life in years	Var 1	Var 2
3	1	0
4	0	1
5	– 1	– 1

Thus, the two variables, Var 1 and Var 2 are used to indicate the 3 levels of life, as per the coding scheme mentioned above.

Similarly the coding scheme for the three levels of the price is as shown as follows:

Price (Rs. Per litre)	*Var 3*	*Var 4*
50	1	0
60	0	1
70	– 1	– 1

Finally, the coding scheme for colour is as shown below:

Colour	*Var 5*	*Var 6*
Green	1	0
Blue	0	1
Cream	– 1	– 1

Thus, 6 variables, that is Var 1–Var 6 are used to represent the 3 levels of life of the paint (3, 4, 5), 3 levels of price per litre (50, 60 & 70) and 3 levels of colour (green, blue and cream). All the six variables are independent variables in the regression run. Var 7 is the rating of each combination given by the respondent, and forms the dependent variable for the regression curve. The recoded input data are shown in Table 1.3.

If the conjoint analysis is run as a regression model, the rating (which is the reverse of ranking) is used as a dependent variable. All combinations from the first to the twenty-seventh are ranked by the respondent. Rank 1 can be considered as the highest rating and given a rating of 27. Rank 2 can be given a rating of 26 and so on. This is not an interval-scaled rating, and should have only ordinal interpretation.

TABLE 1.3 Conjoint Problem Input Data Coded for Regression

var 1	*var 2*	*var 3*	*var 4*	*var 5*	*var 6*	*var 7*
– 1.00	– 1.00	1.00	0.00	1.00	0.00	27.00
0.00	1.00	1.00	0.00	1.00	0.00	26.00
– 1.00	– 1.00	1.00	0.00	– 1.00	– 1.00	25.00
– 1.00	– 1.00	1.00	0.00	0.00	1.00	24.00
– 1.00	– 1.00	0.00	1.00	1.00	0.00	23.00
0.00	1.00	0.00	1.00	1.00	0.00	22.00
– 1.00	– 1.00	– 1.00	– 1.00	1.00	0.00	21.00
– 1.00	– 1.00	0.00	1.00	0.00	1.00	20.00
– 1.00	– 1.00	0.00	1.00	– 1.00	– 1.00	19.00
0.00	1.00	1.00	0.00	0.00	1.00	18.00
0.00	1.00	1.00	0.00	– 1.00	– 1.00	17.00
– 1.00	– 1.00	– 1.00	– 1.00	0.00	1.00	16.00
1.00	0.00	1.00	0.00	1.00	0.00	15.00
– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	– 1.00	14.00

(Table 1.3 contd)

(Table 1.3 contd)

var 1	*var 2*	*var 3*	*var 4*	*var 5*	*var 6*	*var 7*
1.00	0.00	1.00	0.00	0.00	1.00	13.00
0.00	1.00	0.00	1.00	0.00	1.00	12.00
0.00	1.00	0.00	1.00	– 1.00	– 1.00	11.00
1.00	0.00	1.00	0.00	– 1.00	– 1.00	10.00
0.00	1.00	– 1.00	– 1.00	1.00	0.00	9.00
1.00	0.00	0.00	1.00	1.00	0.00	8.00
0.00	1.00	– 1.00	– 1.00	0.00	1.00	7.00
1.00	0.00	0.00	1.00	0.00	1.00	6.00
0.00	1.00	– 1.00	– 1.00	– 1.00	– 1.00	5.00
1.00	0.00	0.00	1.00	– 1.00	– 1.00	4.00
1.00	0.00	– 1.00	– 1.00	1.00	0.00	3.00
1.00	0.00	– 1.00	– 1.00	0.00	1.00	2.00
1.00	0.00	– 1.00	– 1.00	– 1.00	– 1.00	1.00

OUTPUT AND ITS INTERPRETATION

The output of the regression model is shown in Table 1.4. Variables 1 to 6 are treated as independent variables. The column titled 'B' (the regression coefficient column) provides the part utility of each level of attributes.

TABLE 1.4 Multiple regression output for conjoint problem (partial output shown)

Variables in the regression equation	
VARIABLE	*B*
VAR 1	– 7.11
VAR 2	0.11
VAR 3	5.44
VAR 4	– 0.11
VAR 5	3.11
VAR 6	– 0.88

For example, the **life of 3 years** is represented by variable 1 as per our coding scheme. Its utility is equal to – 7.11 (looking under column 'B' of Table 1.4 for variable 1). Similarly the utility for variable 2, representing **life of 4 years** is 0.11. The utility for the 3rd level of life, is not in the table, but is derived from the property of this coding, that all the utilities for a given attribute should sum to 0. Thus, utility for **life of 5 years** should be equal to 7 (– 7.11 + 0.11).

Similarly for price, the utilities of Rs. 50/litre and Rs. 70/litre are given by the numbers 5.44 and – 0.11, as shown against variables 3 and 4 in Table 1.4. but the utility for Rs. 80/ litre is derived from the same property, that the sum of the utilities for different levels of price should sum to 0. Therefore the price Rs. 80/1itre has the utility of 5.33 (5.44 + (– 0.11)).

Finally for colour, green has the utility of 3.11 and blue has the utility of – 0.88. Cream has a derived utility of 2.23 (3.11 + (– 0.88)).

TABLE 1.5 Utilities Table for Conjoint Analysis

Attributes	*Levels*	*Part Utility*	*Range of Utility (Max – Min)*
Life	3 years	– 7.11	= 7.00 – (– 7.11)
	4 years	0.11	= 14.11
	5 years	7.00	
Price	Rs. 50/litre.	5.44	
	Rs. 60/litre.	– 0.11	= 5.44 – (– 0.11)
	Rs. 70/litre.	5.33	= 5.55
Colour	Green	3.11	= 3.11 – (– 0.88)
	Blue	– 0.88	= 3.99
	Cream	2.23	

From the Table 1.5 we can conclude that the life or the number of years the paint lasts is the most important attribute for the customer. There are two indicators for this.

1. The range of utility value is highest (14.11) for the life. (From Range of Utility column)
2. The highest individual utility value of this attribute is at its 3rd level that is, i.e., 7.00.

Both these figures indicate that the number of years the paint lasts is the most important attribute at given levels of attributes. The price/litre seems to be the second most important attribute, as its range of utilities is 5.55. The last attribute in relative importance is the colour, with the utility range of 3.99.

Combination Utilities

The total utility of any combination can be calculated by picking the attribute levels of our choice. For example, the combined utility of the combination 4 years of life, Rs.70/litre, and cream colour is 0.11 + 5.33 + 2.23 = 7.67. If we want to know the best combination, it is advisable to pick the highest utilities from each attribute, and add them. The possible combination is 5 years of life, Rs. 50/litre, and green colour, that is, 7.00 + 5.44 + 3.11 = 15.55. The next best combination is 5 years of life, Rs. 70/litre, and green colour, with the combined utility of 7 + 5.33 + 3.11 = 15.44.

Individual Attributes

The difference in utility with the change of one level in one attribute can also be checked. For the life of 3 years to 4 years, there is increase in utility value of 7.22 units, but the next level, that is, 4 years to 5 years has an increase in utility of 6.89.

Similarly, increase in price from Rs. 50/litre to Rs. 60/litre induces a utility drop of 5.55, whereas from Rs. 60/litre to Rs. 70/litre there is an increase in utility of 5.44.

Finally, colour green to colour blue induces 3.99 units drop in utility. Next, from colour blue to colour cream there is an increase in utility of 3.11.

CASE STUDY 2

Conjoint Analysis*

Problem

The marketing manager of **Reynolds** wants to know how the customer values the various tangible and intangible features offered by its micro-tip pen. He identifies the attributes of his product which are important to customers, then the level for each attribute that the company is willing to design and offer to a customer. These are the following attributes of a micro-tip pen, which are considered to be important.

1. The price of the micro-tip pen.
2. The colour of ink in refill.
3. Diameter of tip of the refill.

The levels of these attributes are:

1. Price—Rs. 5, Rs. 7, and Rs. 10.
2. Colour of ink—blue, black, and red.
3. Diameter of tip—0.25 mm, 0.45 mm, and 0.5 mm

Since we have 3, 3, and 3 levels of the three attributes, we get a total of $3 \times 3 \times 3 = 27$, different combinations of attribute levels.

Ranking

The next stage of this input process was to collect from a respondent his ranking for all the 27 combination of attribute levels. This tabulation is given in the **input data** table. (Table 1.1)

Input

The coding of the attribute levels for the purpose of converting conjoint analysis into an equivalent model is known as 'effects coding'. For example, the price of the pen, Rs. 5, Rs. 7, and Rs. 10 is coded using two variables var 1 and var 2, as follows.

* Prepared by Biswajit Mishra, Imran Ahmed, Vikrant Sharma, and Sharmistha Kundu

Input Data for Conjoint Analysis

TABLE 1.1 Pen Refill Problem

S. No.	*Price (Rs.) 5, 7, 10*	*Refill (Ink Colour) 1 = Blue, 2 = Black, 3 = Red*	*Tip Diameter (mm) 0.25, 0.45, 0.5*	*Rating 27 to 1*
1	5	3	0.5	27
2	7	2	0.5	26
3	7	3	0.45	25
4	5	2	0.5	24
5	7	3	0.5	23
6	5	1	0.5	22
7	5	2	0.45	21
8	7	2	0.45	20
9	10	2	0.5	19
10	5	3	0.45	18
11	10	1	0.5	17
12	7	1	0.5	16
13	10	3	0.45	15
14	10	3	0.5	14
15	7	3	0.25	13
16	5	1	0.25	12
17	7	2	0.25	11
18	5	2	0.25	10
19	5	3	0.25	9
20	10	3	0.25	8
21	10	1	0.45	7
22	10	1	0.25	6
23	7	1	0.45	5
24	5	1	0.45	4
25	7	1	0.25	3
26	10	2	0.45	2
27	10	2	0.25	1

TABLE 1.2 Input Data Coded for Regression

	var1	*var2*	*var3*	*var4*	*var5*	*var6*	*var7*
1	1.00	.00	– 1.00	– 1.00	– 1.00	– 1.00	27.00
2	.00	1.00	.00	1.00	– 1.00	– 1.00	26.00
3	.00	1.00	– 1.00	– 1.00	.00	1.00	25.00
4	1.00	.00	.00	1.00	– 1.00	– 1.00	24.00
5	.00	1.00	– 1.00	– 1.00	– 1.00	– 1.00	23.00
6	1.00	.00	1.00	.00	– 1.00	– 1.00	22.00
7	1.00	.00	.00	1.00	.00	1.00	21.00
8	.00	1.00	.00	1.00	.00	1.00	20.00
9	– 1.00	– 1.00	.00	1.00	– 1.00	– 1.00	19.00
10	1.00	.00	– 1.00	– 1.00	.00	1.00	18.00
11	– 1.00	– 1.00	1.00	.00	– 1.00	– 1.00	17.00
12	.00	1.00	1.00	.00	– 1.00	– 1.00	16.00
13	– 1.00	– 1.00	– 1.00	– 1.00	.00	1.00	15.00
14	– 1.00	– 1.00	– 1.00	– 1.00	.00	1.00	14.00
15	.00	1.00	– 1.00	– 1.00	1.00	.00	13.00
16	1.00	.00	1.00	.00	1.00	.00	12.00
17	.00	1.00	.00	1.00	1.00	.00	11.00
18	1.00	.00	.00	1.00	1.00	.00	10.00
19	1.00	.00	– 1.00	– 1.00	1.00	.00	9.00
20	– 1.00	– 1.00	– 1.00	– 1.00	1.00	.00	8.00
21	– 1.00	– 1.00	1.00	.00	.00	1.00	7.00
22	– 1.00	– 1.00	1.00	.00	1.00	.00	6.00
23	.00	1.00	1.00	.00	.00	1.00	5.00
24	1.00	.00	1.00	.00	.00	1.00	4.00
25	.00	1.00	1.00	.00	1.00	.00	3.00
26	– 1.00	– 1.00	.00	1.00	.00	1.00	2.00
27	– 1.00	– 1.00	.00	1.00	1.00	.00	1.00

Price of the pen in Rs.	*Var 1*	*Var 2*
P5	1	0
P7	0	1
P10	– 1	– 1

Thus 2 variables, var 1 and var 2 are used to indicate three levels of price, as per the coding scheme above.

Similarly, the coding scheme for the levels of the attribute color of ink is shown below:

Color of ink of refill	*Var 3*	*Var 4*
Refill (Ink colour) 1 = Blue, 2 = Black, 3 = Red		
I 1	1	0
I 2	0	1
I 3	– 1	– 1

Finally, the coding scheme for diameter of the tip of the refill is as follows:

Diameters of tip	*Var 5*	*Var 6*
D0.25	1	0
D0.45	0	1
D0.50	– 1	– 1

Thus, six variables, var 1 to var 6 are used to represent the three levels of price of pen (P5, P7, & P10), three levels of color of ink (I 1, I 2, & I 3, where I 1 = blue, I 2 = black, & I 3 = red), and three levels of diameters of tip of the refill (D0.25, D0.45 and D0.50). All the six variables are independent variables in the regression run. Var 7 in the input table is the rating of each combination given by a respondent, and forms the dependent variables for the regression run.

Since the conjoint analysis is run as a regression model, the rating (reverse ranking) is used as a dependent variable. All combinations from the first to twenty-seventh were ranked by the respondent where rank 1 was considered as highest rating and given a rating of 27. Rank 2 was given a rating of 26 and so on. This is not an interval scale rating so it should have only ordinal interpretation. The coded input data used in regression are shown in **Table 1.2.**

ANALYSIS OF OUTPUT

The conjoint analysis was run on a regression model. The output shows that variables 1 to 6 are treated as independent variables and Variable 7 is a dependent variable. The column title 'B' (the regression coefficient column) provides the part utility of each level of attributes.

Multiple Regression

Listwise Deletion of Missing Data

Equation Number 1 Dependent Variable.. VAR7

Block Number 1. Method: Enter

VAR1 VAR2 VAR3 VAR4 VAR5 VAR6

Variables(s) Entered on Step Number

1. VAR6
2. VAR4
3. VAR2
4. VAR5
5. VAR1
6. VAR3

Multiple R	.86686
R Square	.75145
Adjusted R Square	.67688
Standard Error	4.51179

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	6	1230.87452	205.14575
Residual	20	407.12548	20.35627
F = 10.07777		Signif *F* = .0000	

Variables in the equation					
Variable	*B*	*SE B*	*Beta*	*T*	*Sig T*
VAR1	2.012452	1.230598	0.210962	1.635	.1176
VAR2	1.456897	1.230598	0.152724	1.184	.2503
VAR3	– 4.098659	1.230598	– .429656	– 3.331	.0033
VAR4	0.568008	1.230598	.059543	.462	.6494
VAR5	– 6.209770	1.230598	– .631931	– 5.046	.0001
VAR6	– 1.227011	1.214650	– .128095	– 1.010	.3245
(Constant)	14.320881	0.872030		16.422	.0000

End Block Number 1 All requested variables entered.

Price of pen, P5 (Rs. 5) is represented by variable 1 as per our coding scheme. Its utility (from column B) is equal to 2.012452. Similarly, the utility for variable 2 representing P7 (Rs. 7) is 1.456897. The utility for the third level of price (P 10), is not in the table, but is derived from the property of the coding, that all the utility for a given attribute should sum to 0. Thus utility for Pl0 (Rs. 10) should be equal to – (2.012452 + 1.456897) = – 3.469349, or – 3.47.

Similarly, for the colour of ink of the pen, the utilities of I 1 (blue) and I 2 (black) are given by the numbers – 4.098659 and 0.568008, against var 3 and var 4. The utility for I 3 (red) is derived from the same property that the sum of utilities for color of ink should sum to zero. Therefore I 3 has a utility of – (– 4.098659 + 0.568008) = 3.529651, or 3.53.

Finally, for diameter of tip of refill, D0.25 has a utility of – 6.209770 and D0.45 has a utility of – 1.227011. D0.5 has a derived utility of – (– 6.209770 – 1.227011) = 7.436781, or 7.44.

Now we have the utilities for all the levels of all attributes, and we can put them into a table, as follows:

UTILITIES TABLE FOR CONJOINT ANALYSIS

	Attributes	*Level*	*Part Utility*	*Range of Utilities*
1	Price in Rs.	P5	2.00	(Max – Min)
		P7	1.46	= 2 – (– 3.47)
		P10	– 3.47	= 5.47 for price
2	Color of ink	I 1	– 4.10	(Max – Min)
		I 2	0.56	= 3.53 – (– 4.1)
		I 3	3.53	= 7.63 for color of Ink
3	Diameter of tip	D0.25	– 6.21	(Max – Min)
		D0.45	– 1.22	= 7.44 – (– 6.21)
		D0.50	7.44	= 13.65 for diameter of tip.

Now with the part utilities of every level of attribute available to us, so we can come to several conclusions. First, we can conclude that sharpness of tip is the most important attribute for the customer. There are two indicators for this. One, the range of utility values is highest, that is, 13.65, for diameter of tip of the refill. Two, the highest individual value of utility for any level of any attribute is 7.44, for D0.50 (Diameter of tip = 0.50 mm). Both these figures indicate that sharpness of tip is the most important attribute at given level of attributes. The color of ink of the refill is the second most important attribute as its range of utilities is 7.63. The last attribute in relative importance is price of the pen, with a utility range of 5.47.

COMBINATION UTILITIES

We can also pick up one attribute level from each attribute and combine their part utilities to calculate the total utility of the combination. For example P5, I 1, & D0.25 have a combined utility of 2.00 – 4.10 – 6.21 = – 8.31. Similarly P5, I 2, and D0.25 have a combined utility of 2.00 + 0.56 – 6.21 = – 4.65.

If we want the best combination we pick the highest utilities from each attribute, and add them. In our case, P5 + 13 + D0.50 is the most preferred combination with a combined utility of 2.00 + 3.53 + 7.44 = 12.97. The next best combination is P7 + 13 + D0.50, with a combined utility of 1.46 + 3.53 + 7.44 = 12.43.

INDIVIDUAL ATTRIBUTES

We can also check what difference in utility a change of one level in one attribute makes. For example P5 to P7 (price change from Rs. 5 to Rs. 7) induces only a 0.54 units drop in utility, but it gets progressively more at the next stage, that is P7 to P10 has a difference in utility of 4.93.

Similarly, changing the colour of ink from black to blue, that is, I 2 to I 1 costs 4.66 units of utility drop, whereas change in I 3 to I 2 causes a drop in utility by 2.97.

Finally, the sharpness of tip causes a drastic change in utility of 5.99 units from D0.45 to D0.25 and a significant drop in the utility by 8.66 units from D0.50 to D0.45.

CASE STUDY 3

Conjoint Analysis*

Problem

We have taken the example of a consumer durable—a BPL colour television. We first identify the attributes of the product, which are important to customers, and then the levels for each attribute that we are willing to design and offer to a customer. Let us assume that three attributes of a colour television are important:

TABLE 1.1

Price of the Product	*Longevity in Years*	*Dimension of the Product*
Rs. 14000	2 years	14"
Rs. 19000	4 years	21"
Rs. 25000	5 years	25"

1. **Price of the product**—this is the price of the various sizes of the different models.
2. **Longevity in years**—this is the length of the time period for which the product is expected to last.
3. **Dimensions of the product**—this is the length of the television screen measured diagonally in inches.

The levels of the above three attributes are stated in Table 1.1

The objective of conjoint analysis is to determine from a potential customer of colour televisions, how important each attribute is to him. We also want to know, how much utility he derives from given combinations of these levels of attributes. Since we have 3, 3, and 3 levels of the 3 attributes, we get a total of $3 \times 3 \times 3 = 27$ different combinations of attribute levels.

Ranking

The next stage of the input process is to collect from a respondent his ranking for all the 27 combinations of attribute levels. This input data is shown in Table 1.2.

*Prepared by Harakuni Rajiv, Jitender Singh, Siddharth Agarwala, and Vidya Thekke Cherupilli

Input Data for Conjoint Analysis

TABLE 1.2

S. No.	*Price Rs.*	*Inches*	*Longevity Years*	*Rank*
1	14000	14	2	27
2	21000	14	2	26
3	25000	21	4	25
4	14000	14	4	24
5	25000	21	5	23
6	25000	21	2	22
7	14000	14	5	21
8	25000	29	4	20
9	14000	29	5	19
10	25000	14	4	18
11	25000	29	5	17
12	25000	29	2	16
13	14000	29	4	15
14	25000	14	5	14
15	25000	14	2	13
16	21000	29	5	12
17	14000	29	2	11
18	21000	29	4	10
19	21000	29	2	9
20	21000	21	2	8
21	14000	21	5	7
22	21000	14	5	6
23	14000	21	4	5
24	21000	21	4	4
25	21000	14	4	3
26	14000	21	2	2
27	21000	21	5	1

Running Conjoint as Regression Model

It is easy to convert the conjoint analysis input into an equivalent regression model and run it as a regression. The coding of the attribute levels for this purpose is known as 'Effect Coding' and Table 1.6 shows our television example coded for a regression run.

All the 6 variables as shown in Table 1.3 to 1.5 are independent variables in the regression run. Variable 7 is the rating of each combination given by our respondent and forms a **dependent** variable for the regression run.

If the conjoint analysis is run as a regression model, the rating, which is the reverse of ranking, is used as dependent variable. The respondent ranked all combinations from the first to the twenty-seventh. Rank 1 can be considered as the highest rating and given a rating of 27. Rank 2 can be given a rating of 26 and so on.

TABLE 1.3 Coding for Price Levels using 2-dummy Variables

Price of the product	*Var. 1*	*Var. 2*
Rs. 14000 (P14)	1	0
Rs. 21000 (P21)	0	1
Rs. 25000 (P25)	– 1	– 1

Similarly, the coding scheme for the three levels of the longevity of the product are shown below.

TABLE 1.4 Coding for Longevity using 2-Dummy Variables

Longevity	*Var. 3*	*Var. 4*
2 years (L2)	1	0
4 years (L4)	0	1
5 years (L5)	– 1	– 1

Finally, the coding scheme for the dimensions of the different modules are shown below.

TABLE 1.5 Coding for Dimensions using 2-dummy Variables

Dimension (")	*Var. 5*	*Var. 6*
14"	1	0
21"	0	1
29"	– 1	– 1

Conjoint Problem Input Data Coded for Regression

TABLE 1.6

S. No.	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6	VAR 7
1	1	0	1	0	1	0	27
2	0	1	1	0	1	0	26
3	– 1	– 1	0	1	0	1	25
4	1	0	1	0	0	1	24
5	– 1	– 1	0	1	– 1	– 1	23
6	– 1	– 1	0	1	1	0	22
7	1	0	1	0	– 1	– 1	21
8	– 1	– 1	– 1	– 1	0	1	20
9	1	0	– 1	– 1	– 1	– 1	19
10	– 1	– 1	1	0	0	1	18
11	– 1	– 1	– 1	– 1	– 1	– 1	17
12	– 1	– 1	– 1	– 1	1	0	16
13	1	0	– 1	– 1	0	1	15
14	– 1	– 1	1	0	– 1	– 1	14
15	– 1	– 1	1	0	1	0	13
16	0	1	– 1	– 1	– 1	– 1	12
17	1	0	– 1	– 1	1	0	11
18	0	1	– 1	– 1	0	1	10
19	0	1	– 1	– 1	1	0	9
20	0	1	0	1	1	0	8
21	1	0	0	1	– 1	– 1	7
22	0	1	1	0	– 1	– 1	6
23	1	0	0	1	0	1	5
24	0	1	0	1	0	1	4
25	0	1	1	0	0	1	3
26	1	0	0	1	1	0	2
27	0	1	0	1	– 1	– 1	1

Multiple Regression Output for Conjoint Problem

Equation Number 1 Dependent Variable: Variable 8

Method: Enter

Variables Entered on Step Number

1. Var00006
2. Var00004
3. Var00002
4. Var00005
5. Var00003
6. Var00001

Multiple R	0.617187
R Square	0.38177
Adjusted R Square	0.19630
Standard Error	7.11571

Analysis of Variance

	DF	*Sum of Squares*	*Mean Square*
Regression	6	625.333	104.222
Residual	20	1012.666	50.633

F = 2.05837 Signif *F* = 0.1047

TABLE 1.7 Variables in the Equation

Variables	*B*	*SE B*	*Beta*	*T*	*Sig. T*
VAR00001	0.555556	1.936651	0.058238	0.287	0.7772
VAR00002	– 5.22222	1.936651	– 0.547438	– 2.697	0.0139
VAR00003	2.888889	1.936651	0.302838	1.492	0.1514
VAR00004	– 3.22222	1.936651	– 0.337781	– 1.664	0.1117
VAR00005	0.888889	1.936651	0.093181	0.459	0.6512
VAR00006	– 0.22222	1.936651	– 0.023295	– 0.115	0.9098
(Constant)	**14**	**1.369419**	**0**	**10.223**	**0**

Analysis Interpretation

If run as a regression model, the output is shown in Table 1.7. Variables 1 to 6 are treated as independent variables and the column titled 'B', the regression coefficients column, provides a part utility of each level of attributes.

For example, price of P14000 is shown by variable 1 as per our coding scheme. Its utility is equal to 0.555 (looking under column B of Table 1.7 for variable 1). Similarly the utility for variable 2 representing P21000 is – 5.222. The utility for the third level of price, that is, P25000 is not in the table, but is derived from the property of this coding, that all the utilities for a given attribute should sum to zero. Thus, utility for P25000 should be equal to – (0.555 – 5.222) = 4.667.

Similarly, for longevity, the numbers 2.888 and – 3.222 against variables 4 and 5 in Table 1.7 give the utilities of L2 and L4. Therefore the utility for L5 is equal to – (2.888 – 3.222) = 0.334.

Finally, for dimensions, D14 and D21 have a utility of 0.888 and – 0.222 against variables 5 and 6 in Table 1.7. Therefore D29 has a derived utility of – (0.888 – 0.222) = – 0.666.

Now we have the utilities for all the levels of attributes, and we can put them into a table, as follows (rounded off to 1 decimal point).

Utilities Table for Conjoint Analysis

Attribute	*Level*	*Part Utility*	*Range of Utilities (Max. – Min.)*
Prices in Rupees	P14000	0.555	= 4.667 – (– 5.222)
	P21000	– 5.222	= 9.889
	P25000	4.667	
Longevity in Years	L2	2.888	= 2.888 – (– 3.222)
	L4	– 3.222	= 6.110
	L5	0.334	
Dimensions in Inches	D14	0.888	= 0.888 – (– 0.666)
	D21	– 0.222	= 1.554
	D29	– 0.666	

Now with the part utilities of every level of every attribute available to us, we can come to several conclusions. First, we can conclude that price of the colour television is the most important attribute for this customer because the range of utility values is highest (9.889). The durability of the product seems to be the second most important attribute (6.11).

Combination Utilities

We can also pick up one attribute level from each attribute and combine their part utilities to calculate the total utility of the combination, for example, P250000, L4, and D14 have a combined utility of (4.667 – 3.222 + 0.888) = 2.333. If we want the best combination, we pick the highest utilities from each attribute and add them. We observe that P25000 + L2 + Dl4 has the highest combination utility of 8.443.

Individual Attributes

We can also check what difference in utility a change of one level in one attribute makes. For example, P14000 to P21000 induces a 4.667 units drop in utility but it gets increased progressively at the next stage with a 9.889 units increase in utility.

Similar calculations can be made for the different levels of the other two attributes.

The McGraw·Hill Companies

CHAPTER 17 ATTRIBUTE-BASED PERCEPTUAL MAPPING USING DISCRIMINANT ANALYSIS

Learning Objectives

In this chapter, we will

- Introduce another method of drawing perceptual maps, apart from the one in Chapter 14 on multidimensional scaling
- Let the reader use discriminant analysis, already explained in Chapter 11, to draw these maps
- Show through a worked example how product attributes can also be used to draw perceptual maps, instead of paired comparisons between brands
- Discuss the pros and cons of drawing perceptual maps using attribute-based and non-attribute based approaches

APPLICATION AREAS

The application areas for this technique are the same as those discussed under the chapter on multidimensional scaling (MDS). For example, we have a set of six brands in a product category—for example, lipsticks. We would, as marketing managers or brand managers, be interested in finding out how customers perceive these brands. In particular, we would like to know how the brand we are marketing is perceived in relation to other competing brands. We have already learnt how to draw perceptual maps of the six lipstick brands through a paired comparison in the earlier chapter on MDS. But assume we have the ratings on 10 attributes for these six brands of lipstick. Then, we can use another method to draw perceptual maps. This is called the attribute-based approach.

METHODS

Actually, there are two ways to draw these attribute-based perceptual maps. One uses Factor Analysis, and the other uses Discriminant Analysis. Experts differ on which method produces better maps. We will discuss the discriminant analysis based approach in this chapter. As usual, we will do this through a worked-out example, which will clarify the principles, and will be easy to interpret, if you are familiar with the earlier chapters on MDS and discriminant analysis.

In the chapter on discriminant analysis, we had a situation where only two groups were formed (High Risk and Low Risk consumers). In this application, we have as many groups as there are *Brands*. Thus, in this case, we are likely to get more than one discriminant function. For example, with three Brands, you could get upto two functions, and with four brands, you may get upto three discriminant functions. But all these discriminant functions are not necessarily useful. For most applications, we can use the first two or three functions. The discriminant analysis output gives us the Eigen Value of each function and the amount of variance it explains from the original data. The percentage of variance explained by discriminant functions can be used to decide how many discriminant functions we want to use. There is also a significance test available that can tell us if the given function significantly discriminates between the groups (brands). This can also be used to decide how many of the functions to use.

Suppose we decide to use two discriminant functions. These then form the two axes of the perceptual map. If three functions are used, then we will get three perceptual maps, Function 1 vs. Function 2, Function 2 vs. Function 3 and Function 1 vs. Function 3.

The important thing to remember, is that the discriminant functions represent the axes on which the *Brands* are first located, and then, the *Attributes* are also located.

There is, however, one additional job for the reader to do in using this technique. The last perceptual map (or maps) has to be drawn separately using the discriminant analysis output data from SPSS. This can be done using Excel or any other graphics-capable package.

WORKED EXAMPLE

Problem

A chocolate company wants to draw a perceptual map using an attribute-based procedure, of its consumers' perceptions regarding its own brand and two competing brands. Assume that it is Nestle against Cadbury's and Amul. Data was collected from 15 respondents (five consumers of each brand), on five attributes, namely Price, Quality, Availability, Packaging and Taste. The variables are measured using different scales, but a higher value indicates a favourable rating in each variable's measurement.

Please note the data arrangement in this problem. If five respondents had rated all the three brands (which means a sample size of five), the data would still be entered as shown below. This is because the *Brand* is a categorical grouping variable while doing discriminant analysis here.

The input data are shown in Table 17.1. Tables 17.2 to 17.10 show the output of the discriminant analysis from SPSS using the input data.

TABLE 17.1 Input Data for the Problem

Brand	*Price*	*Quality*	*Availability*	*Packaging*	*Taste*
1	12	34	500	5	18
1	11	35	234	4	15
1	10	36	250	4	14
1	13	22	345	5	12
1	12	23	432	3	13
2	10	14	234	2	15
2	11	17	231	3	11
2	15	23	45	4	10
2	13	14	35	3	12
2	12	15	25	2	10
3	10	22	75	4	8
3	12	24	80	4	7
3	13	28	90	5	10
3	11	17	96	2	12
3	11	18	59	2	6

Discriminant Analysis Output

TABLE 17.2 Group Statistics

Brand		*Mean*	*Std. Deviation*	*Valid N (listwise)*	
				Unweighted	*Weighted*
1	PRICE	11.6000	1.14018	5	5.000
	QUALITY	30.0000	6.892002	5	5.000
	PACKAGING	4.2000	.83666	5	5.000
	TASTE	14.4000	2.30217	5	5.000
	AVAILABILITY	352.2000	114.76149	5	5.000
2	PRICE	12.2000	1.92354	5	5.000
	QUALITY	16.6000	3.78153	5	5.000
	PACKAGING	2.8000	.83666	5	5.000
	TASTE	11.6000	2.07364	5	5.000
	AVAILABILITY	114.0000	108.41125	5	5.000
3	PRICE	11.4000	1.14018	5	5.000
	QUALITY	21.8000	4.49444	5	5.000
	PACKAGING	3.4000	1.34164	5	5.000
	TASTE	8.6000	2.40832	5	5.000
	AVAILABILITY	80.0000	14.33527	5	5.000

(Table 17.2 Contd.)

(Table 17.2 Contd.)

Total	PRICE	11.7333	1.38701	15	15.000
	QUALITY	22.8000	7.48522	15	15.000
	PACKAGING	3.4667	1.12546	15	15.000
	TASTE	11.5333	3.22628	15	15.000
	AVAILABILITY	182.0667	151.30266	15	15.000

TABLE 17.3 Tests of Equality of Group Means

	Wilks' Lambda	*F*	*df1*	*df2*	*Sig.*
PRICE	.936	.413	2	12	.671
QUALITY	.418	8.349	2	12	.005
PACKAGING	.722	2.313	2	12	.006
AVAILABILITY	.314	13.131	2	12	.001

Summary of Canonical Discriminant Functions

TABLE 17.4 Eigen values

Function	*Eigen value*	*Percentage of Variance*	*Cumulative Percentage*	*Canonical Correlation*
1	4.749[a]	81.4	81.4	.909
2	1.083[a]	18.6	100.0	.721

[a]First 2 canonical discriminant functions were used in the analysis.

TABLE 17.5 Wilks' Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
1 through 2	.084	24.827	10	.006
2	.480	7.336	4	.119

TABLE 17.6 Standardised Canonical Discriminant Function Coefficients

	Function	
	1	*2*
PRICE	.207	.701
QUALITY	.988	–.454
PACKAGING	–.398	–.293
TASTE	–.136	.986
AVAILABILITY	.999	–.122

TABLE 17.7 Structure Matrix

	Function	
	1	2
AVAILABILITY	.664*	.294
QUALITY	.517*	–.336
PACKAGING	.268*	–.203
TASTE	.431	.668*
PRICE	–.044	.235*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions variables ordered by absolute size of correlation within function.

*Largest absolute correlation between each variable and any discriminant function

Unstandardised Coefficients

TABLE 17.8 Canonical Discriminant Function Coefficients

	Function	
	1	2
PRICE	.143	.483
QUALITY	–189	–.087
PACKAGING	–.385	–.284
TASTE	–.060	.435
AVAILABILITY	.011	–.001
(Constant)	–5.948	–7.486

Unstandardised coefficients

TABLE 17.9 Functions at Group Centroids

	Function	
BRAND	1	2
1	2.745	.123
2	–1.596	1.073
3	–1.149	–1.196

Unstandardised canonical discriminant functions evaluated at group means

TABLE 17.10 Classification Results[a]

			Predicted Group Membership			
		BRAND	1	2	3	Total
Original	Count	1	5	0	0	5
		2	0	4	1	5
		3	0	1	4	5
	Percentage 1	100.0	.0	.0	100.0	
		2	.0	80.0	20.0	100.0
		3	.0	20.0	80.	100.0

[a]86.7% of original grouped cases correctly classified

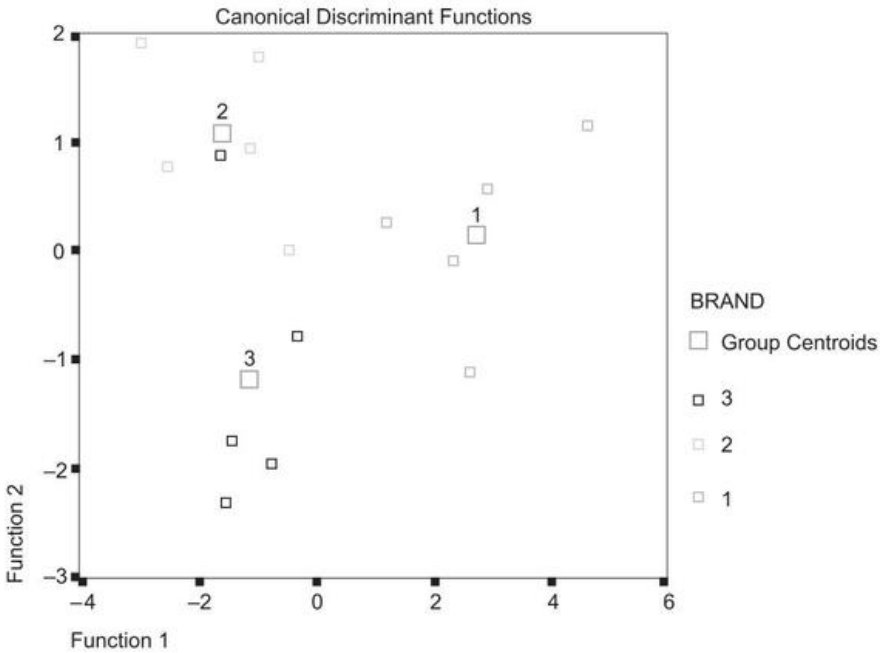


FIGURE 17.1 Plot of the three Brands on Discriminant Functions 1 and 2: Treating Functions as Dimensions

Putting Variables/Attribute Vectors on the Above Map

Figure 17.1 is a part of the SPSS discriminant analysis output. Vectors that represent the original attributes can be located on the above map. If there are more than three brands, we may get more than two dimensions, and may have to draw more than one plot of the above type.

To plot the attributes on the map above, we can use the standardised coefficients of the original variables in the discriminant function. For example, for Taste, the standardised coefficients from Table 17.6 are –.136 and .986 on Dimensions 1 and 2 respectively. So we can locate this point (–.136, .986) on the map, and draw an arrow from the origin to that point. This will be labelled the Taste vector (see Fig. 17.2), and similarly, all other vectors can be located, one for each of the five attributes—Price, Quality, Availability, Packaging and Taste. The length of the arrow represents its effect in discriminating on each dimension. Longer arrows pointing more closely towards a given group centroid (Brand on the map) represent variables most strongly associated with the group (or Brand, in this case). Vectors pointing in the opposite direction from a given group centroid represent lower association with a group (Brand).

Variables with longer vectors in a given dimension, and those closest to a given axis (dimension represented by the discriminant function) are contributing more to the interpretation of that dimension. Looking at all variables that contribute to a given axis (dimension), we can label the dimension as a combination of those variables.

In this case, the interpretation in terms of the variables and their correlation to dimensions 1 and 2 can be found from Fig. 17.2. *This has been drawn separately on Excel using the data from Table 17.6 (standardised discriminant function coefficients of each attribute on each function) and Table 17.9 (centroids for each brand or group, also plotted as Fig. 17.1). The two discriminant functions represent the two axes or dimensions in Fig.17.2.*

As seen from the perceptual map in Fig. 17.2, Nestle, Cadbury's and Amul, the three brands, have their unique positions on the map. In addition, on the same map, we have now plotted values of the attributes on the same two dimensions (each discriminant function represents a dimension). As we can see, Dimension 1 seems to comprise Availability and Quality. This is also evident from the standardised discriminant coefficients for Availability (.999) and Quality (.988) on Dimension 1, from Table 17.6.

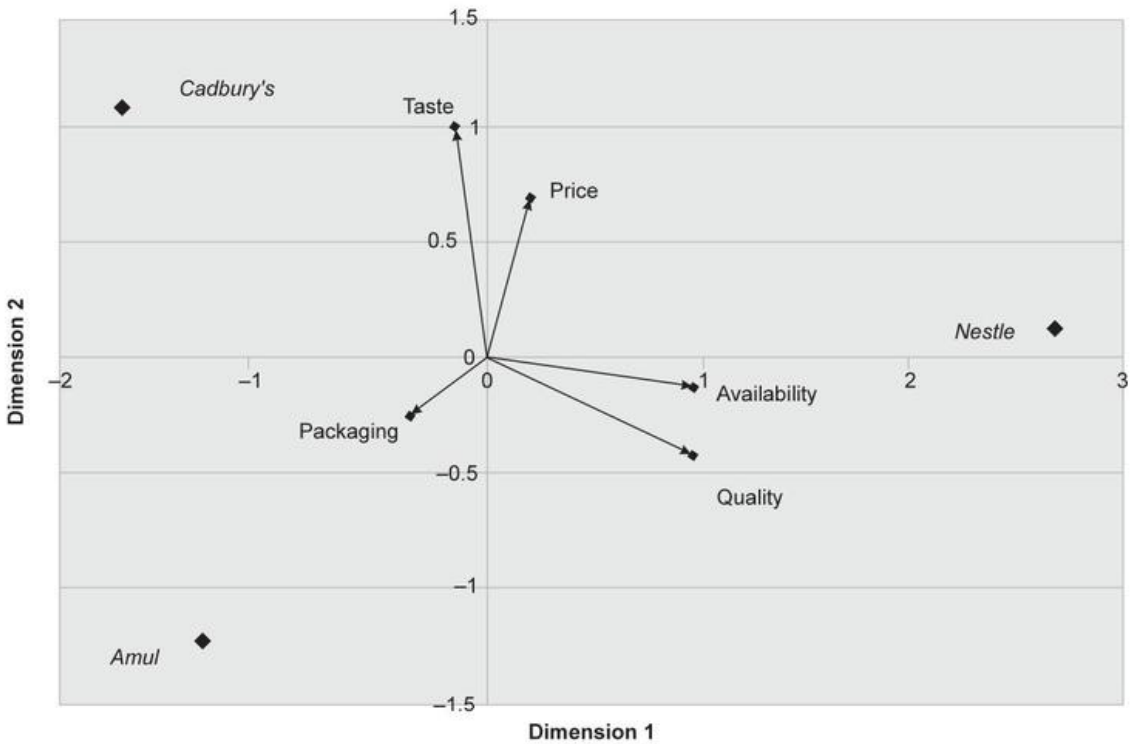


FIGURE 17.2 Perceptual Map of Chocolate Brands and Attributes: Drawn using Excel, from SPSS Output

Dimension 2 seems to comprise Taste and Price, the two vectors (arrows) that are closest to the vertical axis. This is also evident from the standardised coefficients, of .986 and .701 respectively, for Taste and Price on Dimension 2, from the earlier output table.

Packaging is not useful in defining any of the two dimensions, as its arrow is not close to either of the two dimensions.

Brands and their Association with Attributes/Dimensions

Nestle seems to be stronger on Dimension 1 (Availability and Quality), and Cadbury's on Dimension 2 (Taste and to a lesser extent, Price). Amul scores low on both dimensions compared to its competitors.

Attribute vs. Non-attribute based Perceptual Maps: Which are Better?

There are people who are more comfortable with well-defined attributes and easily quantifiable measurements. For them, the author recommends attribute-based mapping techniques similar to those elaborated on in this chapter. The major advantage of this approach is that attributes are well-defined, and do not require interpretation in the final perceptual map. In other words, naming the axes (dimensions) on the perceptual map is easier.

There are some people who are quite comfortable with creative (and many times, equally effective) interpretation from outputs that do not tell you the whole story, unless you fill in some of the colours. For them, the author recommends the non-attribute based approach covered earlier in the chapter on MDS. Here, the dimensions are not pre-defined, and the researcher's knowledge of the brands and product category is required to name the axes (dimensions) on the perceptual map is easier.

Given that both techniques use the computer to produce outputs, they are equally easy to use. But in the author's view, the data in MDS using paired comparisons are intuitively easy for a respondent to give, while attribute-based ratings sometimes get tedious and artificial, particularly when many attributes have to be rated for many brands. Purely from a data collection point of view, the non-attribute based method is simpler, as it involves only overall comparisons between brands, while in the attribute-based methods, measures have to be taken for each attribute for each brand. The final result in terms of the perceptual map is somewhat subjective in MDS and more 'objective' in the case of discriminant analysis.

However, one should not be swayed too much by the labels 'subjective' and 'objective'. Done properly, either of them could prove to be effective. Thus, the choice of method is best left to the researcher and his comfort level and skills.

REFERENCES

1. Steven Struhl, "Multivariate and Perceptual Mapping with Discriminant Analysis", *Quirk's Marketing Research Review*, March 1993 (from www.quirks.com).
2. R. Kenneth Teas and Terry Grapentine, "Is Your Message Getting Across?", *Marketing Research*, Spring 2002, pp. 35–39.

SUMMARY

We have already covered the discriminant analysis technique in an earlier chapter. In this chapter, we extend the same technique, using Brands as groups, to draw perceptual maps to position the brands in space. We also place attribute vectors (of those attributes on which brands are rated) on the same map. This results in a joint space map containing perceived positions of brands and attributes.

This perceptual map is convenient to use, to figure out which attributes are part of which dimensions (axes), and which brands score well on these dimensions/attributes. Arrows representing attribute vectors can be drawn using the standardised discriminant scores on the discriminant functions which serve as the axes in the plot. The plot of attributes is not generated by the SPSS package, but can be easily drawn from the discriminant analysis output, as explained through a worked-out example, and illustrated further through the case studies at the end of this chapter.

Opinion is divided over the use of attribute-based and non-attribute based approaches (such as MDS) to draw perceptual maps for positioning applications. Each approach has its merits, but the decision finally rests on the researcher's comfort level with either method.

ASSIGNMENT QUESTIONS

1. Explain the difference between the attribute-based and non-attribute based approaches to drawing perceptual maps. Which would you prefer, and why?
2. What is the data arrangement for an attribute-based approach using discriminant analysis, if you had to input data into SPSS, for a sample size 200, where all respondents have rated all five brands on six attributes? Explain how many rows, columns of data you would need, and describe these.
3. How do you draw perceptual maps with Attribute vectors and Brands? Does the SPSS package discriminant analysis output contain these?
4. If you get multiple discriminant functions in the output, how would you decide which ones are important to use for analysis?
5. Apart from discriminant analysis, which other technique can be used for attribute-based perceptual mapping?
6. How do you interpret dimensions (axes) in terms of attributes in the attribute-based perceptual maps?
7. If a vector arrow representing an attribute is longer, does it mean the attribute is more important?
8. If a vector representing an attribute is at 45 degrees to the two axes, what does it indicate?

SPSS COMMANDS FOR ATTRIBUTE-BASED PERCEPTUAL MAPPING USING DISCRIMINANT ANALYSIS

1. The commands are exactly the same as given at the end of Chapter 11 (Discriminant Analysis). Please make sure you get the output for discriminant analysis including standardised coefficients, because these will be used in getting the perceptual maps.
2. The *Grouping Variable* is in this case, the *Brand*. If you have three brands, they should be coded 1, 2, 3. The range of values for the Grouping Variable should be defined by clicking on Define Range just below the *Grouping Variable* box.

3. Please note that if 15 respondents have rated all the brands in your survey on six attributes, your sample size is 15. But when you have to do an attribute-based discriminant analysis, the data should be entered in 15x3 = 45 rows, with 15 rows for each brand. The Brand variable should be entered as a nominal scale variable in the first column or last column. The other six columns represent attributes as usual. Refer to the worked-out example in this chapter for further clarification on the input data format.
4. You can also ask for a *Combined Group Plot* on the *Classify* option on the main dialogue box. This plot gives you the position of the brands on the plot of two dimensions (discriminant functions). Attributes are not on this plot.
5. For getting a perceptual map of both *brands* and *attributes* on the same graph, you have to combine outputs from the Group Centroid Table and Standardised Coefficients Table. These data can be used to draw a perceptual map using Excel or any other graphing package.

CASE STUDY 1

Attribute-based Perceptual Mapping Using Discriminant Analysis*

Problem

A consumer durables firm is interested in evaluating the consumer perceptions about television sets of three different brands—Samsung, Thomson and LG. The consumers were interviewed on eight different parameters and were asked to rank them on an 8-point scale.

The data points that were obtained in the survey were as follows:

Price	*Remote*	*Colour*	*Screen*	*Audio*	*Internet*	*Fuzzy Logic*	*Weight*	*Brand*
4	4	4	5	6	3	3	4	1
4	5	3	3	4	5	3	5	1
3	6	4	3	5	5	4	4	1
3	7	2	4	4	4	3	5	1
3	6	3	4	3	6	5	5	1
4	6	2	4	4	2	6	4	1
4	6	3	4	5	4	6	4	1
3	6	2	5	4	5	5	5	1
5	7	4	2	3	3	4	4	1
4	6	4	6	4	2	1	4	1
6	3	4	6	5	4	5	4	1
7	4	4	4	4	4	4	2	1
6	4	4	5	3	3	5	2	1
5	4	4	5	4	4	4	2	1
6	4	5	5	5	4	4	4	1

* Prepared by Rijesh Panicker, Krishna Y Bandaru, Sumit Vats, Sourav Sanyal, Hare Krishna, and Shivam Puri

(Table Contd.)

6	5	2	3	4	3	4	5	1
6	4	3	4	4	3	5	6	1
6	3	4	3	6	4	5	3	1
7	4	5	4	4	3	5	1	1
6	4	3	4	4	3	5	1	1
4	5	2	6	5	3	5	5	2
4	4	5	6	3	4	4	4	2
3	4	5	5	4	4	4	4	2
4	5	4	7	5	4	5	4	2
5	4	4	7	5	3	5	4	2
4	5	5	6	5	5	4	5	2
3	4	4	6	4	4	5	5	2
4	4	3	5	6	4	4	4	2
3	4	5	7	6	3	5	5	2
3	5	4	6	5	4	5	4	2
5	5	6	2	3	5	3	6	2
4	6	7	3	2	6	4	5	2
5	5	6	2	4	6	4	6	2
3	4	7	2	2	5	3	7	2
2	5	6	1	2	6	4	6	2
4	4	6	2	3	5	3	5	2
2	5	6	2	2	6	4	6	2
3	6	7	1	3	5	4	5	2
2	4	6	2	3	6	4	6	2
1	5	7	2	4	5	2	6	2
3	4	6	1	4	5	4	6	3
2	4	7	1	3	4	3	5	3
3	5	6	2	4	4	3	7	3
4	4	7	1	4	3	4	6	3
3	7	6	2	4	4	3	7	3
4	6	7	3	4	4	4	7	3
5	5	6	2	3	3	3	6	3
3	6	6	1	4	4	4	7	3
3	5	7	2	4	3	2	6	3

(Table Contd.)

4	6	6	3	4	4	4	6	3
3	5	6	2	3	5	3	6	3
4	6	7	1	4	4	3	5	3
4	5	6	2	3	5	4	6	3
2	6	5	2	4	4	3	6	3
3	5	5	3	3	5	3	7	3
3	4	6	2	4	5	3	6	3
3	5	6	1	3	4	4	7	3
4	4	7	1	4	5	4	6	3
3	6	6	2	3	5	3	7	3
4	4	7	3	4	4	4	7	3

Methodology

SPSS procedure 'discriminant' analysis was used to produce the tables and Fig. 1.1. Fig. 1.1 was obtained by choosing the 'Combined Group Plot' option on the *classify* dialogue box while doing Discriminant Analysis.

As the output tables below show, two discriminant functions were obtained by using BRAND as the grouping variable. Both functions were used to map the group centroids on a 2 dimensional plot (Fig. 1.1).

Figure 1.2 was obtained by plotting the standardised discriminant function coefficients for function 1 and function 2 for each of the eight brand attributes on top of Fig. 1.1. Thus, a combination plot of Fig. 1.1 and the standardised coefficients resulted in Fig. 1.2.

Tests of Equality of Group Means

Attribute	*Wilks' Lambda*	*F*	*Df1*	*Df2*	*Sig.*
PRICE	.690	12.790	2	57	.000
REMOTE	.964	1.070	2	57	.350
COLOUR	.437	36.753	2	57	.000
SCREEN	.657	14.852	2	57	.000
AUDIO	.930	2.158	2	57	.125
INTERNET	.847	5.154	2	57	.009
FUZZY LOGIC	.842	5.355	2	57	.007
WEIGHT	.479	31.035	2	57	.000

Eigen values

Function	*Eigen value*	*% of Variance*	*Cumulative %*	*Canonical Correlation*
1	3.120	80.3	80.3	.870
2	.765	19.7	100.0	.658

The first two canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*Df*	*Sig.*
1 through 2	.138	106.149	16	.000
2	.567	30.402	7	.000

Standardised Canonical Discriminant Function Coefficients

	Function	
	1	*2*
PRICE	–.402	–.436
REMOTE	–.103	–.228
COLOUR	.964	.304
SCREEN	.146	1.102
AUDIO	.113	–.350
INTERNET	–.353	.642
FUZZY LOGIC	.244	.185
WEIGHT	.709	–.164

Structure Matrix

	Function	
	1	*2*
COLOUR	.643	–.005
WEIGHT	.586	–.146
PRICE	–.359	–.247
FUZZY LOGIC	–.224	.201
AUDIO	–.155	–.036
SCREEN	–.330	.488
INTERNET	.159	.365
REMOTE	.032	–.212

Pooled within-groups correlations between discriminating variables and standardised canonical discriminant functions variables ordered by absolute size of correlation within function.

Canonical Discriminant Function Coefficients

	Function	
	1	*2*
PRICE	–.365	–.396
REMOTE	–.105	–.234
COLOUR	.921	.291
SCREEN	.098	.738
AUDIO	.119	–.368
INTERNET	–.377	.686
FUZZY LOGIC	.272	.206
WEIGHT	.679	–.157
(Constant)	–6.357	–2.677

Unstandardised coefficient

Functions at Group Centroids

BRAND	*Function*	
	1	*2*
1	–2.282	–.420
2	.406	1.189
3	1.876	–.768

Unstandardised canonical discriminant functions evaluated at group means

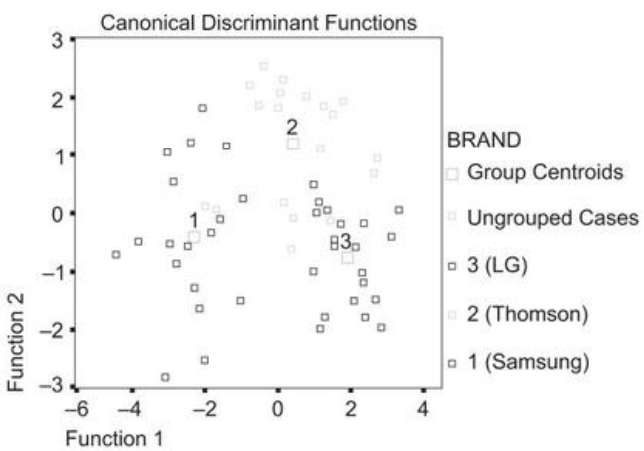


FIGURE 1.1 Plot of the brands on discriminant functions 1 and 2

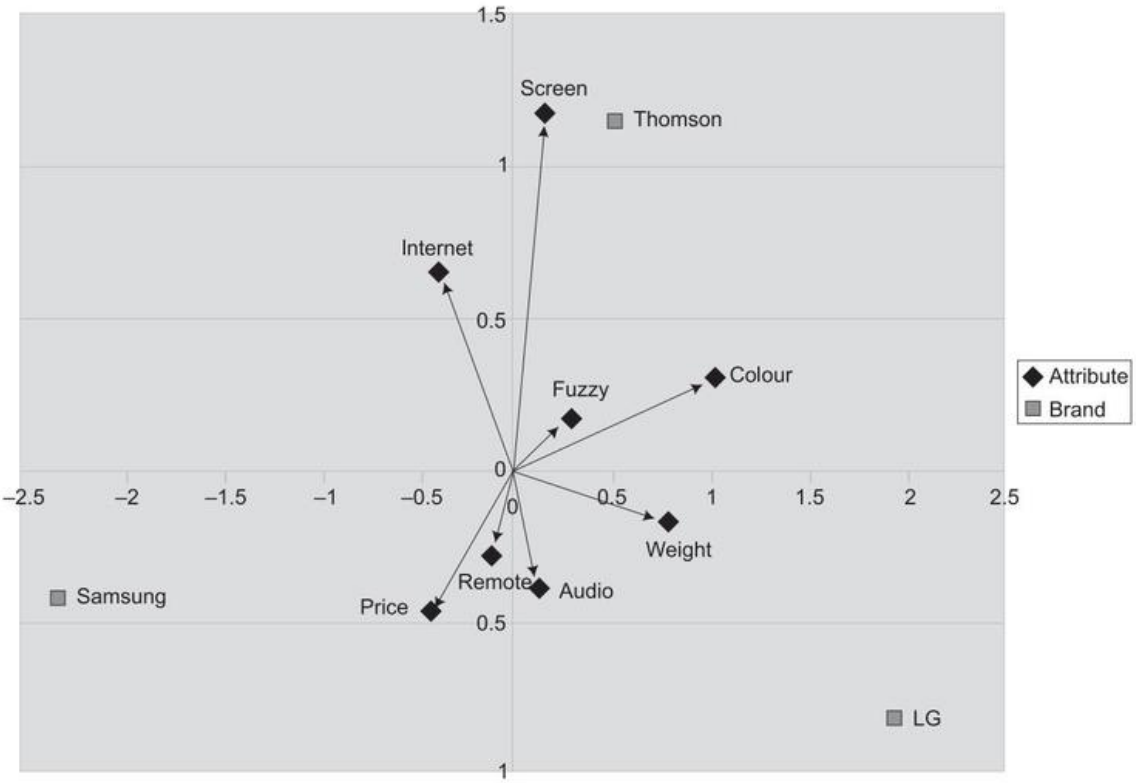


FIGURE 1.2 Perceptual Map of Television brands and attributes

Interpretation

Figure 1.2 represents the vectors for each of the attributes—price, remote, audio, weight, colour, screen, internet, fuzzy logic. These vectors represent the effect of discriminating on each dimension. Longer arrows pointing more closely towards a given group centroid, represent variables most strongly associated with that particular group (or Brand, in this case). Vectors pointing in the opposite direction from a given group centroid represent lower association with the concerned group.

Variables with longer vectors in a given dimension, and those closer to a given axis (dimension represented by the discriminant function) are contributing more to the interpretation of that dimension. Looking at all variables that contribute to a given axis (dimension), we can label the dimension as a combination of those variables.

As seen from the graph (Fig. 1.2), Samsung, Thomson and LG have their unique positions on the map. In addition, on the same map, we have now plotted values of the attributes on the same two dimensions (each discriminant function represents a dimension). As we can see, Dimension 1 seems to be a combination of Colour and Weight (closest to the x-axis). This is also evident from the standardised discriminant coefficients for Colour (.964 and .709 respectively) on Dimension 1, from the output table labelled 'Standardised Canonical Discriminant and Function Coefficients.'

Dimension 2 seems to comprise Screen (size), the vector that is closest to the vertical axis. This is also evident from its standardised coefficient on Dimension 2, from the same table.

Brands and their Associations

While Thomson seems to be stronger in Dimension 2 (Screen size), LG seems to be strong in Dimension 1 (Weight and Colour). Samsung seems to be weak in both the dimensions as compared to its competitors.

CASE STUDY 3

Attribute-based Perceptual Mapping Using Discriminant Analysis*

Problem

The objective of this exercise is to draw a perceptual map using an attribute-based procedure, of consumers' perceptions regarding three brands of lipsticks. The three brands are Revlon, Maybelline and Lakme. Data was collected from 15 respondents (five of each brand), on five attributes, namely colour, price, long lasting nature, packaging and finish. The variables are measured using a five-point rating scale, with a higher value indicating a favourable variable.

Input Data

S. No.	Colour	Price	Long lasting	Packaging	Finish	Brand
1	4	2	4	2	2	1
2	3	3	3	2	3	1
3	3	3	2	1	1	1
4	1	2	3	2	3	1
5	3	2	4	1	1	1
6	2	4	2	4	2	2
7	5	3	2	5	3	2
8	1	5	3	3	2	2
9	2	4	4	4	3	2
10	1	5	3	3	3	2
11	2	2	2	5	4	3
12	1	1	1	3	3	3
13	2	2	3	2	5	3
14	1	4	2	1	4	3
15	2	3	1	3	5	3

* Prepared by Nittu Thomas, Har Amrit Pal Singh, Rajat Shrimal, Sachin Khandelwal, and Tejaswini Tilak

Output

TABLE 1.1 Group Statistics

Brand		Valid N (listwise) Unweighted	Valid N (listwise) Weighted
1.00	Colour	5	5.000
	Price	5	5.000
	Long lasting nature	5	5.000
	Packaging	5	5.000
	Finish	5	5.000
2.00	Colour	5	5.000
	Price	5	5.000
	Long lasting nature	5	5.000
	Packaging	5	5.000
	Finish	5	5.000
3.00	Colour	5	5.000
	Price	5	5.000
	Long lasting nature	5	5.000
	Packaging	5	5.000
	Finish	5	5.000
Total	Colour	15	15.000
	Price	15	15.000
	Long lasting nature	15	15.000
	Packaging	15	15.000
	Finish	15	15.000

Univariate F tests

TABLE 1.2 Tests of Equality of Group Means

	Wilks' Lambda	*F*	*df1*	*df2*	*Sig.*
Colour	0.8235	1.2857	2.0000	12.0000	0.2119
Price	0.4600	7.0435	2.0000	12.0000	0.0095
Long lasting nature	0.6167	3.7143	2.0000	12.0000	0.0555
Packaging	0.5134	5.6875	2.0000	12.0000	0.0183
Finish	0.3822	9.7000	2.0000	12.0000	0.0031

Here the low value of Wilks' lambda indicates higher group differences. Similarly, a low value of degree of significance also indicates higher group differences. However, in this case, the only low values of Wilks' Lambda are for price and finish. But looking at the last column, all attributes except colour seem to be significantly discriminating between the brands.

TABLE 1.3 Eigen values

Function	*Eigen value*	*% of Variance*	*Cumulative %*	*Canonical Correlation*
1	7.09132	70.06355	70.06355	0.93617
2	3.02995	29.93645	100.00000	0.86710

The eigen value is the ratio of the between-groups sum of squares to the within-groups sum of squares. The largest eigen value corresponds to the eigen vector in the direction of the maximum spread of the groups' means. The second largest eigen value corresponds to the eigen vector in the direction that has the next largest spread, and so on. The square root of each eigen value provides an indication of the length of the corresponding eigen vector. Small eigen values result in eigen vectors of essentially no length and account for very little of the total dispersion. The 'percentage of variance' column allows you to evaluate which canonical variable accounts for most of the spread. Here, the first eigen value is able to explain 70% of the variance.

TABLE 1.4 Wilks' Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
1 through 2	0.03067	34.84545	10.00000	0.00013
2	0.24814	13.93754	4.00000	0.00750

This table is used to identify the function which is significant in explaining the differences among the groups. Wilks' Lambda is the proportion of the total variance in the discriminant scores not explained by differences among the groups. Wilks' lambda ranges between 0 and 1. Values close to 0 indicate the group means are different. Values close to 1 indicate that the group means are not different (equal to 1 indicates all means are the same). Here since both the Wilks' Lambda values are close to zero, they are able to explain the differences in the groups. Thus, both the functions are significant.

TABLE 1.5 Standardised Canonical Discriminant Function Coefficients

	Function	
	1	*2*
Colour	0.1210930	–0.4920410
Price	1.2110203	0.2526630
Long lasting nature	0.5012788	–0.4711534
Packaging	1.1494576	0.5798965
Finish	0.7010088	0.6389589

When variables are measured in different units, the magnitude of an unstandardised coefficient provides little indication of the relative contribution of the variable to the overall discriminant function. Standardising the coefficients allows one to examine the relative standing of the measurements. The higher value of the coefficients for a particular attribute on a function indicates the higher loading of the same on that function.

Variable Loading on Functions

TABLE 1.6 Structure Matrix

	Function	
	1	*2*
Price	0.3929*	0.1616
Finish	–0.2370	0.6342*
Packaging	0.2444	0.4160*
Long lasting nature	0.1430	– 0.3955*
Colour	0.0451	– 0.2568*

The structure matrix contains within-group correlations of each predictor variable with the canonical function. This matrix provides another way to study the usefulness of each variable in the discriminant function. For each variable, an asterisk marks its largest absolute correlation with one of the canonical functions. With each function, these marked variables are then ordered by the size of the correlation. In this table, the largest correlation of price exists with function 1 whereas for all other attributes, it exists with function 2.

Centroids of Brands on Functions

TABLE 1.7 Functions at Group Centroids

	Function	
BRAND	*1*	*2*
Revlon	– 0.86899	– 2.12727
Maybelline	3.25287	0.57171
Lakme	– 2.38388	1.55556

This table displays the canonical variable (discriminant function) means by group. Within-group means are computed for each canonical variable. In this example for the first canonical variable, the average discriminant or canonical variable score for Revlon is – 0.86899. This table is used to plot the brands on the attribute plot.

Interpretation

As seen from the graph (Fig 1.1), Revlon, Maybelline and Lakme, the three brands, have their unique positions on the map. In addition, on the same map, we have now plotted values of the attributes on the same two dimensions (each discriminant function represents a dimension). As we can see, Dimension 1 seems to be a combination of Price and Packaging (closest to the x-axis). This is also evident from the standardised discriminant coefficients for Price (1.211) and Packaging (1.149) on Dimension 1, from the earlier output table.

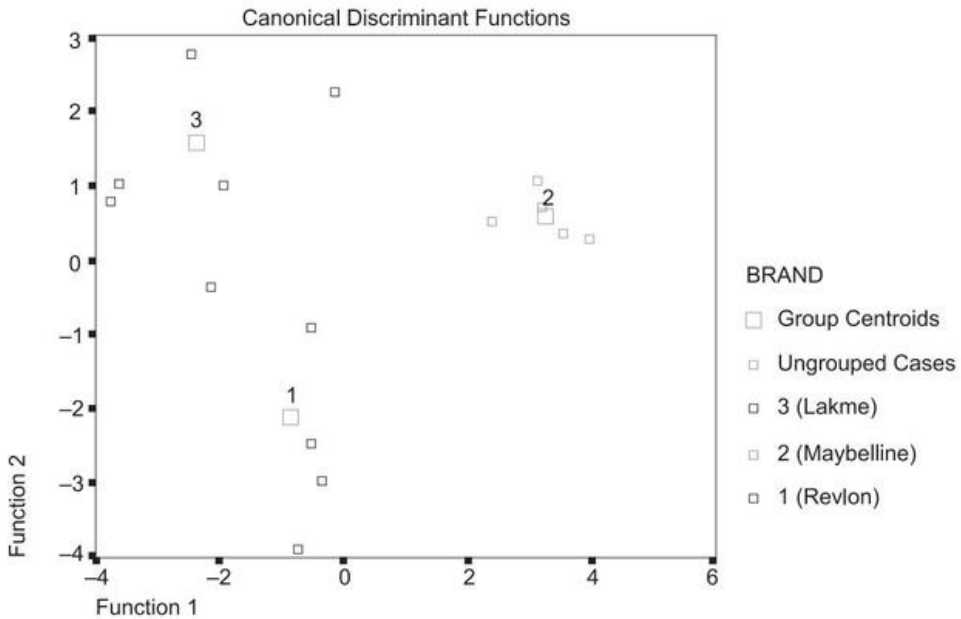


FIGURE 1.1 Plot of Brands on discriminant functions 1 and 2

Dimension 2 seems to comprise mainly Colour, the vector (arrows) that is closest to the vertical axis. This is also evident from the standardised coefficient for colour on Dimension 2, from the earlier output table.

Long lasting nature is not useful in defining any of the two dimensions as its arrow is not close to any of the two dimensions. Also the length of its arrow is small. Finish is also between the two axes, and therefore not very useful in discriminating between the two dimensions.

Brands and their Association with Attributes/Dimensions

Maybelline seems to be stronger on dimension 1 (a combination of price and packaging) and Lakme on dimension 2 (colour). However Revlon seems to score low on both the dimensions compared to its competitors.

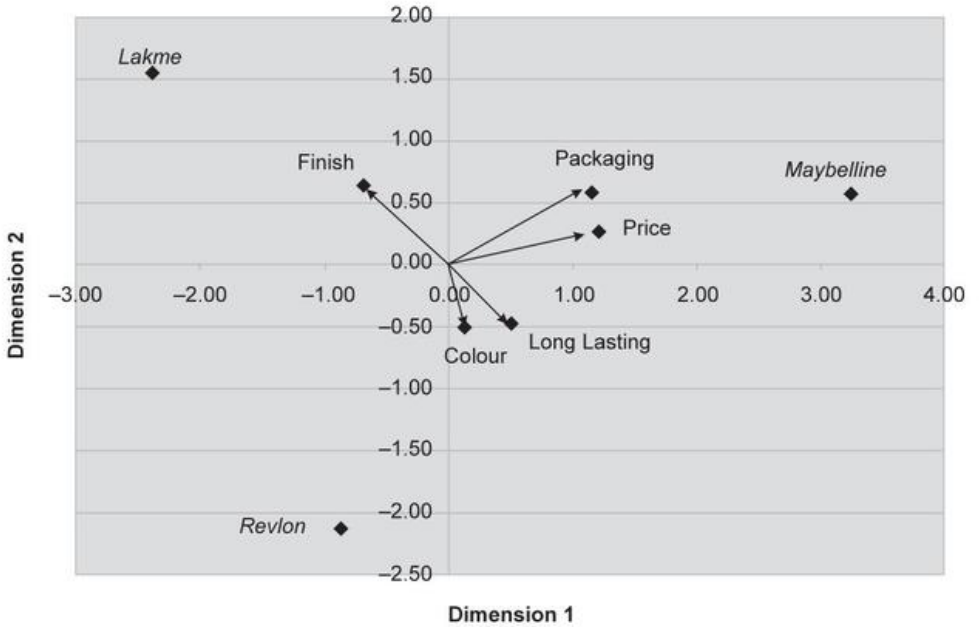


FIGURE 1.2 Perceptual Map of Lipstick Brands and Attributes

CASE STUDY 3

Attribute-based Perceptual Mapping Using Discriminant Analysis*

Problem

A marketing researcher wishes to determine the present position of four leading brands of wrist watches (Casio, Citizen, HMT and Titan) in the minds of the consumer and how these brands measure up on different attributes.

Input Data

The brand of wrist watch is taken as a dependent variable for analysis and has been coded as brand. The four brands of wrist watches chosen for study and their codes are as shown below.

- Casio – 1
- Citizen – 2
- HMT – 3
- Titan – 4

The attributes chosen for analysis and their codes are as follows:

- Digital display – digidisp
- Waterproof feature – watproof
- Provision for alarm – alarm
- Provision for date – date
- Use of jewels in the watch – jewels
- Provision of a metal band – metband

Twenty respondents were asked to rate the four brands listed above (each brand assessed by five respondents) on the above-mentioned attributes on a five point scale (1 – Very bad to 5 — Very good). The data collected is as shown in Table 1.1. This data was entered into the SPSS package and a Discriminant Analysis carried out.

* Prepared by Devmitra Biswas, Parameswaran Guha, Ravi Jhakodia, Shailendra Sahay, Shaleen Prakash, and Shiva Kumar Konda

TABLE 1.1

Brand	*Digidisp*	*Watproof*	*Alarm*	*Date*	*Jewels*	*Metband*
1	4	4	2	3	2	3
1	3	3	3	3	2	2
1	4	4	3	4	3	2
1	4	4	3	4	2	2
1	3	3	2	3	3	3
2	3	3	3	2	4	4
2	2	3	3	4	4	3
2	3	3	2	3	4	4
2	3	4	3	3	3	4
2	2	2	3	2	4	3
3	2	3	2	2	4	4
3	2	3	2	2	3	4
3	3	3	3	2	4	3
3	2	3	2	2	4	3
3	2	2	3	3	4	4
4	3	3	3	3	4	4
4	3	3	3	3	4	4
4	3	3	3	2	4	4
4	4	2	2	4	3	3
4	3	4	3	3	4	3

Output

TABLE 1.2 Group Statistics

BRAND		*Mean*	*Std. Deviation*	*Valid N (listwise)*	
				Unweighted	*Weighted*
1	DIGIDISP	3.60	.548	5	5.000
	WATPROOF	3.60	.548	5	5.000
	ALARM	2.60	.548	5	5.000
	DATE	3.40	.548	5	5.000
	JEWELS	2.40	.548	5	5.000
	METBAND	2.40	.548	5	5.000
2	DIGIDISP	2.60	.548	5	5.000
	WATPROOF	3.00	.707	5	5.000
	ALARM	2.80	.447	5	5.000

(Table 1.2 Contd.)

(Table 1.2 Contd.)

	DATE	2.80	.837	5	5.000
	JEWELS	3.80	.447	5	5.000
	METBAND	3.60	.548	5	5.000
3	DIGIDISP	2.20	.447	5	5.000
	WATPROOF	2.80	.447	5	5.000
	ALARM	2.40	.548	5	5.000
	DATE	2.20	.447	5	5.000
	JEWELS	3.80	.447	5	5.000
	METBAND	3.60	.548	5	5.000
4	DIGIDISP	3.20	.447	5	5.000
	WATPROOF	3.00	.707	5	5.000
	ALARM	2.80	.447	5	5.000
	DATE	3.00	.707	5	5.000
	JEWELS	3.80	.447	5	5.000
	METBAND	3.60	.548	5	5.000
Total	DIGIDISP	2.90	.718	20	20.000
	WATPROOF	3.10	.641	20	20.000
	ALARM	2.65	.489	20	20.000
	DATE	2.85	.745	20	20.000
	JEWELS	3.45	.759	20	20.000
	METBAND	3.30	.733	20	20.000

TABLE 1.3 Tests of Equality of Group Means

	Wilks' Lambda	*F*	*df1*	*df2*	*Sig.*
DIGIDISP	.408	7.733	3	16	.002
WATPROOF	.769	1.600	3	16	.229
ALARM	.879	.733	3	16	.547
DATE	.645	2.941	3	16	.065
JEWELS	.329	10.898	3	16	.000
METBAND	.471	6.000	3	16	.006

TABLE 1.4 Pooled Within-Groups Matrix

		DIGIDISP	*Watproof*	*Alarm*	*Date*	*Jewels*	*Metband*
Correlation	DIGIDISP	1.000	.286	–.100	.230	–.316	–.046
	WATPROOF	.286	1.000	.122	.000	–.086	.075
	ALARM	–.100	.122	1.000	.038	.211	–.274
	DATE	.230	.000	.038	1.000	–.162	–.315
	JEWELS	–.316	–.086	.211	–.162	1.000	.000
	METBAND	–.046	.075	–.274	–.315	.000	1.000

Summary of Canonical Discriminant Functions

TABLE 1.5 Eigen values

Function	*Eigen value*	*% of Variance*	*Cumulative %*	*Canonical Correlation*
1	3.537[a]	76.8	76.8	.883
2	.988[a]	21.5	98.3	.705
3	.078[a]	1.7	100.0	.269

[a]First three canonical discriminant functions were used in the analysis

TABLE 1.6 Wilks' Lambda

Test of Function(s)	*Wilks' Lambda*	*Chi-square*	*df*	*Sig.*
1 through 3	.103	31.847	18	.023
2 through 3	.466	10.675	10	.383
3	.927	1.056	4	.901

TABLE 1.7 Standardised Canonical Discriminant Function Coefficients

	Function		
	1	*2*	*3*
DIGIDISP	–.223	.844	–.680
WATPROOF	–.213	–.215	.349
ALARM	.043	.465	.584
DATE	.031	.416	.606
JEWELS	.663	.351	–.279
METBAND	.589	.412	.242

TABLE 1.8 Structure Matrix

	Function		
	1	*2*	*3*
JEWELS	.756*	.133	–.069
METBAND	.561*	.099	–.051
WATPROOF	–.285*	.084	.269
DIGIDISP	–.518	.702*	.422
DATE	–.312	.443*	.440
ALARM	.019	.332	.593*

Pooled within-groups correlations between discriminating variables and standardised canonical discriminant functions variables ordered by absolute size of correlation within function.

*Largest absolute correlation between each variable and any discriminant function.

TABLE 1.9 Canonical Discriminant Function Coefficients

	Function		
	1	*2*	*3*
DIGIDISP	–.446	1.688	–1.359
WATPROOF	–.348	–.351	.570
ALARM	.087	.931	1.168
DATE	.047	.638	.929
JEWELS	1.398	.739	–.588
METBAND	1.075	.752	.442
(Constant)	–6.362	–13.125	–2.998

Unstandardised coefficients

TABLE 1.10 Functions at Group Centroids

	Function		
BRAND	*1*	*2*	*3*
1	–2.901	–.143	5.866E-03
2	.991	.121	.406
3	1.176	–1.239	–.189
4	.733	1.261	–.224

Unstandardised canonical discriminant functions evaluated at group means

TABLE 1.11 Classification Results[a]

			Predicted Group Membership				
		Brand	*1*	*2*	*3*	*4*	*Total*
Original	Count	1	5	0	0	0	5
		2	0	2	1	2	5
		3	0	1	3	1	5
		4	0	1	0	4	5
	Percentage	1	100.0	.0	.0	.0	100.0
		2	.0	40.0	20.0	40.0	100.0
		3	.0	20.0	60.0	20.0	100.0
		4	.0	20.0	.0	80.0	100.0

[a]70.0% of original grouped cases correctly classified.

Analysis and Interpretation

- The means and standard deviations of the different independent variables across the various brands indicate that the performance of these brands varies greatly across the different

attributes. Also while the performance of the brands is around average for waterproof abilities, use of jewels and provision of a metal band, they are perceived to perform poorly on the other attributes.

- The Univariate ANOVA test results (Table 1.3) indicate that the brands vary significantly across the various attributes. However the importance of the attributes of waterproof abilities and provision of alarm in distinguishing between the brands is found to be low.
- The Pooled within groups matrix (Table 1.4) indicates low correlation among the independent variables. The implies that the attributes chosen for analysis are distinct.
- The tables for Eigen values and Wilks' lambda (Tables 1.5 and 1.6) indicate that the first two functions explain most of the variance in the input data while the contribution of the third is negligible. *Hence these two functions have been used for further interpretation.*
- Using the standardised coefficients of the attributes on the functions 1 and 2 (from Table 1.7), the plot for the different attributes has been plotted as shown in Fig. 1.2. Based on the distances of the attribute vectors from the axis and their lengths, it can be concluded that waterproof ability, use of jewels in the dial and a metal band weight heavily on dimension 1, though the weightage for waterproof ability is relatively lower and in the opposite direction as compared to the other two attributes. The dimension 1 has been interpreted as *aesthetic performance*.

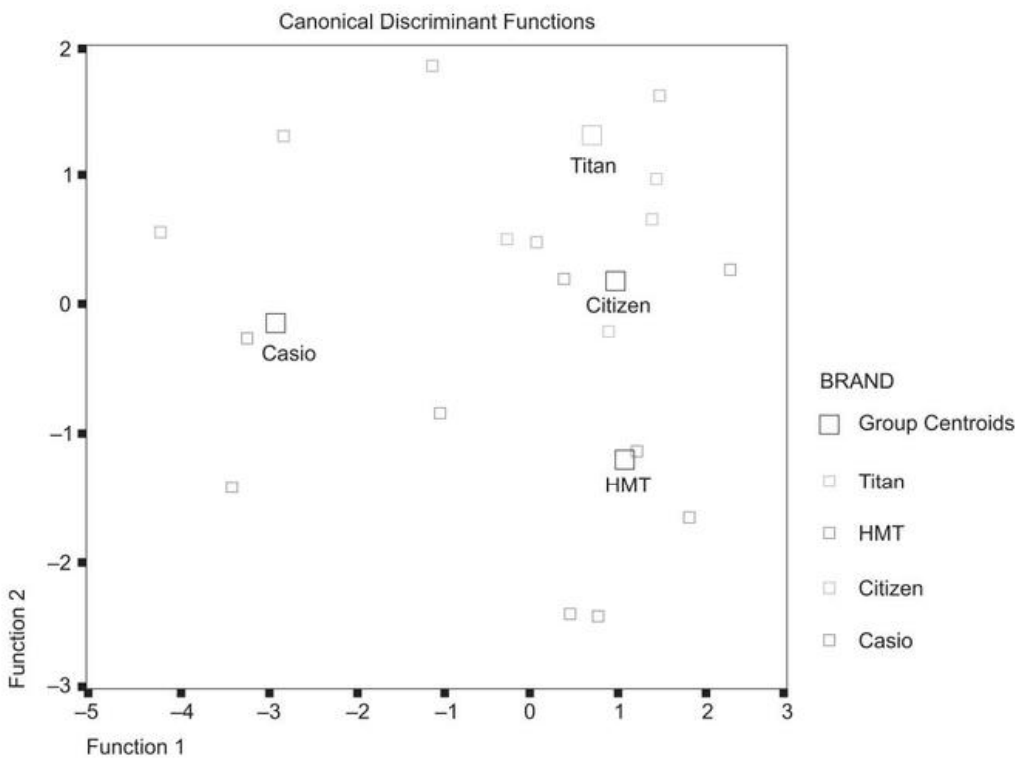
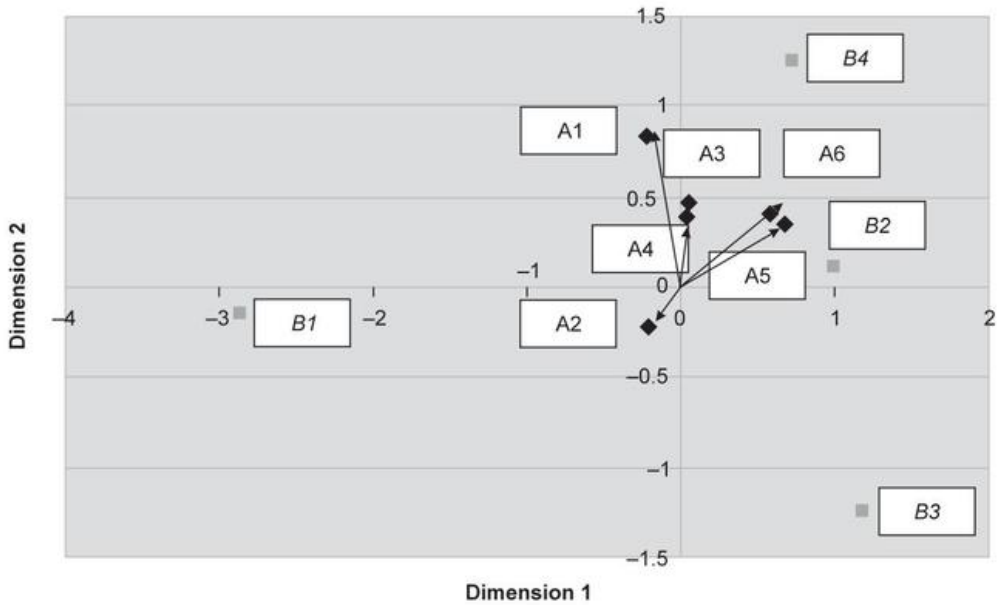


FIGURE 1.1 Plot of Four Brands of Watches



B1 – Casio	A1 – Digital Display	A2 – Waterproof features
B2 – Citizen	A3 – Provision for alarm	A4 – Provision for date
B3 – HMT	A5 – Use of jewels	
B4 – Titan	A6 – Provision of a metal band	

FIGURE 1.2 Perceptual Map of Watch Brands and Attributes

However due to the low and opposite weightage of waterproof ability and no visible link to aesthetic properties, one may also treat this as a separate dimension. The graph also indicates that digital display, provision for alarm and date weigh heavily on dimension 2. Of these, digital display is the strongest attribute while provision for alarm and that for date are close together. Based on these observations, dimension 2 has been interpreted as *technological performance*. The above conclusions are corroborated by the structural matrix indicating the loadings of the different attributes on the two dimensions.

- The positions of the brands under study with respect to these attributes are obtained from the graphs as shown in Fig. 1.2.
  - It appears that Casio is strongly represented by the attribute of waterproof abilities. The arrows of the other attributes point in the opposite direction and therefore do not have a significant influence on the perception of the brand. Hence it can be concluded that Casio is perceived to be mainly a waterproof watch. But the short arrow for this attribute means that it is a less important attribute.
  - Citizen is firmly identified by all attributes except those of digital display and waterproof abilities. However the attributes of metal band and use of jewels in the dial have a greater say in distinguishing the brand. Thus it can be concluded that the respondents do not perceive that Citizen watches offer digital or waterproof features.

- None of the attribute vectors point towards HMT. Hence it can be concluded that none of the attributes contributes to the differentiation of the brand from the rest. The same is indicated by the fact that the group centroid for this brand is equally spaced with respect to both the axes. This indicates that HMT does not occupy a unique position with respect to either dimension.
- It can be concluded that Titan watches occupy a strong position on all the attributes other than the digital display and waterproof abilities. But the provision for alarm and date distinguish this brand more than the other attributes.

The McGraw·Hill Companies

CHAPTER 18 STRUCTURAL EQUATION MODELING (SEM) FOR COMPLEX MODELS (INCLUDING CONFIRMATORY FACTOR ANALYSIS)

Learning Objectives

In this chapter, we will

- Introduce the Structural Equation Modeling technique, along with Confirmatory Factor Analysis
- Illustrate the application of SEM and CFA to a numerical problem
- Explain various significance tests and statistics from the output of SEM and their interpretation

We have covered the techniques of regression and factor analysis in two earlier chapters. This chapter is devoted to Structural Equation Modeling, a combination of these two techniques. There are packages like AMOS (from SPSS), LISREL, STATISTICA and a few others that can be used to perform SEM. The major application of SEM is to test complex relationships simultaneously. It is somewhat like performing several regression analyses and factor analyses together. The academic world has seen a lot of SEM based research in recent years. The commercial world may soon follow.

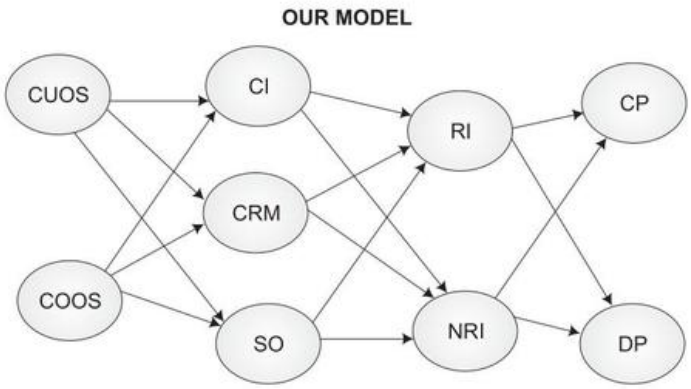
CONFIRMATORY FACTOR ANALYSIS

Researchers may be also interested in other techniques that can be performed in a similar way or as a part of SEM. These are Confirmatory Factor Analysis (CFA) and Path Analysis. Confirmatory Factor Analysis is different from Exploratory Factor Analysis (which we looked at in an earlier chapter) because here, we assign variables to factors before the factor extraction stage and merely confirm whether the hypothesized relationship is true. In other words, factors are predetermined and then their structure is confirmed through CFA.

Both for SEM and CFA, if we find that the given model or structure is not a good fit, we can redesign the model and compare it with the older model. In that sense, this is a technique which allows you to test alternative complex models to explore the nature of relationships between the different variables.

In this chapter, we will explore using very simple diagrams without too much of the complex mathematics, how we can use SEM and CFA. We will start by explaining an actual SEM study with its initial model and result. This deals with multiple links between variables such as Strategy, CRM, Innovation and Financial performance, from left to right in the figure.

The model is illustrated in Fig. 18.1 below:



Expansion of Acronyms
CUOS= Customer Oriented Strategy
COOS= Competitor Oriented Strategy
CI= Competitor Intelligence
CRM= Customer Relationship Management
SO= Service Orientation
RI= Radical Marketing Innovation
NRI= Non-radical Marketing Innovation
CP= Company Performance
DP= Division/SBU Performance

FIGURE 18.1 Proposed SEM Model: Strategy, Market Orientation, Innovation and Performance

This model proposes a complex test of many simultaneous regression models indicated by the arrows. The independent variables are to the left (where the arrows begin) and the dependent variables are where the arrows end. Each path (this is what gives the name path analysis) is a test for one of the several hypotheses being tested simultaneously. The hypotheses in this case are that strategy impacts the three variables Competitor Intelligence, CRM and Service Orientation. These impact Innovation (both Radical and Non-radical, and finally, these have an impact on performance at the company and Division/SBU levels. As you can see, there are multi-stage hypotheses being tested simultaneously. No other technique allows you to do so many tests together. This is one of the major advantages of SEM over usual regression techniques.

All the variables which appear in Fig. 18.1 are the equivalent of Factors in Factor Analysis. The factors are known as Latent variables in SEM models. Each Factor/Latent variable may have multiple observed variables (called manifest variables in SEM) comprising that factor. We will elaborate the factor structure later, but let us first look at the solved model in the Fig. 18.2.

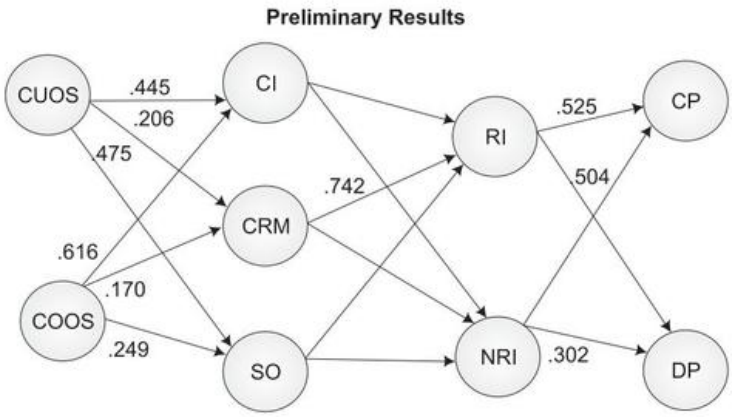


FIGURE 18.2 Results from Testing Proposed SEM Model

Figure 18.2 indicates that some of the hypothesized relationships are significant. These are the arrows with numbers on them. The numbers are like the regression coefficients (here they are called path coefficients). Arrows with no numbers indicate that the path specified by that arrow is not significant. The coefficients have been arrived at after performing SEM using one of the standard SEM packages. (see Table 3 for the output for path coefficients). The data for this study were collected from executives of Indian companies across various industries. The aim of the study (Nargundkar and Vijayakumar, 2005) was to study Market Orientation in Indian companies, and examine its link with performance variables. CI, CRM and SO in the above diagram indicate components of market orientation. Marketing Innovation is also viewed as an intermediate variable, represented by two components, Radical Innovation (RI) and Non-radical Innovation (NRI). This model indicates that CRM practices are contributing to better performance through radical innovation, but non-radical innovations are also linked significantly with performance. The innovations comprise innovations in all four Ps of marketing.

Tests Used in CFA and SEM

Some of the measures used in testing SEM and CFA Models are described below (reference for the following section: website : http://www.statsoft.com , look under SEM)

1. Goodness-of-fit index, GFI (Jöreskog-Sörbom GFI) GFI = F_{ML}/F_O, where F_O is the fit function when all model parameters are zero. GFI varies from 0 to 1, but theoretically can yield meaningless negative values. A large sample size pushes GFI up. Though analogies are made to R-square, GFI cannot be interpreted as percent of error explained by the model. Rather it is the percent

of observed covariances explained by the covariances implied by the model. That is, R^2 in multiple regression deals with error variance whereas GFI deals with error in reproducing the variance-covariance matrix. As GFI often runs high compared to other fit models, some suggest using .95 as the cutoff. By convention, GFI should by equal to or greater than .90 to accept the model. LISREL, AMOS and SEPATH (in Statistica) compute GFI.

2. Adjusted goodness-of-fit index, AGFI AGFI is a variant of GFI which uses mean squares instead of total sums of squares in the numerator and denominator. It, too, varies from 0 to 1, but theoretically can yield meaningless negative values. AGFI > 1.0 is associated with just-identified models and models with almost perfect fit. AGFI < 0 is associated with models with extremely poor fit, or based on small sample size. AGFI should also be at least 0.90.

3. Root mean square residuals, or RMS residuals, or RMSR, or RMR The closer the RMR to 0 for a model being tested, the better the model fit. RMS residuals are the coefficients which result from taking the square root of the mean of the squared residuals, which are the amounts by which the *sample* variances and covariances differ from the corresponding *estimated* variances and covariances, estimated on the assumption that your model is correct. *Fitted residuals* result from subtracting the sample covariance matrix from the fitted or estimated covariance matrix. LISREL computes RMSR. AMOS does also, but calls it RMR.

4. Standardized root mean square residual, standardized RMR (SRMR) The smaller the standardized RMR, the better the model fit. SRMR is the average difference between the predicted and observed variances and covariances in the model, based on standardized residuals. *Standardized residuals* are fitted residuals (see above) divided by the standard error of the residual (this assumes a large enough sample to assume stability of the standard error). SRMR is 0 when model fit is perfect.

5. Centrality index, CI CI is a function of model chi-square, degrees of freedom in the model, and sample size. By convention, CI should be .90 or higher to accept the model.

6. Noncentrality index, DK also called the McDonald noncentrality parameter index, is chi-square penalizing for model complexity. It is computed as ((chisqn-dfn)-(chisq-df))/(chisqn-dfn), where chisqn and chisq are model chi-squares for the null model and the given model, and dfn and df are the corresponding degrees of freedom. To force it to scale to 1, the conversion is exp(-DK/2).

7. Relative non-centrality index, RNI penalizes for sample size as well as model complexity. It should be greater than .9 for good fit. The computation is ((chisqn/n -dfn/n)-DK)/(chisqn/n-dfn/n), where chisqn and chisq are model chi-squares for the null model and the given model, dfn and df are the corresponding degrees of freedom, n is the number of subjects, and DK is the McDonald noncentrality index. There is also a McDonald relative non-centrality index, computed as 1 - ((chisq-df)/(chisqn-dfn)).

Goodness-of-fit Tests Comparing the Given Model with an Alternative Model

This set of goodness of fit measures compare your model to the fit of another model. When none is specified, statistical packages usually default to comparing your model with the independence model, or even allow this as the only option. The independence model is the null model, which is the model in which variables are assumed to be uncorrelated with the dependent(s). Since the fit of the independence model is usually terrible, comparing your model to it will generally make your model look good but may not serve your research purposes.

1. The **comparative fit index, CFI** , is also known as the Bentler Comparative Fit Index. CFI compares the existing model fit with a null model which assumes the latent variables in the model are uncorrelated (the "independence model"). CFI close to 1 indicates a very good fit. By convention, CFI should be equal to or greater than .90 to accept the model, indicating that 90% of the covariation in the data can be reproduced by the given model. The Bentler-Bonnett Index, BBI, is the model chi-square for the given model minus model chi-square for the null model, this difference divided by model chi-square for the null model. BBI should be greater than .9 to consider a fit as good.
2. The **incremental fit index, IFI**, also known as *DELTA2*: By convention, IFI should be equal to or greater than .90 to accept the model. IFI can be greater than 1.0 under certain circumstances.
3. The **normed fit index, NFI**, also known as the *Bentler-Bonett normed fit index*, or simply *DELTA1*. NFI was developed as an alternative to CFI, but one which did not require making chi-square assumptions. It varies from 0 to 1, with 1 = perfect fit. NFI reflects the proportion by which the researcher's model improves fit compared to the null model (random variables). For instance, NFI of .50 means the researcher's model improves fit by 50% compared to the null model. By convention, NFI values below .90 indicate a need to respecify the model. Some authors have used the more liberal cutoff of .80.
4. The **non-normed fit index, NNFI**, also called the *Bentler-Bonett non-normed fit index*, the *Tucker-Lewis index, TLI* (this is the label in AMOS), or *RHO2*, NNFI is similar to NFI, but penalizes for model complexity. NNFI is not guaranteed to vary from 0 to 1. NNFI close to 1 indicates a good fit. By convention, NNFI values below .90 indicate a need to respecify the model. Some authors have used the more liberal cutoff of .80 since TLI tends to run lower than GFI. However, more recently, Hu and Bentler (1999) have suggested NNFI >= .95 as the cutoff for a good model fit. It is one of the fit indexes less affected by sample size.
5. The **Bollen86 Fit Index** is similar to NNFI. It should be greater than .9 for a good fit.
6. The **relative fit index, RFI**, also known as *RHO1*, is not guaranteed to vary from 0 to 1. RFI close to 1 indicates a good fit.

Chi-square Test in SEM

For the chi-square test of goodness of fit of the model as a whole, the chi-square value should not be significant if there is a good model fit: the higher the chi-square, the more the difference between the *model-estimated* and *actual* covariance matrices, hence the worse the model fit. This is a major difference from earlier uses of the chi-square test and should be remembered by first time users of SEM techniques.

There are also some rules of thumb that say the chi-square value divided by the degrees of freedom should be around 2 or 2.5. This takes care of the biases involved with looking at the chi-square value by itself.

Confirmatory Factor Analysis

Using the same problem represented by Fig. 18.1, we can now explain how confirmatory factor analysis (CFA) can be used to test the hypothesis that each of the latent variables in the model is composed of the following manifest (observed) variables- (these questions representing manifest/ observed variables are questions from the questionnaire in Table 18.2.)

1. **Competitor Intelligence** is a latent variable consisting of 3 manifest variables—questions no. 5, 6 and 9 from the questionnaire in Table 18.1. (In the data file given in the appendix, these are variables in column nos. 6, 7 and 10).
2. **CRM** is a latent variable consisting of questions 3, 4 and 8 from the questionnaire (variables in columns 4, 5, 9 in the data file)
3. **Service Orientation** is a latent variable consisting of questions 1, 2 and 7 (columns 2, 3 and 8 in the data file).
4. **Radical Innovation** is a latent variable consisting of the four questions 10, 12, 14 and 16. (column nos. 11, 13, 15, 17 in the data file).
5. **Non-radical Innovation** is a latent variable consisting of the four questions 11, 13, 15 and 17. (column nos. 12, 14, 16, 18 in the data file).
6. **Divisional Performance** is a combination of the three variables in questions 26, 28 and 30. (columns 27, 29 and 31 in the data file).
7. **Company Performance** is a combination of the three variables in questions 25, 27 and 29. (columns 26, 28 and 30 in the data file).

If we were to conduct a CFA to test the above hypothesis regarding the composition of latent variables, we could do it as follows in the software package STATISTICA (assuming data file based on this questionnaire is available).

STATISTICA Commands for CFA

1. With this data file in Statistica active, choose the SEPATH option in Statistica Module Switcher. Then, from the next screen, choose USE PATH WIZARDS option.
2. A choice of two path wizards will appear. Choose **Confirmatory Factor Analysis wizard.**
3. Fill in the latent variable name (say, Competitive Intelligence or CI), followed by selecting the manifest variables (variable 6, variable 7 and variable 10 from the drop down list of variables) for that latent variable, by clicking on **Manifest Var** and selecting relevant variables. For Radical Innovation (you can use a shorter name such as RADINNO) as a latent variable, define the manifest variables as variable nos. 11, 13, 15 and 17 from the variable list representing radical product innovation, radical pricing innovation, radical promotion innovation and radical distribution innovation.

4. Similarly, fill in all the other latent variables you have, with corresponding manifest variables from the questionnaire.
5. Click **OK**.
6. Click **EXECUTE CURRENT MODEL** from dialogue box. The outputs should appear on the screen.
7. In the outputs, go to the **Goodness of Fit Indices** and look for Joreskog GFI under the output labeled '**Other Single Sample Indices**'. This should be more than 0.9 for a good model fit.
8. Under **Model Summary**, look at RMS standard residual. It should be less than 0.5 for a good model. Chi-square should be low, and *not significant*.

SEM Commands using STATISTICA

9. For doing **Structural Equation Modeling** as per Fig. 18.1 with the same variables, we can also generate parameter estimates (path coefficients, similar to regression coefficients) for each specified path under the **Model Summary** output. For doing SEM, choose the **SEM Wizard** in step 2 (above), instead of the CFA wizard. Also, specify exogenous variables and their corresponding manifest variables. Exogenous variables are those which have paths emanating from them. Endogenous variables also have to be specified. These are variables having paths ending at them, not beginning from them. Manifest variables for all the Endogenous variables also need to be specified. The remaining process is similar to that described above.
10. For describing the paths with hypothesized relationships, a screen will appear titled DESCRIBE STRUCTURAL EQUATION PATHS, after you define all the latent and manifest variables. In the example above, CP and DP are endogenous variables, and all others should be defined as exogenous variables. Each variable has some associated manifest variables, as explained above. These should be filled in on the screen. For example, one path is CRM to Radical Innovation. You can select CRM from the left side, Radical Innovation from the right side and click ADD. This path will be automatically specified in the model. Similarly, you can add all the paths that you want to include in the model.
11. After filling all the manifest variables on the screen, click OK. A screen appears which reads as follows—"All Steps Required by SEPATH Wizard to create a model are completed. You can now REPLACE EXISTING PROGRAM WITH A NEW MODEL." You can choose this option and then click **OK**. Again click **OK (Execute Current Model)** from the next screen. This will give you the SEM outputs.
12. Output indicators we need to look at are—a GFI value above .9, an RMS standard residual of .05 or less and significant path coefficients under Model Summary. Chi-square test should be non-significant.

Figure 18.2 shows the results (path coefficients which were significant) for the full model, where the GFI came out as 0.85. The other results for an SEM run on the full model are shown in Table 18.1 below (output taken from Statistica based on data collected from the questionnaire in Table 18.2, for the model in Fig. 18.1):

TABLE 18.1 SEM Output: Single Sample Fit Indices

	Value
Joreskog GFI	0.852228
Joreskog AGFI	0.809421
Akaike Information Criterion	2.90244
Schwarz's Bayesian Criterion	4.00527
Browne-Cudeck Cross Validation Index	2.985637
Independence Model Chi-Square	2442.794
Independence Model df	300
Bentler-Bonett Normed Fit Index	0.790054
Bentler-Bonett Non-Normed Fit Index	0.854552
Bentler Comparative Fit Index	0.878265
James-Mulaik-Brett Parsimonious Fit Index	0.663646
Bollen's Rho	0.750065
Bollen's Delta	0.880488

The results indicate a moderate fit, since the Joreskog Goodness of Fit Index is 0.85. If it had been above 0.9, it would have indicated a good fit. As we can see, the chi-square value is high. It should ideally be low and non-significant, if a good fit is indicated. The researcher can attempt to re-specify the above model and attempt to get a model with a better fit. For an exploratory study, the above values may be perceived as satisfactory.

If we perform only a Confirmatory Factor Analysis, we will get an output similar to the above table. But if we perform an SEM, we will also get path coefficient estimates (Table 18.3 which appears later)

TABLE 18.2 Questionnaire: Market Orientation, Marketing Innovation as Performance Drivers: Extending the Paradigm

Questionnaire taken from a study by Dr. Rajendra Nargundkar, IIM Lucknow, India, and Dr. Gurvinder Shergill, Massey University, New Zealand. Some results from this study were published in the Journal of Global Marketing, Dec. 2005.

Please circle the relevant response on the following questions as they relate to your company (1 = Strongly Disagree, 7= Strongly Agree, and other responses are between these. 4 indicates Neither Agree nor Diagree)

1. We have a working mechanism in place to collect and record customer complaints.

   1 2 3 4 5 6 7

2. We lay equal emphasis on sales and customer service.

   1 2 3 4 5 6 7

3. We measure customer satisfaction at least once every year through a formal research study.

   1 2 3 4 5 6 7

(*Table 18.2 contd*)

(*Table 18.2 contd*)

4. We have implemented CRM initiatives in our company.
   1 2 3 4 5 6 7
5. We have a mechanism in place to regularly collect data on our competitors.
   1 2 3 4 5 6 7
6. We discuss the collected data in regular meetings.
   1 2 3 4 5 6 7
7. Our top management emphasises customer orientation.
   1 2 3 4 5 6 7
8. Our use of I.T. for customer interface is significant.
   1 2 3 4 5 6 7
9. We formulate some responses based on the competitor data collected.
   1 2 3 4 5 6 7
10. Our pricing strategies are innovative in a radical way.
    1 2 3 4 5 6 7
11. Our pricing strategies are innovative in a sustained but non-radical way.
    1 2 3 4 5 6 7
12. Our product related strategies are innovative in a radical way.
    1 2 3 4 5 6 7
13. Our product-related strategies are innovative in a sustained but non-radical way.
    1 2 3 4 5 6 7
14. Our promotion (advertising, sales promotion, sales force, etc.) strategies are innovative in a radical way.
    1 2 3 4 5 6 7
15. Our promotion (advertising, sales promotion, sales force, etc.) strategies are innovative in a sustained but non-radical way.
    1 2 3 4 5 6 7
16. Our distribution (channel, logistics and supply chain) related strategies are innovative in a radical way.
    1 2 3 4 5 6 7
17. Our distribution (channel, logistics and supply chain) related strategies are innovative in a sustained but non-radical way.
    1 2 3 4 5 6 7
18. Our strategies are predominantly customer-driven.
    1 2 3 4 5 6 7
19. Our strategies are predominantly competitor-driven.
    1 2 3 4 5 6 7
20. Our strategies are predominantly driven by our company objectives.
    1 2 3 4 5 6 7

(*Table 18.2 contd*)

(*Table 18.2 contd*)

Company Profile:

21. Your company's area of Business: (Tick one)
    1. FMCG
    2. Durables
    3. Service Business
    4. Business/Industrial Goods
22. Number of years your company has been in business: (Tick one)
    1. Less than 5 years
    2. 6-10 years
    3. 11-20 years
    4. More than 20 years
23. Company Shares are listed on a major stock exchange: (Tick one)
    1. Yes
    2. No
24. Ownership of Company: (Tick one)
    1. Indian, family owned/ Pvt. Ltd.
    2. Indian, Public Ltd.
    3. Public Sector/Govt. Owned
    4. Foreign / MNC
    5. Joint venture
    6. Any other, please specify_____

Please rate the following using a 7 point scale (1=poor, 7=excellent, other points between 1 and 7 appropriately worded, with 4 as Average Performance):

25. Your company's Market Share performance for the last financial year

    1 2 3 4 5 6 7

26. Your SBU/division's Market Share performance for the last financial year

    1 2 3 4 5 6 7

27. Your company's Profit performance last financial year

    1 2 3 4 5 6 7

28. Your SBU/division's Profit performance last financial year

    1 2 3 4 5 6 7

29. Your company's Sales growth in the last year (over the previous year), compared to your major competitors

    1 2 3 4 5 6 7

30. Your SBU/division's Sales growth in the last year (over the previous year), compared to your major competitors

    1 2 3 4 5 6 7

TABLE 18.3 Path Coefficients Output for the Full SEM Model specified in Fig 18.1

(Significant path coefficients at below 0.05 prob. level appear in bold)

Path	*Parameter Estimate*	*Standard Error*	*T Statistic*	*Prob. Level*
(CUS)—>(CRM)	**0.206**	0.060	5.268	**0.000**
(COS)—>(CI)	**0.616**	0.093	6.323	**0.000**
(CUS)—>(SO)	**0.475**	0.051	11.074	**0.000**
(COS)—>(SO)	**0.249**	0.055	1.774	**0.076**
(CRM)—>(RI)	**0.742**	0.931	4.118	**0.000**
(CRM)—>(NRI)	0.564	0.731	3.016	0.303
(CI)—>(RI)	0.323	0.249	-1.290	0.197
(CI)—>(NRI)	0.865	0.163	-1.122	0.262
(SO)—>(RI)	0.234	0.431	-2.496	0.134
(SO)—>(NRI)	0.977	0.342	-2.238	0.250
(RI)—>(CP)	**0.525**	2.330	-1.998	**0.046**
(RI)- >(DP)	**0.504**	2.671	-1.983	**0.047**
(NRI)—>(CP)	0.654	0.199	0.785	0.432
(NRI)—>(DP)	**0.302**	0.214	0.694	**0.048**
(CUS)—>(CI)	**0.445**	0.064	6.332	**0.000**
(COS)—>(CRM)	**0.170**	0.057	2.957	**0.003**

A total of 10 paths have significant parameter estimates, from the above table. They are—

1. Customer Oriented Strategy to CRM
2. Customer Oriented Strategy to Competitor Intelligence
3. Competitor Oriented Strategy to CRM
4. Competitor Oriented Strategy to Competitor Intelligence
5. Customer Oriented Strategy to Service Orientation
6. Competitor Oriented Strategy to Service Orientation
7. CRM to Radical Innovation
8. Radical Innovation to Company Performance
9. Radical Innovation to Divisional Performance
10. Non-Radical Innovation to Divisional Performance

Conclusion from SEM Analysis

The conclusion on the whole from the above analysis is that strategy influences CRM, which in turn affects Radical Innovation, and that impacts performance. But non-radical innovation in marketing also has an impact on company performance. *The other paths are not significant*. Thus, we can see that complex models requiring simultaneous tests of many linear relationships between latent variables can be conducted by using Structural Equation Modeling (SEM). We need to have large sample sizes for testing such models and data should be interval scaled. The relationships between latent variables and the observed or manifest variables is also tested during the SEM. But we can test for the Factor Analysis alone, if we do not need the path coefficients.

A software package capable of doing the analysis (AMOS, LISREL, STATISTICA, etc.) is a prerequisite. Please refer to the individual software for variations on how to perform SEM.

A complete dataset used in performing an SEM example similar to the one discussed in this chapter is given in Appendix 1 to this chapter, immediately after the SUMMARY. It is based on the same questionnaire we have discussed. Those readers who have access to the packages mentioned herein may like to try doing this analysis using the given data. Similar results to the ones presented here may be expected. Alternative model specifications can be tried and the model rerun. For example, some arrows can be removed. Correlations can be assumed between some of the variables. Other changes can also be tried, and the new model compared with the original model. But theory must support any proposed changes.

SUMMARY

We have seen how to use SEM modeling with an example of a complex model with multiple relationship. As a part of the same model, we also saw how the hypothesized relationship between latent variables (similar to Factors in Factor Analysis) and their constituent observed variables can be tested through Confirmatory Factor Analysis.

SEM is somewhat like performing many linear regressions at the same time, along with each of the variables in the regression also being tested as a Factor in factor analysis with pre-hypothesised variables loading on to them. Though some of these relationships can be tested through many simple regressions, the multiple tests done simultaneously add their own value to understanding the larger picture.

This procedure is particularly suited to academic research, but has potential for commercial use in marketing or strategy, as demonstrated in the example given. The dataset for this example, based on the questionnaire given, is included below as an Appendix.

Appendix

Data used in doing SEM

PILOTOR	COMPLAIN	SERVICE	SATISFAC	CRM	COMPDATA	DISCUSS	TOPMGT	INFOTECH	RESPONSE	PRICERAD	PRICENON
Pilot	3.00	2.00	5.00	2.00	1.00	1.00	1.00	2.00	2.00	5.00	3.00
	4.00	3.00	3.00	4.00	5.00	5.00	3.00	3.00	5.00	3.00	3.00
	6.00	5.00	4.00	5.00	5.00	5.00	3.00	4.00	4.00	3.00	4.00
	4.00	3.00	7.00	7.00	2.00	3.00	2.00	5.00	5.00	5.00	5.00
	1.00	4.00	6.00	7.00	2.00	2.00	2.00	6.00	2.00	4.00	3.00
	3.00	5.00	2.00	6.00	1.00	3.00	5.00	5.00	1.00	4.00	4.00
	5.00	4.00	6.00	5.00	3.00	3.00	2.00	2.00	3.00	6.00	5.00
	4.00	1.00	7.00	1.00	3.00	1.00	1.00	2.00	2.00	2.00	3.00
	5.00	6.00	6.00	3.00	4.00	5.00	6.00	6.00	5.00	5.00	4.00
	6.00	5.00	2.00	4.00	3.00	4.00	2.00	5.00	2.00	3.00	5.00
	2.00	1.00	6.00	5.00	6.00	6.00	2.00	3.00	3.00	6.00	3.00
	3.00	5.00	7.00	3.00	1.00	3.00	4.00	3.00	2.00	5.00	3.00
	3.00	4.00	6.00	6.00	5.00	3.00	5.00	6.00	5.00	4.00	2.00
	3.00	2.00	5.00	7.00	3.00	4.00	3.00	4.00	4.00	5.00	5.00
	5.00	5.00	6.00	7.00	2.00	4.00	5.00	3.00	2.00	2.00	5.00
	1.00	4.00	7.00	7.00	5.00	4.00	2.00	4.00	5.00	3.00	2.00
	5.00	5.00	7.00	7.00	6.00	6.00	2.00	4.00	6.00	5.00	6.00
	6.00	2.00	7.00	5.00	2.00	2.00	4.00	3.00	2.00	6.00	3.00
	1.00	4.00	7.00	7.00	2.00	2.00	1.00	4.00	2.00	5.00	3.00
	6.00	4.00	7.00	2.00	3.00	3.00	1.00	4.00	3.00	4.00	2.00
	2.00	2.00	3.00	3.00	3.00	4.00	2.00	4.00	2.00	6.00	5.00
	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	5.00
	3.00	2.00	7.00	7.00	7.00	4.00	2.00	6.00	6.00	5.00	1.00
	5.00	3.00	3.00	7.00	3.00	7.00	2.00	2.00	2.00	1.00	1.00
	3.00	1.00	2.00	3.00	1.00	1.00	2.00	1.00	5.00	4.00	4.00
	3.00	3.00	2.00	6.00	1.00	1.00	2.00	5.00	2.00	2.00	6.00
	2.00	4.00	2.00	5.00	5.00	4.00	3.00	2.00	3.00	1.00	4.00

Pilot28	2.00	1.00	2.00	7.00	2.00	1.00	2.00	6.00	2.00	1.00	6.00
Final1	5.00	4.00	5.00	6.00	7.00	7.00	6.00	5.00	7.00	1.00	5.00
	6.00	6.00	7.00	7.00	6.00	6.00	6.00	6.00	4.00	5.00	6.00
	6.00	2.00	7.00	4.00	6.00	6.00	6.00	6.00	5.00	3.00	6.00
	6.00	6.00	5.00	6.00	7.00	7.00	6.00	6.00	7.00	7.00	3.00
5	7.00	6.00	4.00	5.00	7.00	7.00	6.00	5.00	6.00	6.00	6.00
	5.00	4.00	2.00	4.00	7.00	7.00	4.00	4.00	6.00	3.00	3.00
	6.00	7.00	7.00	6.00	7.00	6.00	6.00	6.00	7.00	6.00	5.00
	6.00	7.00	4.00	5.00	6.00	7.00	6.00	6.00	7.00	5.00	3.00
	7.00	7.00	4.00	4.00	6.00	6.00	7.00	7.00	6.00	5.00	2.00
10	7.00	7.00	2.00	5.00	5.00	4.00	7.00	6.00	5.00	2.00	2.00
	5.00	5.00	5.00	4.00	6.00	6.00	6.00	4.00	4.00	6.00	6.00
	6.00	2.00	4.00	3.00	5.00	7.00	6.00	2.00	4.00	4.00	7.00
	6.00	3.00	4.00	3.00	6.00	6.00	6.00	3.00	4.00	5.00	5.00
	7.00	5.00	7.00	4.00	4.00	3.00	7.00	6.00	3.00	4.00	5.00
15	7.00	3.00	1.00	4.00	1.00	1.00	5.00	3.00	1.00	1.00	1.00
	5.00	4.00	6.00	2.00	6.00	7.00	5.00	7.00	7.00	6.00	2.00
	7.00	5.00	7.00	4.00	4.00	5.00	6.00	5.00	6.00	4.00	5.00
	7.00	7.00	5.00	6.00	7.00	7.00	6.00	6.00	5.00	4.00	6.00
	4.00	5.00	5.00	4.00	7.00	7.00	6.00	4.00	7.00	6.00	3.00
20	6.00	6.00	7.00	4.00	6.00	5.00	7.00	7.00	6.00	5.00	4.00
	5.00	5.00	4.00	4.00	4.00	6.00	6.00	6.00	6.00	6.00	6.00
	6.00	5.00	4.00	6.00	5.00	4.00	6.00	6.00	7.00	5.00	4.00
	3.00	2.00	7.00	3.00	5.00	3.00	6.00	5.00	3.00	4.00	7.00
	4.00	6.00	7.00	5.00	6.00	7.00	5.00	3.00	5.00	6.00	5.00
25	7.00	6.00	6.00	5.00	5.00	5.00	6.00	6.00	6.00	5.00	4.00
	7.00	7.00	7.00	7.00	7.00	5.00	7.00	7.00	7.00	4.00	6.00
	1.00	4.00	5.00	1.00	2.00	2.00	5.00	6.00	6.00	2.00	5.00
	4.00	5.00	3.00	3.00	5.00	3.00	6.00	5.00	4.00	4.00	5.00
	7.00	7.00	7.00	7.00	7.00	4.00	7.00	7.00	7.00	4.00	4.00
30	3.00	5.00	6.00	2.00	3.00	6.00	7.00	5.00	4.00	7.00	3.00
	4.00	5.00	5.00	5.00	7.00	6.00	6.00	6.00	7.00	7.00	3.00
	4.00	7.00	5.00	7.00	7.00	7.00	5.00	5.00	7.00	4.00	5.00
	7.00	5.00	5.00	6.00	2.00	1.00	4.00	4.00	5.00	5.00	5.00
	6.00	6.00	5.00	6.00	7.00	6.00	6.00	5.00	7.00	6.00	5.00

35	2.00	7.00	5.00	5.00	4.00	4.00	6.00	2.00	4.00	5.00	5.00
	2.00	5.00	1.00	7.00	6.00	6.00	5.00	2.00	4.00	5.00	4.00
	5.00	7.00	4.00	4.00	7.00	7.00	7.00	7.00	7.00	6.00	6.00
	6.00	7.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	4.00
	5.00	7.00	5.00	4.00	5.00	5.00	6.00	7.00	7.00	7.00	6.00
40	4.00	3.00	3.00	3.00	3.00	4.00	6.00	6.00	5.00	4.00	5.00
	7.00	7.00	6.00	7.00	6.00	6.00	7.00	6.00	6.00	7.00	5.00
	2.00	4.00	1.00	1.00	4.00	6.00	4.00	1.00	5.00	3.00	4.00
	1.00	2.00	1.00	3.00	5.00	3.00	4.00	1.00	5.00	7.00	6.00
	7.00	3.00	4.00	1.00	4.00	3.00	1.00	3.00	5.00	4.00	6.00
45	1.00	7.00	1.00	1.00	7.00	7.00	7.00	7.00	7.00	....	....
	7.00	4.00	1.00	5.00	6.00	1.00	4.00	4.00	6.00	4.00	5.00
	5.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	6.00	7.00	4.00
	7.00	7.00	4.00	7.00	7.00	6.00	7.00	7.00	6.00	5.00	6.00
	2.00	6.00	4.00	5.00	5.00	4.00	5.00	4.00	6.00	5.00	1.00
50	7.00	7.00	4.00	4.00	7.00	7.00	7.00	5.00	7.00	5.00	....
	4.00	4.00	5.00	6.00	5.00	2.00	5.00	4.00	6.00	6.00	4.00
	7.00	7.00	1.00	2.00	6.00	7.00	4.00	1.00	7.00	....	....
	3.00	4.00	5.00	4.00	4.00	4.00	5.00	5.00	5.00	4.00	4.00
	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	5.00	6.00
55	7.00	7.00	2.00	2.00	7.00	7.00	7.00	2.00	7.00	2.00	3.00
	4.00	7.00	1.00	5.00	6.00	7.00	7.00	4.00	5.00	4.00	4.00
	6.00	7.00	6.00	7.00	7.00	6.00	5.00	7.00	7.00	7.00	3.00
	6.00	5.00	6.00	6.00	7.00	7.00	6.00	5.00	6.00	7.00	5.00
	6.00	6.00	4.00	5.00	5.00	5.00	6.00	7.00	6.00	7.00	2.00
60	4.00	5.00	2.00	3.00	5.00	7.00	4.00	3.00	6.00	5.00	4.00
	7.00	7.00	5.00	4.00	7.00	7.00	7.00	7.00	6.00	6.00	5.00
	7.00	7.00	5.00	1.00	7.00	7.00	7.00	3.00	7.00	7.00	4.00
	3.00	2.00	2.00	2.00	3.00	2.00	2.00	2.00	3.00	3.00	6.00
	4.00	2.00	5.00	3.00	1.00	1.00	4.00	1.00	2.00	3.00	4.00
65	6.00	7.00	5.00	6.00	6.00	7.00	7.00	4.00	5.00	5.00	6.00
	6.00	5.00	7.00	4.00	6.00	7.00	6.00	5.00	6.00	4.00	6.00
	6.00	5.00	7.00	4.00	5.00	5.00	1.00	3.00	1.00	2.00	3.00
	5.00	4.00	1.00	1.00	5.00	5.00	6.00	3.00	5.00	4.00	3.00
	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	5.00	6.00

70	6.00	6.00	6.00	5.00	7.00	7.00	4.00	5.00	7.00	6.00	5.00
	6.00	7.00	1.00	6.00	5.00	6.00	7.00	6.00	6.00	7.00	5.00
	6.00	7.00	1.00	6.00	5.00	6.00	7.00	6.00	6.00	7.00	5.00
	6.00	5.00	4.00	3.00	2.00	2.00	5.00	4.00	3.00	4.00	4.00
	5.00	7.00	4.00	1.00	5.00	6.00	7.00	7.00	6.00	3.00	4.00
75	4.00	4.00	5.00	4.00	4.00	4.00	3.00	5.00	4.00	4.00	4.00
	6.00	5.00	1.00	1.00	5.00	1.00	6.00	6.00	2.00	5.00	5.00
	5.00	7.00	5.00	5.00	4.00	4.00	5.00	3.00	3.00	4.00	4.00
	5.00	3.00	2.00	2.00	2.00	3.00	5.00	2.00	5.00	7.00	4.00
	6.00	6.00	3.00	2.00	6.00	5.00	7.00	4.00	3.00	5.00	3.00
80	5.00	4.00	2.00	5.00	5.00	5.00	6.00	5.00	5.00	4.00	6.00
	3.00	6.00	6.00	5.00	7.00	3.00	6.00	6.00	3.00	6.00	2.00
	2.00	3.00	1.00	1.00	6.00	6.00	3.00	2.00	6.00	2.00	5.00
	5.00	5.00	4.00	4.00	4.00	5.00	7.00	6.00	6.00	7.00	7.00
	6.00	7.00	3.00	5.00	6.00	6.00	6.00	7.00	4.00	5.00	4.00
85	6.00	6.00	1.00	4.00	5.00	6.00	7.00	6.00	7.00	7.00	6.00
	7.00	7.00	7.00	7.00	5.00	7.00	7.00	7.00	5.00	5.00	6.00
	7.00	6.00	7.00	6.00	6.00	7.00	7.00	6.00	5.00	6.00	5.00
	5.00	5.00	6.00	2.00	3.00	2.00	4.00	5.00	3.00	3.00	4.00
	4.00	7.00	7.00	5.00	7.00	1.00	7.00	7.00	1.00	4.00	4.00
90	2.00	4.00	3.00	1.00	1.00	2.00	4.00	4.00	1.00	1.00	1.00
	6.00	4.00	3.00	4.00	4.00	4.00	4.00	2.00	3.00	3.00	3.00
	7.00	7.00	7.00	5.00	3.00	3.00	7.00	7.00	4.00	4.00	4.00
	6.00	7.00	6.00	6.00	5.00	6.00	6.00	5.00	5.00	4.00	6.00
	7.00	7.00	6.00	6.00	6.00	5.00	7.00	7.00	6.00	5.00	6.00
95	5.00	5.00	2.00	5.00	3.00	4.00	4.00	6.00	4.00	2.00	4.00
	6.00	6.00	2.00	1.00	3.00	5.00	6.00	7.00	5.00	4.00	4.00
	1.00	6.00	1.00	11.00	6.00	1.00	7.00	4.00	7.00	7.00	4.00
	1.00	7.00	5.00	7.00	3.00	5.00	6.00	7.00	5.00	5.00	4.00
	1.00	7.00	4.00	1.00	7.00	4.00	7.00	4.00	3.00	6.00	7.00
100	7.00	7.00	7.00	5.00	7.00	7.00	7.00	6.00	7.00	7.00	4.00
	7.00	7.00	7.00	7.00	5.00	6.00	6.00	7.00	6.00	4.00	5.00
	7.00	7.00	7.00	4.00	6.00	7.00	7.00	7.00	7.00	6.00	7.00
	7.00	6.00	4.00	7.00	6.00	7.00	7.00	6.00	6.00	7.00	4.00
	2.00	2.00	3.00	1.00	5.00	6.00	5.00	5.00	6.00	6.00	2.00

105	7.00	6.00	5.00	4.00	4.00	4.00	7.00	5.00	4.00	5.00	4.00
	6.00	5.00	1.00	5.00	5.00	1.00	7.00	6.00	4.00	5.00	4.00
	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	6.00
	7.00	6.00	7.00	4.00	6.00	7.00	6.00	7.00	5.00	6.00	5.00
	5.00	5.00	4.00	4.00	6.00	6.00	4.00	3.00	4.00	2.00	3.00
110	6.00	6.00	7.00	7.00	6.00	5.00	6.00	6.00	7.00	6.00	4.00
	4.00	4.00	3.00	3.00	5.00	5.00	4.00	6.00	5.00	4.00	5.00
	7.00	6.00	7.00	7.00	4.00	7.00	7.00	7.00	4.00	5.00	4.00
	7.00	7.00	5.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	1.00
	5.00	3.00	2.00	1.00	5.00	5.00	4.00	5.00	4.00	3.00	4.00
115	5.00	4.00	3.00	5.00	3.00	1.00	5.00	6.00	3.00	2.00	3.00
	6.00	5.00	5.00	7.00	7.00	6.00	5.00	7.00	6.00	3.00	3.00
	2.00	3.00	1.00	1.00	2.00	2.00	5.00	2.00	3.00	1.00	1.00
	6.00	7.00	4.00	7.00	7.00	6.00	7.00	5.00	6.00	6.00	2.00
	6.00	5.00	6.00	5.00	4.00	3.00	6.00	6.00	4.00	3.00	4.00
120	7.00	7.00	6.00	7.00	7.00	7.00	7.00	6.00	5.00	4.00	4.00
	3.00	6.00	6.00	6.00	7.00	7.00	7.00	5.00	6.00	5.00	5.00
	7.00	7.00	4.00	4.00	5.00	4.00	5.00	1.00	4.00	5.00	5.00
	3.00	4.00	5.00	3.00	4.00	1.00	6.00	3.00	3.00	4.00	4.00
	4.00	4.00	1.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	2.00
125	3.00	3.00	3.00	1.00	1.00	1.00	3.00	2.00	3.00	2.00	5.00
	6.00	4.00	3.00	5.00	5.00	6.00	6.00	6.00	6.00	3.00	3.00
	5.00	2.00	2.00	3.00	5.00	4.00	4.00	4.00	5.00	3.00	3.00
	6.00	3.00	4.00	2.00	5.00	5.00	5.00	3.00	5.00	2.00	3.00
	2.00	3.00	4.00	6.00	5.00	1.00	1.00	7.00	4.00	3.00	5.00
130	1.00	1.00	1.00	2.00	1.00	1.00	2.00	2.00	1.00	1.00	1.00
	7.00	7.00	4.00	4.00	5.00	5.00	6.00	5.00	5.00	4.00	5.00
	3.00	6.00	6.00	1.00	3.00	5.00	6.00	2.00	2.00	1.00	2.00
	1.00	6.00	1.00	1.00	6.00	4.00	1.00	1.00	1.00	2.00	1.00
	7.00	7.00	7.00	4.00	6.00	6.00	7.00	7.00	4.00	5.00	1.00
135	7.00	7.00	7.00	6.00	3.00	3.00	7.00	7.00	7.00	1.00	7.00
	6.00	5.00	7.00	7.00	5.00	7.00	7.00	6.00	5.00	4.00	5.00
	1.00	1.00	6.00	6.00	5.00	4.00	2.00	3.00	5.00	1.00	3.00
	5.00	7.00	4.00	4.00	5.00	5.00	7.00	7.00	6.00	4.00	4.00
	7.00	6.00	3.00	1.00	6.00	6.00	6.00	3.00	6.00	5.00	3.00
140	5.00	5.00	1.00	1.00	3.00	4.00	3.00	7.00	3.00	4.00	4.00
	6.00	6.00	6.00	3.00	6.00	7.00	7.00	5.00	6.00	6.00	7.00
	3.00	4.00	2.00	4.00	2.00	5.00	5.00	3.00	6.00	2.00	4.00

The McGraw·Hill Companies

PRODRAD	PRODNON	PROMORAD	PROMONON	DISTNRAD	DISTNNON	CUSTDRIV	COMPDRIV	OBJDRIV	INDUSTRY	AGE	LISTED
4.00	3.00	3.00	4.00	4.00	3.00	3.00	4.00	2.00	2	2	1
3.00	3.00	4.00	5.00	2.00	2.00	3.00	4.00	3.00	3	4	1
4.00	3.00	5.00	4.00	5.00	3.00	5.00	5.00	4.00	3	4	2
6.00	6.00	6.00	6.00	5.00	5.00	4.00	2.00	4.00	3	4	1
5.00	4.00	7.00	5.00	5.00	4.00	6.00	5.00	4.00	2	2	1
4.00	4.00	5.00	5.00	5.00	6.00	6.00	3.00	4.00	3	2	1
2.00	4.00	2.00	3.00	3.00	4.00	2.00	2.00	4.00	3	2	1
2.00	3.00	3.00	3.00	1.00	2.00	2.00	2.00	1.00	1	4	1
5.00	4.00	5.00	4.00	5.00	4.00	6.00	3.00	5.00	3	2	1
5.00	4.00	6.00	4.00	6.00	4.00	2.00	4.00	2.00	1	4	1
4.00	3.00	5.00	3.00	5.00	3.00	3.00	2.00	2.00	3	4	1
5.00	4.00	5.00	3.00	5.00	4.00	3.00	2.00	2.00	3	1	1
6.00	4.00	4.00	2.00	5.00	3.00	2.00	4.00	5.00	3	2	2
2.00	3.00	4.00	4.00	3.00	4.00	2.00	5.00	5.00	3	1	2
4.00	4.00	7.00	4.00	5.00	3.00	5.00	3.00	3.00	3	3	1
4.00	3.00	6.00	5.00	7.00	6.00	3.00	3.00	2.00	3	4	1
3.00	4.00	7.00	5.00	7.00	6.00	2.00	2.00	4.00	3	4	1
6.00	5.00	3.00	2.00	5.00	3.00	5.00	2.00	2.00	3	4	2
3.00	2.00	5.00	3.00	5.00	3.00	4.00	5.00	2.00	3	4	1
4.00	2.00	1.00	6.00	6.00	6.00	6.00	2.00	3.00	4	4	1
4.00	3.00	4.00	4.00	3.00	3.00	2.00	3.00	2.00	3	4	1
1.00	5.00	1.00	5.00	1.00	5.00	1.00	1.00	1.00	3	4	1
5.00	1.00	4.00	1.00	4.00	1.00	3.00	3.00	3.00	3	4	1
3.00	5.00	2.00	2.00	3.00	4.00	2.00	2.00	3.00	3	4	1
1.00	2.00	2.00	2.00	1.00	2.00	3.00	3.00	1.00	3	1	2
1.00	3.00	2.00	2.00	1.00	6.00	2.00	6.00	1.00	1	1	1
3.00	5.00	2.00	4.00	2.00	4.00	3.00	3.00	3.00	3	1	1
6.00	2.00	1.00	7.00	1.00	7.00	2.00	5.00	4.00	2	1	1
5.00	6.00	2.00	6.00	2.00	6.00	7.00	5.00	6.00	1	4	1
7.00	6.00	4.00	6.00	6.00	6.00	7.00	3.00	6.00	3	4	1
7.00	3.00	4.00	5.00	4.00	5.00	6.00	2.00	5.00	3	3	1
6.00	2.00	4.00	6.00	3.00	7.00	5.00	7.00	7.00	1	2	1
5.00	2.00	6.00	3.00	5.00	3.00	6.00	4.00	7.00	1	4	1

3.00	5.00	3.00	5.00	3.00	4.00	4.00	6.00	5.00	1	4	1
7.00	6.00	6.00	6.00	7.00	7.00	6.00	5.00	6.00	1	4	1
5.00	4.00	6.00	5.00	6.00	4.00	4.00	6.00	5.00	2	4	1
6.00	2.00	6.00	3.00	5.00	2.00	7.00	6.00	7.00	2	4	1
3.00	5.00	3.00	3.00	3.00	3.00	3.00	4.00	4.00	2	4	2
4.00	6.00	4.00	4.00	6.00	6.00	5.00	5.00	1.00	1	4	1
5.00	7.00	5.00	5.00	6.00	4.00	6.00	1.00	7.00	2	4	1
5.00	4.00	4.00	4.00	3.00	3.00	4.00	4.00	6.00	2	4	1
4.00	5.00	7.00	3.00	5.00	5.00	6.00	4.00	7.00	2	3	2
1.00	6.00	3.00	2.00	2.00	2.00	5.00	1.00	7.00	3	4	2
2.00	6.00	6.00	2.00	5.00	2.00	5.00	5.00	7.00	1	4	1
4.00	4.00	4.00	4.00	6.00	2.00	6.00	6.00	7.00	2	3	2
6.00	6.00	4.00	4.00	7.00	....	5.00	2.00	7.00	3	4	1
7.00	2.00	6.00	2.00	2.00	3.00	5.00	6.00	6.00	4	2	2
5.00	6.00	4.00	6.00	1.00	5.00	6.00	5.00	5.00	4	4	1
6.00	6.00	6.00	5.00	6.00	6.00	6.00	6.00	6.00	4	4	1
7.00	5.00	4.00	6.00	5.00	4.00	5.00	6.00	5.00	3	2	1
7.00	1.00	2.00	1.00	5.00	5.00	4.00	7.00	3.00	3	4	2
1.00	2.00	4.00	5.00	3.00	3.00	7.00	7.00	6.00	3	2	2
5.00	4.00	5.00	4.00	5.00	4.00	6.00	7.00	7.00	4	4	1
5.00	2.00	2.00	1.00	1.00	4.00	7.00	4.00	7.00	3	1	1
2.00	3.00	2.00	2.00	4.00	4.00	6.00	5.00	5.00	3	1	2
4.00	5.00	3.00	3.00	4.00	5.00	5.00	4.00	4.00	3	4	1
7.00	4.00	7.00	4.00	7.00	4.00	4.00	4.00	4.00	4	4	1
3.00	3.00	5.00	4.00	5.00	5.00	5.00	2.00	7.00	1	4	1
5.00	4.00	7.00	3.00	7.00	3.00	7.00	6.00	7.00	3	1	2
6.00	3.00	3.00	5.00	7.00	3.00	5.00	3.00	7.00	1	4	2
5.00	5.00	4.00	4.00	4.00	4.00	5.00	5.00	5.00	4	2	1
5.00	5.00	4.00	5.00	5.00	5.00	6.00	5.00	5.00	4	4	1
4.00	4.00	3.00	3.00	6.00	6.00	7.00	3.00	6.00	1	2	2
5.00	4.00	5.00	4.00	6.00	4.00	3.00	5.00	5.00	2	4	1
6.00	3.00	6.00	3.00	4.00	4.00	6.00	6.00	7.00	3	2	2
6.00	4.00	6.00	4.00	4.00	4.00	7.00	5.00	7.00	1	4	2
4.00	3.00	6.00	5.00	3.00	4.00	7.00	6.00	3.00	3	4	2

4.00	4.00	1.00	5.00	4.00	4.00	5.00	5.00	7.00	4	2	2
6.00	5.00	5.00	6.00	6.00	5.00	7.00	6.00	7.00	3	2	1
5.00	4.00	7.00	3.00	6.00	3.00	6.00	6.00	2.00	3	2	2
7.00	6.00	6.00	5.00	4.00	5.00	6.00	7.00	7.00	3	1	1
4.00	6.00	2.00	6.00	6.00	2.00	1.00	2.00	7.00	2	4	1
7.00	....	7.00	....	7.00	....	7.00	4.00	7.00	1	4	1
5.00	5.00	2.00	2.00	4.00	5.00	4.00	7.00	4.00	4	3	1
7.00	6.00	7.00	1.00	7.00	1.00	7.00	5.00	6.00	4	3	1
6.00	6.00	7.00	7.00	6.00	6.00	6.00	5.00	7.00	3	1	1
6.00	2.00	7.00	4.00	5.00	4.00	7.00	4.00	6.00	3	3	2
6.00	....	7.00	....	....	....	7.00	....	7.00	3	1	2
5.00	6.00	6.00	4.00	6.00	4.00	5.00	5.00	6.00	1	4	1
....	....	....	....	....	....	....	1.00	7.00	4	4	1
6.00	6.00	4.00	4.00	4.00	4.00	4.00	7.00	7.00	3	2	2
7.00	6.00	5.00	6.00	7.00	6.00	7.00	6.00	7.00	3	4	1
1.00	2.00	11.00	1.00	2.00	3.00	7.00	1.00	7.00	4	3	1
4.00	4.00	5.00	4.00	4.00	4.00	6.00	6.00	6.00	1	4	1
7.00	3.00	7.00	3.00	7.00	3.00	7.00	6.00	7.00	3	1	1
5.00	5.00	6.00	5.00	6.00	5.00	1.00	7.00	6.00	3	4	2
4.00	4.00	4.00	4.00	4.00	44.00	5.00	3.00	6.00	4	4	1
6.00	4.00	5.00	4.00	7.00	6.00	4.00	7.00	7.00	1	2	2
5.00	5.00	6.00	5.00	6.00	5.00	5.00	5.00	6.00	1	4	1
7.00	4.00	7.00	6.00	4.00	7.00	7.00	7.00	1.00	1	1	2
2.00	6.00	2.00	6.00	6.00	3.00	3.00	5.00	3.00	4	2	2
4.00	2.00	1.00	6.00	1.00	6.00	4.00	2.00	4.00	4	1	2
5.00	6.00	5.00	6.00	5.00	6.00	6.00	7.00	4.00	4	2	2
4.00	6.00	4.00	6.00	4.00	4.00	6.00	5.00	5.00	3	1	2
4.00	2.00	5.00	4.00	3.00	6.00	2.00	6.00	2.00	1	4	2
3.00	5.00	5.00	4.00	5.00	3.00	4.00	6.00	4.00	4	1	2
5.00	7.00	4.00	7.00	4.00	6.00	7.00	5.00	5.00	3	1	2
5.00	5.00	5.00	5.00	5.00	5.00	6.00	4.00	6.00	1	3	2
3.00	5.00	3.00	6.00	3.00	6.00	7.00	5.00	3.00	3	2	2
3.00	5.00	3.00	6.00	3.00	6.00	7.00	5.00	3.00	3	2	2
4.00	5.00	5.00	4.00	2.00	3.00	5.00	3.00	4.00	4	3	2
5.00	4.00	4.00	4.00	5.00	3.00	5.00	3.00	5.00	4	1	1

2.00	5.00	3.00	4.00	4.00	4.00	3.00	5.00	4.00	4	3	2
5.00	5.00	6.00	5.00	2.00	6.00	7.00	3.00	7.00	3	1	2
5.00	6.00	6.00	4.00	5.00	4.00	6.00	6.00	6.00	4	1	1
7.00	4.00	5.00	4.00	3.00	5.00	5.00	4.00	3.00	2	3	2
3.00	4.00	6.00	3.00	4.00	4.00	6.00	5.00	4.00	4	4	1
6.00	6.00	5.00	5.00	6.00	5.00	5.00	5.00	6.00	4	4	1
5.00	3.00	3.00	6.00	2.00	5.00	6.00	5.00	5.00	3	4	1
4.00	4.00	3.00	3.00	2.00	2.00	3.00	2.00	5.00	1	2	1
6.00	....	6.00	....	6.00	....	7.00	7.00	7.00	3	1	2
6.00	4.00	6.00	3.00	....	....	6.00	5.00	6.00	3	1	1
5.00	6.00	2.00	5.00	2.00	4.00	6.00	5.00	6.00	4	4	1
5.00	7.00	6.00	5.00	6.00	6.00	6.00	4.00	6.00	3	3	1
6.00	3.00	4.00	6.00	4.00	4.00	5.00	6.00	3.00	3	2	1
4.00	5.00	3.00	3.00	2.00	2.00	1.00	4.00	6.00	3	3	1
5.00	4.00	7.00	4.00	7.00	4.00	4.00	1.00	7.00	3	4	1
1.00	1.00	2.00	3.00	4.00	4.00	4.00	5.00	6.00	3	2	2
3.00	6.00	2.00	7.00	3.00	3.00	4.00	5.00	5.00	3	1	1
5.00	5.00	4.00	4.00	....	....	5.00	4.00	6.00	3	1	1
5.00	6.00	5.00	6.00	5.00	6.00	7.00	3.00	6.00	4	4	1
5.00	7.00	5.00	6.00	6.00	6.00	7.00	6.00	7.00	3	1	2
2.00	7.00	3.00	6.00	3.00	6.00	4.00	5.00	5.00	3	4	2
6.00	5.00	3.00	3.00	5.00	6.00	7.00	6.00	5.00	3	1	2
7.00	4.00	7.00	4.00	7.00	4.00	7.00	1.00	7.00	3	1	2
6.00	4.00	4.00	5.00	6.00	4.00	6.00	4.00	3.00	3	2	2
5.00	4.00	5.00	4.00	4.00	4.00	7.00	5.00	7.00	3	4	2
7.00	4.00	7.00	4.00	6.00	4.00	7.00	5.00	7.00	3	2	2
4.00	5.00	4.00	5.00	4.00	5.00	6.00	5.00	6.00	3	1	2
4.00	7.00	7.00	4.00	6.00	6.00	7.00	7.00	7.00	3	1	1
7.00	2.00	7.00	4.00	7.00	2.00	6.00	5.00	6.00	1	4	1
2.00	6.00	2.00	5.00	2.00	4.00	6.00	2.00	6.00	3	1	2
4.00	4.00	6.00	5.00	4.00	4.00	5.00	4.00	4.00	4	3	1
5.00	4.00	5.00	4.00	6.00	4.00	5.00	2.00	7.00	3	4	2
7.00	6.00	77.00	5.00	7.00	6.00	7.00	7.00	7.00	3	4	1
5.00	6.00	5.00	6.00	6.00	4.00	6.00	4.00	6.00	3	1	2

3.00	4.00	2.00	2.00	5.00	4.00	6.00	4.00	4.00	1	3	2
6.00	4.00	6.00	4.00	5.00	3.00	6.00	6.00	7.00	3	1	2
5.00	5.00	3.00	3.00	5.00	4.00	5.00	5.00	3.00	4	2	2
7.00	1.00	7.00	1.00	3.00	3.00	4.00	4.00	6.00	3	1	2
7.00	1.00	7.00	1.00	7.00	1.00	7.00	7.00	5.00	4	4	2
3.00	4.00	5.00	3.00	4.00	3.00	5.00	3.00	4.00	3	2	1
1.00	1.00	1.00	1.00	2.00	2.00	4.00	4.00	4.00	3	4	1
3.00	3.00	1.00	1.00	1.00	1.00	5.00	3.00	6.00	3	1	1
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	4	3	2
6.00	2.00	7.00	1.00	4.00	4.00	5.00	5.00	7.00	3	1	2
5.00	3.00	5.00	2.00	5.00	3.00	5.00	3.00	5.00	3	3	1
7.00	6.00	7.00	7.00	7.00	6.00	7.00	7.00	7.00	4	4	1
3.00	4.00	4.00	6.00	5.00	5.00	6.00	6.00	5.00	4	4	2
6.00	5.00	4.00	4.00	7.00	7.00	4.00	5.00	5.00	3	3	1
4.00	4.00	6.00	4.00	4.00	4.00	4.00	7.00	4.00	3	2	1
4.00	2.00	4.00	2.00	4.00	2.00	2.00	4.00	4.00	3	2	2
4.00	5.00	2.00	3.00	3.00	5.00	3.00	5.00	3.00	3	4	1
3.00	4.00	3.00	3.00	1.00	1.00	5.00	4.00	5.00	3	3	2
3.00	3.00	5.00	4.00	4.00	4.00	4.00	5.00	4.00	4	4	1
5.00	4.00	3.00	3.00	3.00	3.00	3.00	4.00	6.00	4	3	2
3.00	5.00	6.00	4.00	3.00	2.00	4.00	2.00	5.00	3	4	2
7.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	3.00	3	2	2
4.00	5.00	5.00	4.00	5.00	4.00	5.00	4.00	5.00	4	4	2
2.00	5.00	2.00	4.00	2.00	3.00	5.00	3.00	5.00	4	2	2
1.00	2.00	7.00	1.00	....	....	....	....	....	4	2	2
7.00	2.00	7.00	2.00	6.00	2.00	7.00	1.00	7.00	3	1	4
7.00	1.00	7.00	1.00	7.00	1.00	7.00	1.00	7.00	3	2	1
4.00	6.00	4.00	5.00	4.00	6.00	6.00	6.00	6.00	4	4	2
2.00	4.00	5.00	4.00	4.00	4.00	3.00	6.00	1.00	3	4	2
4.00	6.00	4.00	4.00	4.00	4.00	7.00	7.00	7.00	3	4	1
6.00	2.00	6.00	2.00	6.00	2.00	3.00	6.00	7.00	3	2	2
3.00	3.00	2.00	3.00	7.00	1.00	7.00	3.00	7.00	4	1	2
7.00	5.00	5.00	6.00	2.00	6.00	7.00	6.00	7.00	1	3	2
5.00	5.00	1.00	2.00	4.00	3.00	4.00	6.00	2.00	4	4	1

TYPE	COSHARE	DIVSHARE	COPROFIT	DIVPROFI	COSALGRO	DVSALGRO	NONRAD4P	RADICL4P
2	4.00	5.0	3.00	5.00	4.00	4.00	3.25	4.00
6	6.00	5.0	5.00	5.00	5.00	5.00	3.25	3.00
6	4.00	3.0	5.00	5.00	4.00	4.00	3.50	4.25
1	5.00	5.0	6.00	6.00	3.00	4.00	5.50	5.50
2	3.00	5.0	1.00	3.00	1.00	4.00	4.00	5.25
2	5.00	6.0	5.00	5.00	4.00	6.00	4.75	4.50
3	2.00	3.0	2.00	3.00	5.00	5.00	4.00	3.25
3	5.00	5.0	6.00	6.00	5.00	5.00	2.75	2.00
2	5.00	4.0	3.00	4.00	5.00	5.00	4.00	5.00
2	2.00	5.0	5.00	5.00	5.00	6.00	4.25	5.00
6	7.00	6.0	6.00	5.00	7.00	5.00	3.00	5.00
1	3.00	4.0	5.00	5.00	5.00	5.00	3.50	5.00
5	4.00	4.0	5.00	5.00	5.00	4.00	2.75	4.75
5	5.00	4.0	3.00	3.00	4.00	4.00	4.00	3.50
6	6.00	7.0	2.00	3.00	5.00	6.00	4.00	4.50
6	6.00	6.0	6.00	5.00	6.00	6.00	4.00	5.00
2	2.00	3.0	6.00	6.00	1.00	3.00	5.25	5.50
2	4.00	5.0	3.00	3.00	3.00	3.00	3.25	5.00
6	6.00	6.0	6.00	6.00	6.00	6.00	2.75	4.50
2	2.00	2.0	1.00	1.00	5.00	5.00	4.00	3.75
6	6.00	6.0	5.00	....	....	....	3.75	4.25
6	5.00	6.0	7.00	6.00	6.00	7.00	5.00	1.00
6	1.00	2.0	1.00	3.00	2.00	3.00	1.00	4.50
2	3.00	2.0	2.00	1.00	3.00	3.00	3.00	2.25
1	4.00	5.0	5.00	5.00	6.00	6.00	2.50	2.00
3	7.00	....	6.00	....	7.00	....	4.25	1.50
2	4.00	4.0	3.00	4.00	4.00	2.00	4.25	2.00
1	7.00	7.0	7.00	7.00	6.00	6.00	5.50	2.25
4	7.00	7.0	5.00	7.00	7.00	7.00	5.75	2.50
4	7.00	5.0	6.00	5.00	5.00	5.00	6.00	5.50
4	7.00	7.0	7.00	7.00	6.00	6.00	4.75	4.50
4	6.00	6.0	7.00	6.00	5.00	5.00	4.50	5.00
4	5.00	4.0	7.00	7.00	3.00	3.00	3.50	5.50

4	3.00	4.0	2.00	3.00	2.00	4.00	4.25	3.00
4	5.00	5.0	5.00	6.00	4.00	5.00	6.00	6.50
2	6.00	6.0	7.00	7.00	7.00	7.00	4.00	5.50
2	7.00	7.0	7.00	7.00	7.00	7.00	2.25	5.50
1	4.00	4.0	4.00	4.00	3.00	3.00	3.25	2.75
2	5.00	5.0	5.00	5.00	5.00	5.00	5.50	5.00
2	5.00	6.0	6.00	6.00	6.00	6.00	5.75	5.00
2	4.00	....	6.00	....	4.00	....	4.00	4.25
2	7.00	7.0	7.00	7.00	7.00	7.00	4.50	5.00
3	5.00	....	3.00	....	5.00	....	2.75	1.75
4	5.00	5.0	5.00	6.00	6.00	5.00	3.00	4.75
4	4.00	4.0	5.00	5.00	4.00	4.00	3.75	4.50
4	7.00	6.0	6.00	6.00	6.00	5.00	5.33	5.25
4	6.00	....	7.00	....	7.00	....	2.50	5.25
2	4.00	6.0	5.00	6.00	4.00	5.00	5.25	3.75
2	6.00	6.0	6.00	6.00	5.00	5.00	5.75	6.00
2	6.00	5.0	7.00	7.00	7.00	6.00	4.75	5.25
4	6.00	7.0	7.00	6.00	7.00	6.00	3.50	4.50
2	6.00	4.0	5.00	4.00	4.00	3.00	3.75	3.50
2	5.00	5.0	6.00	6.00	5.00	5.00	4.00	5.00
1	7.00	7.0	7.00	7.00	7.00	7.00	3.25	3.00
5	3.00	3.0	2.00	3.00	3.00	2.00	3.50	2.50
4	5.00	4.0	6.00	5.00	5.00	4.00	4.50	3.75
3	7.00	5.0	6.00	6.00	7.00	6.00	4.00	6.25
4	4.00	4.0	4.00	4.00	4.00	4.00	3.75	5.00
5	7.00	5.0	6.00	4.00	5.00	5.00	3.25	6.50
1	5.00	....	6.00	....	5.00	....	4.00	5.00
4	3.00	3.0	3.00	3.00	3.00	3.00	4.50	4.50
2	5.00	4.0	3.00	3.00	3.00	4.00	5.00	5.00
1	7.00	....	7.00	....	6.00	....	4.50	4.50
2	5.00	6.0	5.00	5.00	2.00	2.00	4.00	5.25
4	4.00	4.0	5.00	5.00	5.00	5.00	4.00	5.50
1	5.00	6.0	7.00	7.00	7.00	7.00	4.00	5.50
4	6.00	5.0	7.00	6.00	6.00	6.00	4.50	5.00

1	4.00	4.0	4.00	4.00	4.00	3.00	4.50	3.25
2	7.00	7.0	6.00	6.00	6.00	6.00	5.25	6.00
1	5.00	4.0	4.00	4.00	3.00	2.00	3.50	5.25
2	7.00	7.0	5.00	4.00	6.00	5.00	5.50	6.00
1	5.00	2.0	5.00	2.00	5.00	2.00	5.00	4.00
4	7.00	7.0	7.00	7.00	6.00	6.00	....	7.00
2	6.00	6.0	5.00	5.00	6.00	6.00	4.25	3.75
3	4.00	6.0	5.00	5.00	6.00	6.00	3.00	7.00
5	4.00	5.0	3.00	4.00	7.00	7.00	6.25	6.00
2	6.00	5.0	6.00	5.00	7.00	7.00	2.75	5.75
1	5.00	....	5.00	....	5.00	....	....	6.00
2	3.00	3.0	4.00	3.00	3.00	5.00	4.50	5.75
2	5.00	....	3.00	....	6.00	....	....	....
5	6.00	5.0	6.00	5.00	6.00	6.00	4.50	4.50
4	6.00	6.0	7.00	7.00	7.00	7.00	6.00	6.00
4	4.00	4.0	4.00	4.00	7.00	7.00	2.25	4.00
1	5.00	5.0	7.00	7.00	5.00	5.00	4.00	4.25
5	7.00	6.0	6.00	6.00	5.00	5.00	3.00	7.00
2	6.00	6.0	6.00	6.00	7.00	6.00	5.00	6.00
2	4.00	5.0	4.00	4.00	4.00	4.00	13.50	4.75
2	6.00	7.0	4.00	5.00	5.00	5.00	4.50	5.75
4	6.00	6.0	6.00	6.00	5.00	5.00	5.00	5.75
1	7.00	7.0	....	4.00	7.00	7.00	5.25	6.25
6	6.00	5.0	6.00	6.00	6.00	6.00	5.25	3.25
6	7.00	6.0	5.00	5.00	6.00	6.00	4.50	2.25
1	5.00	5.0	5.00	4.00	5.00	5.00	6.00	5.00
1	5.00	5.0	5.00	5.00	5.00	5.00	5.50	4.00
3	4.00	4.0	1.00	2.00	1.00	5.00	3.75	3.50
1	6.00	5.0	5.00	4.00	6.00	6.00	3.75	4.25
4	5.00	7.0	2.00	2.00	7.00	7.00	6.50	4.50
2	7.00	7.0	7.00	6.00	6.00	6.00	5.00	5.25
1	6.00	4.0	6.00	4.00	6.00	5.00	5.50	4.00
1	6.00	6.0	6.00	5.00	6.00	6.00	5.50	4.00
5	6.00	6.0	3.00	4.00	4.00	5.00	4.00	3.75

4	5.00	5.0	4.00	4.00	7.00	7.00	3.75	4.25
5	4.00	3.0	4.00	2.00	4.00	3.00	4.25	3.25
4	6.00	6.0	3.00	3.00	5.00	4.00	5.25	4.50
3	4.00	3.0	4.00	5.00	3.00	3.00	4.50	5.00
1	4.00	4.0	3.00	3.00	4.00	4.00	4.25	5.50
4	6.00	5.0	5.00	5.00	6.00	6.00	3.50	4.50
4	6.00	6.0	5.00	5.00	5.00	5.00	5.50	5.25
3	6.00	7.0	6.00	3.00	6.00	6.00	4.00	4.00
4	4.00	4.0	5.00	6.00	6.00	5.00	3.50	2.75
5	6.00	....	5.00	....	6.00	6.00	7.00	6.25
5	7.00	7.0	6.00	6.00	6.00	6.00	3.67	5.67
4	6.00	6.0	7.00	7.00	7.00	6.00	5.25	4.00
2	6.00	5.0	5.00	5.00	5.00	5.00	6.00	5.50
2	5.00	6.0	4.00	4.00	6.00	6.00	4.50	5.00
2	1.00	3.0	1.00	2.00	1.00	1.00	3.50	3.00
2	7.00	7.0	4.00	4.00	5.00	5.00	4.00	5.75
2	4.00	3.0	4.00	4.00	4.00	4.00	2.25	2.00
5	3.00	4.0	3.00	4.00	5.00	5.00	4.75	2.75
2	4.00	5.0	5.00	5.00	5.00	5.00	4.33	4.33
2	7.00	7.0	4.00	4.00	6.00	6.00	6.00	4.75
1	6.00	6.0	6.00	7.00	7.00	7.00	6.25	5.25
1	6.00	6.0	4.00	5.00	4.00	4.00	5.75	2.50
1	3.00	....	3.00	....	3.00	....	4.50	4.50
1	7.00	5.0	7.00	7.00	7.00	5.00	4.00	7.00
2	4.00	4.0	6.00	7.00	7.00	6.00	4.25	5.25
3	4.00	4.0	2.00	1.00	4.00	2.00	4.75	5.00
2	5.00	5.0	4.00	5.00	6.00	6.00	4.00	6.75
1	....	....	....	....	....	....	5.00	4.00
2	4.00	4.0	4.00	4.00	4.00	5.00	6.00	5.75
2	7.00	6.0	6.00	6.00	7.00	6.00	3.00	7.00
1	3.00	3.0	3.00	3.00	3.00	3.00	4.25	3.00
5	3.00	3.0	4.00	4.00	4.00	4.00	4.25	4.75
3	7.00	....	7.00	....	7.00	....	4.00	5.25
2	7.00	7.0	7.00	7.00	6.00	6.00	5.75	24.50
5	7.00	7.0	4.00	4.00	6.00	6.00	5.25	5.50

1	5.00	4.0	3.00	3.00	4.00	4.00	3.25	3.00
4	5.00	5.0	5.00	5.00	6.00	5.00	3.75	5.75
4	4.00	3.0	3.00	4.00	3.00	3.00	4.25	4.25
2	....	....	....	....	6.00	....	2.25	5.50
2	5.00	5.0	5.00	5.00	6.00	6.00	1.00	7.00
1	2.00	3.0	3.00	3.00	3.00	3.00	3.50	3.75
2	2.00	4.0	2.00	3.00	2.00	3.00	1.75	1.50
2	6.00	3.0	5.00	2.00	6.00	3.00	2.00	2.00
1	6.00	6.0	6.00	6.00	7.00	7.00	1.00	1.00
4	6.00	....	6.00	....	7.00	....	2.25	5.75
2	6.00	4.0	5.00	4.00	4.00	5.00	3.00	4.50
3	1.00	1.0	7.00	7.00	7.00	7.00	5.75	6.25
1	5.00	....	4.00	....	4.00	....	5.00	4.25
2	1.00	3.0	1.00	6.00	1.00	4.00	5.25	5.50
2	6.00	....	4.00	....	5.00	....	4.00	4.50
2	4.00	4.0	4.00	4.00	4.00	4.00	2.00	4.00
3	2.00	2.0	2.00	2.00	2.00	2.00	4.50	2.75
1	5.00	5.0	5.00	5.00	4.00	4.00	2.75	2.50
2	3.00	....	4.00	....	....	....	3.50	3.75
1	6.00	6.0	2.00	2.00	5.00	5.00	3.25	3.25
3	7.00	7.0	7.00	7.00	7.00	5.00	4.00	3.75
3	1.00	1.0	1.00	1.00	1.00	1.00	1.00	2.50
3	6.00	4.0	5.00	4.00	5.00	4.00	4.50	4.50
1	6.00	6.0	4.00	5.00	6.00	6.00	3.50	1.75
2	1.00	1.0	1.00	1.00	1.00	1.00	1.33	3.33
....	5.00	5.0	5.00	5.00	6.00	6.00	1.75	6.25
4	5.00	6.0	6.00	6.00	6.00	5.00	2.50	5.50
3	4.00	4.0	4.00	4.00	3.00	4.00	5.50	4.00
2	5.00	6.0	7.00	7.00	6.00	6.00	3.75	3.00
2	4.00	1.0	4.00	1.00	4.00	1.00	4.50	4.00
1	1.00	1.0	6.00	6.00	1.00	1.00	2.25	5.75
4	5.00	4.0	5.00	6.00	4.00	4.00	2.75	4.00
5	6.00	6.0	6.00	6.00	6.00	6.00	6.00	5.00
2	4.00	4.0	3.00	3.00	5.00	5.00	3.50	3.00

ASSIGNMENT QUESTIONS

1. Compare Exploratory Factor Analysis and Confirmatory Factor Analysis. Which are the applications suitable for each method?
2. Discuss the application of SEM in a marketing problem situation of your choice.
3. Set up a model of SEM and explain the latent variables/constructs. Then come up with a list of 3-4 observed variables for each latent variable in your model.
4. If you can, collect data on interval scale on all the observed variables from a relevant sample, and test your model. Compute at least the GFI, RMS and path coefficients. Determine whether the model is a good fit.

The McGraw·Hill Companies

PART 3 Appendices

- Appendix 1 Industrial Marketing Research
- Appendix 2 Careers in Marketing Research

APPENDIX 1 INDUSTRIAL MARKETING RESEARCH

Industrial marketing research generally receives less attention as compared to consumer research. But it is a fairly active field of research. In post-liberalisation India, industrial marketers are increasingly turning to research in order to gauge customer likes and dislikes, to get inputs for product design or redesign, and to understand satisfaction levels with respect to different elements of their marketing and sales programs. This appendix highlights some of the key elements of doing industrial marketing research and in what ways it differs from consumer research.

DEFINING THE TARGET POPULATION

The first major difference between consumer markets and industrial markets is the difficulty in pinpointing who should be the respondent. In most consumer research, the respondent (and therefore, the target for research) is well-known as a single person, or a household, represented by one decision-maker. In many industrial research situations, it may be multiple persons in a company who are responsible for the purchase, use, and decisions about a product. Take for example, a pump manufacturer who sells to the petroleum refinery. Should he talk to the buyer, Production Manager, Maintenance Manager, or the General Manager, to get inputs into satisfaction levels about his pump? Or does he need to interview all of them? In case the answer is 'all of them', whose answers are more reliable in the case of conflicting opinions being expressed by some of the people mentioned above?

APPLICATIONS

Traditionally, industrial marketing research focussed more on demand estimation. But currently, many of the areas typically reserved for consumer research have been invaded, as the examples which follow, suggest. They are mainly research studies in the field of industrial products/markets.

1. Customer satisfaction measurement by Kirloskar Oil Engines Ltd., a manufacturer of diesel engines.
2. Demand estimation for Tungsten Carbide Cutting Tools, by WIDIA (India) Ltd.
3. Gauging areas of satisfaction and dissatisfaction from customers, by an industrial pump manufacturer.

4. Product redesign including redefining the price-value proposition, by a leading manufacturer of machine tools.
5. Estimation of demand and the factors driving it, by a public sector company considering manufacture of Flue Gas Desulphurisation equipment for pollution control.
6. Testing consumer awareness and preference for Sulphate Resistant Cement by a leading cement manufacturer.
7. Ranking of purchase criteria and competitive rating of different suppliers on these criteria, by a tyre cord manufacturer.
8. Competitive benchmarking by a leading engineering and construction conglomerate.

As these examples show, the boundaries of industrial marketing research are expanding in India. Generally, we would expect the amount of research done to grow in direct proportion to the level of competitive activity in the markets. This has been the case in India after 1991, when major international brands entered, both in consumer and industrial markets.

WHO DOES INDUSTRIAL MARKETING RESEARCH?

Related to the problem of defining the target population in industrial research, is the problem of who will do the data collection. Since the target population usually consists of managers or executives who may be in different functional areas, technically qualified and extremely busy, they need to be met usually with prior appointments. This is unlike most consumer studies where the consumer is more readily accessible at home or in a shopping area. Also, the respondent in most consumer studies is easily substitutable with another. For example, if one toothpaste buyer is not available, the researcher has the option to meet a similar respondent (similar in profile of age, gender, income, and product usage) in the neighbourhood. By contrast, industrial respondents may be limited in number, and the inability to interview one or two may mean that the study is jeopardised or delayed. This is because another respondent company of a similar profile may not be easy to access, if it exists at all. In other words, substitution is not easy for an industrial research respondent.

It could also be the expectation of the industry executive or executives who respond, that their interviewer be knowledgeable about the product for which he is doing research. This is usually not difficult to ensure for the consumer research into soap or detergent, but poses a difficulty in researching an industrial product such as a CNC Lathe, or a drill for mining coal, or a diesel engine for an automobile. A minimum technical knowledge of the product characteristics and its applications is necessary to interview respondents who buy, use or maintain such industrial products.

This brings us back to the question, "Who should conduct industrial marketing research?" There are two dimensions to this problem.

1. Internal Vs External
2. Technically Qualified or others

Internal *Versus* External

Most industrial marketers are wary of competition learning about their activities. Research can be done by the company's own people to prevent anyone else from knowing about it. It could be the company's

sales or marketing people who could do the research. The decision hinges on the cost-benefit analysis. There is usually an opportunity cost involved in company executives doing the research. They are unavailable for their normal activities during the period. Additionally, there is also the question of bias. Since the company executives are likely to be directly affected by the results of the survey, will they do an honest and thorough job of the research?

To make sure that the study is above board, it should be done by an external agency like a marketing research company or a consultant. Even though this leads to a small risk that competitors may learn about what you are doing, some companies prefer to do this. A second benefit of having an external agency do the research is that they have better research expertise, for sampling methodology, interviewing techniques, and reporting in useful formats.

Technical Qualification of Researcher

When industrial research is done externally, the technical qualification of the researcher becomes an issue. Whereas an MBA or an M.Stat. may be able to supervise a consumer research project, some clients for industrial research may insist that an engineer or similarly qualified person do their project. Depending on the product and the kind of respondents targeted, this could be a valid demand. The marketing research company has to evaluate this from the point of view of its own cost, and delivery of value to its clients. If a research executive is assigned to the data collection, the cost of the study goes up, because both the direct cost of executive travel and stay are higher, and so is the opportunity cost of him not being available to work on other projects. In the case of consumer research, an executive could be supervising three to four projects at a time, since he is not involved in field work himself.

If a non-technical person is sent out to the field due to cost considerations, there is a risk of failure which the agency must keep in mind. For various reasons discussed above, a technically qualified person is generally used to do industrial research—either internally or through a marketing research agency.

QUESTIONNAIRE DESIGN

An industrial researcher needs a lot of flexibility in his approach, compared with a consumer researcher. His respondent is usually a factory or office executive at his workplace. The attention span of the respondent in a factory environment is usually short, and the interview may take place in instalments, spread over more than one occasion or visit. Similarly, different people may answer different parts of the questionnaire.

CHECKLISTS

In some industrial research studies, a checklist of questions to be asked is made, instead of a formal structured questionnaire. This allows more flexibility and may bring about a better environment and cooperative frame of mind from the industry-executive respondents. Of course, the answers to checklist-type questionnaires have to be written down in an unstructured format, and the researcher

must later be able to make sense out of the responses recorded. So, the skills required of the interviewer approach those of a good journalist. If the respondent permits, a tape-recorder could be used to record the interview, instead of recording it on paper.

USE OF SECONDARY RESEARCH

Another point of difference between consumer and industrial research is in its use of secondary data. Most industrial buyers belong to trade or industry associations—for example, the Association of Electrical Engineering companies, or the National Association of Software Companies, or the Automobile Component Manufacturers' Association. These associations collect and publish a lot of information about their industries. In addition, the government departments or agencies release periodic reports about major industrial products like cement, steel, and so on. Regular publications like business newspapers, or industrial surveys also contain some information. The Internet also offers websites which contain industry data.

A lot of data, like aggregate demand and supply estimates, market shares, and financial performance of industry participants is thus available from secondary sources. The quality of this data is good, because industrial product producers and buyers are fewer in number (concentrated), and demand for one industrial product can be predicted from another product—either consumer or industrial.

In general, therefore, a secondary data search is useful and meaningful in industrial research tasks.

USE OF INDUSTRY EXPERTS

In many industrial research situations, it is possible to find individuals who are experts in the given field. They could be consultants or retired personnel who have studied the industry and have an idea of the trends it may follow. An industrial researcher can sometimes run into a goldmine of information if such an expert is available and accessible.

ANALYSING GOVERNMENT POLICY

In many industrial products, government policy is a critical factor. Any small charge in the import–export policy or tariff, for example, can affect the fortunes of many industrial products. Steel and chemicals are two examples. Therefore, it is usually necessary to analyse government policy issues at length in any industrial marketing research project, particularly in a changing environment like the Indian one.

FORECASTING DERIVED DEMAND

It is possible to use certain quantitative techniques like regression analysis to forecast demand for one industrial product based on either another industrial product, another consumer product, or a multitude of products. This is because industrial products have demand which is derived from other products or

services. For example, demand for cement may be derived from construction of homes, roads, and commercial properties, or the demand for steel could be linked to shipping, railways, automobile, and construction industries. Such relationships of derived demand are easier to conceptualise and use in industrial research than in consumer research.

To summarise, industrial marketing research is an important area to work in, and it now uses many of the same tools and techniques that consumer research does.

ASSIGNMENT

Review the following case study titled "Marketing Research for Product Redesign: A Case Study of ABC Ltd," and discuss the questions at the end.

Marketing Research for Product Redesign: A Case Study of ABC Ltd.

BACKGROUND

The Company

Set up in 1941 at Chitradurga, a small town in Karnataka, ABC Ltd. was a pioneer in the manufacture of machine tools such as lathes, grinders, and milling machines. Along with the government owned HMT Ltd., earlier known as Hindustan Machine Tools, ABC was a respected name and had a healthy bottom-line until changed circumstances—a global recession, new aggressive competition, and a new liberal import policy by the Indian government—altered it dramatically in the late 1980s.

In 1962, ABC had started a second production unit at Dharwad, 200 km. away from Chitradurga, to cater to export markets. At one time, the Dharwad unit was exporting about 90% of its production. All along, the company had produced mainly conventional machine tools. Slowly, the world had begun to change, with CNC (computer numerically controlled) machines replacing conventional machines in medium and heavy industries. Only the small industries continued to demand conventional machine tools for their in-house or jobbing work.

The Crisis in 1991

Things came to such a pass that in 1991, the Dharwad unit had very few orders on hand. It would have had to shut down, except for one World Bank order for about 700 machines, which came as a godsend. But at the same time, a lot of about 40 machines had been rejected by a German buyer and sent back to the unit.

Another important factor was that ABC Dharwad had started working on the 'Just in Time (JIT)' method at its plant from 1990 onwards. As a result, it was very clear that even if the World Bank order was executed and the rejected machines refurbished and sold, there would still be a bleak order position sometime later—perhaps end 1992 or early 1993.

An additional factor was that the Dharwad unit of ABC, till 1987, had sold in India through a distributor, Botalwala and Sons. After 1987, the company made attempts to set up its own dealer network. Due to its inexperience, this was not done very effectively, and the expected sales did not materialise.

The Planned Shutdown

As a result of the various circumstances described above, ABC Dharwad decided that something drastically different needed to be done. The Executive Vice President in charge of the plant, the General Manager (Business Development), and the Managing Director had a lot of brainstorming sessions on what could be done. Instead of just twiddling their thumbs, waiting for the industry to

revive, or a miracle to take place, the three of them decided to reach for the people whom they were trying to serve—their customers. It was decided that only their customers could tell them what was wrong, and how to set it right. Because, as ABC saw it, some machine tools were still being bought and sold despite the recession—only, they were not their own machines. Therefore, it was decided to make use of the manpower, mainly engineers, who had been idle due to the lack of orders, to conduct an extensive market research. The plant itself was shut down, with a promise to reopen it in three to six months. Workers were paid 90% of their wages, and asked to stay at home until called back. The union reluctantly agreed with this plan.

Research Plan

The major information need identified was of course, "What's wrong with our product?" and "What can we do to make it desirable again?" All available engineers were called together for a briefing. Many of them were not sales or marketing people, and did not have any experience of conducting research. The GM (Business Development) gave them a checklist of all the features of ABC machines, including the current prices, and they were told to spread out and meet as many customers as they could in all the major industrial cities, and industrial estates within these cities, in India. The focus was on small and medium-sized industries, who were the major targets of the unit's products.

Care was taken to assign the engineers to only such states where they could speak the local language, such as Hindi in the northern belt, Tamil in Tamilnadu, and so on. The reason for doing this was that the respondents were likely to be small entrepreneurs and operators fluent only in their native language. An added mission given to this team was to bring in orders for machines according to customer's specifications, and relay them to the factory, so that it could be restarted as soon as possible. A coordination cell to monitor this research activity and incoming orders was set up at the Dharwad factory premises. An international marketing research study was also carried out on similar lines, during the same period. This mainly covered customers in South-east Asia.

The Findings

A lot of data was collected. To simplify data handling, all the researchers were asked to report immediately only three findings:

1. the three most important threats to ABC,
2. the three most important changes ABC needed to make, and
3. the three best opportunities ABC had, if it made these changes.

The quality and design people were called in to do a quick Pareto analysis of the study's findings (80% of the most useful suggestions coming from 20% of the responses). Their major findings were that:

1. ABC products were overpriced.
2. ABC products were overdesigned, with too many features, which the customers did not need or use.
3. Sometimes, they were even oversized. For example, one customer commented that a lathe should fit into a lift (elevator) for easy portability. Many industrial estates are multi-storeyed.

Similar findings were reported from the South-east Asian market.

Product Redesign—A Star is Born

The entire ABC team took upon themselves the challenge of redesigning from scratch their major conventional machine. Initially about 180 orders were obtained through the contacts made by the research team itself. The redesign was done in a record time of 4 months, from June to October of 1993. The same year, the product, now christened as the TURNSTYLE, was displayed at the Thailand trade show, and a distributor was signed up. The same success was repeated in Indonesia after another trade show. Apart from adopting a modular design so that parts could be now used to assemble new products "like a Mechano set", in the words of the GM, Business Development, the total number of different parts were also cut down from a whopping 14,000 to a reasonable 3,500. And these included every nut, bolt, and accessory. Prices on the TURNSTYLE went down from 22–30% on the same specifications, due to the cutting out of unnecessary features and increased production runs (batch sizes). The increased batch sizes had also given ABC better leverage to get low prices from suppliers, and these were passed on to customers. Exports picked up, and from almost nil in 1993, peaked at around 7 crore in 1995–96. An assembly line was also started in Indonesia through a tie up where ABC supplied technical support, without investing any funds. New distributors were signed up in countries like Norway, Finland, and others.

Questions for Discussion

1. What do you see as the strengths and weaknesses of ABC's approach towards its customers from this case?
2. How else can marketing research be utilised by a company making industrial products? List different purposes research can fulfill.
3. What are the advantages/disadvantages of handling customer research internally using company executives as compared to using an external market research company?

APPENDIX 2 CAREERS IN MARKETING RESEARCH

The purpose of this appendix is to give students an insight into the kind of careers this industry can offer, and the skills and aptitude it requires. This is based on the author's own experience of the industry, and discussions with other colleagues who have worked in it.

MAJOR COMPANIES IN MARKETING RESEARCH

The 10 largest marketing research companies in the world, according to the website swcollege.com/marketing/mcdaniel/mcdaniel_student/research_firms.html are:

1. A.C. Nielsen
2. Cognizant Corp.
3. Information Resources, Inc.
4. The Arbitron Co.
5. PMSI/Source Informatics
6. Westat, Inc.
7. Maritz Marketing Research, Inc.
8. NFO Research Inc.
9. The Kantar Group
10. The NPD Group

The largest and most well-known marketing research companies in India are:

1. A.C. Nielsen ORG-MARG
2. IMRB
3. T.N. SOFRES MODE
4. GALLUP MBA

There are many more agencies that are into marketing research in addition to these big companies. A full-service agency does complete projects including field work, tabulation, analysis, and reports/presentation, and some agencies may do only field work, or only computer data entry and analysis. Most big companies in this field have formal or informal collaborations with global research companies.

In addition, many foreign research companies have entered India recently. They usually have an Indian partner to assist them in understanding the special features of the local market and consumers. Then, there are some specialised agencies which do specialised work. For example, J.D. Power and Associates is a global company (no. 20 in the world according to the website quoted above) that specialises in customer satisfaction surveys for automobiles. With the recent spate of automobile models being launched in India, their ratings will assume increasing importance. Other types of research companies which do specialised types of research such as Simulated Test Marketing or Data Mining have also entered India, and could increase their base in times to come.

Many of the frontline agencies do some specialised work such as retail audits, consumer panels, TV audience research, and so on.

With the entry of MNCs into India in all types of industries, the amount of marketing research is likely to grow at a healthy pace. Any competitive industry requires marketing research to keep a company in tune with its customers. Consumer goods companies are typically the highest users of marketing research. But new companies entering the market are also likely to do a lot of research during concept stage, prototyping stage, product testing, promotion decisions, pricing decisions, and even after the launch of a product to track customer satisfaction. All this activity leads to a lot of marketing research being carried out. In the near future, service companies such as banks, credit card companies, phone companies, hotels, airlines will be big users of research. Industrial manufacturing and marketing companies are also likely to increase their spending on research, due to increasing competition.

JOBS IN MARKETING RESEARCH

Even though the marketing research industry is small in sales turnover compared with many traditional ones, we should remember it is a service industry, which does not inflate its figures like say, the advertising industry. In the advertising industry, for example, the billings figure is usually talked about as the turnover. But the revenue to the advertising agency is usually only about 15 percent of the billings. Thus, an advertising agency which shows billings of Rs100 crores may be earning only about Rs15 crores as its commission. This is not the case with marketing research companies. Their billings and revenue are more or less equal.

There are interesting jobs available for people with background in statistics, psychology, management, and so on in marketing research. Even graduates with a flair for doing field work can find jobs in marketing research companies and make a decent living doing it.

Some of the more common types of positions available in a full-service marketing research companies and a brief description of the jobs follows:

Research Executive

Typically, this is an entry level position for an MBA with a marketing major. This is a client liaison job with full responsibility for carrying out a project involving writing a proposal, discussions with the client, design of the study including sampling, questionnaire design, recruiting and supervising the field force through a field supervisor, and supervising the data entry, tabulation, and analysis. Finally,

it ends with report preparation and a presentation to the client. It is very similar to the summer project assignments that some MBA students do for their employers, but usually the scope and responsibilities are a little lighter in the summer jobs.

While doing the study design, the research executive needs to interact with one or more executives from the client's company, and his communication skills have to be of a high order. He has to convince the client of his agency's capability to do a good job, and then deliver on the promise at the end of it.

A research executive's career can progress to a Senior Research Executive, Branch Manager, General Manager, Vice-President and so on in due course, dependent on the organisation structure of the firm. The starting salaries are not as high as in some IT or management consultancies or investment banks, but they are comparable to some of the other regular employers of MBAs.

The work is challenging, for those with a scientific bent of mind, curious about relationships between consumer needs, behaviours, and marketing variables. Sometimes the time pressure is high, and the budget may be limited. A variety of industries, products / services, and client types is assured.

Statistical Analyst

The second major type of job found in marketing research companies is that of a statistician. He could be of a mathematics or applied statistics background. He would supervise the data collection, tabulation, and analysis from a statistical viewpoint. For example, he could assist in sampling, deciding on output formats of tables, and checking the data entry process for possible errors. His major job would be assisting the research executive in the areas of his expertise.

Field Supervisors

The major job of a field supervisor is to recruit, brief, and supervise the field staff. The field staff could be regular workers on daily wage, but they could be housewives, students or part-time workers who do not do this for a regular living, but for more pocket money or for experience.

The field supervisor is usually a graduate, with experience of having worked on the field staff for some time. He is usually familiar with at least the territory or town he lives in, and is energetic and enthusiastic to lead a team of 15 or 20 people in the field everyday. He should preferably be conversant with the local language of the territory he is in charge of. This is desirable because in India, many studies require meeting respondents who converse in the local language, and field staff are recruited for ability to speak the language. The field supervisor also must have the integrity to stick to a given sampling plan and also to complete the quota of given work through his field staff.

In primary research, field work is the most crucial part of the work, and the field supervisor is a key to doing a good job of this. The field supervisor works under the supervision of the research executive in charge of the project. Usually, for multi-centre projects, field supervisors from branch offices of the agency in the cities are utilised. But in some cases, field supervisors may be required to travel to other cities too.

Field Staff

This can be part-time work, project based, or it can be done as a regular job on daily wage basis. Either way, it is an opportunity for people with good communication skills and pleasant personality to work

on marketing research projects on the field. They would be briefed, and mock interviews conducted for the specific questionnaire, before sending them on the field. Field supervisors would be with them to assist with any problems.

SUMMARY OF JOB PROSPECTS

As we have seen in the above discussion, we have a wide variety of jobs. In addition to the above, data entry jobs are available for those with less education, but with good typing skills. All the studies are analysed on the computer these days, and the volume of data to be entered is quite high.

From tabulators to field staff to supervisors, and from statistical analysts to research executives and above, we have a wide variety of jobs available. In addition, qualitative research is an emerging field with a lot of promise. Focus group interviews and other qualitative techniques like projective techniques, depth interviews and so on, require people with psychology, sociology, and similar backgrounds.

GROWTH PROSPECTS

Thus, we see that diverse skills are required of different people in a marketing research agency, not limited to the MBA. But it is the MBA who may have the best chance of growth, due to his all-round view of the job, and because he is the person responsible to the client.

GETTING BUSINESS

In the beginning, getting orders is not the responsibility of a research executive. It is his seniors who do the job. But as he reaches higher rungs in the organisation, he is expected to make sales pitches for an assignment, and build a network of happy clients who give repeat orders or recommend his agency to their friends. In many ways, this job is like a sales job for any other industry. Like in other industries, some enquiries come from clients. Some others have to be generated by making cold calls on potential clients.

It is also possible for qualified people to start a marketing research company as entrepreneurs. Two or three like-minded people are enough, and a lot of the work can be out-sourced—for example, field work, data entry and so on. But in the long-run, clients may prefer a full-service marketing research company.

SUMMARY

Marketing research is a growing industry in India. It has many leading international agencies which either have majority or minority holdings in Indian companies. For example, A.C. Nielsen, VNU Holdings, NFO, Gallup, and T.N. Sofres, are some of the major international players now operating in

India, in association with major Indian firms like ORG-MARG, MODE, MBA and MBL/RCG. Most India-based FMCG (fast moving consumer goods) companies were already heavy users of research. Now, we also have a lot of consumer durables companies (automobiles, refrigerators, TVs, etc.), industrial goods companies and services companies which are increasing spending on marketing research. This creates demand for qualified research professionals.

An MBA from a business school can get entry level openings as a research executive in a marketing research company such as IMRB, ORG-MARG, and so on through open advertisements or campus recruitments. The research executive would handle complete marketing research projects on behalf of clients and can get wide exposure to different product/markets and research methods. Good analytical aptitude, conceptual skills, and communication skills are required.

Statistical analysis including data preparation, coding, processing, and tabulation, and also use of statistical testing and multivariate statistical techniques plays a major role in a research project. Postgraduates in Applied Statistics may find this job of a statistical analyst to their liking, to assist the analysis of data. Their inputs may also be of use in designing samples in a marketing research firm.

Qualitative research specialists are preferably from a psychology or sociology or related background, with at least a postgraduate degree. They must also have a good grasp of research methods, and be able to handle the design and execution of qualitative projects such as focus groups, depth interviews or other specialised types of qualitative research.

Field supervisors require a very good communication and leadership ability, to go on the field for interviews, select and supervise young students, housewives or other field staff, and meet targets or quotas set for them within a deadline. Graduates with a few years experience and command over the local language may be suitable for such positions.

Field staff are generally temporary or part time workers hired for a day or for a few days of work. They are usually students with good communication skills in English and the local language, or regulars who like to do field work under the guidance of a supervisor.

ASSIGNMENT QUESTIONS

1. Prepare a list of jobs available in the marketing research area. List down major skills required for each of these.
2. Make your resume for the marketing research job, which you think you are suitable for.
3. Find out from the Internet, the sales turnover of the ten largest (a) global M.R. companies, (b) Indian M.R. companies.

REFERENCES

Books

1. Green, Paul E., Tull S. Donald and Gerald Albaum. 1999. *Research for Marketing Decisions.* 5th edition, New Delhi: Prentice-Hall of India.
2. Churchill, A. Gilbert Jr. 1995. *Marketing Research—Methodological Foundations.* 6th edition, Orlando: The Dryden Press.
3. Sengupta, Subroto. 1998. *Brand Positioning.* New Delhi: Tata McGraw-Hill.
4. Beri, G.C. 2000. *Marketing Research.* 3rd edition, New Delhi: Tata McGraw-Hill.
5. Ogilvy, David. 1983. In Ogilvy on Advertising. *18 Miracles of Research.* London: Orbis Publishing.
6. Boyd, Harper W., Jr., Ralph Westfall and Stanley F Stasch. 1998. *Marketing Research.* 7th edition, Delhi: AITBS.
7. Kinnear, Thomas C. and James R. Taylor.1991. *Marketing Research.* 4th edition, Singapore: McGraw-Hill International Edition.
8. Luck, J. David and Ronald S. Rubin.1989. *Marketing Research.* 7th edition, New Delhi: Prentice-Hall of India.
9. Hair, Anderson, Tatham, Black, *Multivariate Data Analysis,* 5th Edition, Delhi, Pearson Education.

Articles

1. Sangren, Susie. 1999. "A Survey of Multivariate Methods Useful for Marketing Research." *Quirk's Marketing Research Review* (May).
2. Sabena, Patricia. 1999."10 Trends in Qualitative Research." *Quirk's Marketing Research Review* (December).
3. Struhl, Steven. 1999. "In Pursuit of Software: Major Programs You Won't Want to Miss — SPSS 9.0, SYSTAT 8.0, ANSWERTREE 2.0 and DELTAGRAPH 4.0." *Quirk's Marketing Research Review* (February).

4. Bowers, K. Diane. Winter 1998/Spring 1999. "FAQs on Online Research." *Marketing Research:* 45–48.
5. Annuncio, Charubala. 1993."The Passions of the Mind." *Advertising and Marketing:* 77–80.

Websites

1. www.quirks.com, for articles on marketing research.
2. www.spss.com for details about SPSS software package for marketing research applications and other applications such as CRM (Customer Relationship Management).
3. www.sas.com
4. www.systat.com
5. www.statsoft.com

The McGraw·Hill Companies

INDEX

Advertising research 10
  copy testing 10
  media research 10
Analysis plan 29
ANOVA 208
  case study 1 221
  case study 2 226
  case study 3 235
  completely randomised design 209
  factorial designs 212
  randomised block design 210
Attribute-based perceptual mapping 454
  case study 1 464
  case study 2 471
  case study 3 477

Brand positioning
  using mulidimensional scaling 394
  using discriminant analysis 454
Briefing 28, 111

CRM 16
Careers in marketing research 523
Causal research 23
Chi-squared test 187
  an illustration 188
  indexes of agreement 190
Classification
  using cluster analysis 348
  using discriminant analysis 284
Classification matrix 287
Cluster analysis 348
  case study 1 363
  case study 2 374
  case study 3 382
  hierarchical 349
  K-means 351
Coding 116
Concept research 8
Confirmatory Factory Analysis 485
Conjoint analysis 421
  case study 1 435
  case study 2 440
  case study 3 447
Correlation and regression 242
  case study 1 256
  case study 2 267
  case study 3 273
Consumer panels 26
Contingency coefficient 190
Cross-sectional designs 23
Cross-tabulation with Chi-squared test 186
  case study 1 197

case study 2 200
case study 3 204

Data analysis 119
Data input format 116
Data mining 16
Data processing 115
recoding 191
subsets 192
Data reduction 326
Debriefing 29, 111
Depth interview 39
Descriptive research 22
Discriminant analysis 280
case study 1 293
case study 2 397
case study 3 302
for perceptual mapping 454
Discriminant function 285
Distribution research 9
Double-barrelled questions 61
Duncan's test 214

Effects coding 424
Eigen value 313
Email surveys 15
Emerging issues in marketing research 14
Experimental designs
experiments 25, 209
Exploratory research 22

Factor analysis 326
case study 1 336
case study 2 340
case study 3 344
Factor extraction 327
Factor rotation 327
Field work
control procedures 111
design of field work 109
organising field work 110
quotas 110
Field work plan 24, 28
Focus group 43

Hierarchical clustering 349
Hypothesis testing 119, 123, 125

Indexes of agreement 190
Industrial marketing research 513
Integrated case study 1 for part 1 130
Integrated case study 2 for part 1 141
Integrated case study 3 for part 1 155
Interval scale 55

K-means clustering 351

Lambda (Asymmetric) 190

Marketing information system 4
Marketing intelligence 4
Marketing research 5
ethical considerations in 3
limitations of 11
typical applications 6
who does 6
Multidimensional scaling 394
case study 1 405
case study 2 411
case study 3 416

Nominal scale 51
Non-sampling error 106

Observation 25
Online research 15
Ordinal scale 52

Paired comparisons 57
Percentage calculation 183
in a cross-tab 186
Perceptual maps
attribute-based 454
dissimilarity-based 395
Positioning 394
Presentation 31
Pricing research 8
Primary research 12
Product research 8
Projective techniques 40

Qualitative techniques 39
Questionnaire design 48
Questions
Dichotomous 56
Disguised versus undisguised 54
Multiple choice 56
Open-ended and closed-ended 54
structured and unstructured 53

Rankings 56
Ratings 56
Ratio scale 52
Regression 242
case study 1 256
case study 2 267
case study 3 273
Reliability and validity 62
Report 31
Research designs 21
Research methodology 24
Research objective 21
Retail audit 26
Rotated factor matrix 327, 329

Sample size calculation 91, 96
formula for sample size calculation
when estimating means 92
formula for sample size calculation
when estimating proportions 94
Sampling
population 91
sampling element 90
sampling frame 91
sampling unit 91
Sampling error 106
Sampling plan 27
Sampling techniques 97
cluster sampling 102
convenience sampling 105
disproportionate stratified sampling 100
judgement sampling 105
proportionate stratified sampling 98
quota sampling 104
simple random sampling 98
snowball sampling 105
stratified random sampling 98
systematic sampling 103
Scales of measurement 51
Scheffe's test 214
Secondary data
disadvantages of 37
sources of 36
Secondary research 12
Segmentation
using cluster analysis 348
Semantic differential scale 57
Simple tabulation 181
for ranking type questions 184
integrated case study 1 for part 1 130
integrated case study 2 for part 1 141
integrated case study 3 for part 1 155
tabulating ratings 185
SPSS commands for
ANOVA 220

attribute-based perceptual mapping 462
cluster analysis 362
conjoint analysis 433
correlation and regression 254
crosstabs with chi-squared test 196
data input and t-tests 128
discriminant analysis 292
factor analysis 335
frequency tables 196
logistic regression 314
multidimensional scaling 404
Structural Equation Modeling 485
Variable Transformation 192

T-test
independent sample 123
paired sample 125

Test marketing 41
Tukey's test 214
TV audience measurements 27

Validity 62
Content 62
Criterion 63
Construct 63
Value labels 117
Variables
dependent 181
independent 181
transforming 191
recoding 191
Variable labels 116

Wilk's lambda 288