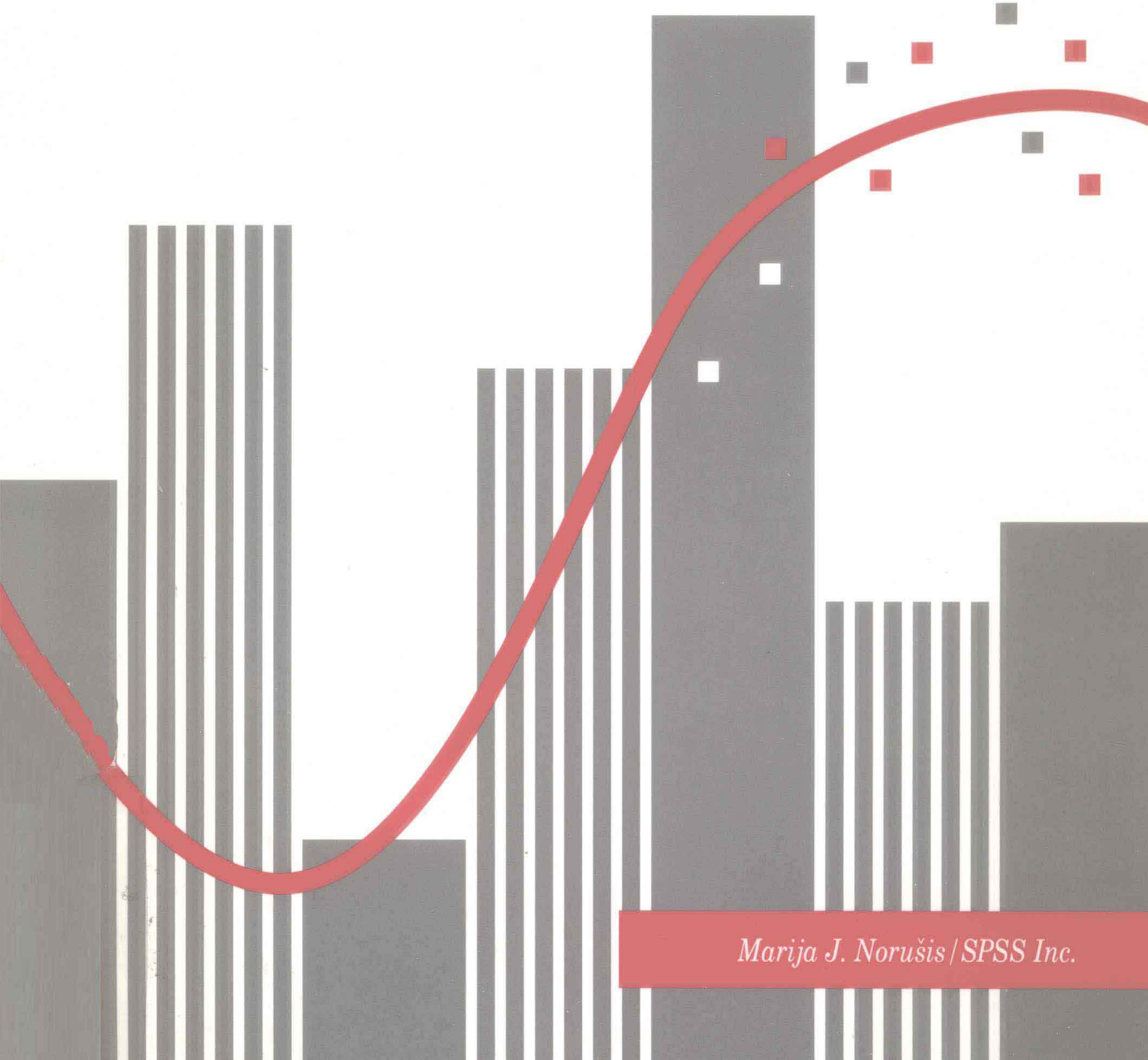
SPSS

SPSS Professional Statistics™ 6.1



Marija J. Norušis / SPSS Inc.

SPSS

SPSS Professional Statistics™ 6.1

Marija J. Norušis/SPSS Inc.

*SPSS Inc.
444 N. Michigan Avenue
Chicago, Illinois 60611
Tel: (312) 329-2400
Fax: (312) 329-3668*

*SPSS Federal Systems (U.S.)
SPSS Latin America
SPSS Benelux BV
SPSS GmbH Software*
[redacted]
[redacted] *SARL*
SPSS [redacted]*portuguesa S. L.*
[redacted]*avia AB*
*SPSS India Private Ltd.
SPSS Asia Pacific Pte. Ltd.*
[redacted]*nc.*
SPSS Australasia Pty. Ltd.

For more information about SPSS® software products, please write or call

Marketing Department
SPSS Inc.
444 North Michigan Avenue
Chicago, IL 60611
Tel: (312) 329-2400
Fax: (312) 329-3668

2 3 4 5 6 7 8 9 0   96 95 94

ISBN 0-13-190125-7

Preface

SPSS® 6.1 is a powerful software package for microcomputer data management and analysis. The Professional Statistics option is an add-on enhancement that provides additional statistical analysis techniques. The procedures in Professional Statistics must be used with the SPSS 6.1 Base system and are completely integrated into that system.

The Professional Statistics option includes procedures for

- Discriminant analysis
- Factor analysis
- Cluster analysis
- Proximity and distance measures
- Reliability analysis
- Multidimensional scaling
- Weighted and two-stage least-squares regression

The algorithms are identical to those used in SPSS software on mainframe computers, and the statistical results will be as precise as those computed on a mainframe.

What's New in SPSS 6.1. The information in this manual is essentially the same as the information in *SPSS for Windows Professional Statistics Release 6.0.* Improvements to the Professional Statistics option of SPSS 6.1 do not affect the information on dialog boxes, the functionality of the statistics procedures, or the syntax. The new operational features of SPSS 6.1 are all available for use with the Professional Statistics option, including the following:

- Toolbar
- Case identification and labeling in scatterplots (discriminant analysis, factor analysis, multidimensional scaling)
- Exporting charts in various formats
- Variable information pop-up window
- Compatibility with Win32s

These features are described in the SPSS Base system documentation.

Installation

To install Professional Statistics, follow the instructions for adding and removing features in the Installation Instructions supplied with the SPSS Base system. (To start, double-click on the SPSS Setup icon.)

Compatibility

The SPSS system is designed to operate on many computer systems. See the installation instructions that came with your system for specific information on minimum and recommended requirements.

Serial Numbers

Your serial number is your identification number with SPSS Inc. You will need this serial number when you call SPSS Inc. for information regarding support, payment, or an upgraded system. The serial number can be found on Disk 2 that came with your Base system. Before using the system, please copy this number to the registration card.

Registration Card

STOP! Before continuing on, *fill out and send us your registration card.* Until we receive your registration card, you have an unregistered system. Even if you have previously sent a card to us, please fill out and return the card enclosed in your Professional Statistics package. Registering your system entitles you to:

- Technical support on our customer hotline
- Favored customer status
- New product announcements

Don't put it off—send your registration card now!

Customer Service

Contact Customer Service at 1-800-521-1337 if you have any questions concerning your shipment or account. Please have your serial number ready for identification when calling.

Training Seminars

SPSS Inc. provides both public and onsite training seminars for SPSS. All seminars feature hands-on workshops. SPSS seminars will be offered in major U.S. and European cities on a regular basis. For more information on these seminars, call the SPSS Inc. Training Department toll-free at 1-800-543-6607.

Technical Support

The services of SPSS Technical Support are available to registered customers. Customers may call Technical Support for assistance in using SPSS products or for installation help for one of the supported hardware environments.

To reach Technical Support, call 1-312-329-3410. Be prepared to identify yourself, your organization, and the serial number of your system.

If you are a Value Plus or Customer EXPress customer, use the priority 800 number you received with your materials. For information on subscribing to the Value Plus or Customer EXPress plan, call SPSS Software Sales at 1-800-543-2185.

Additional Publications

Additional copies of SPSS product manuals may be purchased from Prentice Hall, the exclusive distributor of SPSS publications. To order, fill out and mail the Publications order form included with your system or call toll-free. If you represent a bookstore or have an account with Prentice Hall, call 1-800-223-1360. If you are not an account customer, call 1-800-374-1200. In Canada, call 1-800-567-3800. Outside of North America, contact your local Prentice Hall office.

Lend Us Your Thoughts

Your comments are important. So send us a letter and let us know about your experiences with SPSS products. We especially like to hear about new and interesting applications using the SPSS system. Write to SPSS Inc. Marketing Department, Attn: Micro Software Products Manager, 444 N. Michigan Avenue, Chicago IL, 60611.

About This Manual

This manual is divided into two sections. The first section provides a guide to the various statistical techniques available with the Professional Statistics option and how to obtain the appropriate statistical analyses with the dialog box interface. Illustrations of dialog boxes are taken from SPSS for Windows. Dialog boxes in other operating systems are similar. The second part of the manual is a Syntax Reference section that provides complete command syntax for all of the commands included in the Professional Statistics option. Most features of the system can be accessed through the dialog box interface, but some functionality can be accessed only through command syntax.

This manual contains two indexes: a subject index and a syntax index. The subject index covers both sections of the manual. The syntax index applies only to the Syntax Reference section.

Contacting SPSS Inc.

If you would like to be on our mailing list, write to us at one of the addresses below. We will send you a copy of our newsletter and let you know about SPSS Inc. activities in your area.

SPSS Inc.
444 North Michigan Ave.
Chicago, IL 60611
Tel: (312) 329-2400
Fax: (312) 329-3668

SPSS Federal Systems
Courthouse Place
2000 North 14th St.
Suite 320
Arlington, VA 22201
Tel: (703) 527-6777
Fax: (703) 527-6866

SPSS Latin America
444 North Michigan Ave.
Chicago, IL 60611
Tel: (312) 494-3226
Fax: (312) 494-3227

SPSS Benelux BV
P.O. Box 115
4200 AC Gorinchem
The Netherlands
Tel: +31.1830.36711
Fax: +31.1830.35839

SPSS GmbH Software
Rosenheimer Strasse 30
D-81669 Munich
Germany
Tel: +49.89.4890740
Fax: +49.89.4483115

SPSS UK Ltd.
SPSS House
5 London Street
Chertsey
Surrey KT16 8AP
United Kingdom
Tel: +44.1.932.566262
Fax: +44.1.932.567020

SPSS France SARL
72-74 Avenue Edouard Vaillant
92100 Boulogne
France
Tel: +33.1.4684.0072
Fax: +33.1.4684.0180

SPSS Hispanoportuguesa S.L.
Paseo Pintor Rosales, 26-4
28008 Madrid
Spain
Tel: +34.1.547.3703
Fax: +34.1.548.1346

SPSS Scandinavia AB
Gamla Brogatan 36-38
4th Floor
111 20 Stockholm
Sweden
Tel: +46.8.102610
Fax: +46.8.102550

SPSS India Private Ltd.
Ashok Hotel, Suite 223
50B Chanakyapuri
New Delhi 110 021
India
Tel: +91.11.600121 x1029
Fax: +91.11.688.8851

SPSS Asia Pacific Pte. Ltd.
10 Anson Road, #34-07
International Plaza
Singapore 0207
Singapore
Tel: +65.221.2577
Fax: +65.221.9920

SPSS Japan Inc.
2-2-22 Jingu-mae
Shibuya-ku, Tokyo
150 Japan
Tel: +81.3.5474.0341
Fax: +81.3.5474.2678

SPSS Australasia Pty. Ltd.
121 Walker Street
North Sydney, NSW 2060
Australia
Tel: +61.2.954.5660
Fax: +61.2.954.5616

Contents

Syntax Reference

1 Discriminant Analysis

Gazing into crystal balls is not the exclusive domain of soothsayers. Judges, college admissions counselors, bankers, and many other professionals must foretell outcomes such as parole violation, success in college, and creditworthiness.

An intuitive strategy is to compare the characteristics of a potential student or credit applicant to those of cases whose success or failure is already known. Based on similarities and differences, a prediction can be made. Often this is done subjectively, using only the experience and wisdom of the decision maker. However, as problems grow more complex and the consequences of bad decisions become more severe, a more objective procedure for predicting outcomes is desirable.

Before considering statistical techniques, let's summarize the problem. Based on a collection of variables, such as yearly income, age, marital status, and total worth, we wish to distinguish among several mutually exclusive groups—for example, good and bad credit risks. The available data are the values of the variables for cases whose group membership is known—that is, cases who have proved to be good or bad credit risks. We also wish to identify the variables that are important for distinguishing among the groups and to develop a procedure for predicting group membership for new cases whose group membership is undetermined.

Discriminant analysis, first introduced by Sir Ronald Fisher, is the statistical technique most commonly used to investigate this set of problems. The concept underlying discriminant analysis is fairly simple. Linear combinations of the independent, or predictor, variables are formed and serve as the basis for classifying cases into one of the groups.

For the linear discriminant function to be "optimal"—that is, to provide a classification rule that minimizes the probability of misclassification—certain assumptions about the data must be met. Each group must be a sample from a multivariate normal population, and the population covariance matrices must all be equal. The section "When Assumptions Are Violated" on p. 36 discusses tests for violations of the assumptions and the performance of linear discriminant analysis when assumptions are violated.

The sections "Selecting Cases for the Analysis" on p. 3 through "Sum of Unexplained Variance" on p. 27 cover the basics of discriminant analysis and the SPSS output using a two-group example. Extending this type of analysis to include more than two groups is discussed beginning with the section "Three-Group Discriminant Analysis" on p. 27.

Investigating Respiratory Distress Syndrome

Respiratory distress syndrome (RDS) is one of the leading causes of death in premature infants. Although intensive research has failed to uncover its causes, a variety of physiological disturbances, such as insufficient oxygen intake and high blood acidity, are characteristic of RDS. These are usually treated by administering oxygen and buffers to decrease acidity. However, a substantial proportion of RDS infants fail to survive.

P. K. J. van Vliet and J. M. Gupta (1973) studied 50 infants with a diagnosis of RDS based on clinical signs and symptoms and confirmed by chest x-rays. For each case, they report the infant's outcome—whether the infant died or survived—as well as values for eight variables that might be predictors of outcome. Table 1.1 gives the SPSS names and descriptions of these variables.

Table 1.1 Possible predictors of survival

Variable	Description
survival	Infant's outcome. Coded 1 if infant died, 2 if infant survived.
sex	Infant's sex. Coded 0 for female, 1 for male.
apgar	Score on the Apgar test, which measures infant's responsiveness. Scores range from 0 to 10.
age	The gestational age of the infant measured in weeks. Values of 36 to 38 are obtained for full-term infants.
time	Time that it took the infant to begin breathing spontaneously, measured in minutes.
weight	Birth weight measured in kilograms.
ph	The acidity level of the blood, measured on a scale of 0 to 14.
treatmnt	Type of buffer administered (buffer neutralizes acidity). Coded 1 for THAM, 0 for sodium carbonate.
resp	Indicates whether respiratory therapy was initiated. Coded 0 for no, 1 for yes.

Some dichotomous variables, such as *sex*, are included among the predictor variables. Although, as previously indicated, the linear discriminant function requires that the predictor variables have a multivariate normal distribution, the function has been shown to perform fairly well in a variety of other situations.

In this example, we will use discriminant analysis to determine whether the variables listed in Table 1.1 distinguish between infants who recover from RDS and those who do not. If high-risk infants can be identified early, special monitoring and treatment procedures may be instituted for them. It is also of interest to determine which variables contribute most to the separation of infants who survive from those who do not.

Selecting Cases for the Analysis

The first step in discriminant analysis is to select cases to be included in the computations. A case is excluded from the analysis if it contains missing information for the variable that defines the groups or for any of the predictor variables.

If many cases have missing values for at least one variable, the actual analysis will be based on a small subset of cases. This may be troublesome for two reasons. First, estimates based on small samples are usually quite variable. Second, if the cases with missing values differ from those without missing values, the resulting estimates may be too biased. For example, if highly educated people are more likely to provide information on the variables used in the analysis, selecting cases with complete data will result in a sample that is highly educated. Results obtained from such a sample might differ from those that would be obtained if people at all educational levels were included. Therefore, it is usually a good strategy to examine cases with missing values to see whether there is evidence that missing values are associated with some particular characteristics of the cases. If there are many missing values for some variables, you should consider eliminating those variables from the analysis.

Figure 1.1 shows the SPSS output produced after all the data have been processed. The first line of the output indicates how many cases are eligible for inclusion. The second line indicates the number of cases excluded from analysis because of missing values for the predictor variables or the variable that defines the groups. In this example, two cases with missing values are excluded from the analysis. If cases are weighted, SPSS displays the sum of the weights in each group and the actual number of cases.

Figure 1.1 Case summary

```
          50 (Unweighted) cases were processed.
           2 of these were excluded from the analysis.
              2 had at least one missing discriminating variable.
          48 (Unweighted) cases will be used in the analysis.
```

```
Number of cases by group

                Number of cases
  SURVIVAL  Unweighted      Weighted  Label
         1          26          26.0  DIE
         2          22          22.0  SURVIVE

     Total          48          48.0
```

Analyzing Group Differences

Although the variables are interrelated and we will need to employ statistical techniques that incorporate these dependencies, it is often helpful to begin analyzing the differences between groups by examining univariate statistics.

Figure 1.2 contains the means for the eight independent variables for infants who died (group 1) and those who survived (group 2), along with the corresponding standard deviations. The last row of each table, labeled *Total*, contains the means and standard deviations calculated when all cases are combined into a single sample.

Figure 1.2 Group means and standard deviations

```
Group means

  SURVIVAL        TREATMNT          TIME          WEIGHT          APGAR

         1          .38462       2.88462         1.70950        5.50000
         2          .59091       2.31818         2.36091        6.31818

     Total          .47917       2.62500         2.00806        5.87500

  SURVIVAL             SEX           AGE              PH           RESP

         1          .65385      32.38462         7.17962         .65385
         2          .68182      34.63636         7.34636         .27273

     Total          .66667      33.41667         7.25604         .47917


Group standard deviations


  SURVIVAL        TREATMNT          TIME          WEIGHT          APGAR

         1          .49614       3.48513          .51944        2.77489
         2          .50324       3.70503          .62760        2.69720

     Total          .50485       3.56027          .65353        2.74152

  SURVIVAL             SEX           AGE              PH           RESP

         1          .48516       3.11226          .08502         .48516
         2          .47673       2.71759          .60478         .45584

     Total          .47639       3.12051          .41751         .50485
```

From Figure 1.2 you can see that 38% of the infants who died were treated with THAM, 65% were male, and 65% received respiratory therapy. (When a variable is coded 0 or 1, the mean of the variable is the proportion of cases with a value of 1.) Infants who died took longer to breathe spontaneously, weighed less, and had lower Apgar scores than infants who survived.

Figure 1.3 shows significance tests for the equality of group means for each variable. The *F* values and their significance, shown in the third and fourth columns, are the same as those calculated from a one-way analysis of variance with survival as the grouping variable. For example, the *F* value in Figure 1.4, which is an analysis-of-variance table for *weight* from the SPSS One-Way ANOVA procedure, is 15.49, the same as shown for *weight* in Figure 1.3. (When there are two groups, the *F* value is just the square of the *t* value from the two-sample *t* test.) The significance level is 0.0003. If the observed significance level is small (less than 0.05), the hypothesis that all group means are equal is rejected.