DOCUMENT RESUME

ED 279 731 TM 870 181

AUTHOR	Bessey, Barbara L.; And Others
TITLE	Graphical Methods: A Review of Current Methods and Computer Hardware and Software. Technical Report No. 27.
INSTITUTION	American Institutes for Research in the Behavioral Sciences. Palo Alto, CA. Statistical Analysis Group in Education.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	AIR-87600-2-83-TR
PUB DATE	Feb 83
CONTRACT	300-78-0150
NOTE	62p.
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Computer Graphics; Computers; Computer Software; *Computer Software Reviews; *Data Analysis; *Data Interpretation; *Display Systems; Educational Research; *Equipment Evaluation; Graphs; Literature Reviews; Research Utilization
IDENTIFIERS	National Center for Education Statistics

ABSTRACT

Graphical methods for displaying data, as well as available computer software and hardware, are reviewed. The authors have emphasized the types of graphs which are most relevant to the needs of the National Center for Education Statistics (NCES) and its readers. The following types of graphs are described: tabulations, stem-and-leaf displays, histograms, rootograms, probability plots, bar graphs, box plots, scatterplots, biplots, motion graphics, tables, pie charts, glyphs, faces, maps, and trees and castles. Ways to display distributions and to display relationships between two or three or more variables are discussed. Several software packages are briefly reviewed: SAS, SPSS, BMDP, OSIRIS, EXPAK, Exploratory Data Analysis Package, ABCs of EDA, STATGRAPHICS, S Package, FACES, Trees and Castles, CANDEC, and ORION I. The wider use of inferential types of graphs is recommended for educational data analysis and the presentation of findings in NCES reports such as "The Condition of Education." Hardware discussions include Hewlett-Packard, Tektronix, Digital, and Versatec. Software packages are discussed in terms of use with mainframe computers, or mini and desktop computers: ISSCO Graphics, Hewlett-Packard, Tektronix, SAS/GRAPH, STATGRAPHICS, Intelligent Graphics Systems, and Chart Master. (GDC)

AIR-87600-2/83-TR

AMERICAN INSTITUTES FOR RESEARCH
IN THE BEHAVIORAL SCIENCES
P.O. Box 1113, 1791 Arastradero Rd., Palo Alto, Ca. 94302 ● 415/493-3550

TECHNICAL REPORT 27

GRAPHICAL METHODS:
A REVIEW OF CURRENT METHODS AND
COMPUTER HARDWARE AND SOFTWARE

Barbara L. Bessey
Laurie R. Harrison
David A. Brandt

Statistical Analysis Group in Education
American Institutes for Research
P. O. Box 1113
Palo Alto, California 94302

This work was done under Contract No. 300-78-0150 with the National Center for education Statistics, Department of Health, Education, and Welfare. However, the content does not necessarily reflect the position or policy of either agency, and no official endoresment should be inferred.

February 1983

TABLE OF CONTENTS

GRAPHICAL METHODS: A REVIEW OF CURRENT METHODS AND COMPUTER HARDWARE AND SOFTWARE

Introduction

The National Center for Education Statistics depends heavily on graphical methods to communicate information on the state of education to a very broad and general audience. Graphs play a major role in important NCES publications such as Condition of Education because they, much more than tabulations, are capable of conveying information to a lay audience in a simple, clear, and effective way. Precisely because graphs are relied upon so heavily, it is vital that NCES make use of the most effective and communicative types of graphs in their reports. This paper reviews the literature on graphical methods and the available computer hardware and software. We emphasize the types of graphs that we believe are most relevant to NCES's needs and most effectively convey findings or conclusions to the reader. Finally, we make recommendations concerning the use of graphical methods in the analysis of educational data and in the formal presentation of findings in NCES reports.

The use of graphical methods to display data has waxed and waned over the years. The past decade has witnessed an increase in the development and use of graphical techniques in the analysis and presentation of behavioral data. Both applied and theoretical statisticians have recognized the importance of graphical methods in data analysis, and numerous books and articles devoted exclusively to the development and study of graphical techniques have appeared in the professional literature. This upsurge is due in part to John Tukey's pioneering efforts in this area. His book, Exploratory Data Analysis, published in 1977 (although several prepublication versions were dated much earlier), is the single best-known publication on modern graphical methods to appear in this period.

Tukey and other authors argue that the use of graphical methods should not be confined to the presentation of data in published reports. Rather, graphical methods have an invaluable role to play in the analysis of data. Techniques that he and other statisticians have developed are designed to complement the more formal statistical approaches to data analysis. His book Data Analysis and Regression (written with Fred Mosteller in 1977) illustrates how he believes graphical and statistical methods can be used jointly in the data analysis process. Graphical techniques are viewed as a method for learning about relations among variables rather than simply as a way of displaying tabulations of numbers. These newer techniques are also better suited for presentation graphics, because they are capable of focusing the reader's attention on the most salient features of the data. This paper reviews the methods developed in the last ten years, with special attention given to techniques that appear to be especially relevant to NCES publications such as Condition of Education. A second purpose of this paper is to review the computer software and hardware currently available. These topics are discussed in the following sections:

- types of graphical methods for examining data, and
- computer programs for graphically displaying data.

A final section provides a summary and recommendations as to how these methods may be used. The appendix to this paper contains a description of computer hardware and software specifically designed to produce graphic displays of publishable quality.

Types of Graphical Methods for Examining Data

Graphical methods are extemely useful for uncovering and displaying trends in data, as well as presenting results clearly and effectively. Some graphical methods display the (univariate) distributions of variables, while others plot relationships among two or more variables.

In this section, graphical methods will be presented that display

- o distributions of variables, including simple summary displays;
- o relationships among two variables; and
- o relationships among three or more variables.

Displays of Distributions

Tabulations

Using graphic displays to illustrate tabulations is one of the oldest graphical methods. For example, the traditional tally method counts by fives as follows:

Tukey (1977), however, points out that this method is likely to produce errors when large numbers of events must be counted: It is too easy for an analyst to complete figures accidentally (e.g.,   or   ). Tukey's recommended approach is based on tallies by tens--four dots successively placed in a square represent counts one to four, then four lines connecting them represent counts five to eight, and finally two crossed lines are drawn to represent the counts for nine and ten. While Tukey believes that the occurrence of errors using this method is less than for any other, these new box figures are not universally used. Several examples of Tukey's method of tallying by tens are shown in Figure 1.

The use of computers to represent counts has been explored by Bachi (cited in Wainer & Thissen, 1981). Bachi's graphic rational patterns (GRPs) make use of a system that graphically represents the amount (the more, the darker). The use of GRPs to represent integers from 1 to 100 is shown in the left panel of Figure 2. A comparison of the GRP method with the traditional bar chart is presented in the right panel of Figure 2. For these data, the GRP method requires less space and may be just as informative as the bar chart presented above it.

It is clear from an examination of Figure 2 that these displays must be computer generated to ensure clarity as well as accuracy. Although no computer software is currently commercially available that will generate displays using the GRP method, Wainer and Thissen (1981) note that many of the existing computer graphic systems can produce GRP-like patterns that can then be used in a user-written program to produce the desired plots.

Stem-and-Leaf Displays

Tallies by fives or tens, while appropriate for situations in which a quick grasp of the appearance of the data is sufficient, do not readily lend themselves to an examination of the distributional properties of data. They also are really specialized for hand-accumulation of summary plots rather than for implementation on the computer.

One such descriptive technique, the stem-and-leaf diagram, discussed in detail in Tukey (1977), enables analysts to examine distributional properties of data, including checks for outliers and the calculation of summary statistics. Figure 3 presents a simple set of observations illustrated conventionally using tallies and through use of a simple stem-and-leaf diagram.

In Figure 3, the numbers on the left side of the vertical bar are the "stems;" the numbers on the right side of the vertical bar are the "leaves."

Each observation is represented by a leaf. The label for the stem is the first part of a number followed by each leaf in turn. The asterisks to the right of each stem indicate the number of digits in each leaf. Thus, the second stem line in Figure 3 can be read as five numbers: 11, 14, 16, 16, and 19. As is evident even from this simple example, 27 lines are used to display the tallies in Figure 3, in which only 10 lines are marked, while only 4 lines are needed to represent the identical values expressed in a stem-and-leaf diagram.

For some data, however, a simple stem-and-leaf design (as shown in Figure 3) is inadequate, because the range of values is too large. Consider the display of the 1960 populations of the 50 states shown in Figure 4. Here, state populations range from 230,000 to 16,780,000. Had stems in the entire range been represented, 200 lines would have been needed to depict the data, which would make the display extremely difficult to interpret. Thus, for convenience, the population data are displayed in tens of thousands.

In Figure 4, the leaves are separated by commas, because they consist of one, two, or three digits. The least populous state has a stem of 2*, a leaf of 3, and a population of 23x10,000=230,000. New York's population of 16,780,000 is represented by a stem of 1*** and a leaf of 678. Some states of special interest are named at the right of the display. Implicit in Figure 4 is a logarithmic transformation of the data, which is needed to bring the number of stems down to a manageable and interpretable number. Were the larger populations not compressed by being transformed, there would only be four states in the last 100 lines!

The selection of the number of stems to be used will affect the ease with which summary information can be extracted from the display. For example, the stem-and-leaf diagram shown in Figure 3 should be modified to provide more information, since most of the leaves are crowded onto two stems. One modification is called the <u>stretched</u> stem-and-leaf display, which uses two stems for each starting part--one line for leaves 0, 1, 2, 3, and 4, and the other for leaves 5, 6, 7, 8, and 9. An example of a stretched stem-and-leaf diagram is shown in Panel 1 of Figure 5. Here, the starting parts of each stem are repeated, but the asterisk is shown only for those that include leaves of 0 through 4; dots are used to designate leaves of 5 through 9.

Panel 2 of Figure 5 shows another modification, called the squeezed stem-and-leaf display. This panel shows the actual log-transformation of the popula 'on data included in Figure 4. The display showing three-digit accuracy (on the left) is appropriate for keeping track of the original numbers, but it is too spread out to show the form of the distribution well. The stem-and-leaf display on the right squeezes the leaves onto shorthand stems (* for 0 or 1, t for 2 or 3, etc.), which clarifies the distribution.

Panel 3 of Figure 5 shows yet another variation of the stem-and-leaf diagram. In this example, a mixed-leaf stem-and-leaf display is used: One-digit leaves are used to represent the values -59 to +59, and two-digit leaves are used to represent values greater than or equal to +60 and those less than or equal to -60. By using larger leaves for some digit lengths, only 16 lines are needed, whereas 31 would have been required had all starting parts used the same stem length.

Use of the stem-and-leaf diagrams enables a number of statistical measures of location and spread to be computed. A five-number summary can be calculated for each display--the extreme values, the median, and the 25th and 75th percentiles. The five-number summary for the data shown in Figure 4 is illustrated in Figure 6.

The 25th and 75th percentile values of the distribution are called hinges. The hinges are located midway between the extreme values and the median. In this example, the hinges are the 13th numbers counting in from the top and bottom. The corresponding values, 89 and 432, are recorded inside the box and to the right of the code for hinges "H13."

To the left of the stems in Figure 6 are the cumulative counts--from the top down to the middle and from the bottom up to the middle. For the 50 states, the middle value or median is the average of the 25th and 26th states--or the 25-and-a-halfth entry on the plot (designated as 25h). The median value is 246, which is recorded inside the box and to the right of the code for the median "m25h."

The highest or lowest values, called the extreme values, 23 and 1,678 in this example, are designated as having a depth of "1."

These five-number descriptive summaries--which utilize the median, the hinges, and the extreme values--can also be plotted. The resulting plots (called box-and-whisker plots) are described later in this section. Alternatively, the mean and standard deviation can be plotted. These latter values may be more appropriate if the variables to be displayed are test scores that follow a normal distribution. However, more robust measures of location and spread are applicable to a broader range of variables.

Histogram and Rootogram

The histogram provides a graphical representation of a distribution of frequencies. An important feature of the histogram is that the area of each rectangle is proportional to the frequencies contributing to it. The data are divided into a discrete number of classes or categories, and the number of observations belonging to each class are determined. If the class intervals all have the same size, the heights of the rectangles are proportional to the class frequencies, and it is accepted practice to use heights equal to the class frequencies. If the class intervals are not equal, the heights must be adjusted. Figure 7 displays two histograms--a simple histogram with equal-sized class intervals is shown in Panel 1; and a histogram with unequal-sized class intervals is shown in Panel 2.

Assessing normality. Graphical methods are very well suited for assessing the normality of a distribution. The researcher should, of course, make such an assessment prior to analyzing variables using statistical techniques that assume a normal distribution of the residuals. The methods to be discussed in this section can be used both to assess the normality of the observed variables and to evaluate potential transformations of the data. The methods to be described below are, in general, to be preferred to the presentation of summary statistics, such as the coefficients of skewness and kurtosis, because these statistics do not have the intuitive meaning that a plot of the data does.

To use the histogram for this purpose, one can augment the basic histogram with a normal curve. The departure from normality is examined by comparing the curve to the tops of the histobars. However, as Tukey (1972) points out, using a curve as a standard of comparison is not the optimal procedure. The eye can detect departures from straight lines more accurately than from curves. In particular, the eye cannot easily detect a difference between the normal curve and another bell-shaped curve such as a t-distribution. An even better method is to "hang" the histogram from the fitted normal distribution leaving the discrepancies at the bottom as deviations about a horizontal line.

Since transformed data are more likely to be normally distributed, Tukey (1972) argues that use of transformed data rather than simple frequencies are more likely to follow a normal distribution. For example, square roots of counts are usually better behaved statistically than the counts themselves. When the square roots of the datapoints are plotted, Tukey calls the resulting plot a <u>rootogram</u> to denote both the type of transformation and type of plot. An example of a simple histogram (showing heights of volcanoes) and the same data displayed in rootogram form (that is, the square roots of the heights) are shown in Panels 1 and 2, respectively, in Figure 8.

While the eye may fail to see a definite relationship when looking at the histogram, the rootogram strongly suggests a normal distribution. The question of whether the heights of these volcanoes are normally distributed could be answered by superimposing a normal distribution over the tops of the rootogram and then comparing the goodness of fit. But since using a curve as a standard of comparison is a poor graphical practice, Tukey provides several alternative methods.

Panel 1 in Figure 9 presents the same height data but in the form of a hanging rootogram. Each rootobar in Figure 9 has been "hung" from the fitted normal distribution. The discrepancies from the normal distribution appear as bars that miss or overshoot the horizontal line. Panel 2 shows the same data drawn as a suspended rootogram. Here, the excesses or deficits from normality are plotted on the horizontal line, so that the eye can see visually the departures from normality. Tukey's conclusion that the fit of the normal curve to the heights of these volcanoes is a good one is apparent from Panel 2 of Figure 9 but was not strongly suggested by the simple histogram in Figure 8.

Wainer (1974), in an evaluation of the histogram and rootogram procedures, conducted human performance experiments to determine which technique is the more effective. He found that hanging rootograms are best for detecting skewness and/or kurtosis, followed by hanging histograms and ordinary histograms.

Probability Plots

The probability plot is an alternative to the use of hanging or suspended rootograms for assessing the normality of distributions. With the use of normal probability graph paper, quartiles of the normal distribution (for example, in z-score units) are displayed on the x-axis, and ordered data values are selected for the y-axis such that a plot of a cumulative normal distribution is a straight line. Thus, if a scatterplot is made on such a grid, the plot resembles a straight line to the extent that the data are normally distributed. Deviations from linearity are indicative of nonnormality. An example of a probability plot is shown in Figure 10.

Such graphs are available in several of the BMDP (Biomedical Computer Program) programs designed for displaying univariate data. The program BMDP5D can be used for obtaining histograms and normal probability plots, while the regression programs BMDP1R and BMDP2R (standard and stepwise regression) include normal probability plots as a method of examining residuals. These programs are accessible in SAS (Statistical Analysis System) via the PROC BMDP procedure. This procedure permits the user to access any of the BMDP programs from within SAS just as if they were SAS procedures themselves.

Gnanadesikan and his colleagues (cited in Wainer & Thissen, 1981) present examples that use both quartile (called Q-Q) and percentile (P-P) plots. The term Q-Q plot refers to all possible variations of the probability plot. For example, ordered data may be plotted against quartiles of a theoretical distribution, or the quartiles of one empirical distribution may be plotted against another. The intent is to compare two distributions. If the two distributions are identical, the sameness will be represented by a straight line. In the example shown in Figure 10, the data represent a distribution of 104 energies associated with an individual repeatedly speaking a single word. The distribution is bimodal--one mode is at the lower extreme, the other is around 1.8 on the normal quartile (z-score) scale (noted by the "lump" about two-thirds of the way up the plot). The large number of tied scores at the lower extreme results in the plot beginning higher than expected. This observation means that the distribution does not have a "tail," as would a normal distribution.

P-P plots are plots of percentiles rather than quartiles, but are otherwise similar to Q-Q plots. Wainer and Thissen (1981) note in their review, however, that P-P plots have generally been studied very little.

Bar Graph and Box Plots

Instead of plotting all the datapoints, as we have done in the previous examples, we can plot summary statistics. One simple method for displaying summaries of data is the bar graph, and another method is the box-and-whisker plot. These methods are described below.

Bar graph. The bar graph only displays a measure of location. Figure 11, taken from the 1981 Condition of Education (Chart 2.16, p. 89), presents a bar graph in which the subject matter areas for minimum competency tests are shown along with the numbers of states requiring each.

To provide the reader with more information about the distribution of the data, we can add a measure of spread to the simple bar graph as shown in Figure 12. In this example taken from the Bureau of the Census's Social

Indicators III (p. XXII), the tops of the bars represent the total number (or percent) of 18- to 24-year-olds who reported voting in the 1978 election. The rectangles that are slightly offset from the tops of the bars represent one standard error of measurement. The values that are plotted in a bar graph can also be means or medians, which can be displayed along with their standard errors or another measure of uncertainty. Examples of bar graphs in which means are used will be discussed later in this paper.

Box-and-whisker plots. Summaries of data can also be displayed in the form of a box-and-whisker plot. A simple example is a plot of the five-number summary associated with the stem-and-leaf diagram discussed earlier. In the basic configuration, which is shown in Panel 1 of Figure 13, the lower line of the box (called the lower hinge) is the value represented by the data point at the 25th percentile of the distribution; the upper line of the box (called the upper hinge) is represented by the data point at the the 75th percentile. The median of the distribution is indicated by the line drawn through the center of the box. The "whiskers" extend from the upper and lower hinges to the upper and lower extreme values, respectively. Tukey (1977) provides a variation on the basic configuration in which the end points are identified. In Panel 2 of Figure 13, the whiskers are stopped at the innermost identified values.

As is apparent from Figure 13, much information about the data is shown, although only the summary statistics are plotted. In the next section, more complex variations of the box plot will be displayed that compare several variables plus information about measures of uncertainty.

Displays of Relationships between Two Variables

In the last section, we examined univariate distributions of data. Such plots are valuable in the preliminary stages of research when the most basic features of the data are being evaluated. After this assessment has been completed, associations among variables are typically investigated. Statistical methods for assessing associations among variables, such as the t-test, the analysis of variance, the correlation coefficient, and principal component and factor analysis, are the statistician's most valuable and frequently used tools. Corresponding to these techniques are graphical methods designed to display relations among variables. Although these techniques are much newer than their statistical counterparts, they are no less useful. In this section of the paper, we describe several such methods and evaluate the more primitive graphical methods that are commonly used in their place. Good graphical methods make the associations among them clear and self-evident. Poor graphical methods will often obscure these associations and make the nature of the relationships difficult to see.

Graphical methods commonly employed in this context are the pie chart, simple bar graph, bar graphs that are segmented, and so on. The fact that they are easily accessed through the standard statistical packages is definitely a key to their widespread popularity. While these methods may be appropriate for displaying data in some situations, their value is actually quite limited. Consider the following example as illustrative of this caveat.

It is fairly common to find several graphs presented adjacent to each other on one page to illustrate multiple relationships. For example, four bar graphs, each with segments representing several employment status categories might be presented separately for black and white males and females. To perceive the relationship between sex and employment status, the reader must mentally align the corresponding segments to determine if more males or females are unemployed. For bar graphs that contain many segments or have a few long segments and several very small ones, mental alignment of such segments can be very difficult indeed. The pie chart, which is a circular version of the segmented bar chart, suffers from the same limitation. Clearly, these two types of graphs are not designed to facilitate comprehension of the relationship of a stratification variable to a continuous variable. Instead, they are, more or less, graphical analogues of a simple tabulation of the data. In this section, we will present several techniques that are designed to call attention to the relationship between the variables. In contrast to the segmented bar graph and the pie chart, these techniques are the graphical analogues of the one-way and two-way analysis of variance. They make use of the single most important feature common to formal analytic methods--model fitting. They use numerical techniques to estimate, in an informal way, parameters in a model. The parameters in the models then are used to augment the display of the datapoints in the graph. These displays go beyond the tabulation methods in the sense that they impose a higher level of organization on the data, and they call the reader's attention to that organization.

This section presents ways in which relationships among two variables are obtained and displayed. Some of the methods that will be described are less well known than the tabulation methods, but they are generally more powerful and communicative. Two basic classes of problems are examined. The first set of examples looks at one continuous variable and one or two stratification variables; the second set of examples examines relationships among two continuous variables.

Displays of One Continuous Variable and One Grouping Variable

The stem-and-leaf diagram. One effective use of the stem-and-leaf display is for comparison of several groups of data. The stems of the groups may be presented side by side (as in Panel 1 of Figure 14) or back to back (as in Panel 2 of Figure 14) for ease of comparison. In Figure 14, the land areas of the counties of two states are presented. In Panel 2, the same set of stems is used to display the data for both states (recall that the * represents leaves from 0 to 4, the dot represents leaves from 5 to 9). The "leaves" for Michigan are shown on the left, while the "leaves" for Mississippi are shown on the right. It is apparent from the figure that the county areas in Michigan are clustered between 545 and 595 square miles with a long tail towards the larger areas. Re-expressing these data as logarithms makes the display even easier to read (compare Panel 2 and Panel 3 in Figure 14).

The one-way analysis of variance program, BMDP7D, includes such plots as part of the standard output. Data from each group in the design are plotted side-by-side so that the researcher can easily detect differences in both mean and variance among groups.

Variations of the box plot. The box plot is an excellent method of comparing two or more groups. By plotting summary statistics rather than the datapoints themselves, one obtains a less cluttered but very informative graph. McGill, Tukey, and Larsen (1978) present three modifications to the basic box plot that add information on group size and statistically significant differences among groups. In Panel 1 of Figure 15, the standard box plot is presented for displaying a single month's telephone bills as a function of the number of years lived in Chicago. In this graph, several box plots are presented side by side on the same set of axes. A reader might mistakenly conclude that the overall median of all groups is about $21, when, in fact it is $14. The reason for the misinterpretation is the fact (which is not displayed) that the number of customers in each group differs widely--the ratio between the largest (over 15 years) and the smallest (less than 1 year) groups is over 33:1.

Panel 2 of Figure 15 displays this additional information using the variable-width box plot. Here, the width of each box is proportional to the square root of the number of customers in the corresponding group. The analyst's attention in Panel 2 is immediately drawn to the size differences, which will facilitate making more accurate conclusions about the data. (It is noted by McGill, Tukey, and Larsen (1978) that use of the square root to select box width will tend to minimize the differences in group sizes. If the purpose is to emphasize differences in group size, the analyst may wish to select box widths that are directly proportional to group size.)

The analyst, however, looking at Panels 1 and 2 of Figure 15, may still have difficulties in deciding which of the group medians are significantly different from each other. The box plot in Panel 3 shows the confidence intervals around the medians as "notches." If the notches about two medians do not overlap in the display, the medians are statistically significant at the 95 percent confidence level. As can be seen from Panel 3, none of the first five medians differs statistically from each other. In this example, an advantage will be accrued by combining both variations. The result, called the variable-width notched box plot, is shown in Panel 4 of Figure 15. Here, the first five groups are combined and displayed as one group. The resulting two groups (residence from 0 to 15 years vs. residence over 15 years) are observed to differ significantly, since the notches are non-overlapping. The difference between the groups is about $4, which is the distance between the nearest edges of the notches.

Two examples of box plots are presented here for examination. The first example is taken from a recent paper by Richard Jaeger that appeared in the Winter 1982 edition of Educational Evaluation and Policy Analysis. His data, which are shown in Figure 16, illustrate the regular box plots to show distributions of passing scores on a state's reading and math competency tests recommended by various participant groups. The differences in group sizes might have been better displayed had variable-width box plots been used.

The second example is taken from unpublished data collected by AIR as part of a recently conducted study of state administration in ESEA Title I. Figure 17 displays variable-width notched box plots illustrating state Title I allocations as a function of type of organizational structure. The data displayed in Figure 17 are presented in Table 17. It should be noted that, in two instances, a notch lies ouside a hinge. In these two cases, where the median is very near the hinge, protruding notches have been drawn.

The box plots heretofore discussed have been used to display medians and their associated confidence intervals. The results of many studies, however, are summarized by a series of means. An example of a box plot showing means and their respective confidence intervals is shown in Figure 18. This example is taken from the 1977 Condition of Education (Chart 5.02, p. 104).

The procedure PROC SPLOT in the SAS supplemental library can be used to produce basic box plots for several groups. Also, the SPSS program, MANOVA, contains an option to produce the same type of plot in connection with a multivariate analysis of variance.

Plotting difference scores. In some instances, plotting the differences between the levels of the grouping variable is an effective way of understanding the main effects in the data. Consider as an example the data taken from the 1980 Condition of Education (Chart 2.17, and Table 2.17 on pp. 89 and 88, respectively) as presented in Figure 19 and Table 19, respectively.

Here, the effect of two stratification variables on a continuous variable (consumer knowledge) is displayed as a deviation from the overall mean. (This procedure is directly analogous to the use of so-called "deviation contrasts" in the analysis of variance [Bock, 1975, p. 300]). Differences in performance are plotted, which show visually that males did better on this particular test than females. A similar display showing performance differences by race is also included.

A better way of looking at these data is to examine performance as a function of both grouping variables--sex and race. Looking at the effects of both variables will enable us to determine the effects due to sex alone, race alone, and whether or not the two variables interact. These types of relationships will be examined next.

Displays of One Continuous Variable and Two Grouping Variables

When several grouping variables are involved, good graphical procedures display the effects of the grouping variables. These effects are estimated and then graphed. Two examples are provided here to illustrate techniques for examining data involving two grouping variables. In the first example, the results due to the main effects are displayed graphically. The second example shows how the original level of the continuous variable can be retained while simultaneously displaying the effects of the grouping variables and the residuals from those effects. This type of graph, while more complex, contains a wealth of information and is to be recommended in many situations.

Figure 20 is a segmented bar graph taken from the 1981 Condition of Education (Chart 2.3, p. 63). The accompanying data for the chart are presented in Table 20 (Table 2.3, p. 62). In this example, a continuous variable, namely public and private school enrollment, is stratified by two variables--Region of the Nation (Northeast, North Central, South, and West) and by Metropolitan Status (Metropolitan City, Metropolitan Outside Center City, and Non-Metropolitan). The display in Figure 20 presents three graphs--one for the entire nation, and the remaining two stratified by the two factors of Region and Metropolitan Status. Within each stratum, percentage enrollment is shown for the categories: Public school, Private-religious, Private-unaffiliated, and Not ascertained.

This graph successfully conveys certain features of the data, but it fails to highlight what are perhaps the most interesting aspects of the data. Clearly, the single most obvious fact that the reader notices from inspection is that about 90 percent of the nation's students are in public schools. Most of the space on the graph is actually devoted to expressing a fact that the reader knows already. That is, the overall level of enrollment dominates the graph. Very little space on the graph is available to depict the remaining enrollment categories. Because of this it is relatively difficult to grasp differences in percentage enrollment among strata. Also, except for public schools, the portion of the bar representing each enrollment category is not aligned (i.e., they don't begin at a common location across strata). One consequence of these problems is that it is practically impossible to learn anything interesting about the smallest enrollment category, private non-affiliated schools. Regrettably, the reader can only see that enrollment in these schools is generally small.

Because the most important "message" in the data deals with differences in percentage enrollment as a function of Region and Metropolitan Status, this aspect of the data should be emphasized in the graph. Before graphing, the "main effects" due to Region and Metropolitan Status need to be calculated. One method that can be used for this purpose is called "median polish," which was developed by Tukey (see detailed descriptions in Tukey [1977, see Chapters 10 and 11], Mosteller and Tukey [1977, see Chapter 9], and Velleman and Hoaglin [1981, see Chapter 9]). The procedure is very similiar to a two-way analysis of variance model with the exception that medians are used in place of means. Because medians are more resistant to the effects of outliers and, thus, are more suitable estimates of location for skewed distributions, they are appropriate for a wider range of variables than are means. This method of estimating main effects is described below.

First, the data are displayed in tabular form, as in Table 20a. (To illustrate this method, we are ignoring the "Affiliation not Reported" category). In the right margin on Table 20a, the median percentage for each row is indicated. This statistic is used as an estimator of the "average" percentage enrollment for each region. Our estimates of the "effects" of Region and Metropolitan Status are presented in the row and column margins of Table 20b. Several things have happened.

- First, deviations from the median within each row have been computed (e.g., for the first cell, 80 - 90 = -10).

- Second, the row medians have been replaced by the deviation of each row median from the median of all the row medians (e.g., the median of the row medians for "% Public" is 90.5, so the median deviation for Northeast is -.5). These will be used as estimates of the "effect" of Region.

- Third, the median of the entries in each column is given. These statistics will serve as estimates of the "effect" of Metropolitan Status.

Thus, the row and column marginal statistics from Table 20b are the basic values that are needed for graphing. Figures 20a and 20b graph the resulting "effects" of Region and Metropolitan Status.

From Figure 20a, it is immediately apparent that the West and the South show similar patterns for all three enrollment categories. Thus, it is now as easy to see trends in the Private-unaffiliated category as for the Public category. Similarly, the Northeast and North Central regions show the opposite pattern. A region with an above average Private-religious enrollment has a lower than average Private-unaffiliated enrollment, and vice versa. This feature of the data is not at all apparent from Figure 20.

Figure 20b shows that Metropolitan-city and Non-metropolitan display exactly the opposite pattern. Again, any aspects of the data concerning the small enrollment categories not perceived from Figure 20 can now be seen.

After the "effects" due to factors have been assessed using the above procedure, "effects" due to each combination of factors can be assessed using the following algorithm. The cell entries in Table 20c are deviations of each Table 20b entry from the column median. That is, Table 20c shows residuals after taking out the "effects" due to both Region and Metropolitan Status. These residuals represent the degree to which row and column effects do not adequately predict each cell percentage. Thus, especially large residuals are of interest. For example, in Table 20c it can be seen that Northeast metropolitan cities have an especially low percentage of public school students and an especially high percentage of Private-religious students. This is probably due to the especially high percentage of Catholics in the population and of Catholic schools in New England cities. A trend such as this simply cannot be seen in Figure 20, because it does not examine the "effects" of Region and Metropolitan Status jointly, nor does it give a way of quantifying the idea of "especially large."

To summarize, this method improves upon the segmented bar chart in two ways: First, it quantifies the "effect" of each grouping variable and graphs that effect explicitly, and, second, it considers the effects of the two variables jointly, rather than individually. This permits examination of residuals from both factors.

Plotting level and main effects. The following example adds a third advantage. The original level of the continuous variable is graphed as well as the main effects and residuals. In other words, it is a more unified and elegant way of presenting such data. It is strongly recommended over the segmented bar graph and the pie chart, because it explicitly and clearly displays both the level of the continuous variable and the effects of the grouping variables. In contrast, the tabular methods display original level in a way that obscures the effects of the grouping variables.

In this second example, a single continuous variable, namely hourly starting salaries, is graphed as a function of sex and level of education. Since hourly wage is expressed in meaningful units, it is preferable to retain their values in the plot. These data are taken from the 1982 Condition of Education (Chart 5.24, p. 219) and are displayed in Figure 21; the accompanying data for the chart are presented in Table 21 (Chart 5.24, p. 218). An examination of the figure and the accompanying data suggest that there may be differences in starting salaries due to Sex and Level of Schooling. The precise nature of these effects, however, is not clear. We can again use Tukey's median polish technique to compute the main effects and use a special graphing technique (also developed by Tukey) for displaying these relationships.

The preliminary calculations follow the same procedures as outlined in the previous example. The resulting values are presented in Table 21a:

- The median of each row and the median of the row medians are calculated as shown in Panel 1.

- The row median is subtracted from each entry (e.g., 5.39-5.96=-.57) to obtain the deviations from the median in each row as shown in Panel 2.

- The row median is replaced by the deviation of each row median from the median of all row medians (e.g., 5.96-5.36=.60) as shown in Panel 2.

- The column medians are calculated as shown in Panel 2.

- The column median is subtracted from each entry (e.g., -.57-(-.53)=-.04) to obtain the deviations from the median in each column as shown in Panel 3.

It is this last set of residuals that will eventually be plotted on a graph that maintains the levels for each main effect--here, Sex and Level of Schooling. The key to this technique is the use of two sets of axes. The first set of axes is used to graph the main effects, and the second is used to indicate level and size of residuals. An intermediate stage of the plot, emphasizing the first set of axes, is shown in Figure 21a. In this plot the rectangle, shown in boldface, indicates the predicted hourly wages for each combination of sex and level of education. The datapoints themselves are "hung" from the rectangle and the length of the lines connecting the rectangle to the datapoint indicates the size of each residual. This graph is rotated 45 degrees so that the original level associated with each sex and level of schooling combination can be read using the scales at either side of the plot. That is, the first set of axes is used to indicate the effects of the grouping variables on the dependent variable, and the second set of axes is used to indicate level and size of the residuals. While more details of the graphing method can be found in Tukey (1977, see pp. 377-388 in Chapter 11) and Mosteller and Tukey (1977, see pp. 169-178), the technique can be briefly described as follows:

- Two vertical lines are drawn to represent the overall effect due to Sex (the values obtained from subtracting the overall median of row medians from each row median). Here, the value for females is -.60, the value for males is .60 (the distance between the groups is 1.2, which is the difference between their group medians). Thus, the data say that, on the average, males have an hourly starting salary that is $1.20 more than the average hourly starting salary for females.

- Horizontal lines are drawn to represent the median values for each Level of Schooling.

- The graph is rotated 45 degrees. Now horizontal lines are drawn to represent hourly rates. From each of the 16 intersections (two Sex lines with the eight Level of Schooling lines), the residuals shown in Panel 3 can be plotted vertically on the hourly rate scales. The end points correspond to the hourly rates shown in Panel 1.

In the plot, average starting salary for each combination of sex and level of education is shown. The values can be read using the scales at either side of the graph. In addition, the main effects due to both Sex and Level of Schooling are graphed explicitly. The final plot, emphasizing the second set of axes, is shown in Figure 21b.

In all cases, the starting salaries for females with the same level of education are lower than those for males. While it is true that, for all schooling levels, females tend to be paid less than males, the difference tends to decrease for higher levels of schooling. It is also generally true that starting salaries increase as a function of eduational level. There is one instance when this is not so--the case of more than two years of college and no vocational education. From Figure 21a, the reader can detect this anomaly.

Although this type of graph is more complex than plots of difference scores, it contains more information. The presence of original <u>levels</u> and the <u>residuals</u> from the main effects in the plot makes it easy for the reader to make many kinds of comparisons. For example, for males, any educational level that indicates attendance at some college is associated with a starting salary greater than $5.50 an hour, but, for females, only college graduates make more than an average of $5.50 an hour. This can be seen by simply following the "$5.50 an hour" line across the page. Also, the pattern of residuals suggests that, for females, postgraduate education is associated with higher than expected income, while for males the opposite pattern holds. This may, to some extent, be due to the sex-role stereotyping--and associated differences in salary structure--of working-class jobs for males and females. Male-dominated working class jobs tend to be higher paying than female-dominated working class jobs due to the influence of unions. The effect of unions on the wage structure affects the prediction of the salaries of white-collar jobs in this main-effects model. Figure 21b suggests that an interactive model might be correct--the sex difference in wages differs as a function of educational level.

To summarize, appropriate uses of segmented bar graphs and pie charts are seen as very limited. They do not direct the reader's attention to the effect of one variable on another. The techniques utilized by the modern graphical methods--quantifying the effect of the grouping variables and computing residuals from a model--are, in fact, powerful analytic devices, and their value in graphical methods cannot be overemphasized. Whenever the purpose of a graph is to convey a relationship rather than a tabulation to the reader, such methods described in this section are recommended.

Displays of Two Continuous Variables

In this section we will present ways of displaying two continuous variables. The use of scatterplots to display two continuous variables will be described. These kinds of displays enable us to see the relationship between the variables.

Diagnostic and enhanced scatterplots. The scatterplot is an effective way of displaying the relationship between two continuous variables. An example of a scatterplot is shown in Figure 22.

This example, which is taken from the 1980 Condition of Education, displays personal per-capita income by per-pupil expenditure. It is evident from this plot that the two variables of interest are highly correlated. The labeling of the states enables us to learn even more about the data. For example, while Illinois, Connecticut, and New Jersey all had the same per-capita income, they differed widely in their per-pupil expenditures.

In certain situations, a simple scatterplot, one containing points of identical size, will miss critical information about the data. Such an example is graphed in Figure 23. This graph presents a scatterplot of the proportion of college applicants that are women plotted against the proportion of applicants admitted to the college. It is clear from this graph that the large negative correlation is due almost entirely to the low admission rates for women by the larger departments (box size is determined by the relative number of applicants). Had the data points been of identical size, vital information about the relationship under investigation would have been lost.

Another enhancement to the scatterplot is contributed by Bachi's graphic rational patterns described earlier. Figure 24 uses GRPs to display the frequency of distribution of marriages in the U.S. by age of brides and grooms. The data are categorized first before being plotted. Use of GRPs provides highly accurate information about the counts as well as specific information about the relationship.

Detecting outliers. Devlin, Gnanadesikan, and Kettenring (1975) augment scatterplots with influence contour functions. This is a diagnostic tool for assessing the impact of outliers on correlations. Two illustrations of this technique are shown in Figure 25.

In Panel 1, a computer-generated bivariate normal sample (N=60, r=.836) is plotted with the addition of one outlier (noted by the '1'). This point was moved about the axes as indicated by the contours of the influence functions. The contours indicate approximately how much the correlation would be increased or decreased by removal of the sample point in each location. The plot suggests that the correlation would increase by a little more than

.06 if the point were omitted entirely, and no single observation would affect the correlation by more than .03.

Panel 2 of Figure 25 presents a scatterplot (n=50, r=.730) of the logarithms of sepal lengths by sepal widths for iris plants. Points 16 and 23 have opposite effects on the correlation. Were both points removed, the correlation would remain unchanged. If only one point were removed, the correlation would change by about .025. While Point 42 appears to stand out visually as an outlier, its influence on the overall correlation is relatively small--about .025.

Regression diagnostics. More sophisticated uses of the scatterplot occur in the context of regression analysis. In addition to the normal probability plot of residuals discussed earlier, good regression programs, such as the BMDP series and the new SAS regression procedures in the 1982 version of SAS, contain additional diagnostic plots. Plots of residuals against predicted values are also available, as well as plots of each independent variable against the observed and predicted values of the dependent variable. Also, each predictor can be plotted against the residuals. The following technique, which is a variation on this theme, is an especially useful diagnostic tool.

Larsen and McCleary (1972) define the partial residual plot in the following way: The partial residual plot displays the relationship of a specific independent variable to the dependent variable controlling for the effects of all the other predictors in the equation. That is, the slope of the best fitting straight line through the partial residual plot equals the corresponding partial regression coefficient in the multiple regression equation. Because of this property, the analyst can easily see the effects of outliers and heterogeneity of variance on each regression coefficient in the model. Thus, a set of partial residual plots conveys more specific information than a single "overall" plot of fitted values against residuals. The BMDP series of regression programs contains fac for partial residual plots in addition to the diagnostic plots discussed above.

Figure 26 presents a comparison of the regular residuals versus X_1 (Panel 1), X_2 (Panel 2), and the corresponding partial residual plots

(Panels 3 and 4). Panel 1 does not indicate that any assumptions of least squares regression have been violated with respect to X_1. Panel 2 displays evidence that the variance is increasing with X_2. The partial plots in Panels 3 and 4 confirm these observations. Panels 3 and 4, respectively, also show only a slight positive correlation with X_1 but a definite negative correlation with X_2.

Methods for examining rates of change. In longitudinal studies, one is interested in studying changes over time. A key summary statistic is the average rate of change, obtained from the regression of some dependent variable on time. The graphical analogue to the rate of change is the best fitting straight line.

An example taken from the 1982 Condition of Education to illustrate this point plots per-pupil expenditures (in constant dollars) over time. Figure 27 and Table 27 present the graph and accompanying data (from Chart 2.10 and Table 2.10, pp. 63 and 62, respectively). An examination of the graph suggests that the rate of increase of expenditures is fairly stable but may not be constant. An effective analytic strategy is to fit a simple linear regression model and plot the datapoints together with the regression line and residuals. This was done using the data in Table 27.

From the regression analysis, it was learned that the average rate of change is $58 per year. The plot of expenditures against time is shown in Figure 28. From this graph, it is clear that a simple linear model fits the data well, but the residuals show a systematic pattern. They suggest that the rate of change was faster than average initially and then decelerated. A better statistical fit could be obtained by fitting a more complex regression model, such as a quadratic model, but for purposes of communication with a lay audience, the simple linear model, which reports the average rate of change, together with a description of the patern of residuals may be sufficient.

It is also noted that an alternative indicator of change—percentage change from year-to-year—is not recommended. The percentage change has two major drawbacks. First, it uses the data in an inefficient way by considering successive two-year "pieces" separately. It is preferable to use a single procedure to estimate change from all of the data. The latter is descriptive of change over the entire interval of time under study and, in general, will produce more precise estimates than any procedure that only looks at "pieces" of the data. Second, and more important, the percentage change estimate is confounded with initial level in a way that introduces serious statistical artifacts. The effect of these artifacts is immediately apparent from a comparison of the two plots in Figure 27. The graph of percentage changes is a complex pattern of peaks and valleys and certainly does not convey the fact, apparent from the top graph, that the rate of change is actually quite stable. Other recommended approaches to the measurement of change in multiwave data are discussed in Rogosa, Brandt, and Zimowski (1982).

Assessing changes in the rate of change. Because a complex multiparameter regression equation is difficult to communicate to a lay audience, Tukey (1977) presents a simpler analysis that uses data similar to those examined here. His example plots the population of the U.S.A. from 1800 to 1950 in 50-year increments. The basic plot of these data is found in Panel 1 of Figure 29. Tukey observes that the population increases in the early years were accelerated--possibly at a constant rate per year, while the growth in the later years may have been by about the same number of people each year.

To check the constant rate of growth hypothesis, he converts the population data to logarithms and plots the results (as shown in Panel 2 of Figure 29). A straight line is easily fitted to the data points from the early years (as shown in Panel 4), which confirms the constant rate of growth hypothesis. Since .01 in log is about 2.3 percent in size, we can see that population growth during the nineteenth century increased at a constant rate of about 2.3 percent per year.

Looking again at the (linear scale) data in Panel 1, we see that the later years appear to be marked by linear growth. Panel 3 provides a linear fit for these data. While certain years differ from the expected fit (e.g., 1920-40), a linear model does appear to provide a good explanation for population increases in the later years.

This example is provided to illustrate that, while the initial graphs may provide some information, greater explanatory power can be obtained through additional calculations--calculations that do not need to be complex. In Tukey's example, all of his calculations were done simply by hand and noted on the graphs. This greater amount of information can be presented in the text, even if the decision were made not to graph the best-fitting lines.

Displays of Relationships among Three or More Variables

In this section, we extend our discussion to relationships among three or more variables. Specifically, we will look at displaying three types of data: (a) one continuous variable with three or more stratification variables, (b) interrelationships among several continuous variables, and (c) two or more continuous variables stratified into two two or more groups. The additional problem posed by all three classes of data is that there are now more than two dimensions to be plotted. Thus, some sort of dimension reduction must take place before plotting on a piece of paper.

Displays of One Continuous Variable with Three or More Grouping Variables

An extension of the technique of graphing difference scores that takes into account five factors is used by Levy and Reid in their study of the effect of human brain organization on handedness and handwriting posture (published in the Journal of Experimental Psychology). In their experiment, they examined visual field superiority as a function of sex, handedness, writing posture, and type of test. The display of the findings, which is shown in Figure 30, is an excellent example of how to present multifactor data effectively in two dimensions. Before describing the noteworthy features of the graph, we will first present some of the study's methods and hypotheses in order to provide a foundation for appreciation of this display.

Levy and Reid demonstrate that lateralization of the brain for verbal and spatial ability can be predicted nearly perfectly from a subject's handedness and handwriting posture. Left and right handedness is considered as well as two writing postures: the "normal" posture in which the pen is pointed away from the body, and "inverted" in which the pen is pointed toward the body. Both males and females were tested tachistoscopically on two measures: a syllable test that measures verbal ability and a dot location test that measures spatial ability. Each hemisphere of the brain was assessed individually by presenting the stimulus to each eye separately so that only one hemisphere of the brain could preceive the image. Their results show that females are less lateralized than males (i.e., females have smaller differences in performance on the same test between the left and right hemispheres), but the direction of laterality is the same (i.e., the same hemisphere is superior in both sexes on a particular test). The main hypothesis, which is strongly supported by the data, concerns an interaction between handedness, handwriting posture, and type of test (verbal vs. spatial). This hypothesis is depicted in Figure 31, taken from Levy and Reid (p. 122).

The diagram in Figure 31 illustrates the hypothesized effect of laterality and use of the crossed vs. uncrossed pyramidal tracts in writing. Simply stated, the hypothesis is that mirror-imaged handwriting postures indicate mirror-imaged patterns of brain organization. Patterns of brain organization, indicated on difference scores on the syllable and dot location tests, are plotted as a function of handedness, handwriting posture, and sex in Figure 30.

This graph is particularly effective because the empirical support for the hypothesized mirror-imaged relationship is immediately obvious. From a substantive point of view, it is especially noteworthy that the single right-handed inverted subject had the predicted opposite pattern of lateralization from the right-handed normals. In fact, the graph shows a complex relationship among five factors--handedness, handwriting posture, sex, visual field, and type of test--in a way that calls attention to the hypothesized relationship.

The graph also contains comparative information about the extent of the group differences through placement of the standard errors of the visual field differences as "whiskers" at the ends of the bars. Bars with non-overlapping whiskers are statistically significant. On the Syllable Test, for example, males and females in Group RN did not differ (note overlapping whiskers), while males and females in Group LI and LN did differ. The extent of these differences (p .05 for Group LI and p .01 for Group LN) is visually detected.

For data that do not conform to normality assumptions, Tukey's median polish technique, discussed in the preceeding section, can also be extended to handle cases with three or more factors. The process of removing row and column effects by taking out medians successively uutil zeroes and small residuals remain is described with examples (including graphic displays) by Tukey (1977; see Chapter 13). The procedure for graphing the level, main effects, and residuals discussed earlier can be generalized to three or more factors by plotting two-way "slices" (i.e., collapsing over the remaining ways of classification).

Displays of Interrelationships Among Several Variables

The scatterplot and related techniques discussed earlier are difficult to extend to the case of three or more variables. This, of course, is because three-dimensional configurations are difficult to plot and comprehend; while four and higher dimensional "plots" are not possible to depict. Two basic approaches to the problem of graphing multidimensional data are available. The first class of techniques involves some statistical or graphical dimension reduction, so that most of the multidimensional information is summarized into two (or perhaps three) dimensions. The second class of techniques depends upon a method for looking at several different two- or three-dimensional subspaces ("slices") of the complete multidimensional space using specialized graphics hardware. In the first instance, principal components analysis, or a similar technique, is used to accomplish the dimension reduction; in the latter instance, two-dimensional "slices" of a multidimensional space can readily be graphed. Such graphs, for example, are available in most of the popular factor analysis programs for plotting the factor pattern matrix. In this section, we first review techniques for plotting multidimensional data using statistical methods of dimension reduction. We then discuss a specialized method of looking at three dimensional subspaces of multidimensional data.

Biplot. An interesting and useful method for displaying multidimensional data has been proposed by Gabriel (1971) and is known as the biplot. The method is relevant when the units of analysis have some interpretation (e.g., states, countries, universities). It is less applicable when the units of analysis are randomly sampled and the individual observations have no particular meaning (e.g., a sample of college sophomores). The rows of the data matrix to be graphed correspond to the units of analysis, and the columns correspond to their attributes (e.g., population, finances, enrollment). The basic idea of the biplot is to show which units have which attributes. That is, the units and attributes are plotted jointly. The name biplot refers to this property rather than to the dimensionality of the plot. Two-dimensional biplots are obviously easiest to plot, but three-dimensional biplots are feasible and recommended if three dimensions are needed to summarize the information in the data matrix.

The calculations involved in the dimension reduction, although unfamiliar to most practitioners, are extremely straightforward and can be implemented using SAS with only a few lines of code (PROC MATRIX statements). After describing the method, we present an example of the biplot using some demographic data collected on states. Table 32 contains the control statements used in the example that follows.

The key calculation utilizes a classical method of factoring data matrices known as the singular value decomposition, which is a built-in function in PROC MATRIX. The technique is, in fact, a close relative of principal components analysis, and an alternate algorithm for the same technique that utilizes principal components is described in Everett (1978, pp. 22-24). The remaining calculations necessary to obtain the numbers to be displayed in the biplot are obtained via some straightforward multiplication and division. For an N X q data matrix (e.g., N states by q attributes), one obtains via this technique a set of N two-dimensional vectors, each corresponding to a unit, and a set of q two-dimensional vectors, each corresponding to an attribute. These sets of vectors are then plotted jointly on a two-dimensional axis. This plot enables the reader to associate attributes with units.

In the example presented here to illustrate this method, demographic data (eight variables) are presented for a nationally representative sample of twenty states. The variables are:

1. Total state revenue for education
2. Mean educational per-pupil expenditure
3. Amount of funds for state administration of the ESEA Title I program
4. Number of school districts in the state
5. Population of the state
6. Population density in the state
7. Number of cities in the state greater than 100,000
8. Number of cities in the state greater than 25,000

The corresponding biplot is shown in Figure 33. The vectors corresponding to the variables are drawn, while only the endpoints corresponding to the states are shown.

To interpret the biplot, one first characterizes the dimensions in terms of variables. All but two of the variables (density and mean per-pupil expenditures) cluster together along the X-axis. Their clustering indicates that they are highly intercorrelated. Since these variables are indicators of size of the state, in terms of both population and revenue, the first dimension can be thought of as a general "resources" or "size of state" factor. Thus, it is not surprising that the states at one end of this dimension are New York and California, while Vermont, Alaska, and Delaware are at the other end. The second dimension is defined mainly by the single variable, population density. Massachusetts lies at the positive extreme, while Alaska, of course, lies at the negative extreme at the bottom left quadrant of the plot. The angle between the vector corresponding to density and the ones defining the first factor indicates that density is not correlated with the latter group of variables (i.e., its vector is nearly orthogonal to the others).

To investigate the possibility that a third dimension would convey additional information, a third component was retained from the singular value decomposition. This dimension turned out to be defined by the last two variables, number of cities greater than 100,000 and 25,000, respectively, and, to some extent, by the number of school districts. Thus, this dimension appears to be a measure of urban size. California and Hawaii (which is an unusual case) are at the postive end of this dimension, while Arkansas and Tennessee are at the other end. Because this component was quite small relative to the first two, it did not appear worthwhile to graph.

Motion graphics. In contrast to the above technique, which is widely applicable for purposes of publication, there has been some specialized work on displaying more than three dimensions on a video terminal that depends on costly hardware and software. Thus, these techniques, while useful for exploratory data analysis, are not applicable for purposes of publication. One technique for generating plots of four or more dimensions is called motion graphics, a technique pioneered by Tukey approximately ten years ago.

Tukey's purpose was to study how to use interactive computer graphics in statistics. His first developed system, PRIM-9 (Picturing Rotation, Isolation, and Masking), used real-time motion graphics to explore data up to nine dimensions. This original system, while well received, was not practical for widespread use due to the prohibitive cost of the hardware. The most recent version of the system, ORION I, which was developed by Jerome Friedman and Werner Stuetzle, differs from the earlier PRIM version through the use of modern and relatively inexpensive raster graphics and microprocessor technology. (See Kolata [1982] and Friedman, McDonald, & Stuetzle [1982] for more details of this method.)

The basic requirement for real-time motion graphics is the ability for a computer to compute and draw new pictures fast enough to give the illusion of continuous motion. Five pictures per second is a minimally acceptable rate; a rate of ten to twenty pictures per second provides smoother motion. The selection of particular colors to represent distances enhances the effects. By continuously rotating three-dimensional scatterplots containing 100 to 1,000 points and displaying the moving projection of the object onto the screen, the point clouds appear to move through space. Analysts can use the graphical displays to look for natural groups or clusters occurring in the data, or nonuniformities in the distribution of points.

In an example recently presented in Science (Kolata, 1982), approximately 500 points, which represent the census districts in a metropolitan city, are projected onto a computer screen. Fourteen variables are measured for each district, such as average values for homes, the crime rate, racial composition, student-teacher ratios in public schools, and so on. These 14-dimensional points are projected onto particular three-dimensional spaces. Moving the space around the point clouds results in the points being projected at different angles. An interesting structure in one view can then be followed in other views of the same data.

Multivariate Displays--Two or More Groups

Another approach to the problem of reducing the dimensionality of multidimensional data is to use a graphical rather than a statistical method of dimension reduction. These techniques are especially useful in comparing "profiles" or sets of variables.

Taking two or more variables and mapping them into two dimensions can be done using a number of graphical techniques, such as tables, pie charts, glyphs, faces, and maps. Each of these techniques will be illustrated below. Trees and castles, which are specialized plots used to display cluster analyses, will conclude this section.

Tables. Wainer (1981; Wainer and Thissen, 1981) suggests that properly constructed tables can be an effective way of mapping multivariate data into two dimensions. Reviewing work done by Ehrenberg, he cites several rules for analysts to use to prepare effective displays of multivariate data.

Among these are:

- o the data values should be rounded,
- o the rows and columns should be ordered by some aspect of the data,
- o the data should be spaced and separated on the page to facilitate ease of grouping similar variables, and
- o suitable summary statistics should be added to the display (e.g., column medians) to facilitate group comparisons.

Figure 34 presents a set of data "before" and "after" use of this method.

Note that, in the top panel, the data are difficult to interpret--they are ordered alphabetically, the numbers are presented in varying degrees of accuracy, and so on. In the bottom panel, the data are much easier to interpret--the values are rounded, states are reordered according to one of the variables (here, life expectancy), and column medians are added. Space is left to separate visually Alabama and Mississippi from the rest of the group, since these states are observed to be inferior to the other states on the first five indicators.

Experimental evidence from human performance studies cited by these researchers suggests that tables such as these, which display data sets of modest size and complexity, are extremely effective.

Pie chart. Figure 35 presents a comparison of three methods of displaying meat production in Western Europe (from Wainer and Thissen, 1981). In this example, there are five continuous variables--percentage meat production for each of five types of meat. From the pie chart in Panel 1, it is difficult to compare the types of meats produced by each country. That is, pie charts have the same disadvantages as segmented bar charts, discussed in an earlier section. Panel 3, which presents the same data in profile form, is preferred to the pie chart, because it provides more information. For example, meat production is displayed by country. In addition, for each country, the two types of meat most frequently produced are noted in solid color. While this information could have been extrapolated from the pie chart, it is more readily visible in the profile form.

Glyphs. A central circle with rays coming out of its centroid, in which each ray represents one variable to be displayed and its length represents the value of that variable, is called a star glyph. An example of a star glyph for the data presented in Figure 34 is shown in Figure 36. From these displays, Alabama and Mississippi can again be seen to differ visually from the other states.

Star glyphs provide a visual means of clustering data, but this method cannot handle many variables. Even with seven variables, it is difficult to keep track of each and to interpret why a state appears the way it does.

Faces. Herman Chernoff (1973) proposed that features of the human face could be used to indicate the value of each of several variables. The basic idea is to assign each variable to a feature of the face (e.g., population to length of nose, revenue to size of mouth, etc.).

Use of Chernoff faces allows us to represent many variables by using features of a face. For example, using the data presented in Figure 34, population can be represented by the number of faces per state; the literacy rate can be represented by the size of the eyes (the bigger the eye, the higher the rate); percent of graduates can be represented by the slant of the eyes (the more slanted, the higher the percentage); life expectancy can be represented by the length of the mouth (the longer the mouth, the higher the life expectancy); income by the curvature of the mouth; homicide rate by the width of the nose; and temperature by the shape of the face. These features can be displayed as shown in Figure 37.

Chernoff faces have also been used to summarize complex integrated circuit data. Hahn, Morgan, & Lorensen (1983) added another dimension to the faces--color. Each face depicts a different circuit; the 20 measured performance variables represented on each unit are depicted by the different facial features. Colors are used to indicate the extent to which the units meet the 20 specified production standards. Green facial features are "good" (that is, all 20 measured performance variables meet the design limits), blue features are "marginal" (that is, one to nine of the performance variables are outside the design limits), and red are "defections" (that is, ten or more of the performance variables are outside the design limits).

Chernoff hoped that the face would be an effective means of expressing multivariate data, because the human brain is especially sensitive to subtle variations in the appearance of human faces. However, a disadvantage of the method is that the appearance of the faces depends crucially on the assignment of variable to feature. Different assignment rules may result in different interpretations of the same data. Also, certain features of a face have internalized meanings, some of which trigger emotional reactions, that may be incongruent with the "message" in the dataset. For example, the meaning of a smile or a frown on a human face has a deeply rooted interpretation, and that may not be compatible with the meaning of the variable assigned to "mouth" in the Chernoff face.

Clearly, application of the star glyph and Chernoff faces create interpretation problems of their own. More "down to earth" procedures may be preferred.

Maps. The Bureau of the Census has developed a method that utlilizes vividly colored maps to display two statistical variables on top of two geographical variables. One variable is represented on the map by increasing saturations of blue, with lowest level of saturation being yellow; the second variable using increasing saturations of red, again with the lowest level being yellow. The relationship among the variables is represented by overlaying one color scheme on the other, which results in various color mixtures of yellow, orange, green, and purple. This technique is described in detail by DesJardins (1982). One example, taken from the Bureau's Social Indicators III, depicts average per-pupil costs in education as a percent of the per-capita income. The color key chart, which shows the resulting bivariate mapping, is redrawn in Figure 38 for informational purposes (due to the difficulties in xeroxing colors).

This bivariate graphing method using the two-variable color map has received considerable attention in recent years. Research in the area of human performance has shown that these displays are difficult to comprehend. Several recommendations for improving the displays have arisen from these investigations. One of the proposed improvements concerns the class value assignment of colors. One suggestion proposes substituting white for yellow. This change improves the logic of the color assignment (making the four corners white, dark blue, purple, and dark red) and makes it easier for the user to associate values and colors. Another propoosal involves use of a single color plus shading gradient to facilitate review. Another criticism concerns the impact of the choice of colors. For example, for the same size area, red is perceived as large· than blue. It is also statistically misleading to color states with large land ares and small populations. In one example cited by DesJardins, he points out that the large midwestern areas, which have primarily small populations, tend to overshadow the smaller, more populous states on whatever variable is of interest. The Bureau of the Census is apparently working to enhance multi-level graphics systems, of which the two-variable color maps are one example.

Trees and castles. The use of trees and castles as a graphic method for displaying the results from cluster analysis is described by Kleiner and Hartigan (1981). A hierarchical clustering of the variables is carried out first using the complete linkage method with Euclidean distance between variables to join each pair of clusters. Variables in the same cluster will be assigned to the same part of the tree. The thickness of the trunk is dependent upon the number of branches contributing to it.

One disadvantage of the tree method is that it is difficult to compare values within the same tree, even if they are close together. A castle method, also described by Kleiner and Hartigan (1981), facilitates comparison between variables in the same tree. A castle is a combination tree and profile. A hierarchical clustering of the variables is performed using complete linkages as with the tree method. While 50 variables is the maximum that can be successfully represented by trees, castles can potentially portray many more variables, especially if the clustering combines them into a few well defined groups.

While trees offer a better overall impression than castles (e.g., general size effects, change of clusters over time, and outliers), castles give more detail with a jagged appearance (e.g., extreme values and differences between values are easier to detect).

Figure 39 presents an example of the trees and castles.

Computer Programs for Graphically Displaying Data

In this section, the computer programs that implement some of the techniques discussed in this paper will be summarized. The omission of many graphic displays of statistical data will be singularly apparent. Some of the most advanced, and therefore most valuable techniques have yet to be programmed, and many others are available only in stand-alone programs that do not enjoy wide usage. The major software packages feature the most primitive methods--pie charts, segmented bar charts, and line graphs; only a small subset of the more advanced techniques are included. Among the program packages, BMPD contains the strongest collection of graphical techniques, SAS contains fewer advanced techniques (the primitive methods are very well covered), SPSS contains very few graphical procedures, and OSIRIS contains virtually nothing beyond a scatterplot program. The features of these and other programs are described below.

The appendix of this report will review the hardware and software currently in existence for producing formal presentation graphics. These programs differ from those covered in this section in that they are designed to produce presentation quality graphics only. These programs contain even fewer advanced graphical techniques than the package programs. Thus, they are not suited for exploratory data analysis and offer the researcher only a very limited choice of graphical techniques.

Most researchers and analysts have access to one or more of the commonly used statistical data analysis packages: SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences), or BMDP (Biomedical Computer Programs). A review of these packages yields several, rather disheartening, conclusions for analysts interested in graphically displaying their data:

- These packages have limited facilities for graphing and displaying data. SAS has the "prettiest" displays of the three, but none of these can compare in presentation to the most primitive of the programs discussed in the appendix.

- Many of the graphic analyses described in this report, which provide analysts with more information about their data, are not included in these packages.

- Even the packages that do include some of the features discussed in the previous section of this report have not provided the variations on that feature that enhance its use. For example, none of the stem-and-leaf programs have options for "squeezed" or "stretched" displays. An unfortunate consequence is that nearly all the values of a variable will be graphed onto one stem. Similarly, none of the box-and-whisker plot programs permit variable-width displays or side-by-side comparisons of several groups.

There are, however, a number of other computer programs and systems available at cost that do provide interested users with programs that carry out some of the functions discussed in this report. In fact, many of these specialty programs are specifically designed to provide graphical displays for use in analyzing data. Five such programs are:

- EXPAK (Exploratory Data Analysis Package), a FORTRAN package available from:

  International Educational Services
  Suite 829
  1525 East 53rd Street
  Chicago, IL 60615

  EXPAK is based on recent developments that facilitate exploratory data analysis by displaying data and by computing summary statistics. Some of the routines include stem-and-leaf displays, univariate summary statistics, scatterplots in which the data points are labelled rather than simply identified, three estimates of bivariate correlation, and median polish.

- Exploratory Data Analysis Package, an APL package available from:

  The Analysis Center
  The Wharton School
  3609 Locust Walk C9
  University of Pennsylvania
  Philadelphia, PA 19104

  This package is designed for the interactive analysis of data using the techniques developed by Tukey and others associated with exploratory data analysis. Some of the routines include stem-and-leaf displays, box-and-whisker plots (including variable widths and notches), comparison box plots, scatterplots (including P-P and Q-Q plots), and median polish.

- The ABCs of EDA, a book written by Paul Velleman and David Hoeglin (1981), contains FORTRAN and BASIC subroutine programs of Tukey's (1977) exploratory data analysis techniques. The routines include stem-and-leaf displays (including five-number summaries), box-and-whisker plots (including notches), condensed six-line plots, resistant line (simple regression), nominal data smoothing, coded tables, histograms (including suspended rootograms), and median polish. All of the computer subroutines included in the book are also available in the MINITAB statistical package (Release 81.1 and later).

  Both the BASIC and FORTRAN subroutines are available in machine-readible form from :

  CONDUIT
  P.O. Box 388
  Iowa City, IA 52244

  The FORTRAN subroutines are also available in machine-readable form from:

  International Mathematical and Statistical Libraries, Inc.
  GNB Building, Sixth Floor
  7500 Bellaire Boulevard
  Houston, TX 77036

- STATGRAPHICS, an interactive statistical graphics system in APL, is available from:

  Statistical Graphics Corporation
  P.O. Box 1558
  Princeton, NJ 08540

  STATGRAPHICS is a comprehensive package of general statistical, graphical, and mathematical functions designed to provide graphical displays. It is designed for users that have access to CRT terminals capable of graphic display. Its routines include three-dimensional displays and many of Tukey's exploratory data analysis procedures, such as stem-and-leaf displays, box-and-whisker plots (including notches), pie charts, bar charts, median polish, and partial residual plots.

- S Package, an interactive language and system using the UNIX operating system (of Bell Laboratories) is available from:

  Bell Laboratories
  Computer Information Service
  600 Mountain Avenue
  Murray Hill, NJ 07974

  S Package, which is not externally supported by Bell Laboratories, includes a number of statistical graphics software routines, including three-dimensionl plots, pie charts, histograms, scatterplots (including probability plots), residual plots, non-parametric statistics, and symbol plots (such as star glyphs).

Table 40 compares the features of these eight packages with respect to the analyses and graphical displays discussed in this report. Further information about the last five programs can be obtained from the authors of this report.

In addition to these packages, several special subroutines and programs also exist that handle only one or two analyses or displays. A sampling of these programs includes:

o FACES, a FORTRAN program designed to represent Chernoff's faces, is available from:

  Danny Turner or Eugene Tidmore
  Department of Mathematics
  Baylor University
  Waco, TX 76703

  This program handles up to twelve variables per face.

o Trees and Castles, a FORTRAN program with a separate plot program (GR-Z), is available from:

  Bell Laboratories
  Computing Information Service
  600 Mountain Avenue
  Murray Hill, NJ 07974

  The algorithm for generating the trees and castles is available separately as a FORTRAN program.

o CANDEC, a FORTRAN IV program designed to carry out singular value decomposition for various types of input matrices, is available from:

  K. R. Gabriel
  The Hebrew University
  Jerusalem, Israel

  This program is designed to implement the biplot procedures discussed in Gabriel (1971).

- ORION I, an experimental computer graphics system that uses interactive motion graphics, is available from:

  Jerome Friedman and Werner Stuetzle
  Computation Research Group
  Stanford Linear Accelerator Center and
  Stanford University--Statistics Department
  Stanford, CA 94305

  This program uses real-time motion graphics to display three-dimensional scatterplots on a two-dimensional screen by rotating the objects in space.

- Other computer programs that are designed specifically for graphing multivariate data are listed in Everitt (1978).

Summary and Recommendations

The main body of this paper reviewed many of the "state-of-the-art" techniques that we believe are especially relevant for the graphical presentation of educational data in publications such as Condition of Education. In particular, the section of the paper focusing on graphing the relationship of one or two stratification variables and a continuous variable is directly relevant to a large majority of the graphs in Condition of Education each year. Our conclusion, however, is that the commonly used and widely available techniques for plotting such data--the segmented bar chart and the pie chart--are far less than optimal. In brief, they are descriptive rather than inferential, and, thus, do little more than restate the tabulations that are typically presented adjacent to the graphs.

We recommend the wider use of the kinds of "inferential" graphs discussed in this paper to complement statistical methods of data analysis and for formal presentation. These techniques feature two of the most powerful tools used in statistical methods--model fitting and the examination of residuals--and the modern work in graphical techniques has demonstrated that these two tools are as important in graphical methods as they have proven to be in statistics. We recommend that these types of plots be routinely used in the preparation of NCES reports so that trends in the data can be detected and clearly displayed.

For such recommendations to be feasibly implemented, it is necessary to have these techniques available on the computer, both for use in exploratory data analysis and for formal presentation graphics. As is discussed elsewhere in the paper, several such methods have been implemented only in standalone programs, while others have not been implemented at all. In the latter category, the unified method of plotting the effects of two stratification variables on a continuous variable is the most notable example. Realistically, we are aware that widespread use of state-of-the-art graphical methods depends crucially on their availability in a readily accessable package program. For NCES and most of the rest of the educational research community, this means SAS. We believe that the modern methods discussed here belong in both the "regular" SAS package and SAS GRAPH. The former is necessary to carry out the necessary exploratory data analysis in the preparation of NCES reports, while the latter, of course, is necessary for the preparation of the final, publishable, plots. Such enhancements to SAS and to SAS GRAPH may need some support from public and private agencies, particularly those interested in using such packages for data analysis and displays.

APPENDIX

HARDWARE AND SOFTWARE FOR PRESENTATION-QUALITY GRAPHICS

The presentation of graphics depends on both the equipment and software. Terminals, displays, and self-contained computers allow the user to examine the graphics presentation before hard copy is produced. Most of these pieces of equipment have a one-color view screen; however, newer models are being constructed with color available.

The actual production of tables, charts, and overhead transparencies, depends on the use of printers and plotters. Plotters have color capabilities so that graphics displayed on a terminal in black and white can be produced in color. An important advantage of a color display is that the user can experiment with various color options before producing the hard copy.

Software packages enable the user to create the graphic designs. Almost all software packages are capable of producing pie, line, and bar charts. The individual packages vary in their ability to produce additional graphs, maps, line types, cross hatchings, and curves. Several software packages have the capability to produce intricate schematic drawings or architectural renderings. These are of use to engineers, draftspersons, and architects, but of limited use in displaying statistical results. Many packages are compatible with a fairly wide variety of equipment

This section of the paper has two main subsections: the equipment available for producing presentation graphics, and the software available for generating the graphics. In each section, brief descriptions of the various equipment and software packages available will be given. It must be pointed out that this is in no way an exhaustive list of equipment and software for the production of graphics.

Several graphical software packages make it possible for the user to create plots of publishable quality when used with the appropriate specialized hardware. This is in contrast to the crude diagnostic plots available in the statistical packages (SAS, SPSS, BMDP). This section reviews the specialized hardware and associated software.

Hardware Available for Presentation Graphics

The equipment available for presentation graphics is of two basic types: computers and terminals, and printers and plotters. The former are used to desigu the graphics and allow the user to preview the graphic options. The latter are used to produce the actual hard-copy products.

Computers and Terminals

This section will deal with the products of three main companies: Hewlett-Packard, Tektronix, and Digital Engineering. The first two companies have many different products available and considerable diversity within a product line.

Hewlett-Packard Products

Hewlett-Packard's major graphics terminals are the 2647A ("intelligent graphics terminal") and 2648A ("graphics terminal"). They can be hooked up to Hewlett Packard computers, or to virtually any mainframe computer (including an IBM) using an RS232 interface.

Both of these graphics terminals have independent graphic and alphanumeric display memories. They can plot user-generated tabular data automatically. A menu is provided to lead the user through a question-andanswer session about the data. Once the data parameters are defined, the data are plotted by a single key stroke. The resolution for both terminals is 720 x 360 (viewable and addressable points). The viewing screen is monochrome, but the terminal can be tied into HP and other plotters for color output.

The 2623A is HP's low-cost graphics terminal. Its resolution is somewhat lower than the two previous terminals (512 x 390 viewable and addressable points). It has an optimal built-in graphics printer or can be tied into other peripherals.

HP's desktop computer lines, the 9836A and the 9845 series, have been designed with greater emphasis on the production of graphics. The 9836 desktop computer can provide sophisticated graphics. It is primarily a monochrome system, but can produce color graphics if used in conjunction with a separate color monitor. The resolution is 512 x 390 dots. The 9845 desktop can be selected with a choice of monochromatic or color screens. The 9845 color model produces 4,913 color shades and has a resolution of 560 x 455 viewable and addressable points. There is an internal printer which provides high-resolution hard copy output which precisely replicates the image on the CRT--dot for dot. However, the internal printer produces hard copy in black and white only. For color hard copy, an additional piece of equipment--such as the HP 9872 Eight-Pen Plotter--is required.

In addition, all of the basic Hewlett-Packard computer lines can be used to generate graphics using appropriate software and peripherals. This includes the 80 series computer line, the 125 computer line, and the major 3000 line.

Tektronix Products

There are several Tektronix terminal and display models. The two basic series are the 4010 Series and the 4110 Series. In the 4010 Series, models 4010, 4012, 4014, and 4015 have 1024 x 1024 addressable points and 1024 x 780 viewable points. The 4016 model is somewhat larger and provides 4096 x 4096 addressable points and 4096 x 3120 viewable points. It also can produce straight, dotted, and dashed lines. All of these models are capable of producing four colors on the viewing screen.

The 4113 computer display terminal in the 4110 series has has a color display screen. The 4113 screen can display 8 colors at one time out of a palate of 4016 colors: an optional bitplane is available to expand the number of colors displayed to 16. Colors are selected by specifying hue, lightness, and saturation. The viewable resolution is 640 x 480; however, there is an addressable display space of 4096 x 4096. Sections of the screen may be rotated, scaled, or moved on command. Key-activated zoom and pan allows the user to access the much larger addressable space. The user can also control a graphic cursor via thumbwheels on the keyboard. Solid and dashed lines are available.

Tektronix also produces another color display terminal, model 4027, which has a resolution of 640 x 480 viewable and addressable points. It can display 8 colors at one time out of a palate of 64.

All Tektronix display terminals can be linked to virtually all mainframe computers using an RS232 interface.

Tektronix also has a line of desktop computers. The desktop computer is the 4050 series--the 4051, 4052, and 4054. The 4054 has the greatest graphic capabilities. It has a 19-inch display screen. There are 4000 x 4000 addressable points and 4000 x 3000 viewable points. This model has several graphic and alphanumeric enhancements built in. These are four character sizes, eight language and character fonts, a full screen crosshair cursor positioned by thumbwheels for drawing complex objects, and dot-dash vectors that provide distinct line differentiation and allow the creation of shaded patterns. The user can work interactively with selected parts of a graphic display without affecting the rest of the display, and refreshed objects can be moved around on the screen under thumbwheel or program control.

Other Products

While Hewlett-Packard and Tektronix are perhaps the largest producers of graphics hardware, other manufacturers also produce graphics equipment. For example, Digital Engineering produces Retro Graphics Enhancements for DEC terminals. Once retrofitted, the DEC terminals can operate both as full-featured graphics terminals and as standard alphanumeric terminals. These terminals are compatible with many graphics software packages. The original model (VT 640) has a resolution of 640 x 480, the resolution on the newer model (VT 640S) is 640 x 240.

Graphics Printers and Plotters

Printers and plotters are the devices on which the actual hard copy graphics are produced, once they have been designed using software and a terminal or display. Printers and hard copy units reproduce the image on the screen in grey scale or black and white. Plotters produce graphics in color. For this reason, plotters will be focused on in this section.

Hewlett-Packard and Tektronix produce the widest range of these peripherals. HP and Tektronix product capabilities will be discussed first. However, other producers also have available equipment, and the Printer/Plotter of Versatec, a Xerox Company, will also be described briefly.

Hewlett-Packard Products

Hewlett-Packard produces a variety of graphics plotters. With various adapters provided by Hewlett-Packard, these peripherals can be hooked up to virtually any system. The model 7470A Graphics Plotter can plot 1000 points in a 1-inch line. Lines are plotted at 15 inches per second, and labels and annotations are drawn at speeds of up to six characters per second. There are two built-in pen stalls for easy two-color plotting; for more than two colors the machine can be stopped and new pens inserted. It plots on both paper and transparencies in sizes up to 8-1/2 x 11 inches. The smallest addressable increment for this plotter is .001 inch (.025 mm).

Hewlett-Packard also makes a single pen plotter (7225B) that can be converted to up to a ten-color plotter with manual changing. However, the most sophisticated plotters use eight pens, and Hewlett-Packard has several of these (7220C/7220T, 7221C/7221T, and 9872C/9872T). These plotters handle single sheets up to 11 x 17 inch size or smaller (no minimum). The T models have automatic paper advance and page stacking capability for unattended operation. The smallest addressable increment for these machines is also .001 inch (.025 mm).

Tektronix Products

Tektronix produces two interactive digital plotters--the 4662 and 4663. The 4662 has a maximum 10" x 15" plotting surface. By acquiring option 31 with the machine, it can plot in up to eight colors on paper, acetate for transparencies, or Mylar. Its smallest addressable increment is .005 inch (127 mm), and it has a printing speed of approximately 500 mm per second. It can produce curved lines, and there is a built in joy stick for digitizing. The Model 4663 is a two-pen plotter. Its smallest addressable increment is .001 inch (.025 mm) and can produce output of a larger size (up to 22" x 17"). These products can be hooked up to most systems, including IBM mainframes, using an RS232 interface.

Versatec Products

Versatec Printer/Plotters and Plotters can be linked to most mini and mainframe computers including IBM, Control Data, and Sperry Univac. The V-80 Printer/Plotter prints at a rate of 1,000 lines per minute, up to an 8-1/2" x 11" surface, on paper as well as transparencies. The Versatec Plotter can produce larger hardcopy (up to 72 inches wide). Both can plot 200 dots to the inch. The minimum line width is 7.5 mils with a minimum spacing between lines of 5 mils. These products produce hardcopy in black and white only.

Software Available for Presentation Graphics

The software package instructs the computer to create the desired graphics. The number of colors that a software program can control is dependent upon the capabilities of the plotting device. Eight colors can be programmed for eight-pen plotters, two for two-pen plotters, and so on. However, most software packages can instruct a single or two-pen plotter to stop and notify the user to change pen colors if more than two colors are desired.

Regrettably, these programs are sophisticated only with respect to the aesthetic quality of the plots; the plotting techniques themselves are extremely primitive, even in comparison with the SAS and BMDP statistical packages.

As with hardware, Hewlett-Packard and Tektronix are also major producers of graphics software. However there are several other highly visible software producers. The packages of the following software producers will be treated briefly in this section: ISSCO Graphics, Hewlett-Packard, Tektronix, SAS/GRAPH, STATGRAPHICS, Intelligent Graphics Systems, and Chart Master. They will be discussed in two sections: Software for Mainframe Computers and Software for Mini and Desktop Computers.

Software for Mainframe Computers

ISSCO Graphics (Integrated Software Systems Corporation) Software

Tell-A-Graf. Tell-A-Graf runs on all mainframe computers, and can be connected to 135 different graphic peripherals. It can send output to two devices concurrently. Tell-A-Graf generates a variety of graph types (line, bar, and pie), intricately tailored charts, as well as text pages. It uses the same data entry format for each type of graph so the user can experiment with different graph types until the best presentation is found. Several different graphs may be combined and positioned on the same page. It is designed for users who have little programming experience.

DISPLA. DISPLA is also an ISSCO Graphics product. It produces presentation quality charts and graphs from mainframe computers. It also produces graphics on virtually all graphics peripherals. It can produce line, bar, and pie charts, tables and text, maps, three-dimensional charts, and contour plots in any combination. It can also shade and thicken lines and can smooth curves. There are graph setup routines structured for calendar axes (years, quarters, months, days) as well as linear, logarithmic, semi-log, polar, user-labeled axes systems, or any mixture of these. Axes are automatically scaled and positioned by DISPLA or can be deleted or moved by the user. It has a feature for formula writing through which the user can draw equations with subscripts and superscripts and with standard Greek or mathematical symbols. Additional features include shading between two curves or between a curve and a plot boundary, and a subsystem for producing 3-D curves and surfaces as two-dimensional drawings. The latter capability indicates the axis drawing routines in either Cartesian or spherical coordinate systems.

Tektronix Software

Tektronix Plot 10 provides tools for the graphic and alphanumeric capabilities of Tektronix terminals and displays, which are connected with most computers including IBM mainframes. The Plot 10 software library contains several packages, including a terminal control system, plotter utility routines, easy graphing, an advanced graphing package, an interactive graphing package, and an interactive graphics library. It can be used for graphs and diagrams using solid or dashed lines, where size, shape and format can be dictated. It can also be used for contour maps, interactive design, and complex architectural renderings. Several sets of graphs and data can be displayed on the screen at once, or graphs can be superimposed in the same screen area. The plotter utility routines link the data base and terminal to the 4660 Series plotters for multicolored graphs, charts, maps, and renderings in up to 8 colors. Although designed for Tektronix printers and plotters, it can also be used with those of other manufacturers.

SAS/GRAPH

SAS/GRAPH is a software graphics system for producing color plots, charts, maps, and other displays. It runs on mainframe computers and produces output on almost all graphic peripherals. SAS/GRAPH programs are actually SAS procedures. Data values are entered into SAS data sets before SAS/GRAPH procedures use them. However, SAS/GRAPH is purchased separately and not available as part of the general SAS package.

Bar charts, pies, star glyphs, and bar graphs can be produced. For bar graphs, the user can specify a different color for each bar, and segmented bars can be produced with each subdivision having a different color or pattern. Up to 17 patterns are available; the number of colors depends on the capabilities of the graphics peripheral being used.

Graphs of one variable against another can be produced, and plots for each value of a third variable can be produced on the same graph. Points can be connected by straight lines, or a smoothing technique can be used. Axis colors, point colors, and line colors can be defined separately. Up to 32 solid and dashed line styles are available. Maps, contour plots, and three-dimensional plots can also be produced. All SAS/GRAPH procedures allow the pictures to be saved in a SAS data set that can be recalled. If SAS/GRAPH is used on a Tektronix 4027 terminal, it is possible to change colors in a display by pressing a key on the terminal and taking advantage of the intelligent features built into the software.

Statistical Graphics Corporation (STATGRAPHICS)

STATGRAPHICS is designed to run on IBM 370, 3030, 3081, and 4300 Series machines and other compatible configurations. It is an interactive system where both graphics and statistics are under the user's control. The graphical output can be produced on any high density or color peripheral. The statistical functions are: analysis of variance, basic plotting functions, cluster analysis, descriptive methods, estimation and testing, distribution functions, simulation and random number, forecasting, data input/output, exploratory data analysis, basic draw functions, categorized data functions, multivariable statistics, nonparametric statistics, numerical analyses, sampling, quality control, regression analysis, smoothing, time series analysis, stochastic modeling, experimental design, special math functions, and mathematic programming. The graphics patterns include line graphs, pie charts, and contour maps.

SPSS GRAPHICS

SPSS GRAPHICS is designed for use on most mainframe computers and outputs on 60 output devices, including those of Hewlett-Packard and Tektronix. As with other software packages, it can control up to eight colors, depending on the capacity of the output device. It produces bar graphs, pie charts, and line graphs. All plots can be saved in an SPSS graphics disk file. As with SAS/GRAPH, SPSS GRAPHICS must be purchased separately; it is not available as part of the general SPSS package.

Software for Large Mini, Mini, and Desktop Computers

Hewlett-Packard Software

Hewlett-Packard has developed graphics software to be used with their Series 80, 125, and 3000 computer lines, their desk-top computers, and intelligent graphics terminals. Each of these is briefly described below. All these packages are designed to be used with HP printers and plotters.

HP Series 80 Graphics. Text and pie, bar or line charts can be generated by responding to simple questions. Text and chart labels can be in three typs of letters in nine different sizes. Text can be centered, underlined, or both. Graphs and charts can be prepared using six different types of lines or six types of hatching.

HP 125 Graphics. With this graphics package, the user can turn tabular data into bar charts, pie charts, and line graphs. The system uses a fill-in-the-form style of user interaction. Data can be entered directly from the keyboard or retrieved from stored HP 125 data files. Hard copy graphics can be generated by an HP plotter on standard paper or on overhead transparencies.

For bar charts, the user specifies type of chart (normal, stacked, or comparative), titles, axis labels (up to twelve with five values per label), axis range, colors, shadings (choice of seven), legends, and grid.

For pie charts, the user specifies titles, labels (up to ten), colors, exploded segments, shadings (three plus blank), and optional sort.

For linear charts, the user specifies type of chart (linear, log-x, log-y, log-log), number of columns of data, x columns, y columns, axis range (numeric, months or days), line color, line type (choice of eight), titles, legends, axis label, and grid.

Graphics for HP Desktop Computers and Graphic Terminals: HP 9825, HP 9835, HP 9845, HP 2647. The graphics software package for the HP 9825 desktop computer allows the user to produce function plots and bar graphs. The HP 9876 Printer is required.

The graphics software package for the HP 9835 desktop computer contains routines for histograms, probability plots, time plots, X-Y scatter plots, log-log plots, XYZ plots, and Andrews plots. The HP 9835 external printer and plotter or graphics printer are required.

Software for the HP 9845 produces bar, line, and pie charts. The software can be used to create overhead transparencies as well as charts using the Graphics ROM and HP 9872 Plotter.

Graphics HP 2647 is designed for the HP Intelligent Graphics Terminal. It produces pie, bar, and line charts from numeric data. Flowcharts, diagrams, organization charts, and electronic circuit drawings can also be created. An advantage of using software on the graphics terminal is that it allows the user to preview and modify multicolor charts or transparencies on the terminal screen before sending them to a plotter or printer.

Graphics for the HP 3000: DSG/3000 (Decision Support Grahpics/3000). With DSG/3000, more technically sophisticated users can create charts from information stored in an HP 3000 system or can enter data through the terminal keyboard. Once a graph has been designed, its specifications, including acceptance standards, are saved in a chart file, which is separate from the data. Thus, it is easy to reuse a particular chart on different or revised data. As the data base is updated, so can the charts be easily updated rather than redesigned. DSG/3000 lets the user brouse through this chart file at the press of a button to review or modify graphs already created. The graphs available include line graphs, horizontal and vertical bar charts, pie charts, and scatter plots. Data can be automatically scaled, and arithmatic and statistical functions can be applied to the data and plotted. Chart sizes can be defined by the user, and multiple charts can be placed on one page for easy comparison.

Tektronix Software

The Textronix Plot 50 software supports the Tektronix 4050 Series Desktop Computers. The Plot 50 library contains a statistics package as well as packages for 2-D Drafting, MicroPERT2-Project Management, Picture Composition and Document Preparation, and Technical Data Presentation. The same data can be transformed into different kinds of graphs including line, pie, and bar charts. The statistics package contains subroutines for tests and distributions, analysis of variance, multiple regressions, and nonlinear estimations.

Decision Resources' Chart-Master

Chart-Master has been designed for the Apple II or Apple III desktop computer used in combination with Hewlett-Packard plotters to produce hard copy output. The software can generate bar graphs, pie charts, scatter diagrams, line charts, and text pages.

Table 32

SAS Control Statements Used to Compute Vectors
Used in the Biplot

```
              PROC MATRIX PRINT;

              FETCH Y DATA=STATES;

              N=NROW(Y);

              M=NCOL(Y);

              YMEAN=Y(.,);

              ID=J(N,1,1);

              YBAR=YMEAN@ID;

              YDEV=Y-YBAR;

     ****SINGULAR VALUE DECOMPOSITION ****;

               SVD U Q V YDEV;

              P = U (1:N, 1 2);

              Z = V (1:M, 1 2):

              E = Q (1 2, 1);

              L = DIAG (E);

         ****MATRICES TO BE PLOTTED****;

              G = SQRT (N) # P;

              H = (1 #/SQRT(N)) # (L * Z');

         ***************END*************;
```

Note: The matrix G contains the vectors corresponding to the units of analysis and the matrix H contains the vectors corresponding to the variables.

References

Andrews, H. P., Snee, R. D., & Sarner, M. H. Graphical display of means. The American Statistician, 1980, 34(4), 195-199.

Bock, R. D. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.

Caporal, P. M., & Hahn, G. J. Software for statistical graphics--An overview. In Proceedings of the Third Annual Conference and Exposition of the National Computer Graphics Association, Inc., Volume II. The National Computer Graphics Association, 1982, 1033-1043.

Chernoff, H. The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association, 1973, 68(342), 361-368.

DesJardins, D. L. Multi-level statistical maps in graphic communication. In Proceedings of the Third Annual Conference and Exposition of the National Computer Graphics Association, Inc., Volume I. The National Computer Graphics Association, 1982, 383-392.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. Biometrika, 1975, 62(3), 531-545.

Everitt, B. S. Graphical techniques for multivariate data. New York: North-Holland, 1978.

Friedman, J. H., McDonald, J. A., & Stuetzle, W. An introduction to real time graphical techniques for analyzing multivariate data. In Proceedings of the Third Annual Conference and Exposition of the National Computer Graphics Association, Inc., Volume I. The National Computer Graphics Association, 1982, 421-427.

Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 1971, 58(3), 453-467.

Hahn, G. J., Morgan, C. B., & Lorensen, W. E. Color face plots for displaying product performance. IEEE Computer Graphics and Applications, 1983, 3(1), 23-29.

Jaeger, R. M. An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 1982, 4(4), 461-475.

Kleiner, B., & Hartigan, J. A. Representing points in many dimensions by trees and castles. Journal of the American Statistical Association, 1981, 76(374), 260-269.