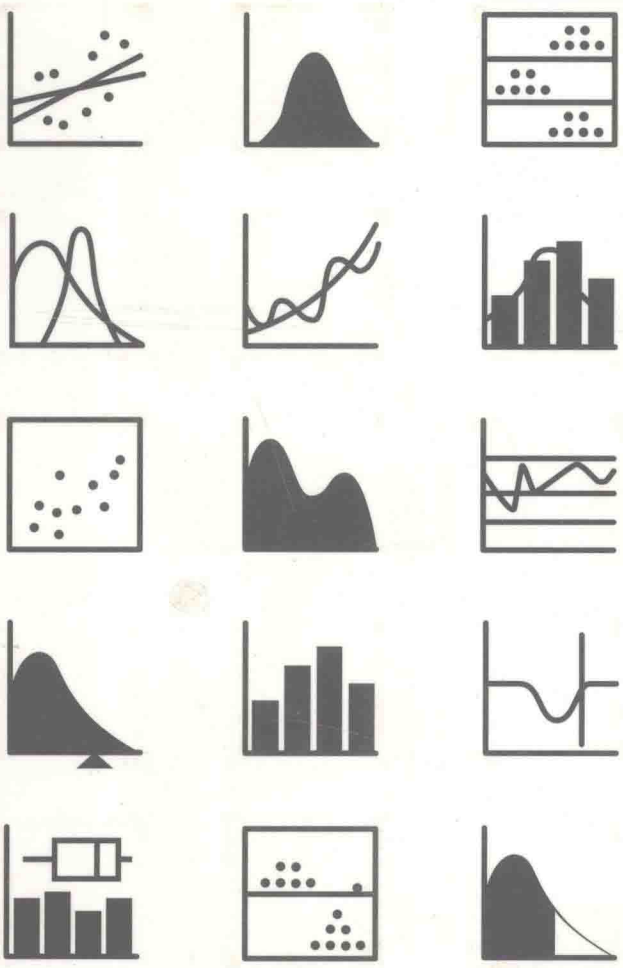
VISUAL STATISTICS

Part I



Doane • Mathieson • Tracy

VISUAL STATISTICS

Part I

David P. Doane
Kieran Mathieson
Ronald L. Tracy
Oakland University

Irwin McGraw-Hill

Boston, Massachusetts Burr Ridge, Illinois Dubuque, Iowa
Madison, Wisconsin New York, New York San Francisco, California St. Louis, Missouri

McGraw-Hill
A Division of The McGraw-Hill Companies

VISUAL STATISTICS PART I

This book is printed on acid-free paper.

1 2 3 4 5 7 8 9 0 MAL/MAL 9 0 9 8 7

ISBN 0–256–20876–X (student edition)
ISBN 0–256–20877–8 (instructor's edition)
ISBN 0–07–292467–5 (student edition manual)

Publisher: *Tom Casson*
Executive editor: *Richard T. Hercher, Jr.*
Marketing manager: *Kurt Strand*
Senior project manager: *Susan Trentacosti*
Production supervisor: *Heather Burbridge*
Senior designer: *Crispin Prebys*
Printer: *Malloy Lithographing, Inc.*

Library of Congress Cataloging-in-Publication Data

Doane, David P.
Visual statistics / David P. Doane, Kieran Mathieson, Ronald L. Tracy.
p. cm.
ISBN 0–256–20876–X (student edition);
ISBN 0–256–20877–8 (instructor's edition)
ISBN 0–07–292467–5 (student edition manual)
1. Statistics--Computer-assisted instruction. I. Mathieson, Kieran II. Tracy, Ronald L. III. Title
QA276.18.D63 1997
519.5'078—dc20 96–35867

http://www.mhcollege.com

To Blythe, my skylark.
–David

To Karen, for all she gives.
–Kieran

To Kathy, with my love.
–Ron

Preface

This textbook began in 1992 when Tracy walked into Doane's office to see if he would be interested in writing a National Science Foundation (NSF) grant proposal to develop software to teach statistical concepts. The collaboration was natural. Each had been teaching for many years, and had evolved definite approaches to teaching statistics. Both had experience writing computer software.

During that fall semester our ideas began to take shape. We believed it was time to bring the computer into the *classroom*. Although students had been doing computer lab projects for decades, the classroom had remained basically unaffected by the computer. We wanted to use computers to teach *concepts* rather than just analyze data. We wanted to stress visual displays rather than numbers, equations, and calculations. While an equation is an explanation to some students, it is a barrier to others. Understanding the full implications of an equation is not a trivial matter for most people, but a series of pictures or a simple animation can illustrate fairly complex ideas in a few minutes. Our proposal asked the NSF to support our efforts to bring the power of computers into the statistics classroom.

In the spring of 1993 we were awarded the NSF grant. Since neither of us had experience in writing programs for the Windows environment, we went looking for an expert. Luckily for us, Oakland University had recently hired such a person. We approached Mathieson who agreed to help us as a consultant. By July it was clear that Mathieson had become an integral part of the project. This marked the beginning of our collaborative efforts.

Our approach is based on four assumptions: (1) For many students visualization is a key to learning; (2) existing statistical packages were designed for data analysis, not visualizing concepts; (3) today's computers have enough power to support visualization; and (4) individual instructors lack the time and resources to design their own software. Our goal was to create learning modules which illustrate concepts that cannot easily be shown mathematically or with packages such as Minitab or Excel. We sought to minimize duplication of capabilities already available in data analysis software.

We wanted to promote learning through active self-discovery (inductive learning) as well as to support the instructor in the traditional lecture/lab setting (deductive learning). A student using a software module with the Learning Exercises in this worktext will be led through experiments that promote self-discovery. Team and Individual Learning Projects pose less-structured problems that continue this process. The software is designed to encourage experimentation. A student can use replication to simulate the "experimental" side of statistics.

These same features can be used effectively by the instructor in the classroom or the computer laboratory. The instructor can use a software module to illustrate a concept (e.g., the relationship between a confidence interval and a test statistic). The concept can then be reinforced with a simulation experiment.

We wanted flexible software. Instructors can omit topics or vary their order. No module requires knowledge of another module. Students can learn in different ways and at different rates. Although each module is aimed at the average learner, each was designed to offer options for those who desire more analytical depth (or who are just adventuresome). We designed modules that allow you to set parameters in several ways. For example, you can manipulate the parameters of some demonstrations using scroll bars (simple level), by typing parameter values (intermediate level), or by choosing a known distribution and its parameters (advanced level).

Every *Visual Statistics* module opens with a virtual Notebook (a familiar ring binder with pages you can turn). The Notebook organizes the options and provides a link with the *Visual Statistics*

worktext and your textbook. The Notebook contains tabs that you can click. These tabs divide the Notebook into different sections. Notebook tabs include Introduction (general idea of what is in the module), Concepts (main topics covered), and More Information tab (references to chapters or other textbooks). Other tabs vary from module to module, such as Examples (real data on a variety of topics), Scenarios (context for realistic sampling simulations), Databases (real data sets), Templates (icons to choose specified distributions or trends), Do-It-Yourself (control panels to manipulate parameters of an experiment), and Data Editor (to enter data or paste data from a spreadsheet to create your own example). You may return to the Notebook while running the program. This allows you to change scenarios, pick a different example, or choose a different set of options.

Every module has its own Help system, similar to that in other Windows programs. It includes definitions, examples, equations, graphics, and hints. You can use the table of contents and use the hypertext capabilities to jump to topics you want to learn more about. You can also do keyword searches using each Help file's extensive index. You will get more from *Visual Statistics* if you learn to use the Help system.

Our software uses familiar Windows visual controls (command buttons, scroll bars, list boxes, option buttons, and so on). Input is mostly from a mouse rather than from the keyboard. This interface will seem quite natural to most people. If you have used other Windows programs such as Word, Excel, or even Solitaire, you already know how to use *Visual Statistics* controls.

Visual Statistics is to be used with a textbook or other course materials. It complements these traditional sources, it does not replace them. However, you'll find topics in these sources easier to understand after you *see* them in *Visual Statistics*.

Visual Statistics consists of 15 software modules and a worktext. Each chapter of the worktext begins with a list of the key concepts the chapter will cover and a short list of learning objectives. Each concept is reviewed and illustrated. A short (about 15 minute) orientation to each module is provided. Past users have told us that this orientation is invaluable. Learning Exercises and Advanced Learning Exercises are provided for each module. These exercises guide you in self-discovery learning. Team and Individual Learning Projects continue this process with open-ended questions. A Self-Evaluation Quiz allows you to test your understanding of the material. Solutions and a guide to answering each question are provided. A Glossary at the end of each chapter provides definitions of terms used in the chapter.

Although we take responsibility for all remaining errors, our project has benefited greatly from field testing, advice, and comments of faculty colleagues and their students at other universities. We thank these individuals for their dedication and interest in our project.

Richard Anderson-Sprecher—University of Wyoming

Mohamed Askalani—Mankato State University

David Booth—Kent State University

Tim Butler—Wayne State University

Patty Collings—Brigham Young University

Peyton Cook—University of Tulsa

Rick Edgeman—Colorado State University

Chris Franklin—University of Georgia

Bob Hordon—Rutgers University

William Jedlicka—William Rainey Harper

Anthony Keys—Wichita State University

Anne B. Koehler—Miami University of Ohio

Gerald Kohers—Sam Houston State University

Benny Lo—Northwest Polytechnic University

Gary Martin—DeVry University - Atlanta

Madhuri Mulekar—University of South Alabama

Sufi Nazem—University of Nebraska - Omaha

Ceyhun Osgur—Valparaiso University

Paul Paschke—Oregon State University

Andy Siegel—University of Washington

Boyd Swartz—Monmouth University

Mack Shelley—Iowa State University

Robert D. St Louis—Arizona State University

Bret Wagner—University of Delaware

Linda Young—University of Nebraska - Lincoln

This project was supported, in part, by the
National Science Foundation
Opinions expressed are those of the authors and not necessarily those of the Foundation

This project would have never begun without the initial support of the National Science Foundation, Department of Undergraduate Education (DUE #9254182). For their belief in our concept, we are grateful.

However, without the support of Irwin/McGraw-Hill, the project might have died with the six modules the NSF supported. Much of this corporate support was due to our executive editor, Richard Hercher. He was an advocate of our project after seeing early versions of our first four modules. As we struggled to meet deadlines, he encouraged and cajoled us. We are indebted to Susan Trentacosti for her able assistance in managing our project through the editorial and production process. We are grateful for the help of Alisa Watson, Marc Mattson, Kurt Strand, Steve Schuetz, and the entire staff of Irwin/McGraw Hill. We also acknowledge the programming assistance of Steve Losey; software testing of Judith Gurney; database research by Cindy Hollifield, Douglas Dotterweich, and Karen Helber; advice on SQC by T.J. Wharton; and numerous comments by Oakland University students.

David P. Doane
Kieran Mathieson
Ronald L. Tracy

Contents

Introduction

Using *Visual Statistics*

OBJECTIVES

- Understand what *Visual Statistics* is all about
- Learn how to install *Visual Statistics*
- Learn how to use the Notebook, the starting point for every *Visual Statistics* module
- Learn about the worktext

What is *Visual Statistics*?

Statistics is a challenging subject for most people. Most of us have to work hard to understand probability distributions, statistical power, and confidence intervals. However, statistics is becoming more important. It's hard to design a product, create an advertising campaign, or play poker well without understanding statistics.

The main problem is that statistical ideas are abstract. When did you last speak with a probability distribution? Have you ever seen a correlation on the shelf at Sears? Or been almost run down by a wild confidence interval (except on an exam)? Formulas and numbers are fine for some people, but most of us learn best when we can see and interact with the thing we're learning about.

That's where *Visual Statistics* comes in. Computers have become more powerful and easier to use, so it's possible to explore statistical ideas in new ways. Graphics and animation can bring concepts to life in ways that formulas, tables, and static diagrams cannot. Rather than imagining what data for a particular situation looks like, *Visual Statistics* lets you create the situation and see it for yourself. You can change the circumstances and watch the result. *Visual Statistics* helps you build your statistical intuition, making it easier to understand the ideas behind data analysis.

It's important to understand that while you can analyze data with some *Visual Statistics* modules, that is not their main goal. Instead, *Visual Statistics* will help you *learn statistics*. Even the most powerful analysis tools, like Minitab, SAS, and SPSS, won't help you if you don't know what to do with them. On the other hand, if you are familiar with statistical ideas, you can do wonders with a simple calculator. The tools you use are less important than how you wield them. *Visual Statistics* will help you use statistical techniques more effectively because you will understand what they can be used for and what their limitations are.

Visual Statistics is not a tutorial, however. For example, the regression module won't explain why regression works or how to calculate regression statistics. Instead, it helps you explore regression, so you understand it better than you would without the visual tools. You'll get more value from your statistics text since you'll be able to experiment with the concepts. It'll be easier for you to *understand* statistics, and with that understanding comes more confidence and less worry.

The *Visual Statistics* suite consists of 15 modules. Each one helps you explore a specific set of statistical ideas. For example, the ANOVA module will help you understand one-way analysis of variance. It is designed for that purpose alone. It's different from the module that helps you understand statistical quality control, since the two topics are different. However, every module shares some characteristics. First, they are all easy to use. Run through the 15-minute orientation for each module, and you should be able to use it effectively. If you have trouble, select Help for assistance. Second, they all use the Notebook. The Notebook introduces the module, and lets you choose how you want to explore the relevant statistical ideas. Different modules have different Notebook options. Some let you examine scenarios, while others allow you to analyze real data sets or enter your own data. The Notebook is described in more detail later in this introduction.

The best way to learn what *Visual Statistics* is all about is to try it. So go ahead. Install *Visual Statistics,* start the Univariate Data Analysis module and play around. You'll see how the same data can look very different, depending on how it is displayed.

Installing *Visual Statistics*

Visual Statistics runs under Windows 3.1 or up, including Windows for Workgroups and Windows 95. You need at least a 33 MHz 486 PC with 8 Mb of memory; however, *Visual Statistics* will run best on a Pentium machine with 16 Mb of memory. *Visual Statistics* takes about 20 Mb of hard disk space. This includes 15 modules and their help files, over 250 data sets, and miscellaneous bits and pieces.

To install *Visual Statistics,* first close down any other programs you are running. Place disk 1 in your 3½" floppy drive (usually A:). Under Windows 95, click the task bar's Start button and select Run.... Under Windows 3.1 and Windows for Workgroups, select File on the Program Manager's menu bar and then select Run. In either case, type a:setup.exe and press Enter. If your floppy drive is not A:, substitute its letter. For example, if it was B:, you would type b:setup.exe. Then simply follow the instructions.

During setup you will be asked where you want to install *Visual Statistics*. You should normally use the defaults. You may be asked to switch to an earlier disk at some point (e.g., to remove disk 2 and insert disk 1). If you have not installed *Visual Statistics* before, you will be asked to type in your name and organization (school, university, etc.). This information will be displayed every time a *Visual Statistics* module is started.

You can remove *Visual Statistics* from your PC (although why would you want to?) using the uninstaller. Under Windows 95, double click the Add/Remove Programs icon in the Control Panel. Select *Visual Statistics* and click the Add/Remove button. Under Windows 3.1 and Windows for Workgroups, double click the icon for the *Visual Statistics* uninstall program in the *Visual Statistics* program group in the Program Manager.

The *Visual Statistics* Notebook

Every *Visual Statistics* module starts with the Notebook. There is a different Notebook for each module, but they all work in the same way. The main purpose of the Notebook is to let you choose the type of data you want to look at. Figure 1 shows the opening page of the Notebook for the Bivariate Data Analysis module.

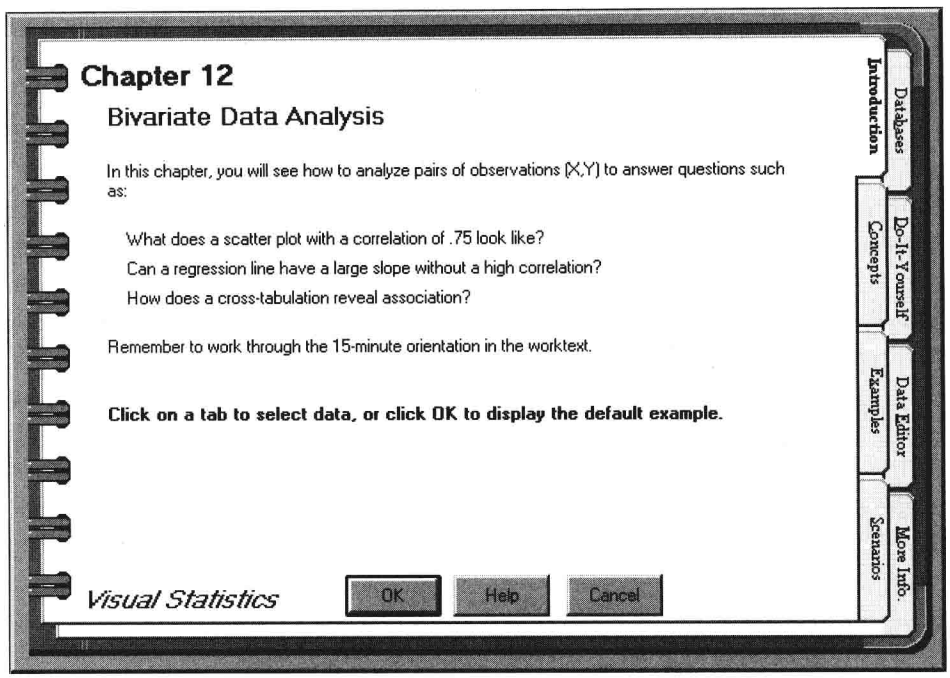


Figure 1: Notebook Opening Page

Navigating through the Notebook

The Notebook is divided into sections, each with its own yellow tab. Each section has one or more pages. To select a section, click on its tab. For instance, to select an example in the Bivariate Data Analysis module, click on the Examples tab. You would see the display shown in Figure 2.

Notice there are several different example categories (Sports, Consumer, etc.). Each page in the section gives examples from a different category. You can pick a category in two different ways. First, you can jump directly to the category's page by clicking on its name. The names are in green and underlined, showing they are hyperlinks. Just as on the Web or in a Windows Help file, clicking on a hyperlink jumps to a different page. Second, you can turn the pages in a section, just like turning the pages in a book. You can move forward and backward through each section by clicking on Previous page or Next page. You can also click on the "folded corners" of the page, as illustrated in the upper and lower right of Figure 3.

Each module has sections titled Introduction, Concepts, and More Information. The Introduction section identifies the module and asks some questions the module will help you answer. The Concepts section lists some ideas the chapter covers. The More Information section identifies related modules, and lists relevant chapters in Irwin/McGraw-Hill textbooks.

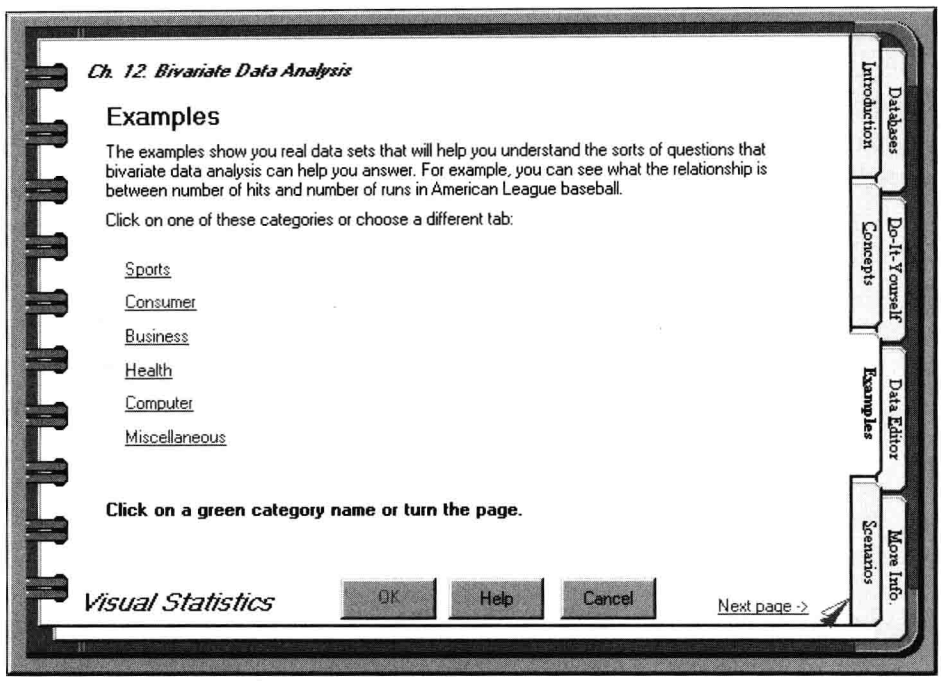


Figure 2: Categories of Examples

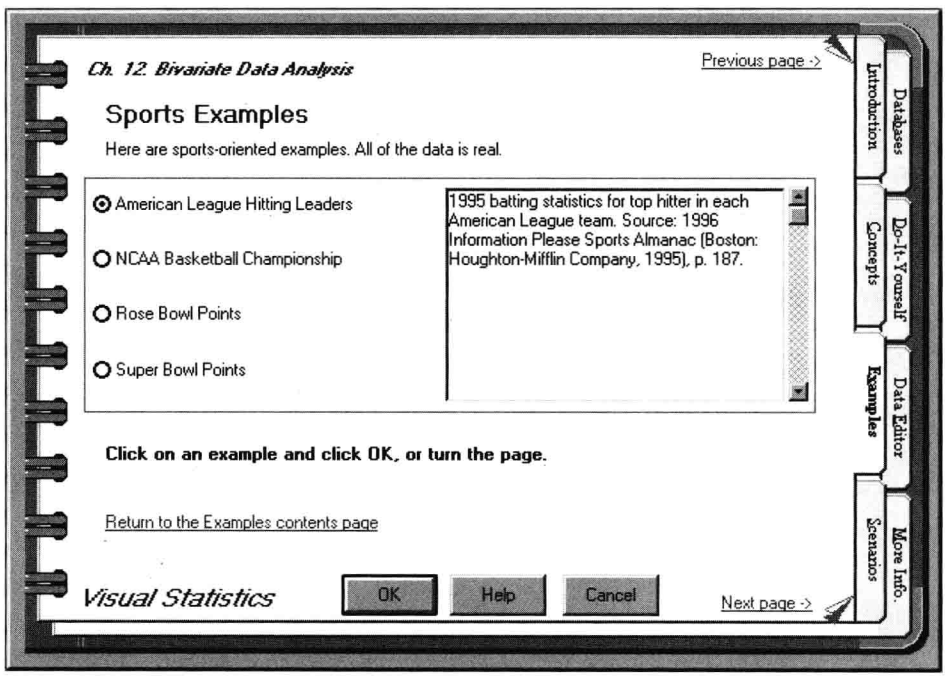


Figure 3: Examples in the Sports Category

Exiting the Notebook

Each page has OK, Help, and Cancel buttons. The OK button closes the Notebook and returns to the module. It's only available when you have selected some data for the module to use. If you click the OK button on the Introduction page when the module first starts, some default data will be chosen for you. The Help button gives you assistance in using the Notebook. The Cancel button closes the Notebook immediately.

Selecting Data Sets

The other sections let you select data to explore. The modules have different sections, depending on the topic. For example, the Bivariate Data Analysis module has a Databases section, but the Continuous Distributions module does not.

The Examples section lists data sets you can examine. It's used in the Univariate Data Analysis and Bivariate Data Analysis modules. The data sets are all real. For instance, Figure 3 shows various sports-related examples. Click on an example's name to select it. You can read a description of the selected example in the yellow area to the right of the example list.

The Databases section is available in the Univariate Data Analysis, Bivariate Data Analysis and Samples modules. It lets you choose from a large number of related data sets supplied with *Visual Statistics*. The data sets are grouped into categories. Figure 4 shows a database page from the Bivariate Data Analysis module. Note the categories Demographics, Economics, Crime, and so on. Click on the + and – signs to expand or collapse the category lists. The yellow area to the right of the data set list describes the selected data set. You select two data sets per analysis in the Bivariate module, but only one in the other modules.

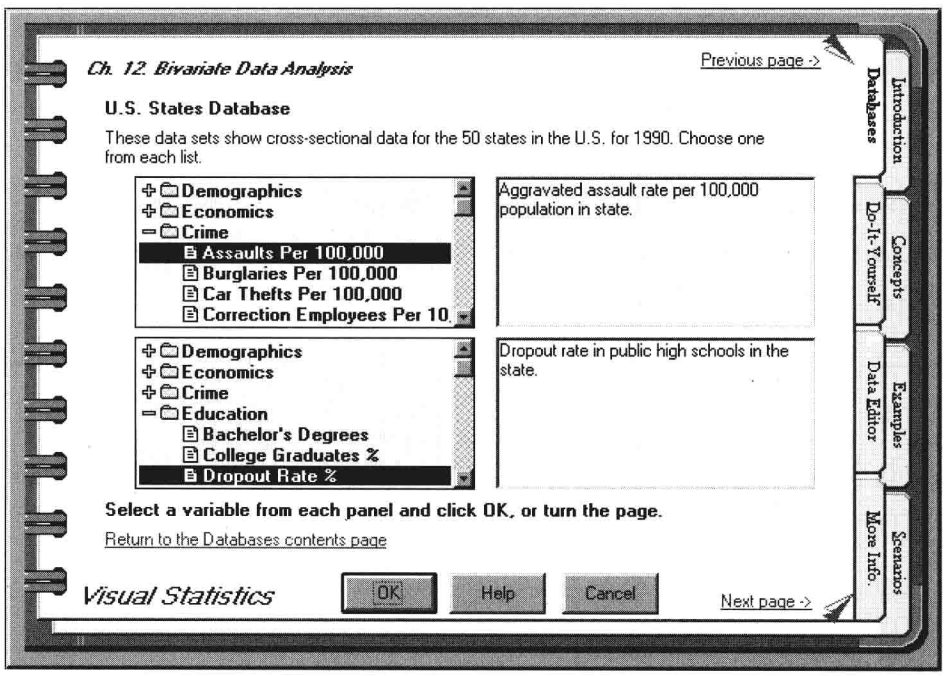


Figure 4: Database Page

The Data Editor

The Data Editor lets you create your own data sets. Figure 5 shows an example from the Bivariate Data Analysis module. The spreadsheet is similar to Excel or Lotus 1-2-3. Each row is one observation (a person, a team, etc.). The first column is a label for the observation (a person's name, a team's name, etc.). It's optional; you don't have to label every point. The other columns are the variables you want to analyze. Use the mouse or arrow keys to select a cell, then type data into it. You can rename each column by typing into the cell above the column (e.g., the cell with "Height" in it). When you are done typing your data, press the Plot button to exit the Data Editor. If you don't want to keep your changes, select Exit and Discard Changes from the File menu.

The Data Editor has other useful options. Use the File menu to save your data to disk and retrieve it later. If you're running *Visual Statistics* on a network, make sure you save your data to a disk you can write to (e.g., a floppy). Use the Edit menu to insert and delete rows, and to move data between programs. You can copy data from the Data Editor and paste it into other Windows applications such as Excel, 1-2-3, Word, or WordPerfect. You can also do the reverse—that is, paste data from another program into the Data Editor. Use the Options menu to sort the data, change its display format, and enter a title for the data set. Select Help in the Data Editor for more information. The Data Editor is available in the Univariate Data Analysis and Bivariate Data Analysis modules.

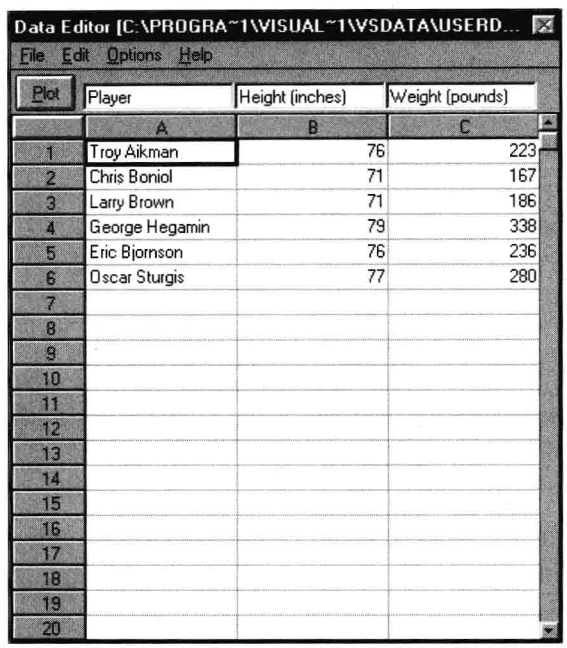


Figure 5: The Data Editor

Scenarios, Templates, and Do-It-Yourself

The Examples, Databases, and Data Editor sections are similar in one respect: They let you examine real data sets. The last three sections are different. They let you experiment with the *process* that generates data sets. For example, suppose you rolled a normal six-sided die twice and got a 1 and a 3. The "1 and 3" is the data set. The "rolled a normal six-sided die twice" is the process that created the data set. If you run the same process again (that is, roll the die twice), you would probably get a different data set. Roll the die twice more, and you get yet another data set. You can ask many questions about this situation. Would the data be the same each time? How different would it be? Suppose you added the numbers together. What is the highest sum you can get? What is the smallest? Are you more likely to get the smallest sum or the largest sum? Are all sums equally likely, or are some values more common than others?

You can learn a lot about statistics by examining these questions. But you can learn even more by varying the process that creates the data. For example, what if you rolled a 10-sided die, as used in *Dungeons and Dragons*? What if you spun a roulette wheel twice? Would the answers be the same?

This is where the Scenarios, Templates, and Do-It-Yourself sections come in. The Scenarios let you choose from processes that have been set up for you. Each scenario generates data that you might see in a particular situation. For instance, one scenario looks at the relationship between the number of times at bat and home runs in major league baseball. It gives you a familiar context for statistical exploration. You can take repeated samples, and see how random chance leads to very different data from identical

situations. Figure 6 shows one of the scenario pages from the Bivariate Data Analysis module. Every module except Univariate Data Analysis has a Scenario section.

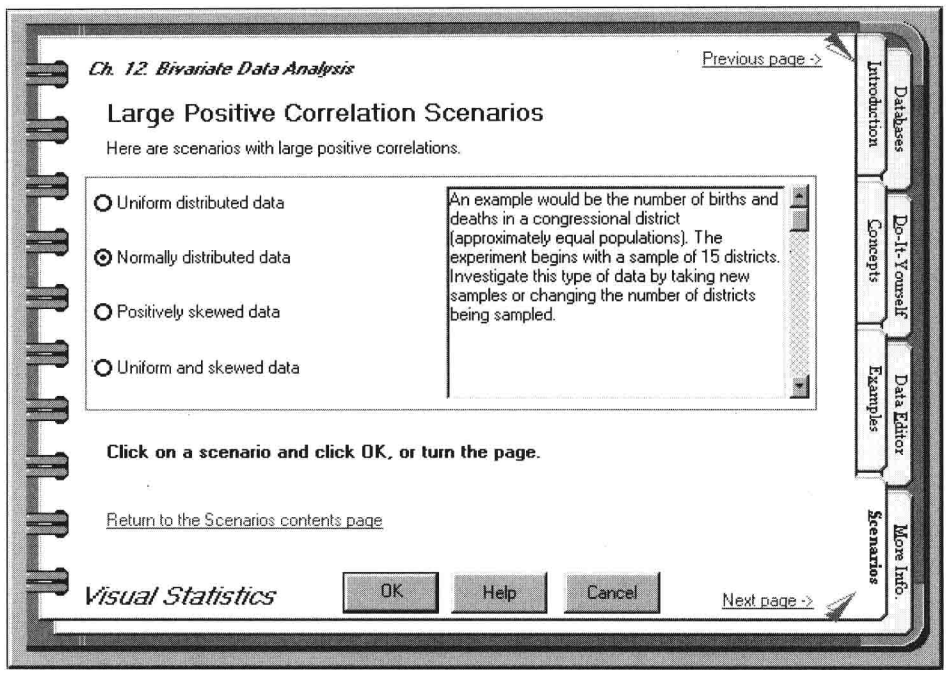


Figure 6: Scenarios

Templates let you generate data conforming to a particular shape. For example, the modules for Distributions of Estimators, Samples, and Time Series all use templates. Figure 7 shows a template screen from the Distributions of Estimators module. Each template represents a probability distribution (you'll learn what that is in your statistics class). To select a distribution to sample from, click on its picture and click the OK button.

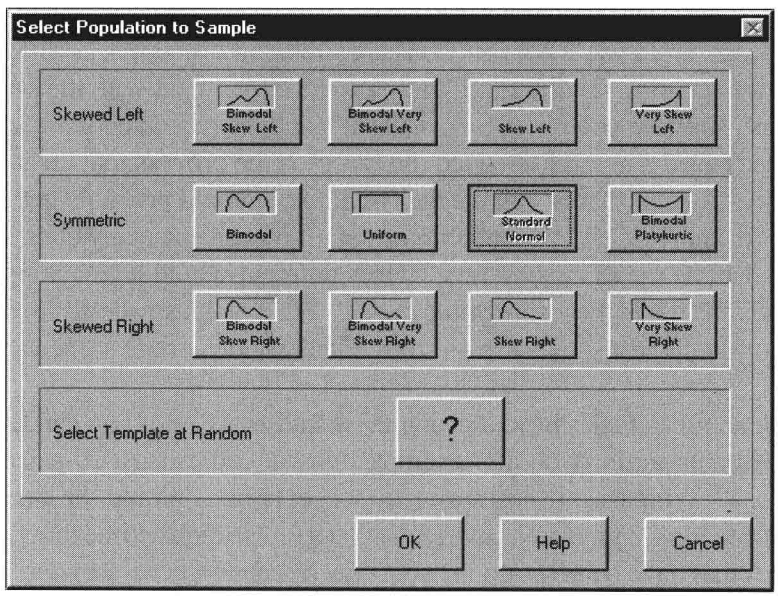


Figure 7: Templates

The Do-It-Yourself section gives you a *lot* of control over the process generating the data. All of the modules except Univariate Data Analysis have a Do-It-Yourself section. Figure 8 shows the Do-It-Yourself controls from the Bivariate Data Analysis module. You can create your own process using the controls for n, ρ, μ, and σ of each distribution. Click the Take Sample button to generate a data set. What the Do-It-Yourself screen looks like depends on what module you are using; see Figure 9 for Do-It-Yourself controls from another module.

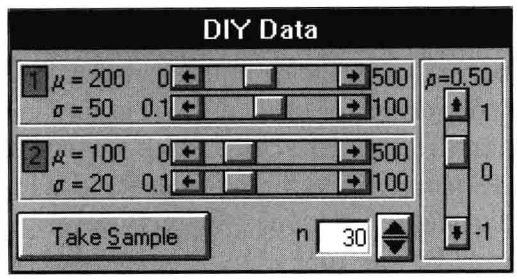


Figure 8: Do-It-Yourself Controls from the Bivariate Data Analysis Module

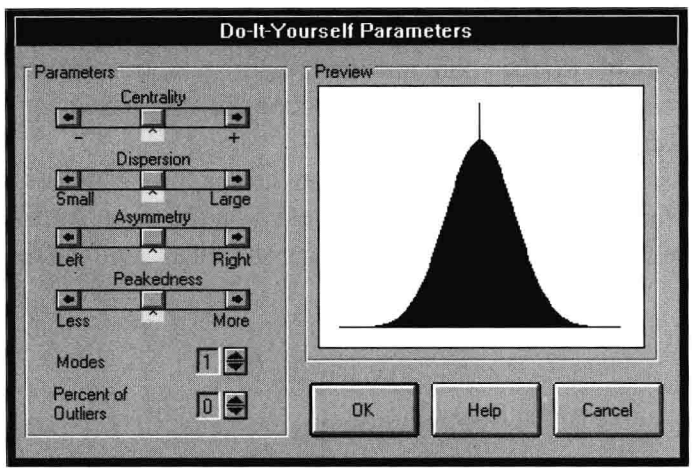


Figure 9: Do-It-Yourself Controls from the Samples Module

Making Choices That Fit Your Needs

You can see that *Visual Statistics* is a powerful tool. There are quite a few modules, some of which have many options. There is something for everyone in *Visual Statistics,* from the novice to the expert. Beginners find it much easier to grasp basic statistical concepts when they use the package. The reason is simple: it's easier to understand something you can actually see than something you have to imagine. However, during our testing of *Visual Statistics,* we found that experts liked using the modules as well. The visualization and experimention capabilities gave them new insights into common statistical problems.

Don't let the suite overwhelm you. Although there's a lot of power available, you don't need to use every option in every module. In fact, you might find yourself using only a few of the features. That's fine if it gives you the level of understanding you want. This worktext will help you get what you need from *Visual Statistics*. In particular, make sure you run through each module's orientation. It only takes a few minutes, and then you'll have a good idea of how to use the module.