January 18, 2002

Dear Customer:

Thank you for your purchase of *A Handbook of Statistical Analyses using S-Plus* (Cat. #C2808) by Brian S. Everitt.

On the reverse are corrections for two figures. We sincerely regret any inconvenience this may have caused you. Please let us know if we can be of any assistance regarding this title or any other titles that CRC Press publishes.

Best regards,

CRC Press LLC
Figure 2.4 Scatterplot matrix for all five variables in the huswif data frame.

Figure 9.4 Scatterplot matrix of variables in the skulls data frame for each epoch.
Contents

Preface

Distributors for S-PLUS

1 An Introduction to S-PLUS
   1.1 Introduction
   1.2 Running S-PLUS
   1.3 The S-PLUS GUI: An Introduction
   1.4 The S-PLUS Command Language: An Introduction
      1.4.1 Elementary Commands
      1.4.2 Vectors and Matrices
      1.4.3 Subsetting Matrices and Vectors
      1.4.4 Other S-PLUS Objects
   1.5 An Example of an S-PLUS Session

Exercises

2 Describing Data and Assessing Distributions: Husbands and Wives
   2.1 Introduction
   2.2 Some Basic Summaries
   2.3 Analysis Using S-PLUS

Exercises

3 Analysis of Variance: Poisoning Rats and Losing Weight
   3.1 Description of Data
   3.2 Analysis of Variance
   3.3 Analysis Using S-PLUS
      3.3.1 Analysis of Variance of Survival Times of Rats
      3.3.2 Analysis of Variance of Slimming Data

Exercises
4  **Multiple Regression: Technological Changes in Jet Fishers**
   4.1  Description of Data
   4.2  Multiple Regression Model
   4.3  Analysis Using S-PLUS
   Exercises

5  **Logistic Regression: Psychiatric Caseness and Mortgage Default**
   5.1  Description of Data
   5.2  Logistic Regression and Generalised Linear Models
   5.3  Analysis Using S-PLUS
      5.3.1  GHQ Data
      5.3.2  Mortgage Default Data
   Exercises

6  **Analysing Longitudinal Data: Beating the Blues**
   6.1  Description of Data
   6.2  Analysing Longitudinal Data
   6.3  Analysis Using S-PLUS
      6.3.1  Summary Measure Analysis of the Depression Data
      6.3.2  Random Effects Models for the Depression Data
   Exercises

7  **Nonlinear Regression and Maximum Likelihood Estimation: Athletes and Geysers**
   7.1  Description of Data
   7.2  Nonlinear Regression and Maximum Likelihood Estimation
   7.3  Analysis Using S-PLUS
      7.3.1  Modelling the Olympic 1500-m Times
      7.3.2  Estimating the Parameters in a Mixture Fitted to the Geyser Eruption Data
   Exercises

8  **Survival Analysis: Motion Sickness and Bird Survival**
   8.1  Description of Data
   8.2  Describing Survival Times and Cox’s Regression
      8.2.1  The Survival Function
      8.2.2  The Hazard Function
   8.3  Analysis Using S-PLUS
      8.3.1  Motion Sickness
      8.3.2  Bird Deaths
   Exercises
9 Exploring Multivariate Data: Male Egyptian Skulls
   9.1 Description of Data
   9.2 Exploring Multivariate Data
   9.3 Analysis Using S-PLUS
       Exercises

10 Cluster Analysis: Low Temperatures and Voting in Congress
   10.1 Description of Data
   10.2 Cluster Analysis
   10.3 Analysis Using S-PLUS
       10.3.1 Clustering Cities in the United States on the Basis of their Year-Round Lowest Temperature
       10.3.2 Classifying New Jersey Congresspeople on the Basis of their Voting Behaviour
       Exercises

11 Bivariate Density Estimation and Discriminant Analysis: Blood Fat Concentration
   11.1 Description of Data
   11.2 Bivariate Density Estimation and Discriminant Function Analysis
   11.3 Analysis Using S-PLUS
       11.3.1 Bivariate Density Estimation
       11.3.2 Discriminant Analysis
       Exercises

Appendix A The S-PLUS Language
   A1 Vectors and Matrices
   A2 List Objects
   A3 Data Frames
   A4 Reading-in Data into S-PLUS
   A5 S-PLUS Functions
   A6 Graphics
   A7 User Functions

Appendix B Answers to Selected Exercises

References
Since the first edition of this handbook was published in 1994 the development of S-PLUS has continued apace, and a flexible and convenient “point-and-click” facility has now been added to supplement the very powerful command language. In addition, many new methods of analysis and new graphical procedures have been implemented. The changes made in this second edition reflect these changes in the software. Most chapters have been completely rewritten and many new examples are included. And, some of the more embarrassing code from the first edition are now excluded. A mixture of the S-PLUS command language and the S-PLUS Graphical User Interface (GUI) is used throughout the book so that readers can become familiar with using both. An appendix gives a relatively concise account of the command language.

It is hoped that this new edition will prove useful to applied statisticians, statistics students, and researchers in many disciplines who wish to learn about the many exciting possibilities for dealing with data presented by the latest versions of S-PLUS, S-PLUS 2000, and S-PLUS 6. All the data sets used in the text are available in the form of S-PLUS data frames from:

www.iop.kcl.ac.uk/IoP/Departments/BioComp/SPLUS.stm

Script files giving the command language used in each chapter are also available from the same address. (Comments given in the text versions are not included in these files.)

Thanks are due to Ms. Harriet Meteyard for her typing of the manuscript and general support during the writing of this book.

B.S. Everitt
June 2001
Distributors for S-PLUS

In the United Kingdom, S-PLUS is distributed by

Insightful
Knightway House
Park Street
Bagshot, Surrey
GU19 5AQ
United Kingdom
Tel: +44 (0) 1276 450 111
Fax: +44 (0) 1276 451 224
sales@uk.insightful.com

In the United States, the distributors are

Insightful Corporation
1700 Westlake Avenue North
Suite 500
Seattle, WA 98109-3044
USA
Tel: (206) 283-8802
Fax: (206) 283-8691
infoQinsightful.com
Web address: www.insightful.com
Dedication

To my daughters, Joanna and Rachel and my grandsons, Hywel and Dafydd
Chapter 1

An Introduction to S-PLUS

1.1 Introduction

S-PLUS is a language designed for data analysis and graphics developed at AT&T’s Bell Laboratories. It is described in detail in Becker et al. (1988), Chambers and Hastie (1993), Venables and Ripley (1997), and Krause and Olson (2000). In addition to providing a powerful language, the most recent versions of the software, S-PLUS 2000 and S-PLUS 6, also include an extensive graphical user interface (GUI) on Windows platforms (this is not available in UNIX). The GUI allows routine (and some not so routine) analyses to be carried out by simply completing various “dialog boxes,” and graphs to be produced and edited by a “point-and-click” approach.

In this chapter we introduce both the GUI and the command line language, although details of the former will be left for the remaining chapters of the book, and of the latter, for Appendix A.

1.2 Running S-PLUS

On a Windows platform S-PLUS is opened by double-clicking into the file (or shortcut for) S-PLUS.exe. The result is an S-PLUS window containing a Commands window and/or an Object Explorer window. During an S-PLUS session, Graphics windows may be opened and often output will
Figure 1.1 Windows seen during a typical S-PLUS session: Command window and script file are shown.

be sent to a Report window, this being opened by changing the default for text output routing in the Options list. The windows seen in a typical S-PLUS session are shown in Figure 1.1. At the top, below the S-PLUS title bar, is the menu bar. On the line below that is the tool bar.

S-PLUS provides a language for the manipulation of ‘objects’ such as vectors and matrices; commands can be typed into the Commands window next to the > prompt, and any resulting output will appear below, also in the Commands window, unless the Report window option has been selected. If a single command extends over one line of input, the > prompt changes to the plus sign, +. The contents of S-PLUS objects may be viewed by simply typing the name of the object.

The Object Explorer window displays objects of the current session by object category. This window can be opened by clicking into
At the end of a session, the user can select which objects created within the session should be saved within the ‘current directory' or database. By default, this is the \_data subdirectory of the directory where the S-PLUS files are located, for example, in C:\Program Files\sp6\_data. The command search lists the current directory under [1].

Since it is usually preferable to keep the data for different projects in different directories, it is a good idea to start an S-PLUS session by setting the directory in which any objects are to be saved and which may contain relevant objects from a previous session. This is done by ‘attaching' the directory at the first position of the search path using the command:

```r
>attach("c:/project/_data",pos=1)
```

Note that forward slashes are used in the directory path rather than the usual backward slashes. Alternatively, two backward slashes may also be used.

### 1.3 The S-PLUS GUI: An Introduction

Use of the GUI involves menus, dialog boxes, and point-and-click graphics. For example, many statistical techniques can be applied in S-PLUS by using the Statistics Menu and then filling in the relevant dialog box. These boxes have many features in common as we shall see throughout the text. As an example we can look at the Linear Regression dialog. This is made available as follows;

- Click on Statistics.
- Select Regression.
- Select Linear.

The resulting dialog is shown in Figure 1.2.

To use the box to carry out a regression analysis would involve filling in the various sections of the box and requesting various options under the Results, Plot, or Predict tabs, as we shall illustrate in detail in Chapter 4.

The GUI approach to producing S-PLUS graphics is extensive and flexible, and can involve either the use of dialog boxes from the Graphics Menu, or the Graphics palettes. For example, to access the Scatter Plot dialog, click on Graph in the tools bar, select 2D and Scatter Plot is highlighted by default. Click OK and the dialog box shown in Figure 1.3 appears.
Figure 1.2 Linear Regression dialog.

Again, to use the box to produce a scatterplot would involve filling out the box appropriately; examples will be given in later chapters. The 2D and 3D palettes are accessed by clicking on

or

respectively, and are shown in Figure 1.4. These can be used after selecting the required data set, to give a wide variety of graphics as we shall illustrate later.
Figure 1.3  Line/Scatter Plot dialog.

Figure 1.4  2D and 3D graphical palettes.
1.4 The S-PLUS Command Language:  
An Introduction

Although many users of S-PLUS will find its relatively recent GUI both convenient and sufficient for the analyses they require, it is the command language that remains the main reason that the software is so powerful and flexible. Familiarity with this aspect of S-PLUS enables customised analyses to be carried out relatively simply as we shall illustrate in later chapters. The differences between the GUI and the command language approaches and the advantages of the latter are nicely summarized in the following remarks made by a statistician who has been involved with computers for over 40 years, John Nelder:

I am very much aware that for the modern student the menu mode is the one preferred, and indeed the only one known. I am, however, not convinced that the menu mode is optimum for all users or for all usages. The freedom of being able to say what you want, instead of responding to given lists, is to me worth having. Imagine how restrictive conversation would be, if instead of making your own points for yourself, you were restricted to pointing at sets of alternatives defined by the person you were talking to. The frustrations would soon become apparent.

In this section we shall simply introduce a few of the most important features of the S-PLUS command language, leaving a more detailed account to be given in Appendix A.

1.4.1 Elementary Commands

Elementary commands consist of either expressions or assignments. For example, typing the expression

\[>42+8\]

in the commands window and pressing return will produce the following output

\[\text{[1]} \quad 50\]

Instead of evaluating just an expression, we can assign the value to a scalar, for example,

\[>x<-42+8\]
\[>x\]
\[\text{[1]} \quad 50\]
1.4.2 Vectors and Matrices

Vectors may be created in several ways, the most common is via the concatenate function, C, which combines all values given as arguments to the function as a vector.

```r
> x <- c(1,2,3,4)
> x
[1] 1 2 3 4
```

(Note that S-PLUS is case sensitive, x and X, for example, are regarded as different objects.)

Arithmetic operations between two vectors return a vector whose elements are the results of applying the operation to the corresponding elements of the original vectors. We can also apply mathematical functions to vectors; the functions are simply applied to each element of the vector.

```r
> x <- c(1,2,3)
> y <- c(4,9,16)
> x * y
[1]  4 18 48
> sqrt(y)
[1] 2 3 4
```

Matrix objects are frequently needed in S-PLUS and can be created by use of the matrix function. For details of this function see Appendix A or use help(matrix) in S-PLUS. (Similar help files are available for all S-PLUS functions.)

```r
> x <- c(1,2,3)
> y <- c(4,5,6)
> xy <- matrix(c(x, y), nrow=2)
> xy

[,1] [,2] [,3]
[1,] 1  3  5
[2,] 2  4  6
```

1.4.3 Subsetting Matrices and Vectors

S-PLUS has two logical values, T (true) and F (false), and a number of logical operations that are extremely useful in choosing particular elements.
from vectors and matrices. (The logical operations are listed in Appendix A.)

We can use a logical operator to assign logical values:

```r
> x <- 3 == 4
> x
[1] F
```

```r
> x <- 3 < 4
> x
[1] T
```

```r
> x <- c(1,2,3,4,5)
> x < 4
[1] T T T F F
```

A logical vector can be used to extract a subset of elements from another vector as follows:

```r
> x[x < 4]
[1] 1 2 3
```

Here the elements of the vector less than 4 are selected as the values corresponding to `T` in the vector `x`.

We can also select elements in `x` depending on the values in another vector `y`:

```r
> x <- c(1,2,3,4,5,6,7,8,9,10)
> y <- c(0,0,6,4,3,1,0,0,1,0)
> x[y == 0]
[1] 1 2 7 8 10
```

### 1.4.4 Other S-PLUS Objects

A number of other important S-PLUS objects are mentioned briefly here and in detail in Appendix A. First let objects that allow other S-PLUS objects to be linked together, for example,

```r
> x <- c(1,2,3)
> y <- matrix(c(1,2,3,4), nrow=2)
> xylist <- list(x,y)
> xylist
```
Note the two alternatives for referring to elements in a list; either the
'double bracket' nomenclature or the $name$ nomenclature can be used.

Secondly, data frames that allow numerical and character vectors to
be bound together are the most useful way of storing sets of data. Creating
a data frame is described in detail in Appendix A, but as a simple example:

```r
> height <- c(50, 60, 70)
> weight <- c(100, 120, 140)
> age <- c(20, 40, 60)
> names <- c("Bob", "Ted", "Alice")
> data <- data.frame(names, age, height, weight)

> data
    names age height weight
  1  Bob 20   50    100
  2  Ted 40   60    120
  3 Alice 60   70    140
```

A data frame can be used in S-PLUS by first 'attaching' it, using

```r
> attach(data)
```

In this way, variables in the data frame can now be conveniently referred
to by name.

```r
> age
[1] 20 40 60
```

The S-PLUS language also provides the facility for creating functions
for specific analyses of interest. Details are given in Appendix A and
examples will be given in subsequent chapters.

© 2002 by Chapman & Hall/CRC
Although commands can be typed into the commands window, it is far more convenient to use a **script file** (*.ssc), which is an ASCII text file that may be opened within S-PLUS to build up and keep a sequence of commands being used to analyse a particular data set, or indeed several data sets. In this way an entire analysis can be repeated at the press of a button if necessary, for example, if a data entry error is detected. The whole script file may be executed by selecting **Script** and **Run** from the menu bar or by pressing **F10**. Alternatively, one or more commands may be selected and run by highlighting the relevant text within the Script file and pressing the triangle insert button,

Script files can be commented by using the hash symbol, #, at the beginning of a line of text; S-PLUS ignores such lines. (In all but this chapter and Appendix A, we shall assume that commands are being run from a script file and, therefore, will dispense with the > before each command seen when using the commands window.) To open a script file, click on **File** in the menu bar, select **New**, and then highlight **Script File** in the list that appears.

### 1.5 An Example of an S-PLUS Session

As with any software, the easiest way to learn about S-PLUS is to use it, and this section attempts to give readers a preview of how S-PLUS is used in practice, which they can follow before reaching the more demanding material in subsequent chapters. Here we shall use both the GUI and the command language approaches to carry out some relatively straightforward analyses of the data shown in **Table 1.1**, which were originally given in Stanley and Miller (1979) and are also reproduced in Hand et al. (1994). (A more-detailed analysis of these data will be made in Chapter 4.)

We shall assume that the data in **Table 1.1** are already available as an S-PLUS data frame object, **jets**. Details of data frames and how they are created from the raw data are given in Appendix A. By typing **jets** in the command window and hitting return its contents will be displayed — see **Table 1.2**. Initially it is sensible to attach the data frame using

```>
>attach(jets)
```

To begin learning about the data we might want to look at some suitable summary statistics for each variable; for this we can use the S-PLUS **summary** function.

```>
>summary(jets)
```
Table 1.1 Data on Jet Fighters

<table>
<thead>
<tr>
<th>Type</th>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
<th>PLF</th>
<th>SLF</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FH-1</td>
<td>82</td>
<td>1.468</td>
<td>3.30</td>
<td>0.166</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>FJ-1</td>
<td>89</td>
<td>1.605</td>
<td>3.64</td>
<td>0.154</td>
<td>0.10</td>
<td>0</td>
</tr>
<tr>
<td>F-86A</td>
<td>101</td>
<td>2.168</td>
<td>4.87</td>
<td>0.177</td>
<td>2.90</td>
<td>1</td>
</tr>
<tr>
<td>F9F-2</td>
<td>107</td>
<td>2.054</td>
<td>4.72</td>
<td>0.275</td>
<td>1.10</td>
<td>0</td>
</tr>
<tr>
<td>F-94A</td>
<td>115</td>
<td>2.467</td>
<td>4.11</td>
<td>0.298</td>
<td>1.00</td>
<td>1</td>
</tr>
<tr>
<td>F3D-1</td>
<td>122</td>
<td>1.294</td>
<td>3.75</td>
<td>0.150</td>
<td>0.90</td>
<td>0</td>
</tr>
<tr>
<td>F-89A</td>
<td>127</td>
<td>2.183</td>
<td>3.97</td>
<td>0.000</td>
<td>2.40</td>
<td>1</td>
</tr>
<tr>
<td>XF10F-1</td>
<td>137</td>
<td>2.426</td>
<td>4.65</td>
<td>0.117</td>
<td>1.80</td>
<td>0</td>
</tr>
<tr>
<td>F9F-6</td>
<td>147</td>
<td>2.607</td>
<td>3.84</td>
<td>0.155</td>
<td>2.30</td>
<td>0</td>
</tr>
<tr>
<td>F-100A</td>
<td>166</td>
<td>4.567</td>
<td>4.92</td>
<td>0.138</td>
<td>3.20</td>
<td>1</td>
</tr>
<tr>
<td>F4D-1</td>
<td>174</td>
<td>4.588</td>
<td>3.82</td>
<td>0.249</td>
<td>3.50</td>
<td>0</td>
</tr>
<tr>
<td>F1F-1</td>
<td>175</td>
<td>3.618</td>
<td>4.32</td>
<td>0.143</td>
<td>2.80</td>
<td>0</td>
</tr>
<tr>
<td>F-101A</td>
<td>177</td>
<td>5.855</td>
<td>4.53</td>
<td>0.172</td>
<td>2.50</td>
<td>1</td>
</tr>
<tr>
<td>F3H-2</td>
<td>184</td>
<td>2.898</td>
<td>4.48</td>
<td>0.178</td>
<td>3.00</td>
<td>0</td>
</tr>
<tr>
<td>F-102A</td>
<td>187</td>
<td>3.880</td>
<td>5.39</td>
<td>0.101</td>
<td>3.00</td>
<td>1</td>
</tr>
<tr>
<td>F-8A</td>
<td>189</td>
<td>0.455</td>
<td>4.99</td>
<td>0.008</td>
<td>2.64</td>
<td>0</td>
</tr>
<tr>
<td>F-104B</td>
<td>194</td>
<td>8.088</td>
<td>4.50</td>
<td>0.251</td>
<td>2.70</td>
<td>1</td>
</tr>
<tr>
<td>F-105B</td>
<td>197</td>
<td>6.502</td>
<td>5.20</td>
<td>0.366</td>
<td>2.90</td>
<td>1</td>
</tr>
<tr>
<td>YF-107A</td>
<td>201</td>
<td>6.081</td>
<td>5.65</td>
<td>0.106</td>
<td>2.90</td>
<td>1</td>
</tr>
<tr>
<td>F-106A</td>
<td>204</td>
<td>7.105</td>
<td>5.40</td>
<td>0.089</td>
<td>3.20</td>
<td>1</td>
</tr>
<tr>
<td>F-4B</td>
<td>255</td>
<td>8.548</td>
<td>4.20</td>
<td>0.222</td>
<td>2.90</td>
<td>0</td>
</tr>
<tr>
<td>F-111A</td>
<td>328</td>
<td>6.321</td>
<td>6.45</td>
<td>0.187</td>
<td>2.00</td>
<td>1</td>
</tr>
</tbody>
</table>

FFD first flight date, in month after January 1940
SPR specific power, proportional to power per unit weight
RGF flight range factor
PLF payload as a fraction of gross weight of aircraft
SLF sustained load factor
CAR a binary variable that takes the value 1 if the aircraft can land on a carrier, and 0 otherwise.

The output resulting from these commands is shown in Table 1.3. (Like many S-PLUS functions, `summary` is generic, meaning that it can be used to process many different classes of data and give results appropriate to each particular class. Further examples will be given in subsequent chapters.)

Summary statistics for the data in `jets` can also be found by using the GUI as follows:

- Click **Statistics**.
- Select **Data Summaries**.
- Select **Summary Statistics**.
Table 1.2  The Jets Data Frame

<table>
<thead>
<tr>
<th></th>
<th>Type</th>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
<th>PLF</th>
<th>SLF</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FH-1</td>
<td>82</td>
<td>1.468</td>
<td>3.30</td>
<td>0.166</td>
<td>0.10</td>
<td>Cannot land</td>
</tr>
<tr>
<td>2</td>
<td>FJ-1</td>
<td>89</td>
<td>1.605</td>
<td>3.64</td>
<td>0.154</td>
<td>0.10</td>
<td>Cannot land</td>
</tr>
<tr>
<td>3</td>
<td>F-86A</td>
<td>101</td>
<td>2.168</td>
<td>4.87</td>
<td>0.177</td>
<td>2.90</td>
<td>Can land</td>
</tr>
<tr>
<td>4</td>
<td>F9F-2</td>
<td>107</td>
<td>2.054</td>
<td>4.72</td>
<td>0.275</td>
<td>1.10</td>
<td>Cannot land</td>
</tr>
<tr>
<td>5</td>
<td>F-94A</td>
<td>115</td>
<td>2.467</td>
<td>4.11</td>
<td>0.298</td>
<td>1.00</td>
<td>Can land</td>
</tr>
<tr>
<td>6</td>
<td>F3D-1</td>
<td>122</td>
<td>1.294</td>
<td>3.75</td>
<td>0.150</td>
<td>0.90</td>
<td>Cannot land</td>
</tr>
<tr>
<td>7</td>
<td>F-89A</td>
<td>127</td>
<td>2.183</td>
<td>3.97</td>
<td>0.000</td>
<td>2.40</td>
<td>Can land</td>
</tr>
<tr>
<td>8</td>
<td>XF10F-1</td>
<td>137</td>
<td>2.426</td>
<td>4.65</td>
<td>0.117</td>
<td>1.80</td>
<td>Cannot land</td>
</tr>
<tr>
<td>9</td>
<td>F9F-6</td>
<td>147</td>
<td>2.607</td>
<td>3.84</td>
<td>0.155</td>
<td>2.30</td>
<td>Cannot land</td>
</tr>
<tr>
<td>10</td>
<td>F-100A</td>
<td>166</td>
<td>4.567</td>
<td>4.92</td>
<td>0.138</td>
<td>3.20</td>
<td>Can land</td>
</tr>
<tr>
<td>11</td>
<td>F4D-1</td>
<td>174</td>
<td>4.588</td>
<td>3.82</td>
<td>0.249</td>
<td>3.50</td>
<td>Cannot land</td>
</tr>
<tr>
<td>12</td>
<td>F1F-1</td>
<td>175</td>
<td>3.618</td>
<td>4.32</td>
<td>0.143</td>
<td>2.80</td>
<td>Cannot land</td>
</tr>
<tr>
<td>13</td>
<td>F-101A</td>
<td>177</td>
<td>5.855</td>
<td>4.53</td>
<td>0.172</td>
<td>2.50</td>
<td>Can land</td>
</tr>
<tr>
<td>14</td>
<td>F3H-2</td>
<td>184</td>
<td>2.898</td>
<td>4.48</td>
<td>0.178</td>
<td>3.00</td>
<td>Cannot land</td>
</tr>
<tr>
<td>15</td>
<td>F-102A</td>
<td>187</td>
<td>3.880</td>
<td>5.39</td>
<td>0.101</td>
<td>3.00</td>
<td>Can land</td>
</tr>
<tr>
<td>16</td>
<td>F-8A</td>
<td>189</td>
<td>0.455</td>
<td>4.99</td>
<td>0.008</td>
<td>2.64</td>
<td>Cannot land</td>
</tr>
<tr>
<td>17</td>
<td>F-104B</td>
<td>194</td>
<td>8.088</td>
<td>4.50</td>
<td>0.251</td>
<td>2.70</td>
<td>Can land</td>
</tr>
<tr>
<td>18</td>
<td>F-105B</td>
<td>197</td>
<td>6.502</td>
<td>5.20</td>
<td>0.366</td>
<td>2.90</td>
<td>Can land</td>
</tr>
<tr>
<td>19</td>
<td>YF-107A</td>
<td>201</td>
<td>6.081</td>
<td>5.65</td>
<td>0.106</td>
<td>2.90</td>
<td>Can land</td>
</tr>
<tr>
<td>20</td>
<td>F-106A</td>
<td>204</td>
<td>7.105</td>
<td>5.40</td>
<td>0.089</td>
<td>3.20</td>
<td>Can land</td>
</tr>
<tr>
<td>21</td>
<td>F-4B</td>
<td>255</td>
<td>8.548</td>
<td>4.20</td>
<td>0.222</td>
<td>2.90</td>
<td>Cannot land</td>
</tr>
<tr>
<td>22</td>
<td>F-111A</td>
<td>328</td>
<td>6.321</td>
<td>6.45</td>
<td>0.187</td>
<td>2.00</td>
<td>Can land</td>
</tr>
</tbody>
</table>

The dialog box shown in Figure 1.5 appears. In the Data Set window choose jets, highlight all but Type in the Variables window and click OK; the results shown in Table 1.4 appear in a Report file which might be printed or copied and pasted into another application.

Perhaps separate summary statistics are required for the class of fighters that can land on a carrier and those that cannot. If so, they can be obtained by highlighting all but CAR in the Variables section of the Summary Statistics dialog and then highlighting CAR and Type in the Group Variables section. This leads to the results shown in Table 1.5. (Other summary statistics, for example, measures of skewness and kurtosis, can be requested simply by clicking on the Statistics tab of the Summary Statistics dialog.)

A t-test for the difference in the population mean values of, say, the variable FFD for planes that can land and cannot land on a carrier can be calculated using the Two-sample t-test dialog which is accessed as follows:
Table 1.3 Summary Statistics for the Jet Fighter Data

<table>
<thead>
<tr>
<th>Type</th>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
</tr>
</thead>
<tbody>
<tr>
<td>YF-107A:</td>
<td>Min.: 82.0</td>
<td>Min.:0.455</td>
<td>Min.:3.300</td>
</tr>
<tr>
<td>XF10F-1:</td>
<td>1st Qu.:123.2</td>
<td>1st Qu.:2.172</td>
<td>1st Qu.:4.005</td>
</tr>
<tr>
<td>FJ-1:</td>
<td>Median:174.5</td>
<td>Median:3.258</td>
<td>Median:4.515</td>
</tr>
<tr>
<td>FH-1:</td>
<td>Mean:166.3</td>
<td>Mean:3.944</td>
<td>Mean:4.577</td>
</tr>
<tr>
<td>F9F-6:</td>
<td>3rd Qu.:192.8</td>
<td>3rd Qu.:6.025</td>
<td>3rd Qu.:4.972</td>
</tr>
<tr>
<td>F9F-2:</td>
<td>Max.:328.0</td>
<td>Max.:8.548</td>
<td>Max.:6.450</td>
</tr>
<tr>
<td>(Other):16</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PLF</th>
<th>SLF</th>
<th>CAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.:0.0000</td>
<td>Min.:0.100</td>
<td>Cannot land:11</td>
</tr>
<tr>
<td>1st Qu.:0.1223</td>
<td>1st Qu.:1.850</td>
<td>Can land:11</td>
</tr>
<tr>
<td>Median:0.1605</td>
<td>Median:2.670</td>
<td></td>
</tr>
<tr>
<td>Mean:0.1683</td>
<td>Mean:2.265</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:0.2132</td>
<td>3rd Qu.:2.900</td>
<td></td>
</tr>
<tr>
<td>Max.:0.3660</td>
<td>Max.:3.500</td>
<td></td>
</tr>
</tbody>
</table>

- Click on Statistics.
- Select Compare Samples.
- Select Two Samples, t test.

Again select the jets data set, highlight FFD as Variable 1 and CAR as Variable 2, then tick the button that shows Variable 2 as a Grouping variable. The results shown in Table 1.6 appear in a Report file. (With such a small sample it may be more appropriate to use the Wilcoxon rank sum test rather than the t-test; we leave this as an exercise for the reader since the steps are essentially identical to those described above.) With the command language, the same results can be found using

>`t.test(FFD[CAR=="Can Land"],FFD[CAR=="Cannot land"])

Graphics are an essential component in the analysis of any data set, and a vast range of graphics are available when using S-PLUS, as we shall see in subsequent chapters. Here, however, we consider only the construction of a simple scatterplot. Using the GUI, we proceed as follows:

- Click on Graph.
- Select 2D Plot.
The **Insert Graph** menu appears. Since **Scatter Plot** is highlighted by default, simply click **OK** to arrive at the **Line/Scatter Plot** dialog. Select the **jets** data set and, say, **SPR** as the *x* column (the ‘‘*x* variable’) and **FFD** as the *y* column (the ‘‘*y* variable’’). Click **OK** to see the scatterplot of these two variables shown here in **Figure 1.6**. A more interesting scatterplot would be one with the points labelled by type of aircraft. This can again be constructed from the **Line/Scatter Plot dialog**.

We first repeat the steps used to obtain **Figure 1.7**, but now also select **Type** in the *z* Columns box, and then click the **Symbol** tab of the dialog. Tick **Use Text As Symbol** and in the **Text to Use** box select *z* column. The height of the plotting symbol might also be increased to, say, 0.15. Clicking **OK** now produces the scatterplot shown in **Figure 1.7**.

Finally, it might be useful to have **Figure 1.8** alongside the corresponding scatterplot in which the points are labelled by whether the aircraft can or cannot land on a carrier. This diagram is obtained as follows:
Table 1.4 Summary Statistics for Jet Fighter Data

*** Summary Statistics for data in: jets ***

$$"Factor Summaries":$$

<table>
<thead>
<tr>
<th>CAR</th>
<th>Cannot land: 11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can land: 11</td>
<td></td>
</tr>
</tbody>
</table>

$$"Numeric Summaries":$$

<table>
<thead>
<tr>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
<th>PLF</th>
<th>SLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>82.00000</td>
<td>0.455000</td>
<td>3.300000</td>
<td>0.0000000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>123.25000</td>
<td>2.171750</td>
<td>4.0050000</td>
<td>0.12225000</td>
</tr>
<tr>
<td>Mean</td>
<td>166.27273</td>
<td>3.944455</td>
<td>4.5772727</td>
<td>0.16827273</td>
</tr>
<tr>
<td>Median</td>
<td>174.50000</td>
<td>3.258000</td>
<td>4.5150000</td>
<td>0.16050000</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>192.75000</td>
<td>6.024500</td>
<td>4.9725000</td>
<td>0.21325000</td>
</tr>
<tr>
<td>Max</td>
<td>328.00000</td>
<td>8.548000</td>
<td>6.4500000</td>
<td>0.36600000</td>
</tr>
<tr>
<td>Total N</td>
<td>22.00000</td>
<td>22.000000</td>
<td>22.0000000</td>
<td>22.0000000</td>
</tr>
<tr>
<td>NAs</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.00000000</td>
<td>0.00000000</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>56.94122</td>
<td>2.367226</td>
<td>0.7529888</td>
<td>0.08665541</td>
</tr>
</tbody>
</table>

- With Figure 1.7 constructed and visible, click on Insert in the toolbar and select Graph.
- Click OK on the Insert Graph menu that appears.
- Select the jets data set and SPR as x and FFD as y.
- Here, however, select CAR as z.
- Click the Symbol tab and repeat the appropriate steps described above.
- Click on OK.

The resulting diagram is shown in Figure 1.8. (It may have been sensible here to have kept the height of the plotting symbol at its default value for the second diagram; we leave this as an exercise for the reader.)

With the command language the scatterplot in Figure 1.7 is obtained from

```r
> plot(SPR, FFD)
```

and the plot in Figure 1.8 from

```r
> plot(SPR, FFD, type="n")
> text(SPR, FFD, labels=as.character(Type))
```

Many other examples of the use of the plot function will be presented in later chapters.

© 2002 by Chapman & Hall/CRC
Table 1.5  Summary Statistics for Jet Fighter Data by Whether or Not Plane Can Land on Carrier

*** Summary Statistics for data in: jets ***

CAR: Cannot land

<table>
<thead>
<tr>
<th></th>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
<th>PLF</th>
<th>SLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>82.00000</td>
<td>0.455000</td>
<td>3.300000</td>
<td>0.00800000</td>
<td>0.100000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>114.50000</td>
<td>1.536500</td>
<td>3.7850000</td>
<td>0.14650000</td>
<td>1.000000</td>
</tr>
<tr>
<td>Mean</td>
<td>151.00000</td>
<td>2.869182</td>
<td>4.1554545</td>
<td>0.16518182</td>
<td>1.921818</td>
</tr>
<tr>
<td>Median</td>
<td>147.00000</td>
<td>2.426000</td>
<td>4.2000000</td>
<td>0.15500000</td>
<td>2.300000</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>179.50000</td>
<td>3.258000</td>
<td>4.5650000</td>
<td>0.20000000</td>
<td>2.850000</td>
</tr>
<tr>
<td>Max</td>
<td>255.00000</td>
<td>8.548000</td>
<td>4.9900000</td>
<td>0.27500000</td>
<td>3.500000</td>
</tr>
<tr>
<td>Total N</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
</tr>
<tr>
<td>NA's</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>51.03724</td>
<td>2.203758</td>
<td>0.5261248</td>
<td>0.07103354</td>
<td>1.200515</td>
</tr>
</tbody>
</table>

CAR: Can land

<table>
<thead>
<tr>
<th></th>
<th>FFD</th>
<th>SPR</th>
<th>RGF</th>
<th>PLF</th>
<th>SLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>101.00000</td>
<td>2.168000</td>
<td>5.970000</td>
<td>0.000000</td>
<td>1.000000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>146.50000</td>
<td>3.173500</td>
<td>4.1550000</td>
<td>0.10350000</td>
<td>2.450000</td>
</tr>
<tr>
<td>Mean</td>
<td>181.54545</td>
<td>5.019727</td>
<td>4.9990909</td>
<td>0.1713636</td>
<td>2.609091</td>
</tr>
<tr>
<td>Median</td>
<td>187.00000</td>
<td>5.855000</td>
<td>4.9200000</td>
<td>0.17200000</td>
<td>2.900000</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>199.00000</td>
<td>6.411500</td>
<td>5.3950000</td>
<td>0.21900000</td>
<td>2.950000</td>
</tr>
<tr>
<td>Max</td>
<td>328.00000</td>
<td>8.088000</td>
<td>6.4500000</td>
<td>0.36600000</td>
<td>3.200000</td>
</tr>
<tr>
<td>Total N</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
<td>11.00000</td>
</tr>
<tr>
<td>NA's</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>Std Dev.</td>
<td>60.75255</td>
<td>2.089900</td>
<td>0.7227926</td>
<td>0.1034527</td>
<td>0.642580</td>
</tr>
</tbody>
</table>

Table 1.6  Results of t-Test for Difference in FFD for Planes that Can and Cannot Land on a Carrier

Standard Two-Sample t-Test
data: x: FFD with CAR = Cannot land, and y: FFD with CAR = Can land
t = -1.2768, df = 20, p-value = 0.2163
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-80.449000  19.35809
sample estimates:
mean of x     mean of y
 151         19.35809
Figure 1.6 Scatterplot of SPR and FFD variables in jets data frame.

Figure 1.7 Scatterplot of SPR and FFD variables in jets data frame with points labelled by type of aircraft.
Figure 1.8 Side-by-side scatterplots for the SPR and FFD variables in the jets data frame.

Exercises

1.1. Investigate the use of the **apply** function to find the means of the numerical values in the **jets** data frame.

1.2. Use the **rep** function to help produce a new two-column data frame for the jet fighter data in which the numerical variable values for all planes are arranged in a single vector with the type of plane rearranged accordingly.

1.3. Use the **boxplot** function to construct box plots of the values of each numerical variable in the **jets** dataframe.

1.4. Explore the use of the **hist** and **density** functions for plotting histograms and calculating probability density estimates for some of the variables in the **jets** data frame.

1.5. Use the **help** function to find out about the **pairs** function and then apply this function to the variables in the **jets** data frame in ways that you think might be useful.
Chapter 2

Describing Data and Assessing Distributions: Husbands and Wives

2.1 Introduction

The data to be used in this chapter consist of five variables recorded on a random sample of 100 married men and their wives. The five variables are

- **husbage**: husband's age in years
- **husbht**: husband's height in mm
- **wifeage**: wife's age in years
- **wifeht**: wife's height in mm
- **husbagem**: husband's age at the time of the marriage

The data are given in Table 2.1. The label NA is used in S-PLUS to denote a missing value, here generally the result of the wife declining to give her age!

We shall use these data to illustrate some of the features of S-PLUS for describing data and for assessing distributions.