Chapter 1
Getting Started

Chapter Table of Contents

Chapter 1
Getting Started

SAS/INSIGHT software is a tool for data exploration and analysis. With it you can explore data through graphs and analyses linked across multiple windows. You can analyze univariate distributions, investigate multivariate distributions, and fit explanatory models using analysis of variance, regression, and the generalized linear model.

This introduction summarizes important features, describes how to use the product, and explains how to learn more about SAS/INSIGHT software.
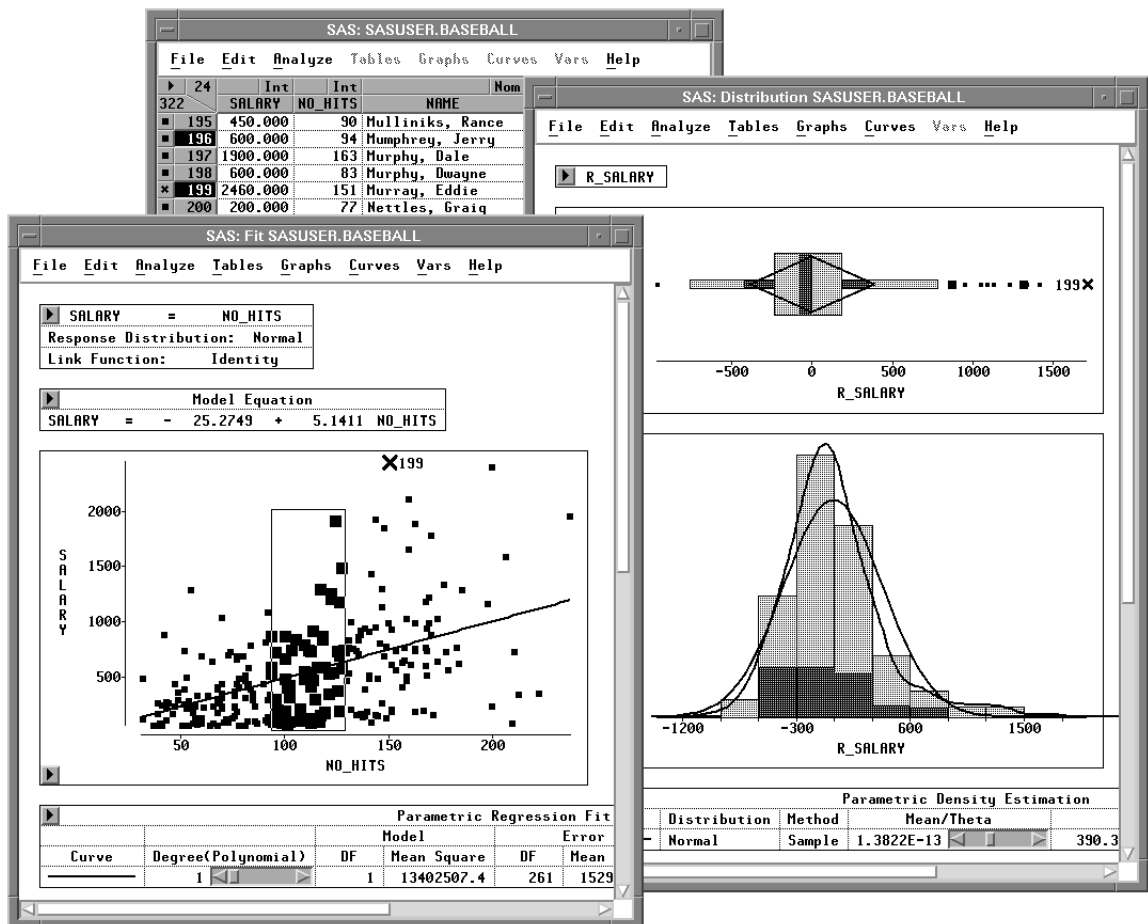

Figure 1.1. Brushing Observations in SAS/INSIGHT Software

Summary of Features

SAS/INSIGHT software provides a comprehensive set of exploratory and analytical tools.

To explore data, you can

- identify observations in plots
- brush observations in linked scatter plots, histograms, box plots, line plots, contour plots, and three-dimensional rotating plots
- exclude observations from graphs and analyses
- search, sort, and edit data
- transform variables
- color observations based on the value of a variable

To analyze distributions, you can

- compute descriptive statistics
- create quantile-quantile plots
- create mosaic plots of cross-classified data
- fit parametric (normal, lognormal, exponential, Weibull) and kernel density estimates for distributions
- fit parametric and empirical cumulative distribution functions
- test hypotheses of completely specified (known parameters) or specific (unknown parameters) parametric distributions based on Kolmogorov's D statistic

To analyze relationships between a response variable and a set of explanatory variables, you can

- fit curves with polynomials, kernels, and smoothing splines
- fit curves with nonparametric local polynomial smoothers using either a fixed bandwidth or loess smoothing
- add confidence bands for mean and predicted values
- fit surfaces with polynomials, kernels, and smoothing splines
- create residual and leverage plots
- fit the general linear model, including classification effects for analysis of variance and analysis of covariance
- fit the generalized linear model, including logistic regression, Poisson regression, and other models

To analyze relationships between variables, you can

- calculate correlation matrices and scatter plot matrices with confidence ellipses for relationships among pairs of variables
- reduce dimensionality of interval variables with principal component analysis
- examine relationships between two sets of interval variables with canonical correlation analysis and maximum redundancy analysis
- examine relationships between a nominal variable and a set of interval variables with canonical discriminant analysis

In addition, you can

- process data by groups
- process multiple data sets
- store option settings to customize SAS/INSIGHT operation
- store results as SAS data sets, SAS/GRAPH catalogs, and text files
- record and submit SAS/INSIGHT statements
- obtain context-sensitive help

Finally, because it is a part of the SAS System, you can use SAS/INSIGHT software to explore results from any SAS procedure. Conversely, you can use any SAS procedure to analyze results from SAS/INSIGHT software.

7

Of Mice and Menus

This section describes how to operate SAS/INSIGHT software and defines terms used in the rest of this book.

Some details depend on your *host*, the specific system of computing hardware and software you use. For example, all hosts present SAS/INSIGHT software in a system of *windows* on the host's *display*, but the appearance of your windows may differ from the figures in this book. You can find more information in the SAS companion for your host and in your host system documentation. On most hosts, you can *point* to objects on the display by using a *mouse*. A mouse is a physical device that controls the location of a *cursor*, a small moveable symbol on the display. The mouse also has *buttons* that work like keys on the computer keyboard. By pointing with the mouse and clicking a button, you can indicate any object on the display. In SAS/INSIGHT software, 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

Objects you can use in SAS/INSIGHT software include variables, observations, values, graphs, curves, and tables. You *select* an object to indicate that it is an object you want to work with. On most hosts, you can select an object by pointing to it and clicking the leftmost button on the mouse. To *click*, press the button down and release it without moving the mouse. Figure 1.2 illustrates the selection of a variable by pointing and clicking.

Figure 1.2. Selecting by Clicking

You can select multiple objects by *dragging* the mouse. To drag, press the leftmost mouse button down, move the mouse across the objects of interest, then release the mouse button. This selects the object at the cursor location when you pressed the mouse button, the object where you released the button, and all objects in between. Figure 1.3 illustrates the selection of three variables by pointing and dragging.

8

Figure 1.3. Selecting by Dragging

When objects are far apart, it is convenient to use *modifier keys* with the mouse button. On many hosts, you can use the **Shift** key to *extend* a selection. In Figure 1.4, the first observation was clicked on, then the one hundredth observation was clicked on while holding down the **Shift** key. This selects the first observation, the one hundredth observation, and all observations in between.

Figure 1.4. Extended Selection

On many hosts, you can use the **Ctrl** key to make a *noncontiguous* selection – that is, a selection of multiple objects not located next to each other. In Figure 1.5, the first observation was clicked on, then the fifth observation was clicked on while holding down the **Ctrl** key. This selects the first observation and the fifth observation without selecting the observations in between.

9

Figure 1.5. Noncontiguous Selection

Some hosts use different modifier keys instead of the **Shift** and **Ctrl** keys, so these names do not appear in the remainder of this book. Instead, the terms *extended selection* and *noncontiguous selection* are used. Using single, multiple, extended, and noncontiguous selection, you can precisely indicate the objects you want to work with.

Choosing from Menus

In SAS/INSIGHT software, operations you can perform include creating graphs and analyses, transforming variables, fitting curves, and saving results. On most hosts, you can choose these operations by *pulling down* a menu from a *menu bar*. To pull down a menu, press the left mouse button and hold it down while you drag the cursor across the menu. Figure 1.6 shows the **Analyze** menu pulled down to create a scatter plot.
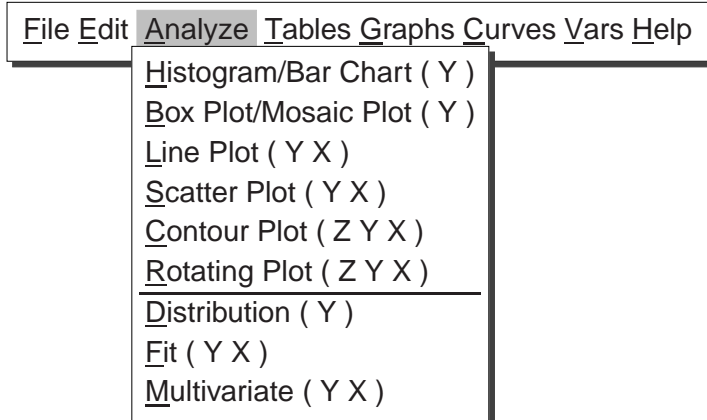

Figure 1.6. **Analyze** Menu

Depending on your host, each window may display its own menu bar or all windows may share a single menu bar. Workstations with large displays usually provide multiple menu bars. Personal computers with small displays may allow only one menu bar.

10

Your host may provide additional choices on the menu bar and within the **File** and **Help** menus. These additional menu choices, if present, are described in the SAS companion for your host.

Pop-up Menus

Pop-up menus enable fast action by providing choices appropriate for the object you point to. Pop-up menus operate on all appropriate selected objects. If no objects are selected, they operate on the object at the cursor location.

Pop-up menus are displayed when you click on *menu buttons* in the data window and in the corners of graphs and tables. On some hosts, you can also display pop-up menus by pressing the right mouse button.

The data window displays a variety of pop-up menus. To display the pop-up menu for data, either click the left mouse button in the upper left corner, as in Figure 1.7, or click and hold the right mouse button anywhere in the data window. See Chapter 31, "Data Windows," for a complete description of the pop-up menu choices in the data window.
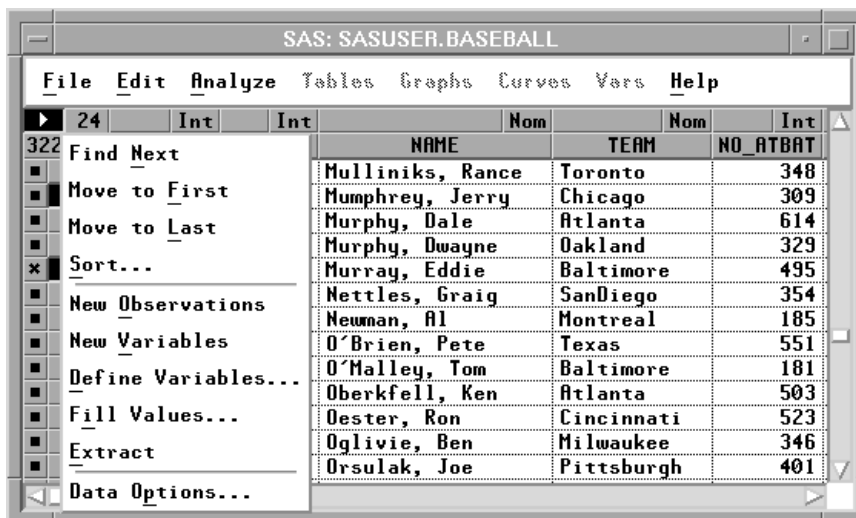
Figure 1.7. Data Pop-up Menu

To display pop-up menus in a graph or table, either click and hold the right mouse button anywhere in the graph or table, or click on the menu button in the corner of the graph or table. Figure 1.8 shows the pop-up menu for a histogram in a distribution analysis.

11
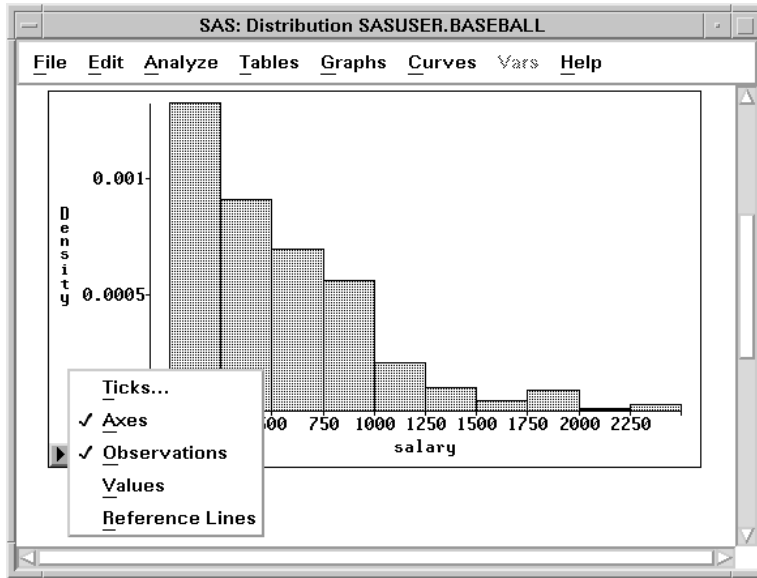
Figure 1.8. Histogram Pop-up Menu

When you are not pointing at a table, graph, or other object, the right mouse button displays the central menu bar, as in Figure 1.9. For more information on pop-up menu choices, see the chapter for the graph or table of interest in the Reference part of this manual.
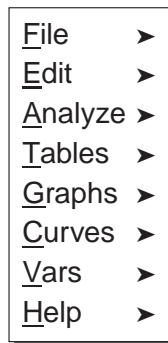
Figure 1.9. Default Pop-up Menu

Menu State Indicators

Menu *state indicators* are either check marks or radio marks. The graphic representation of these marks depends on your host.

Menus with *check marks* always act as *toggles*: they turn a feature on or off. The presence of a check mark indicates the presence of that feature. Toggles are especially useful in graphs, since most graphic features are either on or off.

Menus with *radio marks* do not toggle; they indicate the current state among multiple choices. As with check marks, radio marks help when the current state is not obvious.

For example, the pop-up menu in Figure 1.10 is from a scatter plot. The check marks indicate that axes and observations are displayed and that the marker size is chosen automatically to fit the graph. The radio mark indicates that the current marker size is 4.

```
 ┌──────────────────────────┐
 │  Ticks...                │
 │  ✔ Axes                  │
 │  ✔ Observations          │
 │  Reference Lines         │
 │  Marker Sizes      ➤  ┌──────────────┐
 └───────────────────────│      1       │
                         │      2       │
                         │      3       │
                       • │      4       │
                         │      5       │
                         │      6       │
                         │      7       │
                         │      8       │
                         ├──────────────┤
                         │  ✔ Size to Fit│
                         └──────────────┘
```

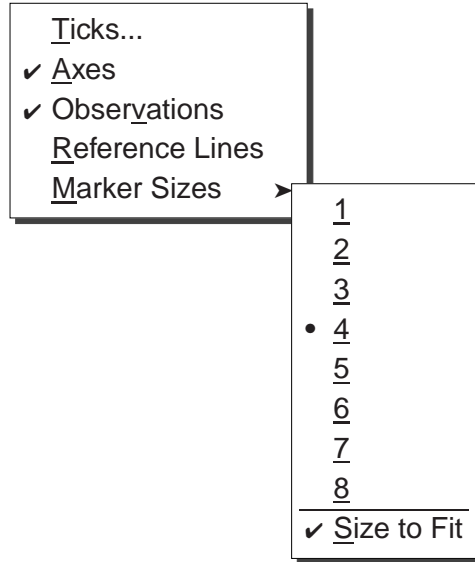**Figure 1.10.**   Scatter Plot Pop-up Menu

Learning More

Using This Manual

The remainder of this manual is divided into two parts: Techniques and Reference.

Techniques are instructional chapters that explain how to accomplish particular tasks. These chapters use sample data sets shipped with the product, so you can read the techniques and follow the steps on your host at the same time. For more information about sample data sets, see the "Sample Data Sets" section in this chapter.

Reference chapters provide comprehensive descriptions of data, graphs, and analyses in SAS/INSIGHT software. Use these chapters to answer specific questions about product features.

If you are experienced with SAS/INSIGHT software or experienced using mice and menus, you may learn most quickly by just invoking SAS/INSIGHT software and exploring its capabilities. Use the Table of Contents and the Index to find specific techniques and reference information.

Conventions

This user's guide employs three special symbols:

$\Longrightarrow$ **This symbol and font marks one step in a technique.**

$\oplus$ **Related Reading:** This symbol and label marks a reference to a related chapter.

† **Note:** This symbol and label marks an important note or performance tip.

This user's guide employs four special typefaces:

- **Bold** is used for steps in techniques.
- *Italic* is used for definitions and for emphasis.
- **Helvetica** is used for words you see on the display.
- `Courier` is used for examples of SAS statements.

Menu items in this user's guide are separated by colons. For example, the **Bar Chart (Y)** item in the **Analyze** menu is written as **Analyze:Bar Chart (Y)**.

Getting Help

Both beginning and expert users can take advantage of SAS/INSIGHT software's context-sensitive help system. To receive context-sensitive help, select any graph or table by clicking on its border. Then choose **Help:Help on Selection**, as illustrated in Figure 1.11. Figure 1.12 shows the context-sensitive help when the Quantiles table is selected.
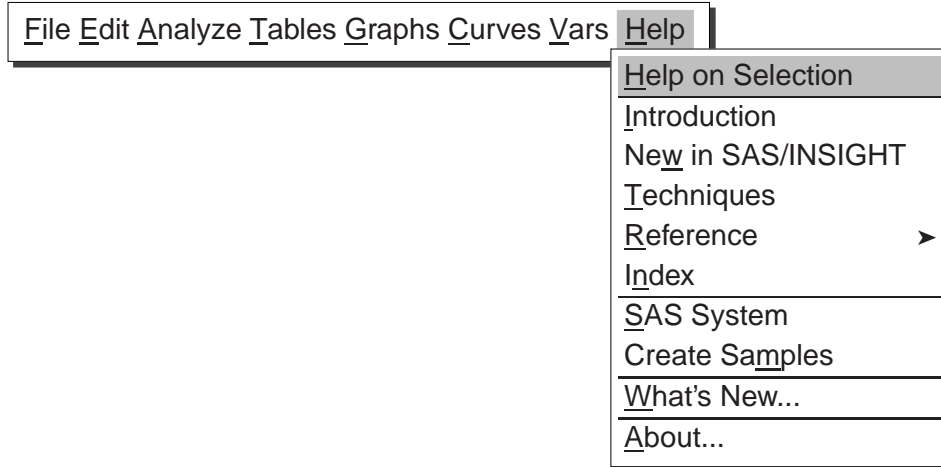
	File Edit Analyze Tables Graphs Curves Vars	Help

<u>H</u>elp on Selection
<u>I</u>ntroduction
Ne<u>w</u> in SAS/INSIGHT
<u>T</u>echniques
<u>R</u>eference ➤
I<u>n</u>dex
<u>S</u>AS System
Create Sa<u>m</u>ples
<u>W</u>hat's New...
<u>A</u>bout...

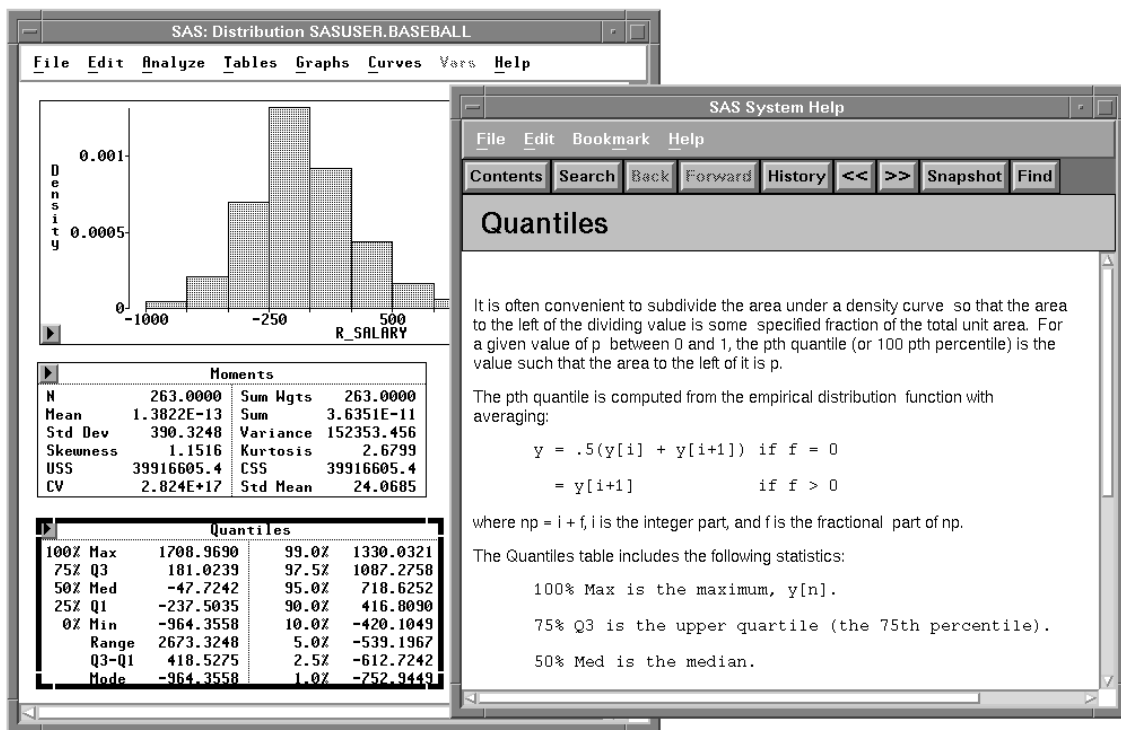**Figure 1.11.** **Help** Menu

Figure 1.12. Context Sensitive Help

You can also get context-sensitive help with the SAS System *Help key*. This key, usually **F1** on your keyboard, displays help on the object at your present cursor position. You can get context-sensitive help in any SAS/INSIGHT data or analysis window by simply placing the cursor on the item of interest and pressing the **Help** key. Within any help window, you can point and click on individual topics to get further information.

15

The **Help** menu entries correspond to parts of this manual. Choose **Help:Introduction** to learn about SAS/INSIGHT software; **Help:Techniques** to learn how to perform a particular task; **Help:Reference** to look up detailed information; or **Help:Index** to see an index of all SAS/INSIGHT topics.
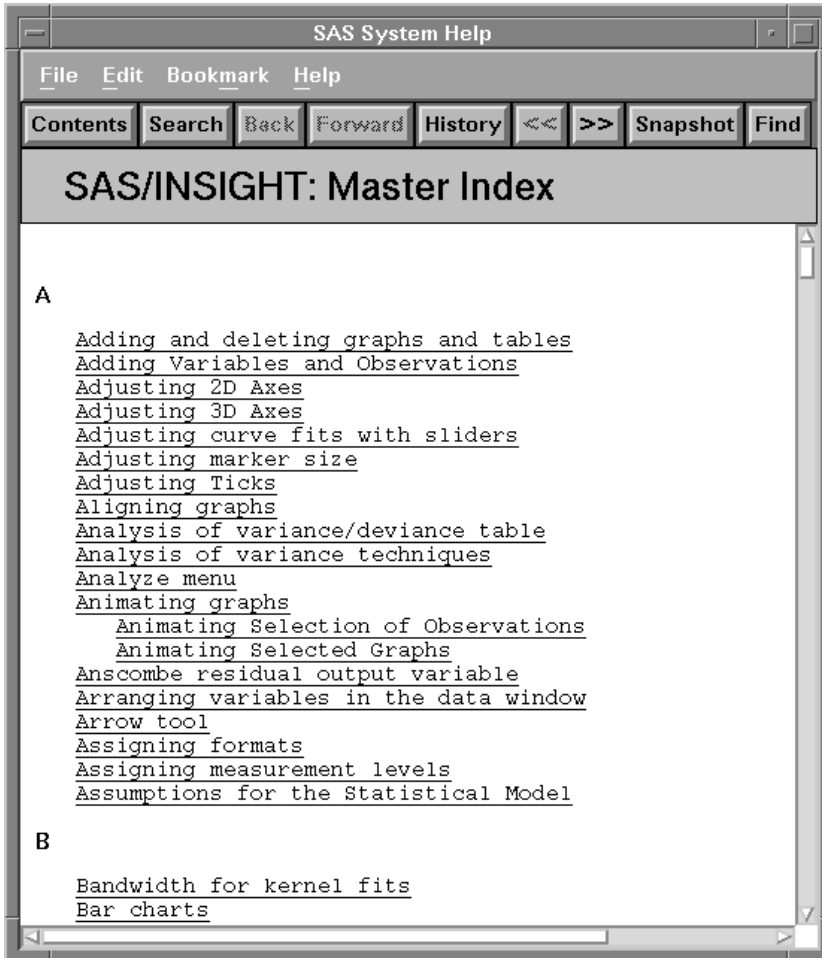

Figure 1.13. Help Index

Choose **Help:SAS System** to see a general index of SAS System topics. Choose **Help:Create Samples** to create sample data sets; examples throughout this manual refer to these data sets. See the following section for more information.

Sample Data Sets

The following sample data sets are included with SAS/INSIGHT software.

The **AIR** data set contains measurements of pollutant concentrations from a city in Germany during a week in November 1989. Variables are

DATETIME	date and hour in SAS DATETIME format
DAY	day of the week
HOUR	hour of the day
CO	carbon monoxide concentration
O3	ozone concentration
SO2	sulfur dioxide concentration
NO	nitrogen oxide concentration
DUST	dust concentration
WIND	wind speed

The **BASEBALL** data set contains performance measures and salary levels for regular hitters and leading substitute hitters in major league baseball for the year 1986 (Collier 1987). There is one observation per hitter. Variables are

NAME	the player's name
NO_ATBAT	number of times at bat in 1986
NO_HITS	number of hits in 1986
NO_HOME	number of home runs in 1986
NO_RUNS	number of runs in 1986
NO_RBI	number of runs batted in in 1986
NO_BB	number of bases on balls in 1986
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 1986
DIVISION	player's division at the end of 1986

17

TEAM	player's team at the end of 1986
POSITION	positions played in 1986
NO_OUTS	number of put outs in 1986
NO_ASSTS	number of assists in 1986
NO_ERROR	number of errors in 1986
SALARY	salary in thousands of dollars

The **POSITION** variable in the **BASEBALL** data set is encoded as follows:

13	first base, third base	CS	center field, shortstop
1B	first base	DH	designated hitter
1O	first base, outfield	DO	designated hitter, outfield
23	second base, third base	LF	left field
2B	second base	O1	outfield, first base
2S	second base, shortstop	OD	outfield, designated hitter
32	third base, second base	OF	outfield
3B	third base	OS	outfield, shortstop
3O	third base, outfield	RF	right field
3S	third base, shortstop	S3	shortstop, third base
C	catcher	SS	shortstop
CD	center field, designated hitter	UT	utility
CF	center field		

The **BUSINESS** data set contains information on publicly-held German, Japanese, and U.S. companies in the automotive, chemical, electronics, and oil refining industries. There is one observation for each company. Variables are

NATION	the nationality of the company
INDUSTRY	the company's principal business
EMPLOYS	the number of employees
SALES	sales for 1991 in millions of dollars
PROFITS	profits for 1991 in millions of dollars

The **DRUG** data set contains results of an experiment to evaluate drug effectiveness (Afifi and Azen 1972). Four drugs were tested against three diseases on six subjects; there is one observation for each test. Variables are

DRUG	the drug used in treatment
DISEASE	the disease present
CHANG_BP	the change in systolic blood pressure due to treatment

The **GPA** data set contains data collected to determine which applicants at a large midwestern university were likely to succeed in its computer science program (Campbell and McCabe 1984). There is one observation per student. Variables are

18

GPA	the grade point average of students in the computer science program
HSM	the average high school grade in mathematics
HSE	the average high school grade in English
HSS	the average high school grade in science
SATM	the score on the mathematics portion of the SAT exam
SATV	the score on the verbal portion of the SAT exam
SEX	the student's gender

The **IRIS** data set is Fisher's Iris data (Fisher 1936). Sepal and petal size were measured for fifty specimens from each of three species of iris. There is one observation per specimen. Variables are

SEPALLEN	sepal length in millimeters
SEPALWID	sepal width in millimeters
PETALLEN	petal length in millimeters
PETALWID	petal width in millimeters
SPECIES	the species

The **MINING** data set contains results of an experiment to determine whether drilling time was faster for wet drilling or dry drilling (Penner and Watts 1991). Tests were replicated three times for each method at different test holes. There is one observation per five-foot interval for each replication. Variables are

DRILTIME	the time in minutes to drill the last five feet of the current depth
METHOD	the drilling method, wet or dry
REP	the replicate number
DEPTH	the depth of the hole in feet

The **MININGX** data set is a subset of the **MINING** data set. It contains data from only one of the test holes.

The **PATIENT** data set contains data collected on cancer patients (Lee 1974). There is one observation per patient. Variables are

REMISS	1 if remission occurred and 0 otherwise
CELL	
SMEAR	
INFIL	
LI	
TEMP	
BLAST	measures of patient characteristics

The **SHIP** data set contains data from an investigation of wave damage to cargo ships (McCullagh and Nelder 1989). The purpose of the investigation was to set standards for future hull construction. There is one observation per ship. Variables are

Y	the number of damage incidents
YEAR	year of construction
TYPE	the type of ship
PERIOD	the period of operation
MONTHS	the aggregate months of service

Choose **Help:Create Samples** to create the sample data sets in your **sasuser** directory. When you have created the sample data sets, turn to the Techniques part of this manual to learn how to enter your data and begin exploring it with SAS/INSIGHT software.

† **Note:** If you have an existing data set in your **sasuser** library with the same name as a sample data set, it will be overwritten if you create the sample.

20

References

Afifi, A.A. and Azen, S.P. (1972), *Statistical Analysis: A Computer-Oriented Approach*, New York: Academic Press, 166.

Campbell, P.F. and McCabe, G.P. (1984), "Predicting the Success of Freshmen in a Computer Science Major," *Communications of the ACM*, 27, 1108–1113.

Collier Books (1987), *The 1987 Baseball Encyclopedia Update*, New York: Macmillan Publishing Company.

Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Lee, E.T. (1974), "A Computer Program for Linear Logistic Regression Analysis," *Computer Programs in Biomedicine*, 80–92.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, London: Chapman and Hall.

Penner, R. and Watts, D.G. (1991), "Mining Information," *American Statistician*, 45(1), 4–9.