MINISTERIO DE INDUSTRIA Y ENERGIA
SECRETARIA DE LA ENERGIA Y RECURSOS MINERALES

ANALISIS METODOLOGICO DE LAS TECNICAS GEOQUIMICAS EMPLEADAS EN PROSPECCION GEOTERMICA

"APENDICE B"

" Introduction to SPSS and BMDP Statistical Programs "

INSTITUTO GEOLOGICO Y MINERO DE ESPAÑA

Julio 1985

APPENDIX B

INTRODUCTION TO SPSS AND BMDP STATISTICAL PROGRAMS

The complete programs may be obtained from:

SPSS Inc., 444 North Michigan Dr., Suite 3000
Chicago, Il 60611

&

BMDP Statistical Software Inc.
1964 Westwood Blvd, Suite 202
Los Angeles, CA 90025

BMDP Statistical Software 1981

W. J. Dixon, chief editor

M. B. Brown
L. Engelman
J. W. Frane
M. A. Hill
R. I. Jennrich
J. D. Toporek

UNIVERSITY OF CALIFORNIA PRESS

BERKELEY · LOS ANGELES · LONDON · 1981

This publication reports work sponsored in part under grant RR-3 and continuation contract number N01-RR-8-2107 of the Biotechnology Resources Branch of the National Institutes of Health. Reproduction in whole or in part is permitted for any purpose of the United States Government

Orders for this publication should be directed to

UNIVERSITY OF CALIFORNIA PRESS
2223 Fulton Street
Berkeley, California 94720

UNIVERSITY OF CALIFORNIA PRESS, LTD.
London, England

Library of Congress Catalog Card Number: 80-53770

Comments on programs or orders for copies of the
program should be addressed to
BMDP Statistical Software
as described in Appendix D.

Manufactured in the United States of America

Contents

Features Common to All Programs

Types of Analyses — Program Descriptions

Appendices

Tables

Preface

This manual describes the capabilities and usage of the BMDP computer programs. These programs provide a wide variety of analytic capabilities that range from plots and simple data description to advanced statistical techniques. The first chapter outlines the organization of the manual and suggests how to use it.

This edition differs in many respects from earlier editions; therefore we recommend that you read Chapter One even if you are familiar with the previous editions.

The first BMD Biomedical Computer Programs manual appeared in 1961, and was followed by numerous editions. Each new edition included new programs with improved features, novel statistical techniques and more robust statistical algorithms. In 1968 we began to develop the English-based Control Language used with the BMDP programs described in this manual. This method of specifying instructions is more flexible than the fixed format used in the BMD programs. In addition, repeated analyses of the same data, or similar analyses of multiple sets of data, can be done by stating a minimum number of Control Language instructions.

In this edition we attempt to integrate an extensive discussion of most program features with examples of how they are used. There are numerous input/output examples, many of which are annotated. Our emphasis is on the use of the statistical techniques and not on the numerical results obtained -- therefore we explain the terms used in the results, but do not repeat the numerical answers in the discussion of the results.

The manual is written for new users of computer programs as well as for experienced statisticians. Chapter One gives an outline of the entire manual and a guide to its use. Chapter Two describes the scope of analytical techniques available, basic terminology and a short description of each program. In chapters four through seven we discuss the specifications common to many programs and explain how they are specified in Control Language. Chapters eight through twenty are devoted to the individual programs. The programs are discussed in relation to the analyses they perform. Each program is extensively illustrated by annotated input/output examples. Difficult formulas and computational algorithms are provided in an appendix.

New chapters are written by the designated authors. Revisions of material appearing in previous editions were done primarily by James Frane, MaryAnn Hill and W.J. Dixon. This process was simplified by the expert editing provided by Morton Brown who edited the previous edition.

This revision could not have been completed without the support of many BMDP staff members. In addition to those mentioned elsewhere we wish to acknowledge the assistance of Noel Wheeler, and Linda Moody.

Ellen Sommers tested and retested the programs to document how the options work and prepared the examples in this manual.

Throughout Janice Cammell provided expert technical editing, improved phrasing, organization and attention to detail. Ching Liu and Avis Williams expertly typed the many drafts of the manuscript and prepared the camera-ready copy. The text in this manual was produced with the aid of a text-editing program produced by the Clinical Evaluation Unit of the Brentwood Veterans Administration Medical Center. Layout was done by Barbara Widawski.

Appendix D contains information on ordering the programs and the computers on which they are available. Although the programs have been under test and development for some time we know difficulties may arise in their use; these difficulties should be reported to us as described in Appendix D. Your comments and criticisms about the programs and this manual will also be appreciated.

In the first edition of the manual 26 programs were described. The 1979 edition contained 36. This edition contains 42 programs. The development of new programs and revision of previously released programs is an ongoing project. New options are continually being added. the new programs in this manual are:

BMDP4F - Two-way and Multiway Frequency Table Analysis - includes all of the essential features of the older BMDP1F, BMDP2F, and BMDP3F programs and adds new features for screening log-linear models. Models can be selected in a stepwise manner. Structural zeros are permitted for both two-way and multiway tables. Cells or strata whose frequencies deviate from the expected frequencies under a proposed model can be identified in a stepwise manner.

BMDP2L - Cox Models for Survival Analysis - analyzes Cox style proportional hazards survival analysis models with covariates. There can be both time-independent and time-dependent covariates. Models can be selected in a stepwise manner by either exact maximum likelihood or by an efficient approximation using partial derivatives. Several plots can be selected to accompany the results.

BMDP1T - Spectral Analysis - performs spectral analysis. Control Language is a series of brief commands to allow easy interactive use. Features include estimation of spectral densities, coherences, variable vs. time plots, lagged scatter plots, and complex demodulation plots. Attention is given to missing values, removal of seasonal means and linear trends and the use of several kinds of filters.

BMDP2T - Interactive Box-Jenkins Analysis, Including Transfer Functions - deals with a general class of time series models that includes ARIMA, intervention, and transfer function components. For the ARIMA component, it allows any number of autoregressive and moving-average factors and each factor may have any number of parameters of any order of lags. Intervention and transfer function components also have dynamic structure similar to the ARIMA components. Plots include individual and multiple time series, autocorrelation, partial autocorrelation and cross correlation functions. Control Language is designed for interactive and batch use. Output is adjusted to the width of a computer terminal and can be performed in a series of steps guided by the user.

BMDP4V - Univariate and Multivariate Analysis of Variance and Covariance, Including Repeated Measures and Cell Weights - expands the analysis of variance capabilities of the older BMDP2V. (It was originally developed as URWAS, the University of Rochester Weighted ANOVA System.) Some of its new features include Greenhouse-Geisser and Huynh-Feldt adjustments to the univariate approach to repeated measures; univariate and multivariate approaches to repeated measures; covariates for both fixed effects and repeated measures; user defined cell weights (cell importance) for hypotheses tested; interactive analysis of submodels; and orthogonalization of effects under user control in order to yield various types of sums of squares for unbalanced designs.

BMDP8M - Boolean (Binary Data) Factor Analysis - performs a factor analysis of dichotomous (binary) data. The analysis in this program differs from that of classical factor analysis (see P4M) based on binary data even though the goal and model (symbolically) appear similar. The goal is to express p variables by m factors where m is considerably smaller than p.

The arithmetic used in the matrix multiplication is Boolean, so the scores and loadings are binary. A case has a score of one if it has a positive response for any of the variables dominant in the factor (those not having zero loadings) and zero otherwise. The program has been found useful in serological and other studies.

BMDP9M - Linear Scores from Preference Pairs - constructs for each case a score that is a linear combination of the variables where the coefficients are based on the judgments (preferences) of experts. The expert compares cases two at a time, stating which of the pair he prefers. The analysis weights the observed variables to best replicate the preference of the judge (whatever subjective or objective information he may use). The expert need not judge all possible pairs of cases.

The score is determined in a stepwise manner weighting the variables by their importance to the expert. Preferences from more than one expert can be analyzed.

Other New Features

Other modules added since the 1977 edition are:

BMDPKM - K-means Cluster Analysis
BMDPLR - Stepwise Logistic Regression
BMDP8V - Mixed-model Anova, Equal Cell Sizes

Important modifications have been made to several programs:

BMDP1D - Option for file sorting
BMDP2D - Option for stem and leaf display
BMDP3D - Robust option for t statistics
BMDP6D - Multiple pairs of variables in bivariate plots permitted
BMDP7D - Capacity to handle more groups in side-by-side histograms, plot of cell means vs. standard deviations
BMDP8D - Completely rewritten with a much more efficient algorithm for computing correlations from incomplete data
BMDP2M - Option for single linkage and related case clustering
BMDPAM - Maximum likelihood estimation of covariance matrices from incomplete data, display of pattern of occurrence of missing values after clustering pattern of missingness by rows and columns
BMDP3R - Specification of function and derivatives for nonlinear regression through BMDP Control Language rather than FORTRAN
BMDPAR - Specification of function for derivative-free nonlinear regression through BMDP Control Language rather than FORTRAN, modifications to permit pharmacokinetic models defined by differential equations
BMDP2V - Cell mean vs. cell standard deviation plot, Greenhouse-Geisser and Huynh-Feldt adjustments to degrees of freedom for repeated measures

Features added to all BMDP modules include:

- Character handling through the transformation processor, e.g., conversion of alphabetic codes to numeric codes
- Easier specification of category codes and names
- Free-format data reader
- Fortran variable format data reading replaced by conversion within BMDP of card images to numbers for enhanced diagnostics and permitting blanks as missing value codes for all computer systems
- Codebook interpretation of input format
- Dynamic storage mechanism to allow easy specification of very large analyses
- CPU usage report, date and time report
- Increased control over output, e.g., for interactive execution
- Output of data files in binary, F, and G format
- Printing of problem title at top of each page
- Improved data printing
- Specification of certain special characters in variable names without the need for enclosing them in apostrophes, e.g., the use of underscore allows easier interface with SAS
- Enhanced portability, especially for CDC and 16 bit machines

ACKNOWLEDGEMENTS

It is impossible to properly credit all who have contributed to the development of these programs. Each program goes through many stages from the initial planning to the final release for distribution. In this manual we list as author(s), where possible, the person(s) who was most instrumental in the development of the program. Generally this person showed originality in the design of the program and also contributed to the statistical methodology required for the analytic technique. At the end of each program we name the designer and programmer. But these credits do not, and cannot, fully cover the contributions of our staff.

Laszlo Engelman designed the basic framework of the BMDP programs, such as the Control Language used to specify instructions, the methods used for transformations and the method of saving data and results between analyses (the BMDP (Save) File). He also programmed many of the subroutines common to all programs and several of the BMDP analyses. For many years he was supervisor of applications programming and supervised the development of many of the programs.

Robert Jennrich proposed and designed many of the programs in regression analysis, analysis of variance and discriminant analysis. He made significant contributions in nonlinear regression (P3R), factor analysis (P4M) and the analysis of variance (P2V and P3V). He supervised Mary Ralston's Ph.D. thesis, which provides the algorithm for the derivative-free nonlinear regression. He is currently involved in planning many other programs.

Jim Frane made significant contributions in the area of multivariate analysis. He designed and programmed several of the analyses at a level that makes the programs more accessible to the user; he is now supervisor of applications programming and is deeply involved in making improvements and program testing.

Alan Hopkins developed the life table programs.

Morton Brown developed the frequency table programs, and John Hartigan (of Yale University) contributed greatly in the area of cluster analysis.

Other staff members, such as Al Forsythe, Ray Mickey and Jerry Toporek made significant contributions to many of the programs.

Many statisticians have made contributions and suggestions for developing the programs: R.L. Anderson in regression and analysis of variance, Virginia Clark and Robert Elashoff in survival analysis, Robert Ling in the method of pictorially representing a matrix, and John Tukey in data analysis. Valuable comments have been received from David Andrews, Peter Claringbold, Cuthbert Daniel, Charles Dunnett, Janet Elashoff, Ivor Francis, George Furnival, Don Guthrie, Henry Kaiser, Sir Maurice Kendall, H.L. Lucas, John Nelder, Shayle Searle, Frank Stitt, Max Woodbury, Karen Yuen, and Coralee Yale. This list is incomplete. Many statisticians have visited the Department of Biomathematics, and/or have discussed the programs with us at conferences. Even the list of references, is, of necessity, incomplete.

Tony Thrall, Lon-Mu Liu and Laszlo Engelman developed the time series programs receiving helpful comments from John Tukey, David Brillinger, Peter Bloomfield, George Box and George Tiao.

During the past three years NIH provided us with an Advisory Committee that considered directions for our research making suggestions for improvements in numerical methods, program design and standards, statistical techniques, exchange of information with statisticians and the statistical computing community.

BMDP has benefited from the work of many conversion centers that have customized BMDP for use on a variety of computer systems. For a complete list of systems contact BMDP at (213) 825-5940. Some of the many who have been responsible for conversions include Rachel Countryman (Burroughs), Eli Cohen (CDC 6000, Cyber), Michael Matzek (DEC 10/20), Gary Anderson (HP 3000), M.M. Barritt (ICL 2900), James Krupp (PDP-11), Rob Charlton (Prime), Malte Sund (Siemens), Robert Byers (Univac 70/90), Ann Coleman (Univac 1100), Bernie Ryan (VAX), Randal Leavitt (Xerox), Lois Secrist (Perkin-Elmer), Aenea Reid (Honeywell 66), H. Koll (Telefunken), and W. Haase (MODCOMP Classic).

W. J. Dixon

1
INTRODUCTION

The BMDP computer programs are designed to aid data analysis by providing methods ranging from simple data display and description to advanced statistical techniques. Data are usually analyzed by an iterative "examine and modify" series of steps. First the data are examined for unreasonable values, graphically and numerically. If unreasonable values are found they are checked and, if possible, corrected. An analysis is then performed. This analysis may identify other inconsistent observations or indicate that further analyses are needed. The BMDP programs are designed to handle all steps in an analysis, from the simple to the sophisticated.

The BMDP programs are organized so the problem to be analyzed, the variables to be used in the analysis, and the layout of the data are specified in a uniform manner for all programs. This permits different analyses of the same data with only minor changes in the instructions.

This manual is arranged by the type of analysis appropriate to the data. Included are chapters on data description and screening, plotting, frequency tables, regression, analysis of variance, multivariate analysis, etc. Each chapter describes the programs that are available to do a specific type of analysis. In the introduction to each chapter the programs are described and contrasted with each other to indicate which is preferred for a specific analysis. Programs in other chapters are cross-referenced if they provide a similar function.

The programs are loosely classified into series:

D:	data description
F:	frequency tables
R:	regression analysis
V:	analysis of variance
M:	multivariate analysis
L:	life tables and survival analysis
S:	special (miscellaneous)
T:	time series

Many programs cross boundaries between two series. For example, multivariate regression belongs to both the multivariate (M) and regression (R) series.

Each program is identified by a three-character code; the first is P (from BMDP) and the last is the series classification. The middle character is assigned when the programming begins; it can be 1-9 or a letter. The order does not indicate increasing complexity, and some numbers do not appear.

For example, the program "Simple Data Description" is labelled P1D since it is the first program in the Descriptive series, and "Nonlinear Regression" is labelled P3R as the third program in the Regression series. Since programs in this manual are described by content, "Multivariate Regression" (P6R) is explained in the chapter on Multivariate Analysis with programs from the M-series.

New programs are continually being developed and released. The first edition of the BMDP manual (Dixon, 1975) contained 26 programs, the 1979 edition thirty-six. New in this manual are:

P4F --	Frequency Tables -- Two-way and Multiway This program replaces P1F, P2F and P3F and expands the capability for analyzing contingency tables
P2L --	Survival Analysis with Covariates -- Cox models
P8M --	Boolean Factor Analysis
P9M --	Linear Scores for Preference Pairs
P1T --	Univariate and Bivariate Spectral Analysis
P2T —	Box-Jenkins Time Series Analysis
P4V —	General Univariate and Multivariate Analysis of Variance, Including Cell Weights and Repeated Measures

In addition, all programs are reviewed and revised in response to suggestions from users, to improve efficiency or to correct errors. For example:

- P3D (t tests) has been expanded to give trimmed t statistics and Levene's test for equality of variances
- P7D now plots an estimate of the cell standard deviation vs. the cell mean and also plots the logs of each cell statistic. The slope of the regression line for the second plot is used to determine a transformation for stabilizing variances
- P8D (Missing Value Correlation) has been rewritten to include a much more efficient algorithm
- PAM (Description and Estimation of Missing Values) has been expanded to include the maximum likelihood method of estimating covariance and correlation matrices from incomplete data
- P2V (Repeated Measures ANOVA) now includes the Greenhouse-Geisser and Huynh-Feldt adjustments to degrees of freedom.

In addition, diagnostic messages for data reading have been enhanced and free-formatted data reading is now available.

Major changes in the programs and novel ways to use them are documented in the newsletter, BMDP Communications. Articles from BMDP Communications that describe ways to use the programs are reprinted in Appendix C.

This manual describes the current status of the programs. Since your facility may have an earlier version of the programs, we have included notes in the manual regarding changes since our November 1978 release. Note that each BMDP program prints a date in the upper left corner of the first page of the output.

An abbreviated manual, the BMDP User's Digest and a BMDP Control Language Pocket Guide are available. See Appendix D.

1.1 A GUIDE TO THIS MANUAL

Introductory Material (Chapters 1 through 3)

The scope of the statistical analyses provided by the BMDP programs is discussed in Chapter 2. Section 2.1 introduces terms used throughout this manual and describes analytical features available in more than one program. The scope of the possible analyses with the BMDP programs is outlined in Section 2.2.

For readers who are using a computer for the first time, Chapter 3 gives annotated examples of simple analyses, describes how to organize your data sheets (research forms), and describes the layout (format) of your data.

Features Common to All BMDP Programs (Chapters 4 through 7)

Chapter 4 describes the English-based BMDP instruction language used to describe the data and specify the analysis. The terminology and notation used throughout this manual are defined here. We recommend that you read the definitions in Section 4.1 even if you are familiar with the BMDP programs. The BMDP instructions used to describe the data and variables are presented in Chapter 5. The free-format data reader described in Chapter 5 is available in all programs except P4D. Methods of transforming and editing data are treated in Chapter 6. Since an analysis often requires multiple steps, the data or results from one program can be saved in a BMDP (Save) File (Chapter 7) and then used in other programs. The use of a BMDP File eliminates having to repeat the description of your data.

Program Descriptions (Chapters 8 through 20)

Chapters 8 through 20 describe methods of analysis and the programs available to perform them. Each chapter begins with an introduction describing alternate methods and programs. This is followed by detailed descriptions of each program. (The features described in Chapters 4 through 7 are not repeated).

Each program description begins with a short abstract and a list of the examples and other options & features. This is followed by one or more examples that illustrate the simplest or most common usage of the program. Each example consists of the BMDP instructions that are required by the program and the results produced by the program. The instructions and results are labelled Example and Output respectively and are identified by the last two characters of the program name and a sequence number within the program, e.g., 2R.3 denotes the third example for BMDP2R. Numbers in circles (e.g., ①, ②, ...) are used to annotate the results; the numbers on the output correspond to circled numbers in the legend or the text.

A list of program options follows the example(s) with page references to where they are discussed. Each option is described; examples are provided for many of them.

The BMDP programs can analyze large amounts of data. Near the end of each program description a statement is made of the largest problem the program can analyze without modification. A more detailed formula for determining the size of problem the program can analyze is given in Appendix B. If your problem exceeds the limit, Appendix B also contains a description of the changes needed to analyze larger problems.

The last section in the program description states the formulas and algorithms that are described in greater detail in Appendix A. The more difficult formulas and computational procedures are collected in the appendix.

Each program concludes with a summary. The summary consists of a table that describes the BMDP instructions used in the program, and provides short definitions and page references to explanations of the program options. These tables can be used as indexes to the program descriptions.

Useful Aids

An index to the manual is included in the last few pages.

On the inside front cover the programs are listed by their three-character identification codes; page references are given to program descriptions and summaries.

On the inside back cover the Control Language common to all programs (and described in Chapters 4-7) is presented in summary form. Opposite the inside back cover, space is left for you to fill in the instructions necessary to begin an analysis on your computer (see Chapters 4-7).

Where to Start Reading

If you are using computers for the first time, we recommend that you start with Chapter 3. Then read Chapters 4 and 5 for a description of the Control Language. And finally, try one of the programs in Chapters 8, 9 or 10.

If you are not familiar with the BMDP programs, but have previously used a computer, we recommend that you read Chapter 2 (Section 2.1 and 2.2) for an overview of the programs and Chapters 4 and 5 for a description of the Control Language before turning to the specific analysis you want to do.

If you are already familiar with the BMDP programs, but not with this manual, we recommend that you skim Chapter 4. If you are already familiar with the program to be used, turn to the summary for the program (see the list of programs inside the front cover); otherwise you can either turn to the chapter describing the type of analysis that you want to do, or to Chapter 2 for an overview of all the programs.

2

DATA ANALYSIS

Using the BMDP Programs

INTRODUCTION

The meaning of "data analysis" is different for each of us, depending on our level of statistical training. Techniques used in data analysis vary from the simplest display of data in a histogram or a plot and the calculation of statistics (such as the mean and standard deviation) to advanced methods of multivariate analysis. To some, data analysis involves a single display or set of computations; to others it involves a sequence of steps, detecting the presence of outliers and inconsistencies, ensuring that assumptions necessary to the analysis are met, etc. Each step may suggest further analyses.

The BMDP programs provide many analytical capabilities -- from elementary to advanced. In this chapter we describe features that are available in more than one program and outline the scope of available analytical techniques.

WHERE TO FIND IT

Chapter Organization	Program Labels
8. Data description	1D, 2D, 4D
9. Data in groups	3D, 7D, 9D
10. Plots and histograms	5D, 6D
11. Frequency tables	4F
12. Missing values	8D, AM
13. Regression	1R, 2R, 9R, 4R, 5R
14. Nonlinear regression and maximum likelihood	3R, AR, LR
15. Analysis of variance and covariance	1V, 2V, 3V, 4V, 8V
16. Nonparametric analysis	3S
17. Cluster analysis	1M, 2M, 3M, KM
18. Multivariate analysis	4M, 6M, 6R, 7M, 8M, 9M
19. Survival analysis	1L, 2L
20. Time series	1T, 2T

2.1 BASIC TERMINOLOGY FOR COMMON FEATURES AND STATISTICS

Each BMDP program is designed to provide a certain analytical capability, such as data description, plots, or regression analysis. Some statistics, plots, or other results are provided by several programs.

In this section we describe some of the common features and introduce terminology that is used throughout this manual. Table 2.1 (p. 7) lists the features and identifies the programs that contain each feature.

Data and Acceptable Values

Data are codes representing characteristics (e.g., sex, eye color), values of measurements (e.g., height, weight) or responses to questions. Each characteristic, measurement or response is called a variable. Each case contains values for all the variables for one subject, animal, or sampling unit. A case may represent such things as responses to questions, outcomes of tests or measurements made on a subject or test animal.

Some values in a case may not be recorded. A value that is not recorded may be left blank or may be recorded with a special code; the blank or special code, whichever is used, is called a missing value code and the unrecorded value is called a missing value. Missing values are excluded from all computations.

In any program you can restrict the analysis of a variable to a specified range by assigning an upper limit (maximum) and a lower limit (minimum) for values of the variables. A value that is greater than the upper limit or less than the lower limit is out of range and is excluded from all computations.

An acceptable value is one that is not equal to a missing value code and is not out of range. A complete case is a case in which the values of all the variables are acceptable (there are no values missing or out of range).

Cases Used When Some Values are Unacceptable (Missing)

All BMDP programs (except P4D) check for values missing and out of range. The treatment of cases depends on the primary purpose of the analysis. For example, P1D -- Simple Data Description -- computes statistics for each variable from all acceptable values for the variable; but P2R -- Stepwise Regression -- uses only complete cases to estimate the linear regression equation.

The BMDP programs include cases in an analysis according to one of three criteria:

A All cases are included (all acceptable values for each variable are used)
B Cases are included only if they have acceptable values for all variables specified in the analysis
C Only complete cases are included (cases that have acceptable values for all variables)

The difference between the above criteria can be explained in terms of an example. Suppose we have three variables, each of which has some unacceptable values. The data for five cases are:

	height	weight	age
1)	67	130	-
2)	64	106	21
3)	65	117	27
4)	67	-	24
5)	-	121	29

We request means for the first two variables only. Using Method A, the mean of each variable is computed from all its acceptable values, whether or not either of the other variables have acceptable values (i.e., four values for each variable). By Method B the means of the first two variables are computed from data in only those cases that have acceptable values for both variables (cases 1, 2 and 3). Method C allows the means to be computed from cases that have acceptable values for all three variables (only cases 2 and 3).

Method C may use fewer cases than either of the other methods. You can specify a list of variables to be checked by Method C with the USE list in the VARIABLE paragraph. Variables that are excluded from the list are not used in any of the computations (not even their means are computed). In Chapter 5 we describe how this list is specified.

The method used by each program is shown in Table 2.1. Three programs, P3D, P8D and PAM allow you to choose explicitly between methods. P8D computes estimates of correlation matrices using all observed values instead of only those values from complete cases. PAM does the same and, in addition, provides estimates to fill in where observations are missing. PAM also has features useful for describing the pattern of where values are missing.

Transformations and Selecting Subpopulations for Analysis

Transformations can be used to replace the value of a variable by its transformed value; e.g., weight by the logarithm of weight. Also, new variables can be created from the observed variables by transformations. For example, if pulse rate is measured before and after exercise, the difference between the two measurements can be a meaningful quantity. This difference can be specified as a new variable and can be used in the analysis. Any number of new variables can be treated as functions of the observed variables. These functions, called transformations, can involve arithmetical operations (e.g., +, -, *, /), powers, trigonometric functions, and complex conditional statements (IF(A/B-C LT 0) THEN D=0.).

You can also select cases to be used in an analysis (e.g., data for males and for respondents in their twenties -- USE = SEX EQ 1 AND AGE GE 20 AND AGE LT 30.).

Methods of specifying transformations and case selection are described in detail in Chapter 6.

How to Define Subpopulations or Groups

Some analyses, such as a t test between two groups or a one-way analysis of variance, require that the cases be classified into groups. The variable whose values are used to classify the cases into groups is called a grouping variable. Groups can be identified as codes (such as sex codes 1 and 2 for males and females) or as intervals (such as age 10-19, 20-29, etc.).

In all programs you can select cases belonging to specific groups (fulfilling certain criteria) by case selection (Chapter 6). The purpose of some analyses is to compare groups, such as in a plot or by a t test or by an analysis of variance. In some programs you can explicitly specify groups to be analyzed (or plotted).

Univariate Statistics

Most programs compute the mean, standard deviation and frequency for each variable. In addition, other univariate statistics are computed in several BMDP programs. We review the definitions of these statistics below.

Let x_1, x_2, ..., x_N be the observed (acceptable) values for a variable in the cases used in an analysis. Then N is the sample size, frequency or count for the variable. The mean, $\bar{x}$, is defined as

$$\bar{x} = \Sigma x_j / N$$

(Other estimates of location, such as the median and more robust estimates, are available in P2D and P7D.)

The standard deviation, s, is

$$s = [\Sigma(x_j - \bar{x})^2/(N - 1)]^{1/2}.$$

The variance is s^2 and the standard error of the mean is $s/\sqrt{N}$. The coefficient of variation is the ratio of the standard deviation to the mean, $s/\bar{x}$. If a variable has a very small coefficient of variation, loss of computational accuracy can result due to the limited accuracy with which a number can be represented internally in the computer.

Many analyses require that the distribution of the data be normal, or at least symmetric. A measure of symmetry is skewness, and a measure of long-tailedness is kurtosis. The BMDP programs compute skewness, g_1, as

$$g_1 = \Sigma(x_j - \bar{x})^2/(Ns^3)$$

and kurtosis, g_2, as

$$g_2 = \Sigma(x_j - \bar{x})^4/(Ns^4) - 3$$

If the data are from a normal distribution, the standard error of g_1 is $(6/N)^{1/2}$ and of g_2 is $(24/N)^{1/2}$. A significant nonzero value of skewness is an indication of asymmetry -- a positive value indicating a long right tail, a negative value a long left tail. A value of g_2 significantly greater than zero indicates a distribution that is longer-tailed than the normal. We recommend that you also examine histograms when using these statistics for they are sensitive to a few extreme values.

The smallest observed (acceptable) value, x_{min}, and the largest, x_{max}, are printed by several programs. The range is $(x_{max} - x_{min})$. The smallest and largest standard scores (z-scores, z_{min} and z_{max} respectively) are also printed by some programs. We define z_{min} and z_{max} as

$$z_{min} = (x_{min} - \bar{x})/s \quad \text{and} \quad z_{max} = (x_{max} - \bar{x})/s .$$

Covariances and Correlations

Covariances and correlations are used in many statistical analyses. The covariance between two variables, x and y, is

$$\mathrm{cov}(x,y) = \Sigma(x_j - \bar{x})(y_j - \bar{y})/(N - 1) .$$

The correlation, r, between two variables is

$$r = \frac{\mathrm{cov}(x,y)}{s_x s_y} = \frac{\Sigma(x_j - \bar{x})(y_j - \bar{y})}{[\Sigma(x_j - \bar{x})^2\Sigma(y_j - \bar{y})^2]^{1/2}}$$

This is also called the product-moment correlation coefficient.

The correlation can also be computed after adjusting for the linear effect of one or more other variables. For example, we may want the correlation between x and y adjusted for z and w (sometimes referred to as correlation at a fixed value of z and w). This is called the partial correlation coefficient between x and y given z and w. It is equivalent to fitting separate regression equations in z and w to x and to y, and computing the correlation between the residuals from the two regression lines.

A multiple correlation coefficient (R) is the maximum correlation that can be attained between one variable and a linear combination of other variables. This is the correlation between the first variable and the predicted value from the multiple regression of that variable on the other variables. R^2 is the proportion of variance of the first variable explained by the multiple regression relating it to the other variables.

Plots and Histograms

An assumption of normality is required by many analyses. The assumption can be assessed by a normal probability plot. The assumption of normality is not usually with respect to the data, but with respect to the residuals, the differences between the observed value and the value predicted by the statistical model. Many programs plot the residuals in a normal probability plot.

Scatter plots of one variable against another are useful in examining the relationships between the two variables; they are also useful in assessing the fit of a statistical model (such as regression). Scatter plots of the data and results are provided in many BMDP programs.

Histograms, or bar graphs, are a basic tool in data screening. They can be used to screen for extreme values or for the shape of the distribution of data. Several BMDP programs plot histograms as part of their analyses.

Table 2.1 indicates which programs produce plots and histograms as part of their analyses. Chapter 10 describes two programs whose primary purpose is to provide plots and histograms in final form.

Printing the Data and Results for Each Case

Many programs can list the data for each case. Some programs print results for each case, such as predicted values and residuals from a regression or scores from a factor analysis. Several programs have special capabilities for listing the data:

- P4D can print the data in a compact card image form
- P4D can also print only cases that contain nonnumeric symbols
- P1D can list all the data so that each column contains all the values of one variable or so that all variables for one case are printed before those for the next case
- P1D can print only cases with missing values or only cases that have values out of range
- P1D can also print the data after sorting the cases according to one or more variables
- PAM can print, in a compressed list, the positions of the missing values and values out of range
- P2M and P4M can print standard scores

The BMDP File

A set of data is usually analyzed many times by BMDP programs. For example, the data may first be examined for extreme values (outliers) and for distributional assumptions; then necessary transformations can be performed, meaningful hypotheses tested, or relationships between the variables studied. The results of an analysis may suggest that further analyses are needed.

All programs can read a data matrix as input. All programs (except P4D) can copy the data into a BMDP File. The BMDP File is a means of storing your data or results from an analysis so you can reuse them more efficiently in other BMDP programs; the File can be created or read by any BMDP program (except P4D). There are several advantages to using a BMDP File:

- data are read efficiently from a BMDP File; the cost of reading a large amount of data from a BMDP File is substantially less than when a format statement is used
- many of the Control Language instructions, specified when the BMDP File is created, need not be respecified for each additional analysis. For example, the variable names, the indicators for missing values and values out of range, codes and names for categories, etc., are stored with the File
- data are stored in the BMDP File after transformations and case selection are performed
- the BMDP File is the only way to store results (such as factor scores, residuals from a regression analysis or a covariance matrix) so they can be analyzed further by other BMDP programs

Table 2.1 shows that all programs (except P4D) can save the data in a BMDP File. Many programs save results, such as predicted values or residuals. Some programs can also save a covariance or correlation matrix or some other matrix of results.

All programs but P4D accept data from a BMDP File (a BMDP File can be created by one program and read by a different program). Several regression and multivariate analysis programs accept the covariance or correlation matrix from a BMDP File, thus saving computer time and cost. P4F accepts multiway tables as input.

Case Weights

Most statistical analyses assume that the error of each observation has a constant variance. When the variance is not constant, the computations of the mean, standard deviation and other statistics are best done by weighting each case by the inverse of the variance. For example, the researcher knows the variance from previous work and includes its inverse as an additional variable for each case.

Case weights can also be used to represent the frequency of an observation when the same observation is made more than once but is recorded in only one case; however, except for the frequency table programs, the sample size will be the number of cases and not the sum of weights.

You can specify case weights in many BMDP programs. The effect of the case weight on the computation of the univariate statistics, covariance and correlation is described below. Let w be the case weight for the jth case. Then

$$\bar{x} = \Sigma w_j x_j / \Sigma w_j$$

$$s = \{\Sigma w_j (x_j - \bar{x})^2 / [(N - 1)\Sigma w_j / N]\}^{\frac{1}{2}}$$

$$\text{cov}(x,y) = \Sigma w_j (x_j - \bar{x})(y_j - \bar{y}) / [(N - 1)\Sigma w_j / N]$$

$$r = \frac{\Sigma w_j (x_j - \bar{x})(y_j - \bar{y})}{\{\Sigma w_j (x_j - \bar{x})^2 \Sigma w_j (y_j - \bar{y})^2\}^{\frac{1}{2}}}$$

where N is the number of acceptable observations used in the analysis with positive (nonzero) weights.

When case weights are not specified, w_j is set to one for all cases and the formulas are identical to the formulas given on p. 5.

Computational Accuracy

The computer represents each number by a binary sequence of limited accuracy. As a result there can be a loss of accuracy in certain types of computation, such as matrix inversion. Loss of accuracy is especially pronounced if a variable has a small coefficient of variation ($s/\bar{x}$) or if a variable has a very high multiple correlation with other variables.

All programs represent data values in single precision. Some programs do computations in single precision. Others do computations in double precision; these programs are the ones whose computations are most likely to be affected by a loss of accuracy if the computations are done in single precision.

2.2 OVERVIEW OF DATA ANALYSIS WITH BMDP

The breadth of techniques available in the BMDP programs is indicated by the chapter titles:

8. Data Description
9. Data in Groups -- Description, t Tests and One-Way Analysis of Variance
10. Plots and Histograms
11. Frequency Tables
12. Missing Values -- Patterns, Estimation and Correlation
13. Regression
14. Nonlinear Regression and Maximum Likelihood Estimation
15. Analysis of Variance and Covariance
16. Nonparametric Analysis
17. Cluster Analysis
18. Multivariate Analysis
19. Survival Analysis
20. Time Series Analysis

In this section we describe some of these techniques in general terms, and explain when the techniques are useful (see also "First Steps",

Table 2.1 Features common to BMDP programs

Chapter	8	9	10	11	12	13	14	15	16	17	18	19	6	20
Program	124 DDD	379 DDD	56 DD	4 F	8A DM	12945 RRRRR	3AL RRR	12348 VVVVV	3 S	123K MMMM	466789 MMRMMM	12 LL	1 S	12 TT
<u>Cases used</u>														
A all cases (all acceptable values)	AAA	AAA	AA	A	AA					AA	A		A	A
B cases with acceptable values for all vars. specified in analysis (see p. 4)		B				B B	B	BBBB			BB	BB		
C complete cases only					CC	CC C	CC	C	C	CC	C C C			C
<u>Analytic capabilities</u>														
T transformations and case selection	TTT	TTT	TT	T	TT	TTTTT	TTT	TTTTT	T	TTTT	TTTTTT	TT	T	TT
W case weights				W	WW	WWW W	WWW	WW		W W	WWWW W			
D double precision					D	D D	DDD	DDDD			DDDD	D		D
<u>Univariate statistics</u>														
$\bar{x}$ mean	$\bar{x}\bar{x}$	$\bar{x}\bar{x}\bar{x}$	$\bar{x}\bar{x}$	$\bar{x}$	$\bar{x}\bar{x}$	$\bar{x}\bar{x}\bar{x}\bar{x}$	$\bar{x}\bar{x}\bar{x}$	$\bar{x}\bar{x}$ $\bar{x}\bar{x}$	$\bar{x}$	$\bar{x}$ $\bar{x}$	$\bar{x}\bar{x}\bar{x}\bar{x}$	$\bar{x}$	$\bar{x}$	$\bar{x}\bar{x}$
$\tilde{x}$ other estimates of location	$\tilde{x}$	$\tilde{x}$												
s standard deviation	ss	sss	ss	s	ss	ssss	sss	ss s	s	s s	ssss	s	s	ss
v coefficient of variation	v				vv	vvvv					vvv			
g skewness and kurtosis	g				g	gg	g				ggg	g		
m min. and max. of the acceptable value	mm	mm		m	m	mmm	mmm	m	m	m	mmm	m	m	m
z z-score for min. and max. values	z				z	zz				z	zzz			
N number of cases (sample size)	NNN	NNN	NN	N	NN	NNNNN	NNN	NNNN	N	NNNN	NNNNNN	NN		NN
G above statistics computed for each group	G	GGG	GG		G	G		GG GG		G	G	G		
<u>Other statistics</u>														
r covariances and/or correlations		rr	r		rr	rrrr			r	r r	rrr			r
p partial correlations						p					p p			p
R squared multiple correlation (R^2 or SMC)					R	RRR					R R R			R
f frequency counts, for each value or category	fff		f	f			f			f		f		
% frequency counts, in %	%		%	%										
λ eigenvalues and/or eigenvectors					λ	λ					λλ λ			
G above statistics computed for each group		GG	GG	G	G	G				G				
<u>Test statistics</u>														
F comparison of group means (t,F)		FFF						FFFFF	F	F	F			
V comparison of group variances		VVV												
χ comparison of group or cell frequencies		χ		χ										
D multivariate comparison of group means (T^2 , D^2 , λ, U)		D						D		D	D			
<u>Plots and graphical displays</u>														
H histograms	H	HH	H											
N normal probability plots			N			NNNNN	NN				N			
S scatter plots of data			S		S	S				S	SSS S			
R scatter plots of results					R	RRRRR	RRR	R		R	RRRR R	R		
O other (see program description)	O	O								OOOO	O	OO		OO
<u>Prints for each case</u>														
D data (after transformations, if any)	D D	D			D	DDD D	DDD	D		D D	DDDDD	D	D	DD
S cases with special values	S S				S									
z standardized scores										z	z			
R residuals and/or predicted values				R		RRR R	RRR	RR R			RR	R		R
F factor, princ, comp. or canon. var. scores						F					FF FF			
M Mahalanobis distances					M	M				M	M M			
O other (see program description)		O			O		O			OOO	OO	OO		O
<u>BMDP File output</u>														
D data (after transformations, if any)	DD	DDD	DD	D	DD	DDDDD	DDD	DDDDD	D	DDDD	DDDDDD	DD	D	DD
R results for each case as part of data					R	RRR R	RRR	R			RRRRRR			R
G codes/cutpoints saved with data	G	GGG	GG	G	G	G	G	GGG			G			
C correlation and/or covariance matrix					CC	CCC				CCC				
O other (see program description)				O	O		O			O	O O O			O
<u>Input from BMDP File</u>														
D data	DD	DDD	DD	D	DD	DDDDD	DDD	DDDDD	D	DDDD	DDDDDD	DD	D	DD
C correlation or covariance matrix						CCC					CCC			
O other (see program description)											O O			O
<u>Input not from BMDP File</u>														
D data	DDD	DDD	DD	D	DD	DDDDD	DDD	DDDDD	D	DDDD	DDDDDD	DD	D	DD
O other (see program description)				O		O	O			O	OOO OO			O

Appendix C.11). Specific program options are presented in the program descriptions. Some of the more advanced techniques, such as maximum likelihood estimation, are discussed only in the program descriptions.

Data Screening and Description

The first step in an analysis is to examine the data for errors and for the appropriateness of assumptions to be used in the analysis (such as normality). If errors remain in the data they can cause a "garbage-in, garbage-out" analysis. Blunders or extreme outliers in the data may need to be removed to achieve a meaningful analysis. The data may need to be transformed to fit the various assumptions (constant variance, normality, etc.) required by the statistical model.

After the original data have been recorded, various descriptive characteristics of the data can be used to detect gross errors in the observations, in coding the data, in including inappropriate cases, etc. A good place to begin screening is to check for

- symbols or characters, such as letters where numbers should be (P4D counts all distinct characters for each column of data, one column at a time); many programs will not run if nonnumeric symbols are in the data used for analysis - error messages are reported, however, by all programs when illegal characters are found.
- outliers or blunders (P2D can be used to obtain a small histogram and frequency counts for all distinct values of each variable)

Listing the cases by one of the methods described on p. 5 may also locate problems in the data.

Outliers can be identified by multivariate screening. For each case P4M prints the Mahalanobis distance squared from the case to the center of all cases. Within group multivariate outliers can be identified by using P7M, which prints the Mahalanobis distance squared from the case to the center of each group. P9R also prints distance measures helpful for identifying unusual cases.

Univariate descriptive statistics are found in most programs, but especially in P2D and P1D. For example, from the cumulative percentiles printed in P2D for each distinct value, you can make summary statements such as, "sixty percent of the patients are in the 50-60 age group, while only seven percent are in their twenties", etc. A stem and leaf histogram is also available in this program.

Data in Groups

In screening, you often need to examine groups (strata or subpopulations) of the data. Unusual data values that are masked in a total population may stand out when the data are separated into groups or strata. Some variables are easily coded into groups, such as sex (males=1, females=2). Continuous variables can be categorized by a grouping variable.

P7D is especially powerful for examining groups; it prints histograms (side-by-side for each group) and statistics for each group; it also provides a choice of one-way or two-way analysis of variance to check group differences. From this output you can identify extreme outliers, obtain an idea of the distribution of data within groups, and examine whether the assumption of normality is reasonable. Heteroscedasticity (lack of constant variance over groups) can also be observed and tested, and may indicate that the input data should be transformed.

An analysis of variance using P7D can indicate whether group differences are large enough to suggest that future analyses should be stratified. P7D also computes ANOVA tests that do not assume equal group variances, plots cell standard deviations vs. cell means and reports Bonferroni probabilities for pairwise tests of cell means. More information on group differences (both univariate and multivariate) can be obtained by using P3D. It yields t statistics, Hotelling's T^2 and Mahalanobis D^2 for each pair of groups; t statistics, based on both pooled and separate variance estimates, are printed in the output. A trimmed t test is available in the 1979 version. The Levene test for equality of variances is in both P3D and P7D.

When the cases are classified by more than one grouping variable or factor, P9D (Multiway Description of Groups) can be used to compute cell frequencies, means and standard deviations. Grouping variables can be suppressed to obtain information about marginal cells. The program tests for the equality of cell frequencies and cell means and for homogeneity of cell variances. These tests are performed on all cells or on specified marginals. Cell means are plotted four variables per page in a compact graphical display scaled by the overall mean and standard deviation. This display is helpful for understanding interactions in more complex ANOVA designs.

Transformations

After screening and describing your data, you should be ready to make decisions regarding transformations. The transformed data can be put directly on a BMDP File ready for easy input into any other BMDP program. Although all programs can perform data transformations, you may need to use P1S, the multipass transformation program, for getting the data transformed and ready for further analyses. P1S can be used when your transformation requires more than one pass through the data.

Plots and Histograms

Many research workers like to see their data in graphical form; <u>scatter plots</u>, for example, are a good way to present information concisely and clearly in final reports. Scatter plots that take advantage of known information can be designed to display unusual cases or outliers -- for example, to show whether or not an individual's systolic blood pressure level is higher than his diastolic level. A scatter plot of these two variables will show if the data coding is mistakenly reversed for some cases. Or in a plot of height versus weight, a case that has a height of 72 inches and 225 lbs. will clearly stand out if the height is mispunched as 52 inches.

A grouping variable can be used in the P6D scatter plot program to provide information about a variable not used as the plot axes. If age, for example, is divided into groups -- less than or equal to 15, 16-35, 36-55 and over 55 -- the letters A, B, C and D are used to represent cases from each age group. When two other measurements for the subjects are plotted, the children may appear in a separate area of the plot, indicating that they should be analyzed separately in later analyses. P6D can also perform a simple regression analysis for the data in a scatter plot. This analysis may indicate whether or not an analysis of covariance should be used later. Variables can be plotted against time of entry into a study to see if observations are independent, or if a drift over time is occurring.

P5D can print a histogram for all the data or for one or more groups, each identified by a different letter. You can specify the scales of the histogram to produce a histogram suitable for a final report.

Normality can be roughly checked by looking at histograms in P7D or P5D. P5D can also print a normal probability plot that provides a better assessment of normality and helps to identify outliers.

Frequency Tables

Cross tabulations are frequently used as a form of final reporting to give a picture of the number of cases in specified categories (or cross-classifications). Tables can be formed from data or from cell frequencies. Tables can also be formed for each level of a third variable (such as separately for males and females). Twenty-three statistics appropriate for the analysis of contingency tables are available in P4F (which includes all the features formerly contained in programs P1F, P2F and P3F).

P4F can test whether rows are independent of columns using the frequencies in all cells. P4F can also test the same hypothesis using any subset of the cells; for example, are rows independent of the columns for all cells, excluding the cells on the diagonal? P4F can also identify cells that contribute heavily to a significant chi-square test of independence.

Multiway frequency tables are formed and analyzed by P4F. A log-linear model can be fitted to the cell frequencies and the fit tested. P4F can be used to select an appropriate model for the data and to estimate the parameters of the model.

Missing Values

All too often the data recorded are not complete and some values are missing. These missing values are usually left blank or coded by a special code called the "missing value code". Missing values, and unusually extreme values that appear to be wrong, are excluded from an analysis.

PAM lists cases containing missing values or data to be excluded from the analysis, computes the percentage of missing data for each variable, and reports special patterns in the data. PAM can also estimate values to replace the missing value code (or excluded values) based upon the data present in the case.

Most regression and multivariate analyses require complete cases; i.e., no missing or excluded values in any case. Many of these analyses can begin from a correlation or covariance matrix. Both PAM and P8D can estimate correlations using cases with some data missing; the correlation matrix can then be stored in a BMDP File and used as input to other programs, including those that require complete data. PAM insures that the resulting correlation matrix is numerically appropriate (positive semidefinite) for a regression or factor analysis; P8D allows you to choose between four methods to compute the correlations.

Regression

A regression analysis studies the relationship between a dependent variable, y, and one or more independent variables, x_i. The linear least squares model with parameters or regression coefficients, β_i, can be written

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e .$$

For simple linear regression (x_1 is the only independent variable in the model), P6D, P1R and P2R can be used. If there are several independent variables, P1R, P2R or P9R can be used to perform multiple linear regression analyses.

P1R, P2R and P9R differ in three important respects:

- the criterion for including independent variables in the multiple linear regression
- the ability to repeat the analysis on subgroups of the cases and to compare the subgroups
- the residual analysis available

P1R includes all the specified independent variables in the multiple regression equation. It computes a multiple linear regression on all the data and on groups or subpopulations. If grouping is requested, P1R first analyzes all cases combined and then analyzes each group separately. After all groups have been analyzed, the regression equations are tested for equality between groups.

P2R computes the multiple linear regression in a stepwise manner. At each step it enters into the regression equation the variable that best helps to predict y or removes the least helpful variable. Several criteria are available for entering or removing variables from the equation (see P2R program description). A stepwise procedure is useful for identifying a good set of predictor variables (separating the most important variables from those that may not be necessary at all), and when sufficient preliminary information regarding the effectiveness of the independent variables is not available. In practical applications the stepwise procedure is often a satisfactory solution.

P9R identifies "best" subsets of independent variables in terms of a criterion such as R^2, adjusted R^2 or Mallows' C_p (described in P9R program description). It also identifies alternative good subsets of the independent variables. P9R computes only a small fraction of all possible regressions to find the numerically best subset.

All three programs print and plot residuals and predicted values. The plots are useful in detecting lack of linearity, heteroscedasticity (lack of constant variance), unusual outliers, gross errors, an unusual subpopulation that should be separated from the analysis, etc. The plots may also indicate that transformations of the data are necessary or that an inappropriate model was chosen.

The residual analysis in P9R is the most extensive of the three. P9R also allows easy cross-validation of the regression model by testing it on a subset of the cases excluded from the analysis.

P4R creates new independent variables, called principal components, that are linear combinations of the original independent variables. These principal components are determined in a way that provides a parsimonious summary of the original variables; a subset of the principal components explains most of the total variance of the original set of independent variables. The program then regresses the dependent variable in a stepwise manner on the principal components; not all the principal components may be used, but useful information is based on all the variables. The regression equations at each step are expressed in terms of the principal components and the original variables.

The relation between an independent and a dependent variable may require terms with higher powers. The model for polynomial regression in P5R is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e \ .$$

P5R reports polynomials of degree one through a specified degree; this helps to determine the highest-order equation necessary for an adequate fit of the data. As higher-order terms are introduced into the model, the fitted regression curve and the original data can be plotted at each step for a visual check on how the fit is proceeding.

Nonlinear Regression

To fit a model where the equation is not linear in the parameters you can use the nonlinear regression programs, P3R and PAR. These are least squares programs appropriate for a wide variety of problems that are not well-represented by equations with linear parameters. Several different functions are available in P3R by simply stating a number, including such functions as sums of exponentials

$$p_1 e^{p_2 t} + p_3 e^{p_4 t} \ ,$$

ratios of polynomials, a combination of sine and exponential functions, etc. If you want a function different from those described in the P3R program description, you can request it by FORTRAN statements in P3R or PAR. In P3R you must also specify the function's partial derivatives.

A special nonlinear model is the logistic function. PLR computes the maximum likelihood estimates of the parameters of

$$E(\frac{s}{n}) = \frac{e^{\beta x}}{1 + e^{\beta x}}$$

where s is the sum of the binary (0,1) dependent variable y ($\sum_n y = s$) and x represents the independent variables. The dependent (outcome) variable records events such as success or failure, response or no response, etc. The independent (explanatory or covariate) variables can be categorical (e.g., sex, treatment, hospital) and continuous (e.g., age, height, blood pressure). The program generates design variables for the categorical variables and their interactions.

Analysis of Variance and Covariance

Analysis of variance is used to test for differences between the means of two or more groups or subpopulations. In a simple one-way analysis of variance each individual (or subject) is classified into one category or group -- for example, in a medical problem patients could be assigned to treatment A, B or C. The patients are grouped by the type of treatment. The model for this one-way design is

$$Y_{ik} = \mu + \alpha_i + e_{ik}$$

where α_1, α_2 and α_3 might represent the effect of treatments A, B and C, respectively, on the dependent variable, Y_{ik}, a blood pressure reading for case k in group i. Programs P7D, P9D, P1V and P2V can be used to test the hypothesis

$$H_0: \ \text{all } \alpha_i = 0$$

that there is no difference between treatments. Group sizes may be unequal in all four of these programs. For each dependent variable analyzed, P7D presents side-by-side histograms that give an excellent visual picture of how the groups differ.

In the medical treatment example above, if the covariate x (age) also affects the dependent variable (blood pressure), the one-way model becomes

$$Y_{ik} = \mu + \alpha_i + \beta(x_{ik} - \bar{x}) + e_{ik} \ .$$

P1V could be used to examine treatment effects after adjusting for the linear effect of age. P1V also allows multiple covariates. It prints an analysis of variance table with F tests for equality of slopes, zero slopes and equality of adjusted group means (which adjusts for the effect of the covariate) and a number of residual plots.

Several factors (or characteristics) may be involved in an analysis of variance model. In a two-way factorial analysis of variance, the individuals in each group are classified by two characteristics, such as sex and treatment. The model can be written

$$Y_{ijk} = \mu + \alpha_i + \eta_j + (\alpha\eta)_{ij} + e_{ijk} \ .$$

Here the α_i's could be treatment effect, the η_j's sex effect and $(\alpha\eta)_{ij}$ a possible interaction between sex and treatment. P7D can be used to analyze these data. The accompanying histograms give additional information.

P2V handles general fixed effects analysis of variance and covariance models. This program can analyze repeated responses, such as the measurements of a subject's blood pressure every day for a week. The repeated responses are called trial factors or repeated measures factors and need not be statistically independent. In the blood pressure example above, time could be a seven-level trial factor (e.g., a subject's blood pressure could be recorded every day of the week). In P2V the usual analysis of variance factors, such as sex and treatment, are called grouping factors to distinguish them from trial factors. The models may have only trial factors, only grouping factors, or both. The groups can contain an unequal number of subjects, but data for each subject must include all observations over the trial factor (a blood pressure reading must be given for each day).

Mixed models are treated by P8V (which requires equal cell sizes) and by P3V (which allows unequal cell sizes and covariates). P4V is a very general program that handles multivariate models, including those with repeated measures and covariates. The user may specify cell weights for use in the definition of model components such as main effects and lower order interactions in factorial models or specify unequal intervals for orthogonal polynomials in response surface analysis.

Nonparametric Statistics

If your data grossly violate the usual analysis of variance normality assumptions, you could try two nonparametric tests in P3S -- the Kruskal-Wallis one-way analysis of variance test, or the Friedman two-way analysis of variance test. Nonparametric tests such as the Mann-Whitney U test, the sign test and the Wilcoxon signed rank test can also be computed with P3S. These tests can be used when the researcher wants to avoid a t test assumption of normality.

Cluster Analysis

Although many research studies involve multivariate observations (many variables observed for each case), sometimes little is known about the inter-relations between variables, between cases, or between variables and cases. In discussing screening and data description, we emphasized that groups or subpopulations should be examined; however, problems often arise when groups are not clearly defined or when it is difficult to see if the data are structured. Clustering is a good technique to use in exploratory or early data analysis when you suspect that the data may not be homogeneous and you want to classify or reduce the data into groups. Clustering performs a display function for multivariate data similar to graphs or histograms for univariate data; it provides a multivariate summary -- a description of characteristics of clusters instead of individual cases.

Three different types of clustering can be performed by BMDP programs: clusters of variables (P1M), clusters of cases (P2M and PKM), and clusters of both cases and variables (P3M). After deciding which program is applicable to your problem, other questions must be answered: How will you measure distances between objects (variables in P1M, cases P2M and PKM)? How will you use the distances to amalgamate or group the objects into clusters? How will you display the resulting clusters? The best answers to these questions are still being developed; investigators have their own preferences as to which distance measure or which amalgamation procedure is best. You may want to try several options given in the program descriptions to see which one provides the best results for your problem.

In both P1M and P2M the clustering begins by finding the closest pair of objects (in P1M, columns, or variables; in P2M, rows, or cases) according to the distance matrix and combining them to form a cluster. The algorithm continues, joining pairs of objects, pairs of clusters, or an object with a cluster, until all the data are in one cluster. These clustering steps are shown in the output cluster diagram, or tree. The correlation or distance matrix can also be printed in shaded form to display pictorially the clusters.

The clustering method in P2M is hierarchical. The procedure in PKM is called k-means and begins with user-specified clusters or with all the data in one cluster: at each step one cluster is split into two. This procedure is useful when you have a large number of cases or when your goal is to divide the cases into homogeneous subsets. PKM provides several ways to standardize the data in order to avoid problems caused by scale differences.

The programs discussed above look for variables to be clustered across all cases or for cases to be clustered (by similarity) across all variables. However, your data may include differences between cases that do not extend across all the variables, or your variables may not cluster across all cases. P3M allows some of the variables (columns) to be clustered as a subset of the cases (rows) and vice versa. This clustering by both cases and variables is represented by a data matrix in the form of a block diagram; rows and columns are permuted and smaller blocks (submatrices) of similar values within the larger block are outlined. This gives a good visual representation of patterns of like values in the data matrix and can be used as a multivariate histogram. P3M is best suited to treat categorical variables that take on a small number of values.

Multivariate Analysis

Cluster analysis is not appropriate for expressing complex functional relationships. For example, if you are interested in describing the inter-relations among your variables, factor analysis may be better suited to your needs, and discriminant analysis provides functions of the variables that best separate cases into predefined groups.

Factor analysis. Factor analysis is useful in exploratory data analysis. It has three general objectives: to study the correlations of a large number of variables by clustering the variables into factors, such that variables within each factor are highly correlated; to interpret each factor according to the variables belonging to it; and to summarize many variables by a few factors. The usual factor analysis model expresses each variable as a function of factors common to several variables and a factor unique to the variable:

$$z_j = a_{j1}f_1 + a_{j2}f_2 + \dots + a_{jm}f_m + U_j$$

where

z_j = the jth standardized variable
m = the number of factors common to all the variables
U_j = the factor unique to variable z_j
a_{ji} = factor loadings
f_i = common factors

The number of factors, m, should be small and the contributions of the unique factors should also be small. The individual factor loadings, a_{ji}, for each variable should be either very large or very small so each variable is associated with a minimum number of factors.

To the extent that this factor model is appropriate for your data, the objectives stated above can be achieved. Variables with high loadings on a factor tend to be highly correlated with each other, and variables that do not have the same loading patterns tend to be less highly correlated. Each factor is interpreted according to the magnitudes of the loadings associated with it. The original variables may be replaced by the factors with little loss of information. Each case receives a score for each factor; these factor scores are computed as:

$$f_i = b_{i1}z_1 + b_{i2}z_2 + \dots + b_{ip}z_p$$

where b_{ij} are the factor score coefficients. Factor scores can be used in later analyses, replacing the values of the original variables. Under certain circumstances these few factor scores are freer from measurement error than the original variables, and are therefore more reliable measures. The scores express the degree to which each case possesses the quality or property that the factor describes. The factor scores have mean zero and standard deviation one.

There are four main steps in factor analysis: first, the correlation or covariance matrix is computed; second, the factor loadings are estimated (initial factor extraction); third, the factors are rotated to obtain a simple interpretation (making the loadings for each factor either large or small, not in-between); and fourth, the factor scores are computed. P4M provides several methods for initial factor extraction and rotation. You can specify the methods to be used or P4M will use preassigned options. The results can be presented in a variety of plots.

P8M, Boolean Factor Analysis, is an alternate technique when the variables are binary or dichotomous.

Canonical correlation analysis. Canonical correlation analysis (P6M) examines the relationship between two sets of variables, and can be viewed as an extension of multiple regression analysis or of multiple correlation. Multiple regression deals with one dependent variable, Y, and p independent variables, x_i. The regression problem is to find a linear combination of the X variables that has maximum correlation with Y. In canonical correlation there is more than one dependent Y variable -- there is a set of them. The problem is to find a linear combination of the X variables that has maximum correlation with a linear combination of the Y variables. This correlation is called the canonical correlation coefficient. Then a second pair of linear combinations, with maximum correlation between this pair and zero correlations with the first pair of linear combinations is found. The number of pairs of linear combinations of the X and Y sets is equal to the number of variables in the smaller set (X or Y). The technique can be used to test the independence of two sets of variables, or to predict information about a hard-to-measure set of variables from a set that is easier to measure. It can also be used to relate a combination of outcome measures to a combination of history or baseline measures. The original and canonical variables can be plotted one against the other in scatter plots.

Partial correlations and multivariate regression. Partial correlations can be computed in P6R; the correlation between each pair of dependent variables is computed after taking out the linear effect of the set of independent variables. For example, if you want to do a factor analysis on several variables (systolic blood pressure, diastolic blood pressure, blood chemistry measurements, income, etc.) but want to remove the linear effect of two variables (age and weight) from the measurements, you can state that the two variables (age and weight) are independent variables and the rest are dependent variables. The resulting partial correlation matrix (of the dependent variables with the effects of age and weight removed) can be stored as a matrix in a BMDP File, and can be used as input in P4M, the factor analysis program.

P6R can be used to regress a number of dependent variables on one set of independent variables. This multivariate regression program gives you a separate regression equation for each dependent variable, squared multiple correlation (R^2) of each independent variable with all other independent variables, R^2 of each dependent variable with the set of independent variables, and tests of significance of multiple regression.

Discriminant analysis. In discriminant analysis, the cases or subjects are divided into groups and the analysis is used to find classification functions (linear combinations of the variables) that best characterize the differences between the groups. These functions are also useful for classifying new cases.

P7M, the stepwise discriminant analysis program, is used to find the subset of variables that maximizes group differences. Variables are entered into the classification function one at a time until the group separation ceases to improve notably (this is similar to the stepwise regression program, P2R, used to find a good subset of variables for prediction). P7M is also used as a multivariate test for group differences (or multivariate analysis of variance); Wilks' lambda (U statistic) and the F approximation to lambda are printed at each step of the output for testing group differences.

A geometrical interpretation of discriminant analysis can be given by plotting each case as a point in a space where each variable is a dimension (has an axis). The points are projected onto a plane or hyperplane selected so the groups are farthest apart, giving a good visual representation of how distinct the groups are (for two groups, the points (cases) are projected onto a line where the groups are farthest apart). P7M presents plots that show such a plane. The X axis is the direction where the groups have the maximum spread; the Y axis shows the maximum spread of the groups in a direction orthogonal to the X axis - this is a plot of the canonical variables.

The canonical variables are related to canonical correlation analysis, which finds the linear combinations of the two sets of variables that are most highly correlated. The first set contains the variables in the classification function; the second set can be viewed as dummy variables used to indicate group membership. The value of the first canonical variable of the classification function set is plotted on the X axis; the value of the second on the Y axis. The coefficients for these canonical variables appear in the output. The coefficients for the second set (dummy variables) do not appear in the output. The eigenvalues and canonical correlations for all canonical variables and the canonical variable scores associated with the first and second canonical variables are also reported.

At each step, P7M uses a one-way analysis of variance F statistic (F-to-enter) to determine which variable should join the function next. At step zero, the standard univariate analysis of variance test is made for each of the variables. The variable for which the means differ most is entered first into the classification function. After step zero, the computed F-to-enter values are conditioned on the variables already present in the function. This is like an analysis of covariance, where the previously entered variables can be viewed as covariates and the nonentered variables are considered as dependent variables.

At each step after a variable is entered, the classification functions are recomputed including the newly entered variable. The number of classification functions is equal to the number of groups. If you have six groups, the values of all six functions are computed for each case and the values are used to compute the posterior probability; each case is assigned to the group in which the value of the posterior probability is maximum. In multiple group discriminant analysis, one function is sometimes stated in the literature for separating each pair of groups. To get this function from P7M, you subtract the classification function coefficients of the the first member from those of the second. At each step, F statistics (the F matrix) that test the equality of means between each pair of groups are given. These F statistics are proportional to Hotelling's T^2 and the Mahalanobis D^2 and give an indication of which group means are closest together and which are farthest apart. After all variables have been entered, the program lists the Mahalanobis D^2 from each case to the center of each group, and the posterior probability of the case assigned to each group. These two bits of information present a good picture of how well (or how poorly) each case has been classified.

The discriminant analysis procedure is successful if few cases are classified into the wrong groups. If a large percentage of the cases are classified correctly (if the posterior probability assigns them to their original group) you know that group differences do exist and that you have selected a set of variables that exhibit the differences. The P7M output presents this classification information in a table of counts indicating how many cases from each original group are assigned to each of the possible groups. A pseudo-jackknife classification table is also printed: for each case a classification function is computed with the case omitted from the computations. The function is then used to classify the omitted case. This results in a classification with less bias. (A classification function can produce optimistic results when it is used to classify the same cases that were used to compute it.)

PLR, Stepwise Logistic Regression, provides an alternative to the multivariate normal model of P7M. When there are only two groups, the all possible subset regression program, P9R, prints alternative functions for each subset which may classify the cases equally as well.

Preference Pairs. P9M, Linear Scores from Preference Pairs, is used to obtain a linear function of one set of variables that reproduces the ordering of cases as established by recorded preferences (stated by expert judges) between selected pairs of cases.

Survival Analysis

The techniques described in Chapter 19 are appropriate when outcome measurements represent the time to occurrence of some event or response (e.g., survival time, or time to disease recurrence). What distinguishes the techniques of this chapter from other statistical methodology is the ability to handle censored (incomplete) data; that is, there are cases for which the response is not observed but the data (time in study) are included in the analysis. This could occur in a study of survival, where an individual may remain alive at the close of the observation period or may drop out before the end.

P1L estimates the survival (time-to-response) distribution of individuals observed over varying time periods. These estimates can be obtained separately for different groups of patients; the equality of the distributions for these groups can be tested by two nonparametric rank tests. Plots of the survival, hazard and related functions can be printed.

P2L provides Cox model survival analysis when there are covariates. Covariates can be selected in a stepwise manner.

Time Series Analysis

The primary distinguishing feature of time series analysis, as opposed to other types of statistical analysis, is the assumption that cases of data represent measurements or observations made at equispaced points along some linear dimension. Usually the underlying linear dimension is time, as in the record of a subject's blood pressure taken every second over a period of time. However time is occasionally replaced by some other dimension. For example, the thickness of thread from a certain manufacturing process might be measured each millimeter along the length of the thread. This would constitute a 'time' series in which length replaces time as the underlying linear dimension. Nevertheless, we follow conventional terminology and use the word 'time'.

A basic goal of time series analysis is to characterize the way in which the data vary over time. The a priori assumption, common in most other types of statistical analysis, that cases are statistically independent, is here relaxed. We allow that cases may be correlated, assuming that the correlation between cases depends on the time interval separating them. In addition, we allow for the presence of a trend in the data. Thus the trend of increasing commodity prices over the last decade might be represented by a straight line, or by an exponential curve. So the estimated trend and autocorrelation function are one possible characterization of a time series. Other characterizations and more elaborate models are also possible.

The BMDP package includes two programs for time series analysis: BMDP1T employs the frequency domain approach, while BMDP2T uses the time domain approach. We may describe the frequency domain approach as representing the data by a superposition of sinusoidal waves at different frequencies. A central function of BMDP1T is to enable the user, by means of printer plots and accompanying printout, to identify the groups of frequencies contributing most of the overall variability of the data. In the time domain approach, we seek a model from a family of parametric time domain models that is simple and yet captures the variability of the data. BMDP2T uses the iterative model building procedure of Box and Jenkins, consisting of tentative model identification, parametric estimation, and diagnostic checking or residual analysis. Once a satisfactory model has been reached, the user may request BMDP2T to forecast future values of the time series. Both programs have further capabilities which are described in Chapter 20.

SPSS

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES

SECOND EDITION

NORMAN H. NIE
Department of Political Science
and
National Opinion Research Center
University Of Chicago

C. HADLAI HULL
Computation Center
University of Chicago

JEAN G. JENKINS
National Opinion Research Center
University of Chicago

KARIN STEINBRENNER
National Opinion Research Center
University of Chicago

DALE H. BENT
Faculty of Business Administration
and
Computing Services
The University of Alberta

McGRAW-HILL
BOOK COMPANY

New York
St. Louis
San Francisco
Auckland
Düsseldorf
Johannesburg
Kuala Lumpur
London
Mexico
Montreal
New Delhi
Panama
Paris
São Paulo
Singapore
Sydney
Tokyo
Toronto

Library of Congress Cataloging in Publication Data
Nie, Norman H
SPSS: statistical package for the social sciences.
1. SPSS (Electronic computer system) 2. Electronic data processing—Social sciences. I. Title. II. Title: Statistical package for the social sciences.
HA33.N48 1975 029.7 74-16223
ISBN 0-07-046531-2

STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES

8 9 0 W C W C 7 9 8

This book was set in Times Roman by Creative Book Services, division of McGregor & Werner, Inc. The editors were Kenneth J. Bowman and Matthew Cahill; the designer was Joseph Gillians; the production supervisor was Thomas J. LoPinto. New drawings were done by J & R Services, Inc.
Webcrafter, Inc., was printer and binder.

CONTENTS

9 DATA-SELECTION CARDS 127

10 RETRIEVING DATA AND DESCRIPTIVE INFORMATION FROM SPSS SYSTEM FILES: THE FILE RETRIEVAL (LIST AND WRITE) CARDS 133

11 ADDING TO, DELETING FROM, AND REARRANGING DATA IN SPSS SYSTEM FILES: THE FILE-MODIFICATION CARDS 149

18
BIVARIATE CORRELATION ANALYSIS: PEARSON CORRELATION, RANK-ORDER CORRELATION, AND SCATTER DIAGRAMS 276

19
PARTIAL CORRELATION: SUBPROGRAM PARTIAL CORR 301

APPENDIXES

PREFACE

Statistical Package for the Social Sciences (SPSS), the system of computer programs described in this volume, represents nearly a decade of systems design and programming and documentation on the part of the authors and others. When the first edition of this manual was published in 1970, SPSS was being used at approximately 60 installations. It is now being used at nearly 600 installations, including conversions to almost 20 different operating systems and computers. The current version of SPSS has almost double the amount of statistical procedures and data-management facilities as were documented in the first edition. Needless to say, we are thankful for the opportunity to continue developing SPSS and to serve an ever-increasing audience.

SPSS has been a success, much more so than the authors ever expected. We feel that the major reason for this success is that SPSS was developed through the close cooperation of three types of specialists: practicing social science researchers, computer scientists, and statisticians. At each stage in the development of SPSS we have attempted to satisfy these criteria:

1 That the statistical procedures be mathematically and statistically correct
2 That the program design and code be computationally efficient
3 That the logic and syntax of the system parallel the way in which social scientists approach data analysis
4 That the system provide statistical procedures and data-management facilities tailored to the particular needs of empirical social researchers

Without the contribution of experts in each field SPSS could not have effectively satisfied these goals.

Vying for importance in the success of SPSS has been our continuing concern for documentation. While many of our users are very sophisticated social scientists, they are not necessarily sophisticated statisticians and computer specialists, nor should they be expected to have the time or training necessary to master these fields. We feel that SPSS documentation must be comprehensive enough to enable students and researchers alike to use SPSS accurately and efficiently. We hope that even someone who has no experience with computers will be able to run SPSS jobs successfully by using this manual. Furthermore, since the SPSS system is used in many social science methods courses, we feel it is important that the manual not only explain how to invoke the statistical procedures but also provide brief discussions of the various statistical techniques available. Of course, in attempting to satisfy these various needs the manual has grown to be a very large book indeed. However, we hope that the organization of the manual as well as the inclusion of an index will enable users to find what they need conveniently.

Finally, the availability within SPSS of a wide variety of data-management facilities and statistical procedures simplifies the process of data analysis. It avoids many difficult, time-consuming, and generally unrewarding tasks involved in using a variety of single-purpose computer programs, each with its own idiosyncratic control-card syntaxes and input data formats. The researcher thus spends less time as a data-preparation clerk and more time as a social scientist analyzing substantive results.

The development of SPSS began in 1965 at Stanford University as a result of the frustration felt by several of the authors in trying to serve the research and teaching needs of the political science department and the Institute for Political Studies by a library of single-purpose data analysis programs. We and the users suffered the annoyance of having to learn how to operate many different programs and endured the time-consuming task of transferring data and results between essentially noncompatible programs. Furthermore, the documentation for many of these programs ranged from cryptic to nonexistent. That and the fact that the programs were written in many different languages made them extremely difficult to maintain. Finally, while many statistical analysis programs were available, we were forced to write our own programs to perform even the simplest recoding, data transformation, file editing, and other routine housekeeping chores that are essential in social science data analysis. As a result of our experiences we began to design an integrated system that would automate the routine tasks of data processing and around which a series of statistical programs could be built. The basic data-modification, file-handling, and data-description facilities were programmed, and as time permitted statistical analysis procedures were added.

The system developed into SPSS as it was described in the first edition of the manual. Since that time we have followed our plan of incremental development. New facilities have gradually been added and many flaws in the original design have been corrected. SPSS users have submitted suggestions and even complete procedures which have been incorporated into the system. The update manuals that we published each year finally became so unwieldy that we were compelled to provide this second edition of the manual so that SPSS would once again be fully described in a single volume.

The current version of SPSS still has a number of deficiencies, but we feel that it meets a great many of the needs of social science data analysis. SPSS has always been considered an open-ended system, and we have again included a programmer's guide to SPSS (Appendix I) in the hopes that users who wish to add statistical procedures to the system will do so. Because of the design of SPSS, all future programs incorporated into the system can take complete advantage of the capabilities for file maintenance and data handling which exist in the package; this should be an incentive to those contemplating the addition of other statistical programs.

For some time now we have felt that one of the most serious drawbacks of the current SPSS system lies not in the lack of a wider variety of statistical techniques but in that it operates only as a batch program. We and others have had a great deal of success running SPSS through a text-editing and remote batch-entry system, enabling the user to prepare and enter all jobs from a remote terminal and to retrieve and print small jobs directly on the terminal. However, though this adds convenience, it in no way substitutes for a true conversational statistical analysis package. Consequently, the SPSS project staff is now developing a conversational version of SPSS.

The difference between performing analysis in a batch system and a conversational system can be compared to the difference between interacting with another person by letter and by phone. With a conversational system the researcher is brought into close contact with the data being analyzed; results from a statistical procedure are returned on the researcher's terminal in a few seconds and may be used immediately to guide the next step of the analysis. This facilitates the iterative process of investigative research in which ideas are tested on the data, the results suggest new ideas and modifications of old ideas, the new ideas are then tested, and so forth. A train of thought which could take days (and many trips to a perhaps distant computer center) to explore in a batch system can be explored in a single interactive session. Thus the researcher's concentration and interest are focussed on the relationships being investigated. At the same time that a conversational system facilitates pursuing a single line of thought, the limited printing speed on remote terminals discourages the grand fishing expeditions and the production of excessive amounts of output that plague research in a batch environment. Finally, a good conversational system is somewhat self-teaching; it can guide the user in formulating the correct requests, and the almost instant feedback that facilitates the use in performing statistical analysis is also invaluable in identifying syntax errors and enabling the user to correct them immediately. In short, we feel that a conversational statistical analysis program could help revolutionize empirical social research.

Conversational SPSS, in addition to including conversational statistical analysis procedures, will feature interactive data- and file-definition capabilities, expanded data-transformation facilities, and ultimately the capacity to process hierarchical and other complex file structures. Although we are very enthusiastic about the prospects for conversational data analysis, we do not expect it to completely supplant batch processing. While a conversational system is ideal for investigative research, hypothesis testing, and rapid retrieval of a small number of specific statistics, it cannot be used efficiently for lengthy routine jobs and/or processing files that contain a very-large number of data cases. Thus we expect that researchers will wish to use both the batch and conversational versions of SPSS. This will be reflected in the design of conversational SPSS as well as the evolution of batch SPSS—we are attempting to make the two systems useful as a team. Conversational SPSS will be capable of reading and producing batch-SPSS-format system files, and the control statement syntax for conversational SPSS will be as compatible as possible with batch-SPSS conventions.

We are all very excited about the future of SPSS, and we thank you, the users, for the opportunity to continue its development. Your support, enthusiasm, and constructive criticism have been both impetus and reward to us. We dedicate this book to you in the hope that it will make data analysis an easier task, and that in some small way, through your research, we are contributing to the development of the social sciences.

ACKNOWLEDGMENTS

For the past five years the SPSS project has been located at the National Opinion Research Center at the University of Chicago. Throughout this critical period of our development NORC has provided us with financial, organizational, and intellectual support. This support has indeed been generous and has been a major factor responsible for our rapid growth and development. We can think of no other organization which would have provided a better home for us. In particular, we are grateful to Norman Bradburn and James Davis, the past and present directors of NORC, and Associate Director Shelby Orrell, without whose backing we might have floundered in the early years when our financial situation could only be described as precarious. We also wish to thank the University of Chicago Computation Center for donating computer time and allowing C. Hadlai Hull the flexibility to continue his work on the SPSS project.

The name of Patrick Bova, the librarian of NORC, is very familiar to many SPSS users. Mr. Bova has organized and directs the distribution of the SPSS system and all associated literature, and while maintaining our financial records, had thus far managed to keep us both solvent and honest. We are all very grateful for his efforts in what has become a very difficult, complicated, and time-consuming effort. The SPSS project is also very fortunate in the recent addition of ViAnn Beadle, who has joined the staff as user consultant. Ms. Beadle gives competent and conscientious attention to whatever questions and problems users may have. She has taken an enormous burden off our programming staff, providing us with the time for, among other things, the completion of this manual.

Special thanks must be given to Jae-On Kim, who has contributed greatly to SPSS over a period of years; Professor Kim helped design the factor analysis and analysis of variance routines, wrote several chapters of this manual, and regularly consults with us about statistical matters. We also wish to thank William R. Klecka, who contributed to the development of the discriminant function analysis procedure and wrote the associated chapter; William C. Mitchell, who designed the multiple-regression processing algorithms; and Bill Reynolds of the University of North Carolina, who designed the SCATTEGRAM procedure. For various improvements and corrections we are thankful to David Muxworthy of the University of Edinburgh, David Specht of Iowa State University, Jonathan B. Fry and Michael Klein of Akron University, Alfred J. Tuchfarber of the University of Cincinnati, Philip Burns of the University of Illinois, Cleve Moler of the University of Michigan, Harold W. Gugel of General Motors, and William L. McKeown of Stanford University. Martin L. Levin of Emory University and Kisun Han of DUALabs have developed SOS, the SPSS-Override System, which enables SPSS to process hierarchical files of the structure found in the Public Use Sample files.[1] Finally, we thank all those installations and users who have contributed suggestions and procedures. Even those which have not been implemented have provided us with ideas and thus have influenced the development of SPSS.

We are very appreciative of the work of Susan Hull, who prepared all of the more than 500 control-card and output examples appearing in this manual, and who implemented the integer version of the BREAKDOWN procedure. We must add that she performed these tasks as a volunteer, yet was as dedicated and painstaking as any other member of the staff. Another large contribution to the completion of this manual was made by Karin Donker, who typed and retyped until she nearly went mad, but who nevertheless cheerfully worked evenings and weekends when we had deadlines to meet.

We are indebted to the Vogelback Computing Center at Northwestern University for continued support of SPSS and in particular to James Tuccy of that center. He not only manages the CDC-6000 version of SPSS but also designed and programmed a number of the facilities described in this manual, including the CANCORR, DISCRIMINANT, ONEWAY, and T-TEST procedures. Space prevents us from individually mentioning each of the other SPSS conversions and the conversion managers, but we are grateful to them, as, we are sure, are their users.

We also express our appreciation to those at Stanford University who supported SPSS during its early development there: the political science department and the Institute of Political Studies, Professor Sidney Verba, Director of the Cross-National Program on Political and Social Change, and Professor Heinz Eulau and Kenneth Prewitt of the City Council Research Project. We thank the Stanford Computer Center for donating computer time. We are also grateful to the many students and researchers both at Stanford and at the University of Chicago who patiently used test versions of SPSS.

A system like SPSS is, of course, never developed in a vacuum, and we would like to acknowledge the contributions made to our general thinking by several previous systems: Data Text, developed at Harvard University, and BMD, developed at the University of California at Los Angeles, were especially significant. The factor analysis subprogram was adapted from a program developed at the University of Alberta. The Guttman Scale subprogram borrowed heavily from a program originally designed by Professor Ronald Anderson. The output format and some of the table statistics used in subprogram CROSSTABS were directly borrowed from the Data Text system. While we wish to acknowledge our great indebtness to those who developed these programs, we alone are responsible for whatever errors or mistakes are present in their implementation.

Norman H. Nie
C. Hadlai Hull
Jean G. Jenkins
Karin Steinbrenner
Dale H. Bent

[1] SOS is currently operational for IBM 360/370 and Univac Series 70 computers. Further information may be obtained from Kisun Han of DUALabs, Suite 900, 1601 North Kent Street, Arlington, Virginia 22209.

in selecting appropriate statistical procedures. An introduction to the general features and operation of the SPSS system is presented along with several examples in Sec. 1.3. The attention of the reader is drawn especially to Sec. 1.4 where suggestions for the use of this text are given for persons with greater or lesser experience in the use of computers for data analysis.

1.1 COMPUTERS AND THE PROCESS OF INDUCTIVE SOCIAL RESEARCH

Whether the intent of the researcher is to construct broad-gauge or middle-range social theory or simply to describe social reality, or as in most actual research endeavors, to do a little bit of each, the intellectual process of inductive social research ideally takes the following form:

1 The researcher begins with a set of ideas concerning the operation of certain aspects of social reality. This involves the isolation of variables at the conceptual level and the formation of some general notions concerning their interrelationships and causal effects upon each other.
2 An empirical data base is generated (or located from existing data files), containing indicators of the conceptual variables in which the researcher is interested.
3 The researcher then formulates more concrete hypotheses concerning what pattern of interrelationships should be found in the empirical indicators if the original ideas about the operation of social reality were correct.
4 The data are then analyzed using one or more of a variety of statistical techniques in order to determine whether the expected pattern of relationships can actually be discerned in the data.

Most often it is then discovered that at best the actual patterns in the data only partially reflect the original conceptions. There then begins an iterative process in which original conceptions are altered in light of the empirical findings and further analysis is performed to test these ideas. The data suggest new ideas, which in turn, suggest new analyses. This iterative process is continued in the hope that the researcher will be able to reach an understanding of interrelationships and cause and effect as they are reflected in the patterns of the data.

The computer has become an indispensable tool in this process of inductive social research. First, efficiently designed computer programs facilitate the movement back and forth between the researcher's ideas and the findings from the data, making this process both quick and painless. Second, such programs operating on high-speed computers have yielded an explosion of statistical capability. This has meant not only the efficient production of traditional tools such as crosstabulation tables, but the availability of complex multivariate statistical techniques such as regression analysis. Third, these programs have made it possible to test social theory with data files containing large numbers of cases and variables, which heretofore were virtually impossible to handle.

The research process is particularly facilitated when the researcher is able to use a unified *system of programs* which performs most of the different statistical techniques necessary, and which shares a common set of conventions regarding the way in which the user interacts with the programs. If well designed, the system permits the user to execute a sequence of tasks with a minimum of manual intervention, data handling, etc. The SPSS system is such a set of related programs for the manipulation and statistical analysis of many types of data, with a particular emphasis on the needs of the social sciences. Subsequently, we will refer to the programs of the system as *subprograms*, or *procedures*. Once the user has entered the raw data into the system, the computer can be instructed to carry out a variety of related tasks in any sequence the circumstances dictate. It is not necessary for the user to reenter the data at any time, since the system will store and retrieve the appropriate data when required.

While an attempt has been made to include in the SPSS system a number of the most commonly used procedures in social science data analysis, it is possible to retrieve data from the system so that it can be used with some other program. Also, SPSS itself can be extended by experienced computer programmers to include procedures which have not already been provided.

SPSS provides a set of common conventions for use of its various subprograms. This set of conventions constitutes a simplified language, corresponding closely to the natural language a social scientist might use to describe the operations to be performed on the data.

1.1.1 STATISTICAL ANALYSIS ON COMPUTERS: USE AND ABUSE

As we have suggested, statistical systems like SPSS can be important tools in the research process because they provide simple and rapid access to the researcher's data and make available a wide variety of statistical techniques. However, precisely because of their power they may be easily abused as shown in the following two ways:

1 *Ease of access often means overaccess.* Modern computing packages have made it so easy to produce large amounts of information on a single pass of a data file that even the experienced researcher is tempted to go on "grand fishing expeditions," substituting the crudest form of empiricism for the careful interaction of concepts, hypotheses, and data analysis. With SPSS, for example, the following crosstabulation statement will produce one crosstabulation table:

```
CROSSTABS       TABLES = A BY Z
```

At the same time literally hundreds of tables may be introduced by the following card:

```
CROSSTABS       TABLES = A TO Z BY A TO Z
```

Here a separate table will be generated for every combination of variables between A and Z; this might mean hundreds of tables which, while taking the computer only a few seconds to produce, would require a lifetime to examine with any degree of care or thought. Essentially, there are two inherent problems here that significantly impede the development of any meaningful type of social research. First, there is the problem of overproduction of information. How does one look at 500 crosstabulation tables? Where does one begin to cut into such a mountain of information or begin to digest the significance of even a small portion of it? Second, if one generates enough information, one is bound by the laws of probability to have some statistically significant findings. For example, a 100 × 100 correlation matrix generated from a file of totally random numbers produces almost 5,000 unique correlation coefficients. Within this correlation matrix there will likely be 5 large correlation coefficients significant at the .001 level, 50 at the .01 level, and 250 at the .05 level. The best rule-of-thumb here is to request only those tables, coefficients, etc., for which you have some theoretical expectations based upon the hypotheses in your research design.

2 *Uninformed use of the available statistical techniques.* The wide dissemination of statistical packages such as SPSS, containing large numbers of complex statistical procedures, have, almost overnight, made these techniques available to the social science community. There is little doubt that social scientists are using them, and there is equally little doubt that in many instances statistical techniques are being utilized by both students and researchers who understand neither the assumptions of the methods nor their statistical or mathematical bases. There can also be little doubt that this situation leads to some "garbage-in, garbage-out" research. The statistical procedures in SPSS have little ability to distinguish between proper and improper applications of the statistical techniques. They are basically blind computational algorithms that apply their formulas to whatever data the user enters. For example, if so directed, subprogram FACTOR will perform a factor analysis on nominal data as long as it is numerically coded. However, as is emphasized in the next section, entering nominal variables into such statistical procedures as factor analysis will almost always produce meaningless results, though this may not be immediately apparent from looking at the printed output.

The general rule is that a user should never attempt to use a statistical procedure unless he understands both the appropriate procedure for the type of data and also the meaning of the statistics produced. We do not mean to say that a researcher should not use factor analysis if he personally cannot invert a matrix or extract an eigenvalue. On the other hand, we are equally convinced that the basic ideas of principal components and factors as best linear combinations of variables and the general geometric mechanisms behind factor rotation are an absolute requisite for the successful use of factor analysis in research.

Associated with each statistical procedure described in this manual is a basic introduction to the statistical fundamentals underlying the technique. In general, the more complex the technique the longer the introduction. Researchers wishing to use a technique with which they are not totally familiar should read carefully these introductions, and they should also refer to the basic statistical texts cited in these chapters.

1.2 USING THE STATISTICAL CAPABILITIES OF SPSS

The statistical techniques used in the social sciences differ not only in the nature of the research questions that they are designed to answer, but also in the nature of the data to which they may be applied. The *level of measurement* of each of the variables in the user's data set is the most basic information that a researcher must have before selecting the statistical techniques that will be applied to the data.

1.2.1 A NOTE ON LEVELS OF MEASUREMENT

When data are being collected, the process of assigning a value or score to the observed phenomenon constitutes the process of *measurement*. The rules defining the assignment of an appropriate value determine the *level of measurement*. The different levels are distinguished on the basis of the ordering and distance properties inherent in the measurement rules. Knowledge of these rules and their implications is important to the user of statistics because each statistical technique is appropriate for data measured only at certain levels. The computer does not know what level of measurement underlies the numbers it receives, and will process whatever numbers are fed into it. Thus, it is up to the user to determine whether a particular technique is suitable for his or her data.

The traditional classification of levels of measurement was developed by S. S. Stevens (1946). He identified four levels: nominal, ordinal, interval, and ratio. This remains the basic typology that every user of statistics should know. Other variations exist, however, and several issues concerning the statistical effect of ignoring levels of measurement are still being debated by social scientists. Attention will be paid to these matters at the end of this section.

1.2.1.1 Nominal-Level Measurement

The nominal level of measurement is the "lowest" in Stevens' typology, because it makes no assumption whatever about the values being assigned to the data. Each value is a distinct category, and the value itself serves merely as a label or name (hence, "nominal" level) for the category.[1] No assumption of ordering or distances between categories is made. For instance, the city where a person was born is a nominal variable. There is no inherent ordering among cities implied by such a variable. Although we could order cities according to their size, density, or degree of air pollution, those are quite different concepts from "place of birth." When we attach numeric values to nominal categories, we are using numbers merely as symbols that are easily read by the computer. The properties of the real number system, for example, being able to add and multiply numbers, etc., cannot be transferred to these numerically coded categories. Therefore, statistics that assume ordering or meaningful numerical distances between the categories should not be used.

1.2.1.2 Ordinal-Level Measurement

When it is possible to rank-order all of the categories according to some criterion, then the ordinal level of measurement has been achieved. For instance, the classification of social classes

[1] Keep in mind that any valid measurement scheme requires that the assignment rules are inclusive and mutually exclusive. That is, each possible case can be assigned to *one and only one* distinct value.

as working, middle, and upper is ordered according to status. Each category has a unique position relative to the other categories, that is, it is lower in value than some categories and higher than others unless, of course, it happens to be the lowest or highest category. Furthermore, knowing that middle class is higher than working class and that upper class is higher than middle class automatically tells us that the upper class is higher in the ordering than the working class. However, we do not know *how much* lower the middle class is, relative to the upper class. All we know is that it is lower; we do not know the *distance*. The characteristic of ordering is the sole mathematical property of this level, and the use of numeric values as symbols for category names does not imply that any other properties of the real number system can be used to summarize relationships of an ordinal-level variable.

1.2.1.3 Interval-Level Measurement

In addition to ordering, the interval level of measurement has the property that the distances between the categories are defined in terms of fixed and equal units. A thermometer, for instance, records temperatures in terms of degrees, and a single degree implies the same amount of heat whether the temperature is at the lower or the upper end of the scale. Thus, the difference between 30 and 31°F is the same as the difference between 80 and 81°F. The important thing to note is that an interval scale does not have an inherently determined zero point. In the Fahrenheit and Centigrade systems, zero degrees is determined by an agreed-upon definition. Neither implies the absence of heat. Consequently, interval-level measurement allows us to study *differences* between things but not their *proportionate* magnitudes. That is, it would be incorrect to say that at 80°F twice as much heat is present as at 40°F.

In social research, it is very difficult to find true interval-level measures. Usually, if distances between categories can be measured by some fixed unit, a natural zero point can also be established. Yet, a great many statistics assume no more than an ordinal level of measurement. What must be kept in mind is that statistics developed for one level of measurement can always be used with *higher-level variables, but not with variables measured at a lower level.* The median, for example, assumes an ordinal level of measurement, but it can be used legitimately with interval- or ratio-level scales; it cannot, however, be applied to variables measured at the nominal level.

1.2.1.4 Ratio-Level Measurement

The ratio level of measurement has all of the properties of an interval scale with the additional property that the zero point is inherently defined by the measurement scheme. Thus, when we measure physical distances, whether we use feet or meters, a zero distance is naturally defined: It is the absence of any distance between two objects. This property of a fixed and given zero point means that ratio comparisons can be made, as well as distance comparisons. For example, it is quite meaningful to say that a 6-foot-tall man is twice as tall as a 3-foot-tall boy.

Since ratio-level measurements satisfy all the properties of the real number system, the numbers employed to describe the cases are more convenient symbols. Any mathematical manipulations appropriate for real numbers can also be applied to ratio-level measures. Although this level of measurement is common in social research, very few statistics require all of its properties; however, it is important to remember that all statistics requiring variables measured at the interval level are also appropriate for use with variables measured at the ratio level.

1.2.1.5 The Special Case of Dichotomies

A *dichotomy* is a variable with only two possible categories or values, such as sex (male or female). While some dichotomies are based on a natural ordering (passing a course versus failing it), many have no inherent basis on which either category could be judged superior, preferable, larger, etc. Yet, any dichotomy can be treated as though it were an interval-level measure and in some cases even a ratio-level variable.

Although a rank order may not be inherent in the category definitions, either arrangement of the categories satisfies the mathematical requirements of ordering. (It does not matter which end of a ranking is considered "high" and which is "low.") The requirement of a distance measure based on equal-sized intervals is also satisfied because there is only one interval naturally equal to itself. Consequently, a dichotomy can be treated as either a nominal, ordinal, or interval-level measure, depending upon the research situation.

1.2.1.6 Other Typologies for Levels of Measurement

A simpler scheme than Stevens' is to divide variables into *quantitative* and *qualitative* types. Quantitative variables are those for which a fixed unit of measurement is defined —essentially, variables at the interval and ratio levels. These are the variables for which the most powerful and sophisticated techniques have been developed. Qualitative variables, then, are all others—namely, those at the nominal or ordinal level.

Coombs (1953) has expanded upon Stevens' four-level typology by adding two more levels. The *partially ordered* level falls between nominal and ordinal. It applies to situations where an ordering can be defined between some of the categories but not over all of them. Of greater interest to social scientists is the *ordered metric* level. Falling between the ordinal and interval levels, an ordered metric consists of ordered categories where the relative ordering of the intercategory distances is known even though their absolute magnitude cannot be measured. For example, consider the rating of a person's ability to read a foreign language as (A) no ability, (B) able to read with the assistance of a dictionary, and (C) able to read without assistance. Although there is no way to ascertain the distances between A, B, and C, it could be argued that B is closer to C than to A, because a type B person can translate and understand written material in that language. Thus, we could rank the distances between the categories as BC being the smallest, followed by AB, with AC as the largest.

Abelson and Tukey (1959) argue that the proper assignment of numeric values to the categories of an ordered metric scale will allow it to be treated as though it were measured at the interval level. Labovitz (1970) goes further by arguing that, except for extreme situations, interval statistics can be applied to *any* ordinal-level variable. He argues, "Although some small error may accompany the treatment of ordinal variables as interval, this is offset by the use of more powerful, more sensitive, better developed, and more clearly interpretable statistics with known sampling error." Statistical purists disagree with some or all of these suggestions, but more and more data analysts are following them, especially when the research is exploratory or heuristic in nature. Whatever position the user adopts, it remains his responsibility to select an appropriate statistic and to interpret the results in light of the nature of the data.[1]

There is no unique method for classifying the different types of statistical procedures included in SPSS. One distinction, based on levels of measurement, is between parametric (or quantitative) and nonparametric (or qualitative) statistics. Nonparametric statistical procedures require few assumptions about the distribution or level of measurement of the variables and may be applied to nominal and ordinal data. The parametric procedures, on the other hand, theoretically require more stringent assumptions concerning the distribution of the data (usually an assumption of normality), and they are designed primarily for data at an interval or ratio level of measurement. While the statistical procedures in SPSS can be catalogued according to this rubric (for example, Spearman versus Pearson correlation, *n*-dimensional crosstabulation versus partial correlation and multiple regression, Guttman scaling versus factor analysis, etc.), these assumptions are so often violated (often with justifiable reasons) during the process of data analysis that their utility is questionable.

1.2.2 STATISTICAL PROCEDURES IN SPSS

SPSS contains many of the most common statistical procedures employed by social scientists, but it is by no means exhaustive of the many useful procedures that have been

[1] For an extended argument against strict and blind adherence to rules linking specific statistics to particular levels of measurement, see Labovitz (1972). He also argues that the level of measurement for a concept can often be improved by reconceptualizing the way in which it is operationalized (measured)

invented for social research or that have come from other fields to the social sciences. The choice of statistical procedures in SPSS has been determined by our examination of the amount of use they receive in day-to-day statistical analysis and, of course, by the exigencies of time and resources.

In presenting the statistical procedures contained in SPSS, we will start with those that the researcher often begins with and then proceed through the various types of procedures according to increasing level of complexity and sophistication. No single research endeavor would normally employ all or even a large number of these procedures, but it will often be the case that at least one procedure from each of the groups will be employed at some point during the analysis.

1.2.2.1 One-Way Frequency Distributions, Measures of Central Tendency and Dispersion

In most types of social science research, the first task of data analysis is to examine the distributional characteristics of each of the independent and dependent variables under investigation. SPSS contains two statistical procedures for this purpose. (1) CONDESCRIPTIVE calculates numerous common measures of central tendency and dispersion for interval-scale variables that assume a large number of values. (2) FREQUENCIES calculates similar types of descriptive statistics and generates tabular reports of absolute and relative simple-frequency distributions for use with variables that assume only a limited number of values. An example of the type of variable for which CONDESCRIPTIVE is appropriate would be income measured in dollars, which can assume a continuum of values. FREQUENCIES would be applicable to a measure of income when the information has been grouped, such as, \$0–\$3,000, \$3,001–\$5,000, \$5,001–\$10,000, \$10,001+. The latter procedure can also produce descriptive frequency distributions for nominal variables, such as religious affiliation, race, or political party affiliation.

Both CONDESCRIPTIVE and FREQUENCIES can produce statistics such as the mean, mode, minimum, maximum, standard deviation, variance, skewness, kurtosis, and range, at the user's discretion. FREQUENCIES can also be used to produce histogram plots, and allows the user to select from a variety of tabular formats for distribution tables. CONDESCRIPTIVE will optionally punch or write the standardized values of variables, *Z* scores, for all cases in the file. These standardized variables can be reentered and merged with the variables in the SPSS file on a subsequent run.

1.2.2.2 Table Displays of Relationships Between Two or More Variables

After the researcher understands the characteristics of each of the variables, he normally begins to investigate sets of relationships. One or more procedures for examining relationships will be selected depending upon the characteristics of the variables and the purposes of the researcher. The researcher may choose correlation analysis or some form of table display such as those discussed in this section, particularly if the variables are nominal or ordinal and are classified into a limited number of categories.

SPSS procedure CROSSTABS permits the user to produce two-way to *n*-way crosstabulations of variables and to compute a variety of nonparametric statistics based on these tables. CROSSTABS produces a sequence of two-way tables displaying the joint frequency distribution of two variables. The frequency counts can be expressed as a percentage of the row total, column total, table total, or any combination thereof. The statistics available to measure the degree of association of the two variables based on the distribution of frequency counts in the table include chi-square, Cramer's *V*, Kendall's tau *B* and *C*, gamma statistics, and Somer's *D*. For *n*-way crosstabulations, a sequence of such two-way tables is produced, one for each two-dimensional subsection of the *n*-dimensional table.

Another technique for examining the relationship between two or more variables in a table format is provided by the BREAKDOWN procedure. This procedure, which requires that the dependent variable be at least ordinal in scale, compiles the means, standard deviations, and variances of a criterion or dependent variable for each desired subgroup in a sample or population. In many respects this operation is analogous to crosstabulations of the type produced by

CROSSTABS, only in this case, each mean and standard deviation summarizes the distribution of a complete row or column of a crosstabulation table. Also in this case, the means, etc., of each group within groups are available on a single table. The user may enter up to six variables into a single BREAKDOWN table. BREAKDOWN optionally computes a one-way ANOVA table including a test for linearity.

1.2.2.3 Bivariate Correlation Analysis and Scatterplots

Correlation analysis provides the researcher with a technique for measuring the linear relationship between two variables and produces a single summary statistic describing the strength of the association; this statistic is known as the *correlation coefficient*. SPSS has two programs for computing correlations. PEARSON CORR produces zero-order or product-moment correlation coefficients (Pearson's R) that are best suited for normally distributed data with an interval scale. NONPAR CORR, suitable for ordinal data with a larger number of categories than would be appropriate for crosstabulation tables, enables the user to compute either Spearman or Kendall rank-order correlation coefficients, or both. Both PEARSON CORR and NONPAR CORR can produce correlations for selected pairs or lists of variables as well as complete matrices of coefficients. The output from both subprograms provides the correlation coefficient, the number of observations upon which the correlation was based, and the level of statistical significance of the coefficient. In addition, each procedure provides for the output of correlation matrices that may be used when applying multivariate statistical techniques.

Though bivariate correlation analysis provides a single summary statistic describing the relationship between two variables, there are numerous instances when the researcher may wish to examine such a relationship in greater detail. Subprogram SCATTERGRAM provides this capability by producing a scatterplot diagram of the relationship between two variables. The total correlational pattern may thus be visually inspected. In addition to the plot itself, the Pearson correlation, and the standard error of estimate, the regression intercept and slope are also available at the user's request.

1.2.2.4 Partial Correlation

Partial correlation provides a single measure of association (the partial-correlation coefficient) describing the linear relationship between two variables while adjusting or controlling for the effects of one or more additional variables. In this respect, partial correlation is analogous to n-dimensional crosstabulation for continuous variables. First- to nth-order partial-correlation coefficients can be obtained for any set of variables with the PARTIAL CORR procedure. This program can operate on raw data or from matrices of simple correlation coefficients, such as may be produced by a previous run of PEARSON CORR or NONPAR CORR.

Up to five orders of partials can be simultaneously computed for any set of variables, and the user has total control over the orders and the partials to be computed. Output from this procedure includes the partial-correlation coefficients, and the level of statistical significance and degrees of freedom for each partial. The zero-order correlations, means, and standard deviations of the variables may also be obtained. The user may also optionally request the output of zero-order correlation matrices for further computation.

1.2.2.5 Multiple Correlation and Regression

Multiple regression is an extension of the bivariate correlation coefficient to multivariate analysis. Multiple regression allows the researcher to study the linear relationship between a set of independent variables and a dependent variable while taking into account the interrelationships among the independent variables. The basic goal of multiple regression is to produce a linear combination of independent variables which will correlate as highly as possible with the dependent variable. This linear combination can then be used to "predict" values of the dependent variable, and the importance of each of the independent variables in that prediction can be assessed.

A variety of multiple-regression calculations can be accomplished with the use of the REGRESSION procedure. This subprogram, like PARTIAL CORR, can operate either on raw data or a matrix of correlation coefficients. The user can perform the regression upon a fixed number of variables or, using a forward-selection stepwise technique, allow the variables to be introduced into the computation sequentially depending upon their explanatory power. REGRESSION also allows the user to perform a regression procedure midway between these two extremes; he can allow the program to choose the order of introduction of the variables from a certain set, then force certain other variables into the calculation, then proceed stepwise for a period, and so forth. This flexibility, together with the ability of SPSS to transform variables, allows the user to handle most polynomial and dummy-variable multiple-regression applications with relative ease. Output from the program includes both the standardized and nonstandardized regression coefficients, their standard error, and the significance level of the coefficients. Multiple r, r^2, and the significance of the regression equation are also computed at each stage. The user can also obtain written or punched zero-order correlation matrices.

Subprogram REGRESSION also permits the user to write or punch-out a full set of residuals for each individual case in the file for any set of regression equations. The residual can then, in any subsequent run, be entered into SPSS as a new variable or group of variables in the analysis.

1.2.2.6 One-Way to *n*-Way Analysis of Variance and Covariance

Analysis of variance is a statistical technique that assesses the effects of one or more categorical independent variables *(factors)*, measured at any level upon a continuous dependent variable that is usually assumed to be measured at an interval level. Conceptually, the cases are divided into categories based on their values for each of the independent variables, and the differences between the means of these categories on the dependent variable are tested for statistical significance. The relative effect upon the dependent variables of each of the independent variables, their combined effects and interactions, may be assessed.

Analysis of variance is in many ways similar to multiple regression and, in fact, the SPSS *n*-way ANOVA procedure is based on the least-squares general-linear hypothesis approach. Analysis of variance differs from regression insofar as it relaxes the restrictions on levels of measurement of independent variables and provides a convenient way for examining the interaction effects of specific combinations of independent variables. In analysis of covariance, the set of independent variables includes both categorical and continuous variables; the continuous variables *(covariates)* are assumed to be linearly related to the dependent variable.

SPSS subprogram ANOVA performs *n*-way analysis of variance with up to five factors and will adjust for up to five covariates. It can handle factorial designs that are unbalanced and contain some empty cells. In addition, the ANOVA subprogram can present the results of analyses of variance and covariance in multiple classification analysis (MCA) format. In addition to the general *n*-way analysis of variance procedure (ANOVA), SPSS contains the very detailed one-way analysis of variance subprogram ONEWAY. ONEWAY includes many special optional features including tests for trends across categories of the independent variable, user-specified a priori contrasts, a posteriori contrasts, and homogeneity of variances.

1.2.2.7 Discriminant Analysis

With *discriminant analysis* a researcher calculates the effects of a collection of interval-level independent variables on a nominal dependent variable (classification). Linear combinations of independent variables that best distinguish between cases in the categories of the dependent variable are found.

SPSS subprogram DISCRIMINANT calculates and prints discriminant-function coefficients and classification-function coefficients. All independent variables may be entered into the discriminant functions, or, if the user chooses, DISCRIMINANT will operate in a stepwise mode, entering variables in the order of their explanatory power. The user may control both the number of discriminant functions generated and the number of variables entered.

Discriminant scores, the probability of membership in each category of the dependent variable, and the predicted category may be calculated and printed for each case in the file. The discriminant scores and the predicted category for each case may be punched or written on an output file and may be reentered into SPSS on a subsequent run for further analysis.

1.2.2.8 Guttman Scaling

All the statistical procedures previously discussed, with the exception of those used to examine the characteristics of individual variables, represent different methods for examining, explaining, and predicting the relationship between one or more independent variables and a dependent variable. In this section and the two sections following, we discuss procedures contained in SPSS for locating underlying continuums or variable sets from a larger groups of variables.

Guttman-scale analysis is a means of analyzing the underlying operating characteristics of three of more items in order to determine if their interrelationships meet two special properties that define an acceptable Guttman scale—unidimensionality and cumulativeness.

In the SPSS GUTTMAN SCALE procedure the scales are computed by the Goodenough technique. Each item to be included in a scale may have up to three cutting points, and on an individual scale the item is computed for all possible combinations of cutting points specified. The order of items may be automatically determined by the subprogram according to the proportion of the respondents who "fail" or "reject" items. Alternatively, the user may personally fix the order of items.

In addition to the basic table giving the frequencies, errors, and scale types, the user may request a number of statistics that will aid him in evaluating the scales. Included in the available statistics are: (1) the coefficient of reproducibility, (2) the minimum marginal reproducibility, (3) the percent improvement, and (4) the coefficient of scalability. All these statistics help the user to determine the quality of the scale. Interitem correlations and part-whole correlations may be requested also.

1.2.2.9 Factor Analysis

Factor analysis is a much more generalized procedure for locating and defining dimensional space among a relatively large group of variables. Because of the generality of factor analysis, it is difficult to present a capsule description of its functions and applications. The major use of factor analysis by social scientists is to locate a smaller number of valid dimensions, clusters, or factors contained in a larger set of independent items or variables. And viewed from the other side, factor analysis can help determine the degree to which a given variable or several variables are part of a common underlying phenomenon.

The SPSS FACTOR procedure can begin with either raw data, a correlation matrix, or a factor matrix. The methods of factoring available are principal-component analysis, alpha factoring, principal-axis factoring, Rao's canonical factoring, and image factoring. The factoring procedure can be controlled by specifying the number of iterations to be performed, if applicable, the number of factors to be extracted, if applicable, or the minimum value of an eigenvalue for which a factor will be extracted. Following the factor-extraction phase, rotations may be performed. The types of rotations that may be used are varimax, equimax, quartimax, and the direct oblimin method of oblique rotation.

Factor-scale scores for each observation in the file can also be produced, and these scales (which can be written or punched on an output medium of the user's choice) can be added as new variables to the user's file on subsequent runs.

1.2.2.10 Canonical Correlation

Multiple regression is a technique for examining the relationships among several independent variables and a single dependent variable. Factor analysis is a technique of data reduction that locates fewer underlying dimensions (higher-order variables) out of a larger pool of variables in which no distinction has been made between independent and dependent variables. *Canonical correlation* is in some respects a combination of the two alternate multivariate techniques. It contains data reduction capabilities similar to factor analysis, but, having required the user to divide the variables into two sets, also assesses the relationship between the two sets of factors (called *canonical variates).*

In this way, the researcher is able to conveniently simplify and analyze the relationship between a large number of independent variables and a large number of dependent variables. More precisely, canonical correlation analysis takes as its basic input two sets of variables, each of which can be given theoretical meaning as a set, and extracts linear combinations of the variables within each set; each linear combination maximally correlates with a corresponding linear combination from the other set. These linear combinations are the canonical variates and come in associated pairs. Thus, the higher-order dimensions are created not on the basis of accounting for the maximal variance within one set of variables (as in factor analysis), but on the basis of accounting for a maximum amount of the relationship between the two sets of variables.

Input to the SPSS CANCORR procedure can be either raw data or a correlation matrix. The user may specify the number of pairs of canonical variates to be extracted and the significance level required for extraction. The procedure automatically outputs the canonical correlations, along with tests of their statistical significance, and the coefficients of the canonical variates. CANCORR will optionally punch or write the values of the canonical variates for all cases in the file. These variates can be reentered into SPSS as new variables on a subsequent run.

We have described the principal statistical procedures available within the SPSS system. It is important to realize, however, that these procedures can be executed in any sequence, or repetitively in the course of a single run or session with the computer. Thus the user may elect to perform some crosstabulations, do a multiple regression, and then do some correlations upon the same file of data in a single run. Also, the procedures described share the general capabilities of SPSS for file handling, variable manipulation, and so forth, so that they constitute a sequence of steps available to the user in any order that makes sense in the context of the problem. In Sec. 1.3 we discuss some of the general capabilities of SPSS that are available in conjunction with any statistical procedure the user may specify.