Doing It with SAS: A Supplement to Applied Statistics for Engineers and Scientists [1] [2]

J. D. Petruccelli B. Nandram M. Chen

[1]Development of these materials was partially supported by the National Science Foundation's Division of Undergraduate Education under Grant DUE 9254087.

Contents

Introduction

This document contains detailed instructions of how to use SAS to do the kinds of analyses documented in the text **Applied Statistics for Engineers and Scientists**. Each of the first fifteen chapters is keyed to one chapter of the text, and references material in that chapter. For example, Chapter 1 begins by describing where to find the electric usage data from Chapter 1 of the book and how to use those data and SAS to generate Figures 1.1 and 1.3 in the text.

Many of the analyses described in the text use a graphically-oriented component of SAS called SAS/INSIGHT. [1] An introduction to SAS/INSIGHT is found in Appendix A.

We have written a collection of SAS macros to facilitate the use of SAS in statistical applications and in labs. The macros are included on the accompanying computer disk. The uses of these macros are described where appropriate in this supplement. Two tables in Appendix B catalogue these macros, their functions, and the SAS components, in addition to base SAS, that are needed for their execution. To run any macro, read the desired *.mac file into the SAS program editor and submit it. A window will prompt you for any necessary input.

The data sets used in the text are also found on the accompanying computer disk. These data sets are provided in two formats:

- SAS transport file format. The data sets in this format can be converted to SAS data sets using the SAS procedure *proc copy*. In addition, users of JMP software[2] can read data sets in the SAS transport format directly into the JMP program. The SAS export files are identified by the suffix .xpt.
- ASCII file format. ASCII files are text files. The data files in this format have a header containing the variable names, and are space-delimited. The ASCII files are identified by the suffix .dat. Missing values in these files are denoted by periods.

Throughout this supplement we will give only one-part names for all SAS data sets. This is the same as assuming the data sets are in your work, or temporary storage, area. It may be that your instructor or those responsible for the computer facility at your school have put these data sets into a SAS data library. Then, you will have to use a two part name for the data sets. At WPI, for example, all data sets are in a SAS data library named SASDATA. If WPI students want to access the ELECE data set in this SAS library, they must use the name SASDATA.ELECE.

[1] SAS and SAS/INSIGHT are registered trademarks of SAS Institute, Inc.
[2] JMP is a registered trademark of SAS Institute, Inc.

Chapter 12

Doing It with SAS: Chapter 12

12.1 Data Sets

Name	Description
SF	Surface finish data, Example 12.1
SF31	Unreplicated surface finish data, Example 12.2
SF32	Surface finish data with center points, Example 12.2
WASH	Washing test scores, Example 12.3
PLUGS	Sparkplug removal times, Example 12.4
PLANES	Paper airplane flight times, Example 12.5

12.2 Analysis of Unreplicated 2^k Experiments

The SAS macro EFFECTS computes the effect estimates for an unreplicated 2^k design, and produces a plot showing the effects and the values of MOE and SMOE. Two SAS files are created. The first, whose name you specify at the prompt "DATA FILE TO STORE OUTPUT", contains the response, factors and interaction terms. The latter are labeled I12, I13, I123, etc. The second file, called DRANK contains the quantities effect name (EFFECT), effect estimate (ESTIMATE), normal quantile (QUANTILE), and effect label (LABEL).

Normal Quantile Plot of Effects

To obtain a normal quantile plot of the effects, you should open DRANK with SAS/INSIGHT and plot QUANTILE versus ESTIMATE, including LABEL as a label variable. To do this, choose *Analyze:Scatter Plot (Y X)* from the menu bar on the data window. A dialog window will appear. In this window, select QUANTILE as the Y variable, ESTIMATE as the X variable, and LABEL as the label variable. Click on "OK" to do the plot. When the plot appears and you resize it, you can click on any of the estimated effects appearing on it to see the name of the effect being estimated.

Residuals and Fitted Values

To obtain the residuals and fitted values, take the following steps:

1. From SAS/INSIGHT access the file you have named to store the response, factors and interaction terms.

2. Fit the model you desire by choosing *Analyze: Fit(Y X)* and choosing the response as Y and the desired factors in the model as Xs. The residuals and fitted values will automatically be created and placed in variables with names R_*name* and P_*name*, where *name* is keyed to the response variable name. For example, R_FINISH and P_FINISH might be created for the surface finish data.

You can then plot the residuals versus any variable you desire.

12.3 Analysis of Replicated 2^k Experiments

The macro CEFFECTS is the analogue of the macro EFFECTS for 2^k experiments with replicated center points. CEFFECTS works very much like EFFECTS: it computes all interaction variables and outputs them along with the responses and factors to a SAS file of your choice, and it computes the quantities effect name (EFFECT), effect estimate (ESTIMATE), normal quantile (QUANTILE), and effect label (LABEL) and puts them in the SAS data file DRANK. It also computes a test for curvature, which EFFECTS does not.

Normal Quantile Plot of Effects

This is obtained as for the unreplicated design.

Residuals and Fitted Values

This is done essentially as for the unreplicated design, except that you must exclude the center points from the fit. To do this, select the center points in the data window, and then choose *Edit: Windows: Exclude in Calculations*. After this, proceed as for the unreplicated design.

12.4 Interaction Plots

The interaction plot shown in Figure 12.3 was produced by the SAS macro IPLOT. The data are found in the SAS data set SF. To generate Figure 12.3, you should answer the prompts for input as follows:

1. The response variable is Y.

2. There are 2 main effects. The first is A, the second B.

3. The variable on the horizontal axis is A.

4. The variable showing the vertical levels is B.

IPLOT can also be used for plotting higher way interactions, as shown in Figure 12.9. You must first run EFFECTS (or CEFFECTS). From the EFFECTS (or CEFFECTS) input window, choose a data set to contain the values of the response, main effects and interactions. For present purposes call it OUT. When EFFECTS (or CEFFECTS) has run, call IPLOT. Input the name OUT as the data set in IPLOT. As stated in the chapter, there are many ways to display a three way interaction. The plot in Figure 12.9 was produced as follows:

1. The response variable is FINISH.

2. There are 3 main effects. The first is LEAD, the second FEED, and the third DWELL.

3. The variable on the horizontal axis is I12 (meaning the interaction of the first two variables: LEAD*FEED).

4. The variable showing the vertical levels is DWELL.

12.5 Transformations

The transformations discussed in Section 12.12 are easily available in SAS/INSIGHT from the data window by choosing *Edit:Variables* from the menu bar.

12.6 Restrictions

Some nice features have been implemented into the macros EFFECTS, CEFFECTS and IPLOT, but these require some restrictions on what can be done automatically in them. Three that you should be aware of are:

1. A maximum of 7 factors can be accommodated.
2. As usual, SAS variable names must be 8 letters/characters or less. However, when there are 5 or more factors, the total number of letters/characters in the names of all main effects is restricted. For 7 factors there can be no more than 34, for 6 factors there can be no more than 35 and for 5 factors there can be no more than 36 total letters/characters in the main effect names.
3. For 7 factors, the MOE/SMOE plot is in two parts.

Chapter 13

Doing It with SAS: Chapter 13

13.1 Data Sets

Name	Description
SF32	Surface finish data with center points, Example 13.1
MOLD	EVA ring data, Example 13.7
HANGER	Picture hanger data, Example 13.9
HANGERR	Reduced picture hanger data, Example 13.9

13.2 Obtaining a Design of a Given Resolution

Suppose we want to obtain a 2_V^{5-2} design (if possible). Call up the macro DESIGN2. A window will appear which will prompt you for the number of factors (tell it 5), the desired names of the factors (tell it A, B, C, D and E) the size of the fraction (tell it 4), the number of blocks (tell it 1), the maximum size interaction to display in the alias structure (tell it 5), and the name of a SAS data set to contain the design points. SAS will give you a design of maximum possible resolution.

Now look at the SAS OUTPUT window. An orthogonal array will be displayed, consisting of the main effects (labeled A-E), and a column of ones for blocks. Ignore the latter for now. **This array can be used to run the experiment, as the order of its runs has been randomized.** Now scroll upward in the window. The aliasing structure will be displayed. (note that SAS uses "0" instead of "I" to denote the identity).

The orthogonal array has also been output to the SAS data set you specified. When you run the experiment, you can use SAS/INSIGHT to enter the responses in this data set, and save the results for further analyses.

13.3 Blocking in 2^{k-p} Designs

To incorporate blocks into the 2^{k-p} design, run the macro DESIGN2 as above and simply input the number of blocks you want at the appropriate prompt. Try this now for a 2_{III}^{5-2} design with two blocks. The variable "BLOCK" in the orthogonal array in the output tells to which block each treatment combination is assigned. The aliasing structure in the output shows which effects the blocks (denoted "[B]") are confounded with. Here they are AC, BD, ABE, and CDE. In terms of the orthogonal array, those terms with a "+" in the product of the A and C columns are assigned to one block, the terms with a "-" are assigned to the other block. This is the design for the EVA ring data shown in Table 13.9

of the text, if we take A to be Mold Temperature, B to be Screw Speed, C to be Hold Pressure, D to be Probe Temperature and E to be Hold Time.

13.4 Using EFFECTS and CEFFECTS with 2^{k-p} Designs

You may use the macros EFFECTS and CEFFECTS to obtain estimates in 2^{k-p} designs. However, **you must input only k − p of the k main effects**. You can then determine the estimate of confounded effects by using the aliasing structure of the design. For example, suppose you want to run a 2^{6-2} design with factors A, B, C, D, E and F. You use the macro DESIGN2 to generate the design shown in Table 13.1.

A	B	C	D	E	F
-1	1	1	1	1	-1
-1	-1	1	1	-1	-1
1	-1	1	-1	1	-1
1	1	1	1	1	1
1	-1	-1	-1	-1	1
-1	-1	-1	-1	-1	-1
1	1	1	-1	-1	-1
-1	-1	-1	1	1	1
1	-1	1	1	-1	1
1	-1	-1	1	1	-1
1	1	-1	-1	1	1
-1	1	-1	-1	1	-1
1	1	-1	1	-1	-1
-1	-1	1	-1	1	1
-1	1	1	-1	-1	1
-1	1	-1	1	-1	1

Table 13.1: *Orthogonal Array for* 2^{6-2} *Design*

You then run EFFECTS, inputting the number of factors as 4 and naming these as A, B, C and D. Table 13.2 shows how the output from EFFECTS giving the computed effects would appear. As can be seen, they are named as main effects or interactions of A, B, C and D. In order to determine effects involving E and F you will have to consult the aliasing structure, which is displayed in Table 13.3.

From the aliasing structure, we can see, for example, that the effect for E is the same as the BCD interaction which will appear on the EFFECTS output. Similarly, the effect for F will be found as the ACD interaction, and so on for any other effect of interest.

Note: It is possible to choose a set of $k-p$ main effects which have some interactions that are aliased with main effects resulting in EFFECTS or CEFFECTS producing estimates of 0. If this happens, choose another $k - p$ main effects. Experience shows that sticking to the first $k - p$ main effects as inputs to EFFECTS or CEFFECTS avoids this problem.

OBS	EFFECT	LABEL	ESTIMATE	MOE	SMOE
1	A	a	2.50	0.077864	0.15807
2	B	b	-0.50	0.077864	0.15807
3	C	c	-2.75	0.077864	0.15807
4	D	d	-0.75	0.077864	0.15807
5	I12	a*b	1.00	0.077864	0.15807
6	I123	a*b*c	0.25	0.077864	0.15807
7	I1234	a*b*c*d	0.50	0.077864	0.15807
8	I124	a*b*d	-0.75	0.077864	0.15807
9	I13	a*c	-0.25	0.077864	0.15807
10	I134	a*c*d	-1.00	0.077864	0.15807
11	I14	a*d	0.75	0.077864	0.15807
12	I23	b*c	0.75	0.077864	0.15807
13	I234	b*c*d	1.00	0.077864	0.15807
14	I24	b*d	-1.25	0.077864	0.15807
15	I34	c*d	0.50	0.077864	0.15807

Table 13.2: *Effect Estimates for* 2^{6-2} *Design from Macro EFFECTS*

Aliasing Structure

0 = A*B*E*F = A*C*D*F = B*C*D*E
A = B*E*F = C*D*F = A*B*C*D*E
B = A*E*F = C*D*E = A*B*C*D*F
C = A*D*F = B*D*E = A*B*C*E*F
D = A*C*F = B*C*E = A*B*D*E*F
E = A*B*F = B*C*D = A*C*D*E*F
F = A*B*E = A*C*D = B*C*D*E*F
A*B = E*F = A*C*D*E = B*C*D*F
A*C = D*F = A*B*D*E = B*C*E*F
A*D = C*F = A*B*C*E = B*D*E*F
A*E = B*F = A*B*C*D = C*D*E*F
A*F = B*E = C*D = A*B*C*D*E*F
B*C = D*E = A*B*D*F = A*C*E*F
B*D = C*E = A*B*C*F = A*D*E*F
A*B*C = A*D*E = B*D*F = C*E*F
A*B*D = A*C*E = B*C*F = D*E*F

Table 13.3: *Aliasing Structure for* 2^{6-2} *Design*

Chapter 14

Doing It with SAS: Chapter 14

14.1 Data Sets

Name	Description
CAM1	Cam data 2^2 design, Example 14.1
CAM2	Cam data CCD design, Example 14.1

14.2 Creating a Central Composite Design

The macro CCDGEN will give you a range of Central Composite Designs to choose from for any desired number of factors. Input consists of the number of factors. Output, which is written to the SAS Output Window, consists of the types of designs available and instructions on how to generate them. As an example, the Table 14.1 displays output from CCDGEN when the number of factors is input as 3.

This output shows two basic CCDs. The first is the standard design with 8 corner points, 9 center points and 6 star (here called axial) points. The "axial extreme" is the coded value of a (see Section 14.6) at which the star point is located. Note that in that section it was stated that $a = \sqrt{3}$ (=1.732) would give a rotatable design. Here, the design is optimized using other considerations than just rotatability, but the result is still nearly rotatable.

The second design involves blocking and will not be considered further here.

The commands below the heading "%adxccd() parameters to construct:" tell how to generate the design and have it output to the Output Window and stored in a SAS data set. So that, if you want to store the output in the SAS data set "dataset" (remembering that this name should begin with "sasuser." to be permanent), submit the command

%adxccd(dataset,3,8,9,1.6818); from the SAS Editor Window.

14.3 Analyzing a Central Composite Design

Once you have the data for a CCD in a SAS data file, you may fit a response surface model using the macro RSCOMP. Input to RSCOMP is self-explanatory. Output is written to the input window and consists of the fitted model, significant effects (at the .05 level), stationary point (in coded units), eigenvalues, eigenvectors, and the estimated response at the stationary point.

```
     Number of
    Runs in the
     Factorial        Number of       Axial         Total Number
      Portion       Center Points    Extreme          of runs
    -----------   ----------------   -------   --------------------
1.        8        9                 1.6818     23
2.        8        6 = ( 2*2) +  2   1.6330     20 = ( 2*  6) +8


                    %adxccd() parameters to construct:
                   ----------------------------------------
              1.  %adxccd(*data set name*,3,8,9,1.6818)
              2.  %adxccd(*data set name*,3,8,2/2,1.6330,3)


  For blocked designs, equations give

              Number of   Number in each      Number in
    Total = ( factorial *   factorial   )  +   axial
               blocks         block            block
```

Table 14.1: *Output from Macro CCDGEN*

14.4 Doing Lab 14-1 with SAS

The macro QUADGEN will prompt you to input values for x_1 and x_2, and will output the value of the response, y. Use it to attempt OFAT optimization. Later, you can use the macro SURFPLOT will produce a contour plot and a 3-D plot of the response surface. Use these plots to see how well the OFAT optimization did.

Appendix A

An Introduction to SAS/INSIGHT

SAS/INSIGHT is an environment for interactive analysis of data. Its focus is on interactive graphics: graphics which the user can modify at the screen. An example of this is the ability to click on a data point (an unusual observation, for example) on a plot and have it identified with its corresponding observation number. Or to reverse this process, a subset of the data points on an existing plot (say all males) could be easily highlighted. SAS/INSIGHT also has many data-handling and data-analytic capabilities to complement its graphical capabilities.

This very brief introduction covers only the barest essentials of SAS/INSIGHT. Its goal is to get beginners up and running in the SAS/INSIGHT environment, and to provide a guide to some basic tasks.

A.1 Invoking SAS/INSIGHT

To access SAS/INSIGHT, select the "Globals" entry from the menu bar on any of the three windows SAS automatically brings up: PROGRAM EDITOR, LOG, or OUTPUT. Then select the "Analyze" and "Interactive Data Analysis" entries in succession. Try this now. A small window entitled "SAS: SAS/INSIGHT: Open" will appear on your screen. We will call all activities you perform in SAS/INSIGHT from the time this window appears until you exit SAS/INSIGHT, a **session**. The box before you is the initial dialog box. By pressing the "Open" button at the bottom, you may read an existing SAS data set into SAS/INSIGHT. You will be asked to do this later in this tutorial. To begin, however, you will be asked to create your own SAS data set using SAS/INSIGHT. To begin this process, click on the "New" button.

A new data window, entitled something like "SAS: WORK.A" will appear. This means that the SAS data set you will be creating will be found in the SAS data library "WORK", which is a storage area for temporary SAS data sets (data sets that will be erased when you exit from the current SAS session). The data window is divided into a number of rows and columns of rectangles. Each rectangle, which we will call a **cell**, will hold one piece of data. The upper left cell should be highlighted, which indicates that it is **selected** and ready to accept data entry. You will begin entering data soon. First, however, a few details about getting around.

A.2 Choosing from Menus

In SAS/INSIGHT, operations you can perform include creating graphs and analyses, transforming variables, fitting curves and saving results. These operations are chosen by **pulling down** a menu from a **menu bar**. The menu bar is located at the top of SAS/INSIGHT windows (the one on the data window has the items **File Edit Analyze Help**). To pull down a menu, click on the item of interest from the menu bar. A pop-up menu will appear. Continue holding the mouse button down while you drag it down the pop-up menu until you reach the desired item. If another pop-up menu appears, continue

holding the mouse button down and drag to the desired item. Release the mouse button when you have arrived at the desired operation.

For example, select the "Help" item on the menu bar. A pop-up window will appear. Drag the mouse down the items to "Reference →". Another pop-up window will appear. Move the pointer to the first item, "Data", and release the mouse button. This activates a help window which explains about the data windows in SAS/INSIGHT. The sequence of steps by which you brought up this data window can be written in shorthand and *italicized* as *Help:Reference:Data*. This shorthand and *italicized* notation will be used in the rest of this tutorial to describe how to move through the menus.

If you find you have made a mistake and don't want the pop-up menu you've opened, click on some neutral area of the window, such as blank space on the menu bar.

A.3 Help

You have just seen an example of the extensive help system available in SAS/INSIGHT (and, for that matter, all of SAS). You should use this resource both to learn more about SAS/INSIGHT and to try to figure out what to do when you are stuck. Specifically, *Help:Introduction* gives an overview of SAS/INSIGHT, *Help:Techniques* tells how to perform various tasks, *Help:Reference* will give you detailed information and *Help:Index* contains a list of all help topics. Take a cruise through these help windows; it will be worth your while.

There is also **context-sensitive help** available. For example, if you are displaying a bar chart (a subject considered later in this tutorial) and you want some question answered about bar charts, you can put the pointer on the bar chart and press the F1 key on the keyboard.

This tutorial will not attempt to duplicate the information found in the help windows. Instead it will focus on some of the features in SAS/INSIGHT which are unique or particularly easy to use.

A.4 Creating New Data

For this section of the primer we will assume that a project team consisting of three professors has just run a set of helicopter drops as described in Lab 1-2. If you aren't yet familiar with the helicopter drop, it consists of timing how long it takes a paper helicopter to drop a specified distance. The experiment requires someone to release the helicopter (the RELEASER) and someone to record the time it takes the helicopter to hit the floor (the RECORDER). The resulting data need to be entered into SAS:

RELEASER	RECORDER	TIME
Moe	Larry	2.15
Moe	Larry	1.34
Moe	Larry	2.47
Curley	Larry	.90
Curley	Larry	2.97
Curley	Larry	6.01
Moe	Curley	2.16
Moe	Curley	.
Moe	Curley	1.91
Larry	Moe	1.96
Larry	Moe	1.93
Larry	Moe	2.11
Curley	Moe	1.96
Curley	Moe	5.21
Curley	Moe	.36
Larry	Curley	5.96
Larry	Curley	2.33
Larry	Curley	2.88

If your team has already run the helicopter drops in Lab 1-2, you should follow along in this section but enter your team's data instead of the above data.

Now begin entering the data. Click on the upper-leftmost cell in the data window to select it for the first data value. Type "Moe" <*enter*> (note: <*keyname*> means press the key named *keyname* on the computer keyboard. On some computers the *enter* key has the name *return*.) The name "Moe" should appear in the selected cell as you type, and <*keyname*> should select the next cell down. Now in succession type "Moe" <*enter*>, "Moe" <*enter*>, and "Curley" <*enter*>. You are on your way to entering the data!

Defining Variables

You may have already noticed that the letters "Nom" appeared at the top of the first column, and below them the letter "A". A is the name SAS has given the first variable, and Nom indicates it is a **nominal variable**. A nominal variable is one which "names". Because the values you have input consist of letters, SAS has concluded (correctly) that the first variable is nominal. We want to name the first variable "RELEASER". To do this click on the triangle in the upper left corner of the data window (right below the "File" entry at the top of the window) with the left mouse button (always select with the left mouse button unless told otherwise). A popup menu will appear. Click on the menu entry "Define Variables...". A "SAS: Define Variables" dialog box will appear. Click on the "A" to the right of "Name:", enter the name "Releaser" (without the quotes), and click on the "OK" button. The name of the variable will now be "RELEASER".

Notation

Before we go on, two things. First, a word about notation. In what follows, we will denote the triangle you first clicked on with the symbol ▷. As we go through this tutorial, this triangle button will appear in a variety of windows and locations, but no matter where it appears, it will be referred to as ▷. Thus two mouse selections you used in changing the name of the variable would be described as "choose ▷*: Define Variables...*".

Second, a few comments about the data window. The window should now have four names entered under the variable named RELEASER. Notice the number 1 is to the right of the triangle and the number 4 is below it. The first tells the number of variables (columns) in the data set (there is only RELEASER) and the second tells how many observations. The left column contains small squares. These are the symbols used in plotting. The column to the right of these contains the observation number of each observation.

Finishing Entry of the Funnel Data

Now enter the rest of the data. You may continue entering the rest of the releaser names as you have been doing, or you may click on any cell to enter the value of a single observation, or you may enter rows of data. Let's try the latter. Click on the cell at the upper left containing the first data value you entered. Now press <*Tab*>. The next cell to the right should be highlighted. Enter "Larry". Tab over once more, enter "2.15" and press <*enter*>. Now enter "1.34", and press <*shift-tab*> (i.e. hold down the "shift" and "tab" keys simultaneously). This will enter the "1.34" and move one column to the left. You may now enter "Larry", press <*Enter*> to move one row down, and continue. You may use this or the column entry you began with to complete entry of the data, or you may devise some other method of your own.

When you have finished data entry, name the second and third variables RECORDER and TIME. Notice that RECORDER is a nominal variable, but TIME is an interval variable, which is the default for numerical measurements.

A.5 Creating a SAS Data Set

So far, the data you have entered are accessible only to SAS/INSIGHT and only during this session. If you exit INSIGHT the data will be lost. However, you can save these data in a SAS data set.

SAS data sets contain data and information about data such as variable names. They are created by SAS and are readable only by SAS. There are both temporary and permanent SAS data sets. Temporary data sets disappear after you finish your SAS session. They are stored in a library called WORK. Permanent data sets are stored in SAS data libraries in your directory, and may be accessed later. The default data library is SASUSER. Many SAS data sets have been created and stored for your use in the data library SASDATA.

To save your data to a SAS data set, from the data window choose *File: Save: Data*. A dialog box will appear offering you your choice of libraries to save to and allowing you to choose a name for the data set. If you want to create a temporary data set, select the library WORK. If you want to create a permanent data set, select the library SASUSER. In either case, call the data set FUNNEL.

A.6 Entering an Existing SAS Data Set

It may be that you want to use SAS/INSIGHT to analyze data in an existing SAS data set. Data from an existing SAS data set are entered into SAS/INSIGHT through the initial dialog box, which is automatically brought up when entering SAS/INSIGHT. The initial dialog box may also be accessed if you are already in SAS/INSIGHT, by choosing *File: Open*. Whichever method you use, bring up the initial dialog box now.

To enter a SAS data set into SAS/INSIGHT, click on the name of the library where the data set resides and then on the data set name. One or both these actions may involve scrolling the names in a window. To scroll, place the pointer on the slider bar, hold down the left mouse button, and move the mouse. You can scroll more slowly by clicking with the left mouse button on the arrows at the top or bottom of the scroll bar.

For this tutorial, we are going to analyze a data set supplied with SAS, called BASEBALL. Your instructor will have to tell you the library in which this data set resides. Select this library and then the data set BASEBALL. A data window containing this data set will appear. Use your mouse to enlarge this window and view its contents.

This data set consists of performance measures and salary levels for regular hitters and leading substitute hitters in major league baseball for the year 1986 (a year that will live in infamy for all Red Sox fans). The variables are:

NAME	The player's name
NO_ATBAT	Number of at bats
NO_HITS	Number of hits
NO_HOME	Number of home runs
NO_RUNS	Number of runs
NO_RBI	Number of runs batted in
NO_BB	Number of bases on balls
YR_MAJOR	Years in the major leagues
CR_ATBAT	Career at bats
CR_HITS	Career hits
CR_HOME	Career home runs
CR_RUNS	Career runs
CR_RBI	Career runs batted in
CR_BB	Career bases on balls
LEAGUE	Player's league at the end of the 1986 season
DIVISION	Player's division at the end of the 1986 season
TEAM	Player's team at the end of the 1986 season
POSITION	Position(s) played
NO_OUTS	Number of putouts
NO_ASSTS	Number of assists
NO_ERROR	Number of errors
SALARY	Salary in $1000's

You may access more than one SAS data set from SAS/INSIGHT at the same time. However, as you may have noticed, when the data window appeared, the initial dialog box window disappeared. To enter other data sets, choose *File: Open*. The initial dialog box will reappear to allow you to access another data set.

A.7 Exiting SAS/INSIGHT

To close any SAS/INSIGHT window, choose *File:End*. When a data window is closed, all windows generated from that window are also closed. When you have closed all data windows, you exit SAS/INSIGHT.

A.8 Selecting and Choosing

In SAS/INSIGHT, all operations you may want to perform are listed in **menus**. So to perform any task, you point with the mouse and click the buttons to select objects and choose operations from menus.

Selecting Objects

You select an object to indicate that it is an object you want to work with. Objects you can select in a data set in SAS/INSIGHT include variables (such as NAME or NO_ATBAT in the baseball data set), observations (such as all data for Wade Boggs), and individual values (such as Bill Buckner's number of errors). You can also select the results of analyses you conduct in SAS/INSIGHT, such as graphs, curves and tables. Selected objects become highlighted on the display.

To select an object move the pointer to it with the mouse and **click** (i.e. press and then release) the leftmost mouse button . To select multiple objects, **click and drag** by pressing and holding the left mouse button down while moving the pointer across the objects of interest, then releasing the mouse button. This selects all objects touched by the pointer while the mouse button was held down.

Try these techniques now on the baseball data. Select the variable NAME by clicking on it. Select observation 2 (Alan Ahsby) by clicking on the number 2 next to Alan's name. Select Andre Dawson's number of hits by clicking on the 141 in the appropriate box. Select the observations for the first 6 players by clicking and dragging in the leftmost column.

When objects are far apart, it is convenient to use **modifier keys** with the mouse button. The shift key can be used to make an **extended selection**. For example, to select the observations for the first 100 players, click on the number 1 next to Andy Allenson's name, scroll down to player 100 (Eddie Milner), and click on the number 100 while holding down the shift key.

To make a **non-contiguous** selection, use the Ctrl key in a similar way. For example, select the variables NAME, NO_HITS and CR_HOME by clicking on any one of them first, then on a second while holding down the Ctrl key, and again on the third while holding down the Ctrl key. Try it yourself.

De-selecting Objects

As you've noticed, selecting another object de-selects previously selected objects.

A.9 Manipulating Data

In this section, you will learn several features of SAS/INSIGHT for data manipulation.

Arranging Variables or Observations

You can easily change the order in which the variables appear in the data window. For example, you can move the variable SALARY from its position at the far right of the baseball data set to the leftmost position. There are two ways to do this:

1. You may select ▷*:Move to First* which, as long as a variable is not already selected, will bring up a dialog box containing the names of all variables. Scroll down the list in this box, click on "SALARY" and then on "OK".

2. You may first select the variable SALARY in the data window and then select ▷*:Move to First*. In this case no dialog box will appear. This also works with any pre-selected observation.

▷*:Move to Last* will reverse this operation. These methods also work with several variables or observations: just select the desired variables or observations.

Sorting Observations

Sorting observations by values of a variable is easy in SAS/INSIGHT. As an example, suppose you want to sort the data according to player's salary. To do this, scroll to SALARY using the horizontal scroll bar at the bottom of the data window. Select the variable "SALARY". Now click on ▷*:Sort*. The data are now in order of ascending salary. Note that the "."s in the data set stand for missing data. (You could also have done this without selecting "SALARY" first. Then a dialog box would appear and you would select "SALARY" from it.).

Finding Observations

Sometimes you want to find observations that share some characteristic. For example, I know you all want to find all the Red Sox players in this data set. To do this, click on *Edit:Observations:Find*. A dialog box will appear. Select the variable TEAM from the left box, "=" from the center box, and "Bos." from the right box, then click on "OK". Now all the Red Sox players are highlighted.

You can do a bit more. By selecting ▷*:Find Next* the Red Sox player closest to the top will be put at the top of the data set, and the order of observations will be maintained. By selecting ▷*:Move to First*, all the Red Sox players will be moved to the top of the data set, but of course the order of the observations will be changed.

Transforming Data

You can transform variables to create new variables in SAS/INSIGHT. For example, though there is no batting average variable in the BASEBALL data set, you can easily create one as follows (For you non-fans, batting average is the number of hits divided by the number of at bats):

1. Choose *Edit:Variables:Other*. A dialog box will appear.
2. In the box with the variables list click on NO_HITS to select it, then click on the "Y" button. "NO_HITS" should appear in the box below it.
3. Next click on NO_ATBAT to select it, then click on the "X" button. "NO_ATBAT" should appear in the box below it.
4. Click on the "Y/X" under "Transformation:".
5. Select a reasonable name for the new variable. BA is good.
6. Now click on "OK". The variable for batting average will appear last in the data window.

While SAS/INSIGHT does not have any formal editing facilities (the SAS editor or a system editor is what we recommend for that task), you can easily change individual data values. Suppose we don't like Mike Schmidt's .0500 batting average and want to change it to .3500. To do this select Mike's batting average and then ▷*:Fill Values*. A pop-up window will ask for the value you want to put in that cell. Type in .350 and hit "OK".

A.10 Graphing Data

SAS/INSIGHT's strength is its ability to create sophisticated graphical displays. To introduce you to get you SAS/INSIGHT's capabilities, we'll consider the simplest graphical display, the **histogram**. A histogram is a graphical summary of a data set which creates a number of subgroups of the data based on the value of the variable being plotted. One bar is drawn over the range of values in each subgroup. The height of the bar drawn over a subgroup is proportional to the number of data points in that subgroup.

Draw a histogram for each of the variables SALARY and BA. To do this, select these two variables in the data window. Choose *Analyze:Histogram/Bar Chart (Y)* from the menu bar. A window containing two histograms will appear. Enlarge this window now. The graphs will remain small.

To enlarge the graphs, choose *Edit:Windows:Renew* from the menu bar of the graph window. A dialog box will appear: click on "OK".

You can move and change the size and/or shape of the graphs using the mouse. To move a graph, click with the left mouse button anywhere (except at a corner) on the side of the frame enclosing the graph. Then, still holding mouse button down, move the frame to a new location. Release the mouse button when the frame is where you want it. To enlarge (or shrink) the graph, click on a corner of the frame. As you move the mouse, the frame will change shape. Release the mouse button when the graph is the right size. With a little practice, you'll get quite good at this.

Incidentally, now would be a good time to try out the context-sensitive help facility in SAS/INSIGHT. Put the pointer on one of the histograms and press the F1 key. This will bring up a help window about histograms.

Customizing the Bar Chart/Histogram

SAS/INSIGHT will automatically choose the number of groups and the group boundaries on the histogram. You can customize the histogram by altering both the number of groups and/or the group boundaries, as follows:

1. Select *Edit:Windows:Tools* from the menu bar of the bar chart window. A window will pop up containing three icons at the top, a palette of colors below it and a number of buttons with different symbols below that.

2. Click on the icon shaped like a hand. This is the **move tool**.

3. Click on the bar chart. This will change the widths of the bars depending on how far the hand is from the base of the bars: clicking close to the base gives greater width; clicking far from the base gives smaller width.

4. Similarly, the place where the first bar starts can be changed by varying the horizontal position of the move tool.

A good way to see how the appearance of the bar chart can be changed is to hold down the left mouse button while moving the move tool all around the bar chart. Try this now. Does it help you to get a better picture of the data?

You can more precisely specify the positions of the bars in the bar chart by first selecting the variable being bar-charted by clicking on its name in the bar chart window and then choosing ▷*:Ticks*, where ▷ is found in the lower left corner of the bar chart window. The resulting dialog box allows you to specify the minimum and maximum of the axis as well as the starting and ending location of the bars (first and last ticks) and bar width (tick increment).

By choosing the vertical variable before choosing ▷*:Ticks*, you can control the look of the vertical axis.

Identifying Observations

This feature demonstrates some of the power of SAS/INSIGHT. Suppose you want to look at the data in the leftmost bar of the bar chart for SALARY. To do this, click on that bar. You will notice that not only does that bar become highlighted, but parts of the bar chart for BA do as well. Now look at the data window. You'll notice that the observations of the players whose salaries are displayed in the leftmost bar of the bar chart are also highlighted. This illustrates two things. First, you can select observations by clicking on locations on graphs. Second, when you select a subset of observations, the selection is displayed on all relevant windows in SAS/INSIGHT. To de-select, just click on an empty region of the barchart window. Try this now.

You can do this in reverse as well. Go to the data window and select observations 1-10. These will become highlighted in the data window and on your graphs.

Deleting Graphs from a Window

To delete a graph, first select it by putting the cursor outside the graph frame and clicking and dragging the cursor inside the frame. The graph will become highlighted. Then choose *Edit:Delete*. The window will disappear.

Plots Broken Down by Groups

Suppose you want to compare the bar charts of batting averages for American and National Leagues. This is easily done as follows. Choose *Analyze:Histogram/Bar Chart (Y)*. From the resulting dialog box, select BA and click on the "Y" button. Next select the variable LEAGUE and click on the "Group" button. Click on "OK". Separate bar charts for each League should appear side by side in the resulting window.

Be careful in comparing them, though! The scale of their axes won't be the same. A neat way to fix this is to put one directly below the other, being careful to align the boxes. Now choose *Edit:Windows:Align*. The axes will now line up for easy comparison. Do you detect any differences in batting averages between the two leagues?

Scatter Plots

A scatter plot or X-Y plot is a graph of bivariate data which plots the X variable on the horizontal axis and the Y variable on the vertical axis. As an example, suppose you are interested in whether there

was a relation between a player's salary and his batting average. The best way to see any relationship is to plot SALARY (Y) versus BA (X). To do this, choose *Analyze:Scatter Plot (Y X)* from the menu bar of **either** the data window or the barchart window. A dialog box will appear. Select BA as the X variable by clicking on BA in the variables box on the left and then clicking on the "X" button at the upper right. Select SALARY as the Y variable by clicking on SALARY in the variables box and then clicking on the "Y" button. Select NAME as the label variable by clicking on it in the variables box and then clicking on the "Label" box. Then click on "OK". The scatter plot will appear. Enlarge the window and renew the plot as desired.

Do you see a pattern to the data? Are there any unusual points? To find out who they are, click on any of those points on the plot. The player's name will appear because that is the label you gave the data. Who were the most underpaid players in terms of batting average? The most overpaid?

Perhaps you want to find which variables among NO_RBI, CR_RBI and SALARY were most related. You can use SAS/INSIGHT to produce a **scatterplot array**. In the data window select the variables NO_RBI, CR_RBI and SALARY. Then from the menu bar choose *Analyze:Scatter Plot (Y X)*. Enlarge the window as desired and renew the plot. Check out the results. Smooth, huh? What do you conclude about the relationships between pairs of these variables?

A.11 Examining Data

You can also examine data that you see in graphs. As an example, go back to the scatterplot of SALARY versus BA. Choose an unusual observation and double click on it. A window will appear with the values of all variables for this observation. You can do the same for groups of observations. You can obtain the same results by single clicking on the observation(s) and choosing *Edit:Observations:Examine*.

Edit:Observations:Examine is also useful in examining data for observations chosen by *Edit:Observations:Find*. For example, you can look at the records of all Red Sox players by choosing *Edit:Observations:Find*, selecting the variable TEAM from the left box, "=" from the center box, and "Bos." from the right box, then clicking on "OK". Now choose *Edit:Observations:Examine* to get the data on all the Red Sox.

A.12 Slicing

Slicing is a dynamic technique for viewing subsets of data based on a range of values for one variable. For example, to see how BA is related to SALARY and NO_RBI, look again at the two scatter plots you produced in the previous section.

Create a rectangular **brush** by clicking in the middle of the point cloud on the SALARY by BA scatter plot, holding the left mouse button down, and moving the mouse to create a rectangle. When you release the mouse button, all points in the brush are selected and will become highlighted on both graphs. Now move the brush by clicking in it and dragging. As the brush moves, different observations are selected in both graphs. Now to see how the relation between SALARY and NO_RBI changes for changing BA values, make the brush long (in the SALARY direction) and thin (in the BA direction) and move it left to right or right to left on the SALARY by BA scatter plot.

To make the effect more dramatic, choose *▷:Observations* and then drag the brush. Now only the selected observations will appear. One final feature you should be aware of that's also kind of fun is that if you release the mouse button while still dragging the brush, it will continue to move on its own.

A.13 Marking Observations

You can assign markers to use for displaying observations in scatter plots, boxplots (which you'll learn about later) and rotating 3-D plots (for which you're on your own). The markers appear with each observation in the data window. You can assign markers for observations you select, and you can let SAS/INSIGHT assign markers automatically based on the value of a variable. You can control the size of the markers in any plot.

Marking Individual Observations

To see how to mark individual observations, create a scatter plot of NO_RBI versus NO_HITS. Select an observation that interests you by clicking on it. If the SAS:Tools window is not already open, Choose *Edit:Windows:Tools* (if you choose *Edit:Windows* and see a highlighted square to the left of Tools, the SAS:Tools window is already open). A SAS Tools window will appear. Click on the shape of the marker you want to denote the chosen observation. The marker will change to the shape you choose in all graphs and in the data window.

Marking by Nominal Variable

A nominal variable is a variable whose values stand for names of categories. LEAGUE, DIVISION, TEAM, and POSITION are all nominal variables. SAS/INSIGHT can assign markers based on the value of a nominal variable. Let's mark the National and American League players separately in the NO_RBI versus NO_HITS plot. To do this, select LEAGUE in the data window and click on the multiple marker button at the bottom of the SAS: Tools window.

Marking by Interval Variable

You can also assign markers based on the value of an interval variable (i.e a variable whose values stand for numerical quantities, such as BA and NO_HITS). Let's assign markers in the NO_RBI versus NO_HITS plot based on SALARY. To do this, select SALARY in the data window and click on the multiple marker button at the bottom of the markers window. A different marker will be assigned to the players in the upper, middle and lower third of SALARY values.

Adjusting Marker Size

You can adjust the marker size on the plot by choosing $\triangleright$*:Marker Sizes*. Try a few sizes to find one you like.

A.14 Coloring Observations

If you are using a color monitor, coloring the markers different colors may be a more effective strategy than changing marker shapes. (Although for printing purposes, different shapes of markers show up better).

Basically, coloring observations proceeds in the same way as marking observations. The same SAS:Tools window used in marking is also used in coloring, so make sure it is open.

Coloring Individual Observations

To see how to color individual observations, create a scatter plot of NO_RBI versus NO_HITS. Select an observation that interests you by clicking on it. From the SAS:Tools window click on the color you want to denote the chosen observation. The color will change to the shade you choose in all graphs and in the data window.

Coloring by Nominal Variable

Let's color the National and American League players separately in the NO_RBI versus NO_HITS plot. To do this, select LEAGUE in the data window and click on the multiple color button (the rectangular colored button) at the bottom of the colors.

Coloring by Interval Variable

Let's assign colors in the NO_RBI versus NO_HITS plot based on SALARY. To do this, select SALARY in the data window and click on the multiple color button. A different color will be assigned to the players in the upper, middle and lower third of SALARY values.

A.15 Hiding Observations

You can adjust the range of data displayed and show subsets of the data by hiding observations. To illustrate the procedure, display the scatter plot of SALARY versus BA. We would like to investigate this relationship for each league on the same scatter plot (note that we could generate two separate scatter plots by using the variable LEAGUE as a group variable). We need to select the players from the National and American Leagues separately. A clever way to do this is to generate a bar chart of the variable LEAGUE. By clicking on the bar for the American League, all American League players are selected. Do this now.

To look at the scatterplot of SALARY versus BA for just National League players, choose *Edit:Observations:Hide in Graphs*.

Now look at the data window. De-select the selected observations by clicking on the upper left data cell of the data array. Notice that the previously selected observations now have no markers at all in the far left column. This says that these observations are hidden in all graphs (notice that the bar chart of LEAGUE has only the National League bar).

To make the observations visible in the graphs again, first choose *Edit:Observations:Invert Selection*, which de-selects all selected observations and selects all de-selected observations. Since all observations were de-selected just prior to this, all observations are now selected. If you now choose *Edit:Observations:Show in Graphs*, all observations will appear in the the graphs.

A.16 Toggling the Display of Observations

You can show subsets of the data by toggling the display of observations. This causes observations to be displayed only when they are selected. To illustrate this, create two scatter plots: one of SALARY versus BA, and the other of SALARY versus NO_RBI, by choosing *Analyze:Scatter Plot (Y X)*, and assigning SALARY the Y role and BA and NO_RBI the X role.

You will now create a toggle on the value of LEAGUE as follows:

1. Choose from the lower left of one of the scatterplots ▷*:Observations*. All the observation markers will disappear from the two scatter plots.

2. Choose *Edit:Observations:Find* from the data window.

3. In the dialog box, select LEAGUE from the variables list. Select the value you wish to display first: American or National League. Click on "OK".

Both scatterplots will now display the data for the league you selected. To toggle between the two leagues, choose *Edit:Observations:Invert Selection*. Each time you do this the data displayed will change to the other league. By doing this quickly, you can detect differences between the leagues.

To undo the toggling, choose ▷*:Observations* again. Click on an empty area of the graph window to de-select.

A.17 Printing Window Contents

All SAS/INSIGHT output seen at the screen is written to the SAS/INSIGHT windows. To print the contents of a window, choose *File:Print*. Unless you select certain objects, the entire window contents (i.e. what you see in the window) will be printed. Objects which do not appear in the window will not be printed. If you select a subset of the objects that you see in the window, only they will be printed.

A.18 Saving Data

The SAS data sets you read into SAS/INSIGHT are not affected by any modifications you may have made during your SAS/INSIGHT session. You can, however, save the data modified in SAS/INSIGHT to a SAS data set. The resulting data set will contain:

- All data values and variables as they currently appear in the data window.
- All observation states, including color, marker shape and show/hide.

To save the baseball data set as it currently exists in SAS/INSIGHT, choose *File:Save: Data*, and from the resulting dialog select the library where you want the data set stored (usually WORK if you want it to be temporary and SASUSER if permanent). You should also choose a data set name.

A.19 Connection with SAS

SAS/INSIGHT accesses the same SAS data sets common to all SAS modules. Therefore any output written to a SAS data set by SAS/INSIGHT can be accessed by other SAS modules and vice-versa. Also, SAS/INSIGHT can be run simultaneously with other SAS modules such as SAS/CALC.

There is one caution, however. If a data set is open in SAS/INSIGHT, other SAS programs may be unable to access or write to it. In this case a good strategy is to save a copy of the data set to a temporary data set as outlined in Section A.18, and use one for analysis in SAS/INSIGHT and another for all other SAS analyses.