Figure 13.12. INVESTIGATING RELATIONSHIPS: Multidimensional Data Analysis

Program 10 in: *Against All Odds: Inside Statistics*

This program completes the presentation of data analysis by showing the use of computing technology and by a case study that uses many of the tools you have learned to this point. The assignment for the Program is a review assignment, and the self-test is a sample examination that covers Programs 2 to 9.

The video begins with an example of a study that analyzes data on many variables to get a picture of environmental stresses in Chesapeake Bay. The variety and abundance of creatures living in the Bay's sediments, the level of salt and dissolved oxygen in the water, and other variables are measured at many locations over several years. Numerical and graphical tools for data analysis help turn this mass of numbers into conclusions about the effects of human activities on the Bay.

Data analysis in practice is carried out with the aid of *statistical software* on a computer. One function of software is to do calculations quickly and accurately. Another is to prepare graphs quickly for our inspection. Yet another is to store and manipulate data sets too large to be practical for analysis with a calculator. Modern computing facilities extend these basic functions in two important ways. First, they are *interactive*, so that the computer responds immediately to your command; this enables you to inspect the result and take the *next* step in the data analysis with better information. Second, statistical computing now features *graphical output* on a high-resolution screen. The text emphasizes the importance of *graphing* data; modern computing allows you to make more elaborate graphical displays *immediately* and to *interact* with them.

The video shows graphics for statistical analysis as practised at Bell Communications Research. The first general principle illustrated is the effectiveness of human-machine interaction. The computer calculates and graphs very quickly, while the human eye and mind see patterns and draw conclusions in ways that are beyond the power of machines. This is why interactive graphics are so important in statistical computing. Some of the displays, such as scatterplots, are familiar to you. But now we can display *several* scatterplots when more than two variables are present, and can *link* points corresponding to the same cases on several plots. The software allows you to move a *brush*, a rectangle on the screen whose size and shape you can choose, over one scatterplot. The matching points on the other plots are automatically highlighted. This allows you to see *relationships* between more than just two variables.

Brushing is one way the computer can help us grasp *multivariate data*, in which more than two variables are measured on each unit. Earlier programs showed statistical methods for a single variable (Programs 2 to 5), for a variable changing over time (Programs 6 and 7), and for relationships between two variables (Programs 8 and 9). A look at new computer-aided methods for inspecting many variables at once completes this development.

A computer screen, like a piece of paper, is two-dimensional. A scatterplot for two variables uses both dimensions. How can we display data with more than two dimensions? Several scatterplots display the variables two at a time, and brushing helps us see multivariate relations. Another method is to try to present a multivariate scatterplot directly. A scatterplot of three variables, for example, is a cloud of points in space. To make the third dimension visible on a flat screen, computing systems use *colour* or *motion*. Notice in the video how motion in particular makes a three-dimensional pattern apparent. Modern statistical software allows you to roam around a three-dimensional scatterplot until you find a viewpoint that reveals the nature of the relations. The video illustrates this by looking at the epicenters of earthquakes in the Fiji Islands, where the third dimension is depth beneath the earth. The proper viewpoint shows that the epicenters mark the boundaries between two moving plates on the earth's surface, whose collision causes the earthquakes. The same idea can be applied to data with four or more dimensions. A scatterplot of two of the variables is a particular two-dimensional view; it takes more experience to grasp changing two-dimensional views as the computer changes viewpoints.

Another way to present multivariate data on a flat screen or piece of paper is to use a representation unrelated to scatterplots. Many such graphs have been invented. The video looks at one chosen because it is easy to understand: *faces*. Each variable controls one feature of a cartoon human face, so that different relations between the variables result in faces with different expressions. The human eye and mind can sometimes use the faces to group cases that are similar and to distinguish them from other cases. For example, very good forged currency has been distinguished from genuine currency by making several measurements and representing them as a face.

Blank page