PROGRAM ABSTRACTS/ALGORITHMS

A BASIC program for exploratory data analysis

DAVID S. MALCOLM

Fordham University, New York, New York

Although the introduction of computers has made a substantial contribution to statistical data analysis, it has not occurred without some loss. Because computations are rarely done by hand, we no longer spend as much time "looking" at the data. Unusual distributions, odd scores, and other aspects of the data may go unnoticed. Tukey and others (Hartwig & Dearing, 1979; Hoaglin, Mosteller, & Tukey, 1983; Tukey, 1977) have argued that we do not adequately explore and inspect data. It is too easy to assume that data are normally distributed, that the requirements for a parametric test have been met, or that a set of bivariate data is linear, and then proceed to apply the standard statistical analyses without having verified this important information.

The goal of the present exploratory data analysis program is to provide tools that facilitate examination or exploration of the data and assist one in choosing next steps in the data analysis. The tests provided in this program fall into three main categories: (1) tests designed to promote the exploration of a data set, (2) parametrically related tests, and (3) resistant statistical tests based on the ordinal scale properties of the data set.

Program Description

Exploratory data routines. A variety of descriptive statistics is included to provide a good representation of the characteristics of the data set. In addition to the usual parametric measures of mean, standard deviation, skew, and range, the ordinal measures of quartiles or hinges, spreads, minimum, maximum, and outliers are provided. These ordinal measures are often called *resistant* measures, because they are much less sensitive to extreme or deviant values than are the usual parametric ones. A stem-and-leaf display (Tukey, 1977) presents a rank-ordered list of the data while at the same time it provides a display of the data similar to a histogram.

Graphic methods include histograms, scatterplots, and box-and-whisker displays. The box-and-whisker display (Tukey, 1977), a graphic representation of the data set based on ordinal measures, gives a full picture of the data distribution. Outliers, data that lie outside the normal range of the data set, are also shown. The user sets the

scale for the histogram and box-and-whisker displays to facilitate comparisons among different variables.

Parametric related tests. Several procedures included in this program are directly related to testing whether a data set meets the requirements for a parametric analysis. Homogeneity of variance is tested according to Bartlett's procedure (Sokol & Rohlf, 1981). The normality of a distribution may be tested in two ways. The Kolmogorov-Smirnov test can be performed against an expected normal distribution with a choice of either grouped or ungrouped data (Rohlf & Sokol, 1981; Sokol & Rohlf, 1981). Also, a method that plots the cumulative frequency distribution on a probability scale allows a graphical determination of the skewness, kurtosis, and number of modes of a distribution (Sokol & Rohlf, 1981).

For both least squares and resistant regression (see below), displayed scatterplots include both regression lines. Residual plots for both types of regression allow a visual determination of the variability remaining after the linear component is removed, and they permit a quick check on the degree of nonlinearity, if any, present in the data. If the relationship appears to be nonlinear, one or both of the variables may be transformed to improve the linearity of the relationship.

A wide variety of transformations is included. Transformations may be used to rescale a data set, to increase the linearity of a relationship, or to normalize a data set, permitting the use of parametric statistics. The transformations are divided into linear ones, which do not alter the shape of the distribution, and nonlinear ones, which do. The linear include the z-score transformation and other transformations of the form $a + bX$, where a and b are specified by the user. These may be used in conjunction with a nonlinear transformation to prescale the data to make optimum use of the nonlinear transformation.

The nonlinear transformations include those which effect only one or both tails of a distribution. The convex and concave serve to reduce negative and positive skewed distributions, respectively. If both tails of the distribution show too much spread, they may be compressed with the sine or odd root transformation. Conversely, both tails of the distribution may be spread out equally with the folded log or folded root transformations (Hartwig & Dearing, 1979). These transformations should be used to normalize the shape of a distribution prior to the use of parametric tests.

Resistant or ordinal procedures. In addition to the ordinal descriptive procedures described above, two other statistical routines based on medians are included. Commonly used methods for smoothing, or removing, irregularities from a set of data are based on parametric

Correspondence may be addressed to David S. Malcolm, Division of the Social Sciences, College at Lincoln Center, Fordham University, New York, NY 10023.

statistics (e.g., the moving average method based on the mean; see Hamberg, 1983). This routine, provided for use when the mean is not an appropriate measure, is based on taking running medians in groups of three and then groups of five, to smooth the data (Hartwig & Dearing, 1979).

An alternative procedure to the traditional least squares regression is also provided. It is based on dividing the data into "slices" and computing a regression line from the medians within each slice (Hartwig & Dearing, 1979). This procedure has in common with other ordinal statistics the advantage of reducing the impact on the analysis of extreme values in the data set. As the program computes the regression with both methods, a comparison may be made to determine the best method for a particular data set.

Data-handling routines. All data to be analyzed by the program must be entered and saved on disk. Once stored on disk, the data file may be edited. It may also be displayed in subject or variable order. The output of the program may be displayed on the screen or printed. Data modified by the program may be displayed and saved to the disk. All of the routines will handle missing data, with the user assigning the representative value.

The program also allows for the importation of data sets prepared by other statistical packages, including Epistat, NH Analytica, and StatSoft. Data may also be imported from many programs (e.g., word processing) that will write an ASCII output file.

Program Language and Hardware Requirements

The program, written in GW-BASIC on a Tandy 1000SX, has been used successfully on IBM PCs and Zenith com-

patibles. It should function correctly on any PC-compatible computer. The program requires GW-BASIC or BASICA, 256K of memory, and a CGA card with a color or monochrome monitor. Hercules graphics are not supported. Printed output is available, and the graphics screens may be printed with a dot-matrix printer with graphics capability.

Availability

A listing of the program, a cross-reference listing, and additional information and notes on the program are available from David S. Malcolm, Division of the Social Sciences, College at Lincoln Center, Fordham University, New York, NY 10023. A copy of the program on disk is available if a 5.25-in. disk is provided in a mailing container with return postage.

REFERENCES

- HAMBERG, M. (1983). *Statistical analysis for decision making* (3rd ed.). New York: Harcourt Brace Jovanovich.
- HARTWIG, F., & DEARING, B. E. (1979). *Exploratory data analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-016). Beverly Hills, CA: Sage Publications.
- HOAGLIN, D. C., MOSTELLER F., & TUKEY, J. W. (Eds.) (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.
- ROHLF, F. J., & SOKOL, R. R. (1981). *The statistical tables*. San Francisco: W. F. Freeman.
- SOKOL, R. R., & ROHLF, F. J. (1981) *Biometry* (2nd ed.). San Francisco: W. F. Freeman.
- TUKEY, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

(Manuscript received January 13, 1989;
revision accepted for publication May 24, 1989.)