

INTRODUCTORY STATISTICS GUIDE

The SPSS logo is rendered in a bold, blue, sans-serif font. It is set against a white, cloud-like background that has a soft blue glow around it. The logo is positioned in the upper left quadrant of the cover.

MARIJA J. NORUŠIS

01 7956
FELIX
BOOKS & GIFTS
993

C8
N 889

9760335

JENKINS S

NEB FFE
1394>13085
EP>>413.10

Introductory Statistics Guide

贈閱



OTHER QUALITY
USED BOOK
SA BOOKSTORE

Marija J. Norušis



E9760335

SPSS Inc.
Suite 3000
444 North Michigan Avenue
Chicago, Illinois 60611

For more information about the SPSS^{XTM} system and other software produced and distributed by SPSS Inc., please write or call

Marketing Department
SPSS Inc.
444 North Michigan Avenue
Chicago, IL 60611
(312) 329-3500

SPSS^X, SPSS, and SCSS are the trademarks of SPSS Inc. for its proprietary computer software. No material describing such software may be produced or distributed without the written permission of the owners of the trademark and license rights in the software and the copyrights in the published materials.

SPSS^{XTM} Introductory Statistics Guide

Copyright © 1983 by SPSS Inc.

All rights reserved.

Printed in the United States of America. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a data base or retrieval system, without the prior written permission of the publisher.

7 8 9 0 SEMSEM 8 9 8 7 6

ISBN 0-918469-20-1

(previously ISBN 0-07-046549-5)

Library of Congress Cataloging in Publication Data

Norušis, M. J. (Marija J.), date
SPSS^X introductory statistics guide.

Bibliography: p.

Includes index.

I. SPSS (Electronic computer system) I. SPSS Inc.
II. Title. III. S.P.S.S.-X introductory statistics
guide.

HA32.N68 1983 001.64'2 83-5456

Preface

Through and through the world is infested with quantity: To talk sense is to talk quantities. It is no use saying the nation is large—How large? It is no use saying that radium is scarce—How scarce? You cannot evade quantity. You may fly to poetry and music, and quantity and number will face you in your rhythms and your octaves.

—Alfred North Whitehead

Quantity is as inescapable today as it was in Whitehead's time. Even those outside technical professions face a plethora of numbers when they look at a newspaper. The purpose of data analysis is to make it easier to deal with quantity—to simplify and summarize data and to illuminate patterns that are not immediately evident.

THE SPSS^x SYSTEM

The SPSS^x Batch system is a comprehensive tool for managing, analyzing, and displaying data. A broad range of statistical analyses and data modification tasks are accomplished with a simple, English-like language. Results can be easily obtained with minimal understanding of computer intricacies.

This manual is intended for novice users of the system and introduces only the basic features and procedures: descriptive statistics, measures of association for two-way tables, tests for equality of means, nonparametric procedures, and bivariate and multiple regression.

Similarly, while this text includes instructions for entering and defining data for analysis, for managing data files, and for transforming, selecting, sampling, and weighting data, it does not attempt to cover the full range of data and file management facilities available in SPSS^x. For those who want to extend their use of the system beyond the scope of this introduction, documentation can be found in *SPSS^x User's Guide* (McGraw-Hill Book Company, 1983). The computational methods used are described in *SPSS^x Statistical Algorithms*, available from SPSS Inc. But the system and its documentation are continually being extended. Before obtaining other manuals, check with your computation center for information about the current release of SPSS^x being used there and the documentation for that release.

USING THIS TEXT

This manual is designed to be a supplement in courses that integrate the teaching of statistics and computing. The first two chapters discuss preparation of a data file and the fundamentals of an SPSS^x job. Each subsequent chapter describes a problem and the SPSS^x output useful for its solution, followed by information about the SPSS^x commands needed to obtain the analysis.

Exercises at the end of each chapter reinforce and extend the material in three main areas: syntax, statistical concepts, and data analysis. Answers for questions on syntax and statistical concepts are given in Appendix A. The data analysis exercises provide an opportunity to formulate hypotheses, create the SPSS^x commands needed to carry out the analysis, and run those jobs using one of four data files distributed with the SPSS^x system. Those data files are described in Appendix B. Consult the SPSS Coordinator at your installation for information about using the files.

The last chapter contains a brief guide to the features of the SPSS^x system described in this manual, including instructions for running the procedures. Descriptions of SPSS^x commands in Chapters 1–12 are deliberately brief, and users should make a habit of extending their knowledge of SPSS^x facilities by consulting Chapter 13 as they work through this text. Many exercises require information from Chapter 13.

ACKNOWLEDGMENTS

Most of the SPSS Inc. staff have participated either in designing and preparing this manual or in creating and maintaining the system it documents. In particular, Elisabeth Adams, Doug Chene, Bob Gruen, Pam Hecht, Nancy Morrison and Keith Sours contributed substantially to the writing of the operations sections and preparation of this book. Sue Shott also assisted in various phases of the preparation.

I am also grateful to the reviewers and users of the first edition of this book for many helpful comments and suggestions, and to Harry Roberts, Harry Davis, and Richard Shekelle for permission to use and distribute the data files. Finally, I wish to thank Rasa, Irena, Linas, and Egle who advise me on everything.

—Marija J. Norušis

Contents

Preface xi

Chapter 1 From Paper into a File 1

- 1.1 Cases, Variables, and Values 1
- 1.2 Identifying Important Variables 2
- 1.3 Recording the Data 2
- 1.4 Coding the Variables 2
- 1.5 An Example 4
- 1.6 DESIGNING FORMS 6
- 1.7 The Data File 6
- EXERCISES 7

Chapter 2 The SPSS^x Job 9

- 2.1 PREPARING SPSS^x COMMANDS 9
- 2.2 Commands and Specifications 9
- 2.3 DATA DEFINITION 10
- 2.4 Describing the Data File 10
 - 2.5 Locating the Data
 - 2.6 Specifying the Number of Records
 - 2.7 Choosing Variable Names
 - 2.8 Indicating Column Locations
 - 2.9 Specifying Multiple Records
 - 2.10 Types of Variables
 - 2.11 Indicating Decimal Places
- 2.12 The DATA LIST Table 13
- 2.13 Variable and Value Labels 14
- 2.14 Identifying Missing Values 15
- 2.15 THE LIST COMMAND 15
- 2.16 IN-LINE DATA 16
- 2.17 THE ACTIVE FILE 16
- 2.18 THE SPSS^x SYSTEM FILE 16
- 2.19 WARNINGS AND ERROR MESSAGES 17
- 2.20 OTHER SPSS^x COMMANDS 18
- EXERCISES 18

Chapter 3 Blue Mondays: Data Tabulation 21

- 3.1 A FREQUENCY TABLE 21
- 3.2 Visual Displays 22
- 3.3 What Day? 23
- 3.4 Histograms 23
- 3.5 Screening Data 24
- 3.6 USING AN SPSS^x SYSTEM FILE 25
- 3.7 RUNNING PROCEDURE FREQUENCIES 25
- 3.8 Bar Charts 26
- 3.9 Histograms 26
- 3.10 Missing Values in Tables and Statistics 26
- EXERCISES 27

Chapter 4 Telling the Whole Truth and Nothing But: Descriptive Statistics 33

- 4.1 EXAMINING THE DATA 33
- 4.2 Percentile Values 35
- 4.3 SUMMARIZING THE DATA 36
- 4.4 Levels of Measurement 36
 - 4.5 Nominal Measurement 4.6 Ordinal Measurement 4.7 Interval Measurement
 - 4.8 Ratio Measurement
- 4.9 Summary Statistics 37
 - 4.10 Measures of Central Tendency 4.11 Measures of Dispersion
- 4.12 The Normal Distribution 39
 - 4.13 Measures of Shape 4.14 Standard Scores
- 4.15 Who Lies? 41
- 4.16 STATISTICS AVAILABLE WITH PROCEDURE FREQUENCIES 41
- 4.17 Percentiles 41
- 4.18 RUNNING PROCEDURE CONDESCRIPTIVE 42
- 4.19 Z Scores 42
- EXERCISES 43

Chapter 5 Lost Letters in Cities and Towns: Crosstabulation and Measures of Association 47

- 5.1 CROSSTABULATION 47
- 5.2 Cell Contents and Marginals 48
- 5.3 Choosing Percentages 49
- 5.4 Adding a Control Variable 49
- 5.5 GRAPHICAL REPRESENTATION OF CROSSTABULATIONS 50
- 5.6 USING CROSSTABULATION FOR DATA SCREENING 51
- 5.7 CROSSTABULATION STATISTICS 51
- 5.8 The Chi-Square Test of Independence 52
- 5.9 Measures of Association 54
- 5.10 Nominal Measures 54
 - 5.11 Chi-Square-Based Measures 5.12 Proportional Reduction in Error
- 5.13 Ordinal Measures 57
- 5.14 Measures Involving Interval Data 58
- 5.15 RUNNING PROCEDURE CROSSTABS 59
- 5.16 Recoding Data 60
- 5.17 Entering Crosstabulated Data 61
- EXERCISES 62

Chapter 6 Breaking Down Discrimination: Describing Subpopulation Differences 67

- 6.1 SEARCHING FOR DISCRIMINATION 67
- 6.2 Who Does What? 68
- 6.3 Level of Education 68
- 6.4 Beginning Salaries 70
- 6.5 Introducing More Variables 71
- 6.6 RUNNING PROCEDURE BREAKDOWN 71
- EXERCISES 72

Chapter 7 Consumer Surveys: Testing Hypotheses about Differences in Means 75

- 7.1 TESTING HYPOTHESES 75
- 7.2 Samples and Populations 76
- 7.3 Sampling Distributions 76
- 7.4 Sampling Distribution of the Mean 77

- 7.5 THE TWO-SAMPLE T TEST 79
 - 7.6 Significance Levels 80
 - 7.7 One-Tailed vs. Two-Tailed Tests 80
 - 7.8 What's the Difference? 81
- 7.9 USING PROCEDURE CROSSTABS TO TEST HYPOTHESES 81
- 7.10 INDEPENDENT VS. PAIRED SAMPLES 82
 - 7.11 Analysis of Paired Data 82
- 7.12 HYPOTHESIS TESTING: A REVIEW 83
 - 7.13 The Importance of Assumptions 83
- 7.14 RUNNING PROCEDURE T-TEST 84
- EXERCISES 85

Chapter 8 Making Sales Click: Correlation and Scatterplots 89

- 8.1 READING A SCATTERPLOT 89
 - 8.2 Examining Relationships 91
- 8.3 THE CORRELATION COEFFICIENT 92
 - 8.4 Some Properties of the Correlation Coefficient 95
 - 8.5 Calculating Correlation Coefficients 94
 - 8.6 Hypothesis Tests about the Correlation Coefficient 94
 - 8.7 Correlation Matrices and Missing Data 95
 - 8.8 Choosing Pairwise Missing-Value Treatment 95
- 8.9 THE REGRESSION LINE 96
 - 8.10 Prediction 96
 - 8.11 Goodness of Fit 97
 - 8.12 Further Topics in Regression 98
- 8.13 RUNNING PROCEDURE SCATTERGRAM 98
- 8.14 RUNNING PROCEDURE PEARSON CORR 99
- 8.15 SCATTERGRAM AND SELECTING CASES 100
- EXERCISES 103

Chapter 9 What's Your Proof? One-Way Analysis of Variance 109

- 9.1 DESCRIPTIVE STATISTICS AND CONFIDENCE INTERVALS 109
- 9.2 ANALYSIS OF VARIANCE 110
 - 9.3 Partitioning Variation 110
 - 9.4 Testing the Hypothesis 111
- 9.5 MULTIPLE COMPARISON PROCEDURES 111
 - 9.6 The Scheffé Test 112
- 9.7 EXPLANATIONS 113
 - 9.8 Tests for Equality of Variance 113
- 9.9 RUNNING PROCEDURE ONEWAY 113
 - 9.10 ONEWAY and Other SPSS^x Commands 114
- EXERCISES 115

Chapter 10 Beauty and the Writer: Analysis of Variance 117

- 10.1 DESCRIPTIVE STATISTICS 117
- 10.2 ANALYSIS OF VARIANCE 118
 - 10.3 Testing for Interaction 119
 - 10.4 Tests for Sex and Attractiveness 121
- 10.5 EXPLANATIONS 121
- 10.6 EXTENSIONS 121
- 10.7 RUNNING PROCEDURE ANOVA 122
 - 10.8 ANOVA and Other SPSS^x Commands 122
- EXERCISES 123

Chapter 11 Fats and Rats: Distribution-Free or Nonparametric Tests 125

- 11.1 THE MANN-WHITNEY TEST 125
 - 11.2 Ranking the Data 126
 - 11.3 Calculating the Test 126
 - 11.4 Which Diet? 127
 - 11.5 Assumptions 127
- 11.6 NONPARAMETRIC TESTS 127
 - 11.7 The Sign Test 128
 - 11.8 The Wilcoxon Signed-Ranks Test 128
 - 11.9 The Kruskal-Wallis Test 129
 - 11.10 The One-Sample Chi-Square Test 129
 - 11.11 The Rank Correlation Coefficient 130
- 11.12 RUNNING PROCEDURE NPAR TESTS 131
 - 11.13 The One-Sample Chi-Square Test and Freefield Input 131
- 11.14 RUNNING PROCEDURE NONPAR CORR 132
 - EXERCISES 132

Chapter 12 Statistical Models for Salary: Multiple Linear Regression Analysis 135

- 12.1 INTRODUCTION TO REGRESSION STATISTICS 135
 - 12.2 Outliers 136
 - 12.3 Choosing a Regression Line 136
 - 12.4 The Standardized Regression Coefficient
 - 12.5 From Samples to Populations 138
 - 12.6 Estimating Population Parameters 12.7 Testing Hypotheses 12.8 Confidence Intervals
 - 12.9 Goodness of Fit 140
 - 12.10 The R^2 Coefficient 12.11 Analysis of Variance 12.12 Another Interpretation of R^2
 - 12.13 Predicted Values and Their Standard Errors 143
 - 12.14 Predicting Mean Response 12.15 Predicting a New Value 12.16 Reading the Casewise Plot
 - 12.17 Searching for Violations of Assumptions 146
 - 12.18 Residuals 12.19 Linearity 12.20 Equality of Variance 12.21 Independence of Error
 - 12.22 Normality
 - 12.23 Locating Outliers 151
 - 12.24 When Assumptions Appear To Be Violated 151
 - 12.25 Coaxing a Nonlinear Relationship to Linearity 12.26 Coping with Skewness
 - 12.27 Stabilizing the Variance 12.28 Transforming the Salary Data
 - 12.29 A Final Comment on Assumptions
- 12.30 MULTIPLE REGRESSION MODELS 154
 - 12.31 Predictors of Beginning Salary 154
 - 12.32 The Correlation Matrix 12.33 Partial Regression Coefficients
 - 12.34 Determining Important Variables 155
 - 12.35 BETA Coefficients 12.36 Part and Partial Coefficients
 - 12.37 Building a Model 157
 - 12.38 Adding and Deleting Variables 12.39 Statistics for Variables Not in the Equation
 - 12.40 The 'Optimal' Number of Independent Variables
 - 12.41 Procedures for Selecting Variables 160
 - 12.42 Forward Selection 12.43 Backward Elimination 12.44 Stepwise Selection
 - 12.45 Checking for Violation of Assumptions 163
 - 12.46 Interpreting the Equation 163
 - 12.47 Statistics for Unselected Cases 164
 - 12.48 Problems of Multicollinearity 164
 - 12.49 Methods of Detection 12.50 SPSS^x and Multicollinearity
- 12.51 RUNNING PROCEDURE REGRESSION 166
 - 12.52 Building the Equation 166
 - 12.53 Requesting Statistics 166
 - 12.54 Residuals Analysis 167
 - 12.55 Additional Subcommands for the Equation 168
 - 12.56 REGRESSION and Other SPSS^x Commands 169
 - EXERCISES 170

Chapter 13 SPSS^x Command Reference 175

- 13.1 DATA DEFINITION COMMANDS 175
 - 13.2 The BEGIN DATA and END DATA Commands 175
 - 13.3 The DATA LIST Command 176
 - 13.4 File Definition on DATA LIST 13.5 Variable Definition on DATA LIST
 - 13.6 The FILE HANDLE Command 177
 - 13.7 The FILE LABEL Command 178
 - 13.8 The FILE TYPE—END FILE TYPE Structure 178
 - 13.9 The GET Command 180
 - 13.10 The RENAME Subcommand 13.11 The DROP Subcommand 13.12 The KEEP Subcommand
 - 13.13 The MAP Subcommand
 - 13.14 The MISSING VALUES Command 181
 - 13.15 The SAVE Command 182
 - 13.16 The RENAME Subcommand 13.17 The DROP Subcommand 13.18 The KEEP Subcommand
 - 13.19 The MAP Subcommand
 - 13.20 The VALUE LABELS Command 183
 - 13.21 The VARIABLE LABELS Command 184
- 13.22 UTILITY COMMANDS 184
 - 13.23 The COMMENT Command 184
 - 13.24 The DISPLAY Command 184
 - 13.25 The DOCUMENT Command 185
 - 13.26 The EDIT Command 185
 - 13.27 The FINISH Command 186
 - 13.28 The FORMATS Command 186
 - 13.29 The INFO Command 186
 - 13.30 The N OF CASES Command 188
 - 13.31 The NUMBERED Command 188
 - 13.32 The PRINT FORMATS Command 188
 - 13.33 The SUBTITLE Command 189
 - 13.34 The TITLE Command 189
 - 13.35 The WRITE FORMATS Command 189
- 13.36 DATA TRANSFORMATION COMMANDS 190
 - 13.37 The COMPUTE Command 190
 - 13.38 Arithmetic Operations 13.39 Numeric Functions 13.40 Generating Distributions
 - 13.41 The COUNT Command 192
 - 13.42 The DO IF—END IF Structure 192
 - 13.43 Logical Expressions
 - 13.44 The IF Command 194
 - 13.45 The RECODE Command 194
 - 13.46 The SAMPLE Command 195
 - 13.47 The SELECT IF Command 196
 - 13.48 The SORT CASES Command 196
 - 13.49 The SPLIT FILE Command 196
 - 13.50 The TEMPORARY Command 197
 - 13.51 The WEIGHT Command 197
- 13.52 PROCEDURE COMMANDS 198
- 13.53 ANOVA 199
 - 13.54 The STATISTICS Command 200
 - 13.55 Limitations 200
- 13.56 BREAKDOWN 201
 - 13.57 General Mode 201
 - 13.58 Integer Mode 202
 - 13.59 The VARIABLES Subcommand 13.60 The TABLES Subcommand
 - 13.61 The CROSSBREAK Alternative Display Format
 - 13.62 The STATISTICS Command 203
 - 13.63 The OPTIONS Command 203
 - 13.64 Limitations 203
- 13.65 CONDESCRIPTIVE 204
 - 13.66 The Variable List 205
 - 13.67 Z Scores 205
 - 13.68 The STATISTICS Command 205
 - 13.69 Missing Values 206

- 13.70 CROSSTABS 207
 - 13.71 General Mode 207
 - 13.72 Integer Mode 208
 - 13.73 The VARIABLES Subcommand 13.74 The TABLES Subcommand
 - 13.75 Cell Contents 209
 - 13.76 Other Options 209
 - 13.77 The STATISTICS Command 210
 - 13.78 Limitations 210
- 13.79 FREQUENCIES 211
 - 13.80 The VARIABLES Subcommand 211
 - 13.81 General vs. Integer Mode
 - 13.82 The FORMAT Subcommand 212
 - 13.83 Table Formats 13.84 The Order of Values 13.85 Suppressing Tables 13.86 Index of Tables
 - 13.87 Bar Charts and Histograms 213
 - 13.88 The BARCHART Subcommand 13.89 The HISTOGRAM Subcommand
 - 13.90 The HBAR Subcommand
 - 13.91 Percentiles and Ntiles 215
 - 13.92 The PERCENTILES Subcommand 13.93 The NTILES Subcommand
 - 13.94 The STATISTICS Subcommand 215
 - 13.95 Missing Values 216
 - 13.96 Limitations 216
- 13.97 LIST 216
 - 13.98 The VARIABLES Subcommand 217
 - 13.99 The CASES Subcommand 217
 - 13.100 The FORMAT Subcommand 217
- 13.101 NONPAR CORR 218
 - 13.102 Specifying the Design 218
 - 13.103 The OPTIONS Command 219
 - 13.104 Limitations 219
- 13.105 NPAR TESTS 220
 - 13.106 One-Sample Tests 221
 - 13.107 One-Sample Chi-Square Test 13.108 Kolmogorov-Smirnov One-Sample Test 13.109 Runs Test
 - 13.110 Binomial Test
 - 13.111 Tests for Two Related Samples 223
 - 13.112 McNemar Test 13.113 Sign Test 13.114 Wilcoxon Matched-Pairs Signed-Ranks Test
 - 13.115 Tests for k Related Samples 224
 - 13.116 Cochran Q Test 13.117 Friedman Test 13.118 Kendall Coefficient of Concordance
 - 13.119 Tests for Two Independent Samples 225
 - 13.120 Two-Sample Median Test 13.121 Mann-Whitney U Test
 - 13.122 Kolmogorov-Smirnov Two-Sample Test 13.123 Wald-Wolfowitz Runs Test
 - 13.124 Moses Test of Extreme Reactions
 - 13.125 Tests for k Independent Samples 226
 - 13.126 k -Sample Median Test 13.127 Kruskal-Wallis One-Way Analysis of Variance
 - 13.128 Other Options and Statistics 226
 - 13.129 Limitations 227
- 13.130 ONEWAY 227
 - 13.131 Specifying the Design 228
 - 13.132 The POLYNOMIAL Subcommand 228
 - 13.133 The CONTRAST Subcommand 228
 - 13.134 The RANGES Subcommand 229
 - 13.135 User-Specified Ranges
 - 13.136 Missing Values 230
 - 13.137 Limitations 230
- 13.138 PEARSON CORR 230
 - 13.139 Specifying the Design 231
 - 13.140 Two-Tailed Significance Levels 231
 - 13.141 Missing Values 232
 - 13.142 Limitations 232

13.143	REGRESSION	232
13.144	The VARIABLES Subcommand	233
13.145	The DEPENDENT Subcommand	233
13.146	The Method Subcommands	234
13.147	VARIABLES Subcommand Modifiers	235
	13.148 The MISSING Subcommand	13.149 The DESCRIPTIVES Subcommand
	13.150 The SELECT Subcommand	
13.151	Equation Control Modifiers	237
	13.152 The CRITERIA Subcommand	13.153 The STATISTICS Subcommand
	13.154 The ORIGIN Subcommand	
13.155	Analysis of Residuals	239
	13.156 Temporary Variables	13.157 The RESIDUALS Subcommand
	13.158 The CASEWISE Subcommand	13.159 The SCATTERPLOT Subcommand
13.160	The WIDTH Subcommand	241
13.161	SCATTERGRAM	242
13.162	Specifying the Design	242
13.163	Scaling	243
	13.164 Setting Bounds	
13.165	Integer Scaling	243
13.166	Random Sampling	243
13.167	The STATISTICS Command	243
13.168	Missing Values	243
13.169	Limitations	244
13.170	T-TEST	244
13.171	Independent Samples	244
	13.172 The GROUPS Subcommand	13.173 The VARIABLES Subcommand
13.174	Paired Samples	245
13.175	Independent and Paired Designs	245
13.176	One-Tailed Significance Levels	246
13.177	Missing Values	246
13.178	Limitations	246

Appendix A Answers to Exercises 247

Appendix B Codebooks 262

Bibliography 272

Index 274

Chapter 1 From Paper into a File

Statistical software packages such as SPSS^x are used to analyze information. This information, called data, has many sources, such as public opinion surveys, laboratory experiments, and personnel records.

Sometimes the information to be analyzed is already stored in a form that can be processed by a computer—for example, on a disk or magnetic tape. Data from large-scale studies, such as those done by the United States Census Bureau or the National Opinion Research Center, are distributed in machine-readable form. In these situations, all that is required to analyze the data is a directory which describes the way in which the data are recorded and stored and the necessary commands to access the data.

However, often the information does not reside on a machine-readable medium. Instead, the data are stored in file folders in personnel offices, in patient medical charts, or in some other form that a computer cannot read. Before this information can be analyzed by a computer program, it must be entered onto cards, disk, or tape. This chapter examines the steps necessary to prepare data for analysis.

1.1 CASES, VARIABLES, AND VALUES

Consider Table 1.4, which contains data for five cases from a study designed to identify factors associated with coronary heart disease. In the Western Electric study, 2,017 men with no history of coronary heart disease were followed for 20 years, and the occurrence of coronary heart disease was monitored (see Appendix B for further information about this study). Much information was obtained for each participant at the beginning of the study and at various points during it. Table 1.4 contains only a very small subset of the data available for each man. Each line in the table represents a *case*, or observation, for which *values* are available for a set of *variables*.

For the first case, employee John Jones, the value for the age variable is 40 and the value for the height variable is 68.8 inches. The same variables—age, family history, first cardiac event, height in inches, day of death, cholesterol level—are recorded for all cases. What differs are the actual values of the variables. Each case has one and only one value for each variable. “Unknown” and “missing” are acceptable values for a variable, although these values require special treatment during analysis.

The case is the basic unit for which measurements are taken. In this analysis, the case is an employee of Western Electric. In studies of political opinion or brand preference, the case is most likely the individual respondent to a questionnaire. A case may be a larger unit, such as a school, county, or nation; it may be a time period, such as a year or month in which measurements are obtained; or it may be an event, such as an auto accident.

For any single analysis, the cases must be the same. If the unit of analysis is a county, all cases are counties, and the values of each variable are for individual counties. If the unit is a state, then all cases are states and the values for each variable are for states.

1.2 Identifying Important Variables

A critical step in any study is the selection of variables to be included. For example, an employee can be described using many variables, such as place of residence, color of hair and eyes, years of education, work experience, and so forth. The variables that are relevant to the problem under study must be chosen from the vast array of information available. If important variables are excluded from the data file, the results will be of limited use. For example, if a variable such as years of work experience is excluded from a study of salary discrimination, few—if any—correct conclusions can be drawn. All potentially relevant variables should be included in the study since it is much easier to exclude unnecessary variables from analysis than to gather additional information.

1.3 Recording the Data

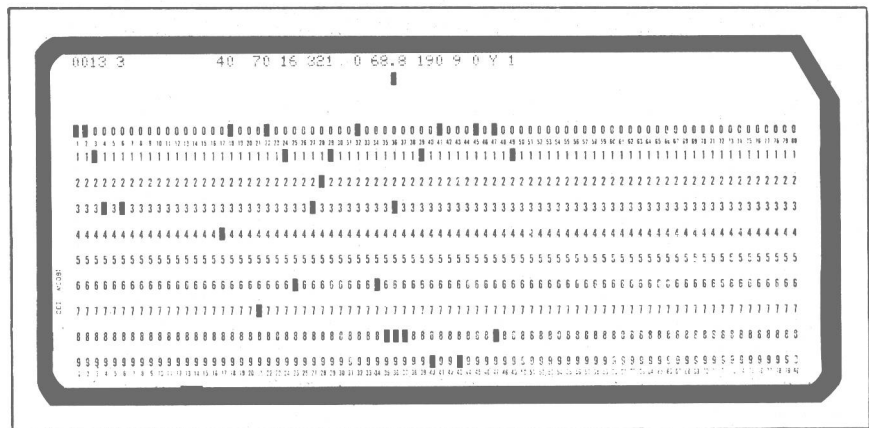
Once the variables have been selected, you must decide how they will be recorded. Do you need to record the actual date of birth or can you simply record the age in years? Is it sufficient to know if someone is a high-school or college graduate or do you need to know the actual number of years of education? It is usually a good idea to record the data in as much detail as possible. For example, if you record actual ages, cases can be grouped later into age categories. But if you just record each case as over 50 years or under 50 years of age, you can never analyze your data using any other age categories.

1.4 Coding the Variables

To enter observations into a data file, the values of the variables must be typed onto punched cards using a keypunch or entered directly into a disk file using a terminal. Punched cards are rectangular pieces of stiff paper on which 80 characters of information can be entered. Each character is represented by the position of holes in the card. Figure 1.4 shows a typical punched card. A terminal is like an electric typewriter; the information entered with it is directly stored on a disk.

One way to simplify data entry is to assign numbers or symbols to represent responses. This is known as *coding* the data. For example, instead of typing "Yes" or "No" as the values for the family history variable, you can use the codes Y and N. If only numbers are included in a coding scheme it is called *numeric*. If letters or a mixture of numbers, letters, and special symbols are chosen, the code is termed *alphanumeric*. By coding, you substantially decrease the number of symbols that you need to type, especially for variables whose values are originally recorded as words (such as state names).

Figure 1.4 A punched card



Coding schemes are arbitrary by their very nature. The family history variable could also be coded 0 for no and 1 for yes. All that is necessary is that each possible response have a distinct code. For example, coding the states by their first letter is unacceptable since there are many states that begin with the same letter. Maine, Massachusetts, Michigan, Maryland, Minnesota, Mississippi, Missouri, and Montana would be indistinguishable.

It is usually helpful to have one variable that uniquely identifies each case. For the Western Electric employee data, that variable could be the name of the individual. But, since names are generally long and not always unique, an ID number can be used as an identifier. This identifier can help you easily locate the records for cases with unusual values or missing information. Without the identifier, there is no quick way to find the correct age for an employee with a value of 12 in the data file.

Table 1.4 Excerpt from uncoded data for Western Electric study

Name	First event	Age	Diastolic BP	Education	Cholesterol	Cigarettes
John Jones	Nonfatal MI	40	70	B.A.	321	0
Clark Roberts	Nonfatal MI	49	87	11th grade	246	60
Paul Buttons	Sudden death	43	89	High school	262	0
James Smith	Nonfatal MI	50	105	8th grade	275	15
Robert Norris	Sudden death	43	110	Unknown	301	25

Height	Weight	Day of week	Vital10	Family history	Incidence of CHD
68.8	190	None	Alive	Yes	Yes
72.2	204	Thursday	Alive	No	Yes
69.0	162	Saturday	Dead	No	Yes
62.5	152	Wednesday	Alive	Yes	Yes
68.0	148	Monday	Dead	No	Yes

1.5 An Example

Consider the coding scheme in Table 1.5. Figure 1.5a contains data for the first three cases from Table 1.4 coded according to this scheme. Once the data are coded, a format for arranging the data in a computer file must be determined. Each punched card or line of type (if data are entered from a terminal) is known as a *record*. Each record is composed of columns in which the numbers or characters are stored. Punched cards have a maximum record length of 80 columns. Records stored on tape or disk can be longer. Two decisions that must be made are how many records will be needed for each case and in what column locations each variable will be stored.

Table 1.5 Coding scheme for employee data form

VARIABLE	CODING SCHEME
ID	no special code
FIRST CHD EVENT	1=No CHD 2=Sudden death 3=Nonfatal myocardial infarction 4=Fatal myocardial infarction 6=Other CHD
AGE	in years
DIASTOLIC BP	in mm of mercury
EDUCATION	in years
CHOLESTEROL	in milligrams per deciliter
CIGARETTES	number per day
HEIGHT	to nearest 0.1 inch
WEIGHT	in pounds
DAY OF WEEK	1=Sunday 2=Monday 3=Tuesday 4=Wednesday 5=Thursday 6=Friday 7=Saturday 9=Unknown
VITAL10	status at 10 years 0=Alive 1=Dead
FAMILY HISTORY OF CHD	N=No Y=Yes
CHD	0=No 1=Yes

Figure 1.5a Coded data

CASEID	FIRSTCHD	AGE	DBP58	EDUYR	CHOL58	CGT58	HT58	WT58	DAYOFWK	VITAL10	FAMHXCVR	CHD
13	3	40	70	16	321	0	68.8	190	9	0	Y	1
30	3	49	87	11	246	60	72.2	204	5	0	N	1
53	2	43	89	12	262	0	69.0	162	7	1	N	1

Figure 1.5b shows a listing of a file in which one record is used for each case. The column locations for the variables are also indicated. The ID number is in columns 1–4; first event is in column 6; age is in columns 17–18; diastolic blood pressure is in columns 20–22; years of education is in columns 24–25; cholesterol level is in columns 27–29; number of cigarettes smoked per day is in columns 31–32; height is in columns 34–37; weight is in columns 39–41; day of death is in column 43; status at 10 years is in column 45; family history of coronary heart disease is in column 47; and incidence of coronary heart disease is in column 49. The numbers are positioned in each field so that the last digit is in the last column of the field for the variable. For example, an ID number of 2 would have the number 2 in column 4; leading blanks or zeros occupy the beginning columns. This is known as *fixed-column format*. (Freefield input is discussed in Chapter 11.) The decimal point for the height variable is included in the file. However, it does not need to be included since SPSS^x commands can be used to indicate its location. If the decimal point is included, it occupies a column like any other symbol.

Figure 1.5b File with one record per case

0	0	1	1	2	2	3	3	4	4	5	
1	5	0	5	0	5	0	5	0	5	0	Columns
13	3			40	70	16	321	0	68.8	190	9 0 Y 1
30	3			49	87	11	246	60	72.2	204	5 0 N 1
53	2			43	89	12	262	0	69.0	162	7 1 N 1
				.							
				.							

When there are many variables for each case, more than one record may be necessary to store the information. In Figure 1.5c, the first case (CASEID 13) occupies two records. The first record contains codes for first cardiac event, age, diastolic blood pressure, education, cholesterol, and cigarettes smoked. The second record contains codes for height, weight, day of death, status at 10 years, family history of coronary heart disease, and incidence of coronary heart disease. Each record contains the case ID number in columns 1–4 and a record identification number in column 50. It is usually recommended that you enter the identification number and record number onto all records for a case. You can then easily locate missing or out-of-order records.

Figure 1.5c File with two records per case

0	0	1	1	2	2	3	3	4	4	5	
1	5	0	5	0	5	0	5	0	5	0	Columns
13	3	40	70	16	321	0					1
13	68.8	190	9 0 Y 1								2
				.							
				.							

It is important to allocate a sufficient number of columns for each variable. For example, if only two columns are used to record a weight variable, only weights less than 100 pounds will fit. Always allocate the maximum number of columns that you might need. Don't worry if your observed data do not actually require that many columns.

All data files considered in this manual are *rectangular*. That is, all cases have the same variables and the same number of records per case. Some data files are not rectangular. The same variables may not be recorded for