INTRODUCTORY STATISTICS GUIDE

MARIJA J. NORUŠIS

Introductory Statistics Guide

[redacted]

SPSS X™

Marija J. Norušis

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

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

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

7 8 9 0 SEMSEM 8 9 8 7 6

ISBN 0-918469-20-1
(previously ISBN 0-07-046549-5)

Library of Congress Cataloging in Publication Data

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

Bibliography: p.
Includes index.
1. SPSS (Electronic computer system) I. SPSS Inc.
II. Title. III. S.P.S.S.-X introductory statistics guide.
HA32.N68 1983 001.64'2 83-5456

Preface

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

—Alfred North Whitehead

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

THE SPSS^X SYSTEM

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

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

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

USING THIS TEXT

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

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

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

ACKNOWLEDGMENTS

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

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

—Marija J. Norušis

Contents

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

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

Chapter 13 SPSS^{X} Command Reference 175

Chapter 1 From Paper into a File

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

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

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

1.1 CASES, VARIABLES, AND VALUES

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

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

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

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

1.2 Identifying Important Variables

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

1.3 Recording the Data

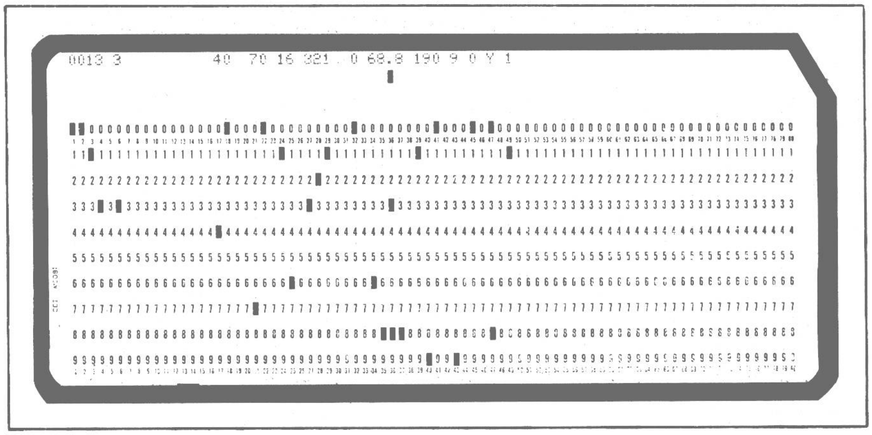
Once the variables have been selected, you must decide how they will be recorded. Do you need to record the actual date of birth or can you simply record the age in years? Is it sufficient to know if someone is a high-school or college graduate or do you need to know the actual number of years of education?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

1.4 Coding the Variables

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

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

Figure 1.4 A punched card



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

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

Table 1.4 Excerpt from uncoded data for Western Electric study

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

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

1.5 An Example

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

Table 1.5 Coding scheme for employee data form

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

Figure 1.5a Coded data

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

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

Figure 1.5b File with one record per case

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

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

Figure 1.5c File with two records per case

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

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

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