PC World, Sept. 1984, Vol. 2, 186-195.

Statistical Quintet

Over 40 statistical programs are now available for the PC. Here are five new packages for under $500 each.

Nancy Goodban and Kenji Hakuta

"Statistics" is a general term that covers a broad range of methods used to process, analyze, and interpret data. Various types of statistical analysis are used by businesses, professionals, and hobbyists. A baseball statistician calculates the batting average of a baseball player, a civil rights lawyer explores whether a particular company pays higher salaries to its male employees than to female employees who do similar work, and an insurance company determines the risk factors of extending coverage to personal computer owners. Statistics are used by political scientists, market researchers, engineers, and anybody else who needs to analyze data.

At least 40 statistical packages are currently available for the IBM PC, and the number is rapidly growing. Here we review five that each offer a broad range of commonly used statistical procedures for less than $500: *AbStat 3.3, Crisp 83.1, Microstat Release 4.0, SL-Micro,* and *Systat.*

This evaluation considers four major dimensions: each program's ability to manipulate data, the types of statistical procedures it provides and the quality of the output it produces, the accuracy and speed of its computation, and its ease of use.

Ratings of each package in terms of these criteria are listed in Table 1. Since the science of statistics, like other specialized fields, is inundated with jargon, some of the terms used in this article are defined in "A Statistical Glossary," and basic statistical concepts are outlined in "A Statistical Primer."

Data Manipulation

Statistical packages should enable the user to input, manipulate, and store raw data. All the packages reviewed except *SL-Micro* allow the user to enter data directly at the console. *SL-Micro* is run in batch mode from command files created on a word processor.

A statistical program should be able to read data from an ASCII file created by another applications program and to write data to disk in a form that can be used by other packages. This would make it possible, for example, to use a mainframe to manipulate or summarize large data sets, to import data from PC spreadsheet or data base management programs, or to read files containing experimental lab data stored directly on disk (see "Lotus in the Lab," *PCW,* Vol. 2, No. 2). Communications capability is also necessary if you have to upload the data to a mainframe for additional analysis or present it in an alternative fashion with a spreadsheet package on the PC. All five packages discussed here can read and write ASCII data files, but they vary greatly in how easily and flexibly they do so.

For users familiar with the use of format statements for input variables, the packages that most easily accept ASCII files from disk are *SL-Micro,* which requires variable formats almost identical to those of the mainframe package *SPSS,* and *Systat,* which is similar to the mainframe package *SAS. Crisp* and *Microstat* are also straightforward; they prompt the user for the variable list and format. *AbStat* provides a utility to read data from an ASCII file, but the program has some undocumented bugs.

> All five packages can read and write ASCII data files.

All the packages reviewed except *SL-Micro* enable you to write a file to disk so that it can be read internally by the program. *Systat* offers the most flexibility in writing data to disk in ASCII format; you can output data in any format desired. Writing

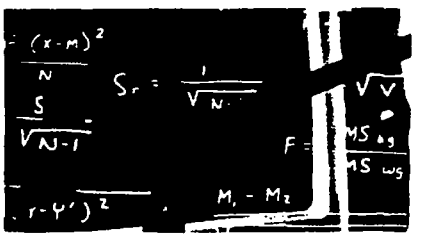
output files is also relatively smooth in *AbStat* and *Microstat*. *Crisp*'s "file export" utility works, but it is inconvenient; *Crisp* presents a menu and requires the use of function keys to specify variable locations and lengths. It is easy to make a mistake, and if an illegal value is typed in, it is stored in memory and registered as the answer to the next question.

Once data is available to the program, you need to be able to transform and edit it. Transformation is a general term that covers a number of computations that change the values of variables. For maximum flexibility a package should offer conditional statements, such as IF SEX = F THEN GROUP = 1, for transforming data as well as the ability to recode values and compute new variables. *Systat* contains its own version of *BASIC*, allowing you to write and save the full range of transformations available in the language. *AbStat*, *Crisp*, and *Microstat* let you transform and save new variables easily using a number of basic logical and arithmetic functions. *SL-Micro* allows transformed variables to be saved but stores the transformed data in a special file. To use the transformed data file, you have to edit the command file to specify the location and variable layout of the special file. In order to write new transformations, you must first erase or rename the special file.

Only *SL-Micro* and *Systat* accept character data, such as *M* and *F* for male and female. The other packages require all data to be entered as numbers. *Crisp* is able to read only 8 digits, including a decimal point, so larger numbers are truncated.

Crisp, *SL-Micro*, and *Systat* allow you to assign variable names such as AGE or SEX and to refer to a variable by name. *Microstat* assigns all variables both a number and a name; the variables must be referred to by number, but the program can display the variable numbers and names for reference.

It is important to select a package that can handle the largest data sets anticipated. *AbStat* is limited to 64 variables, and *Systat*, to 75 variables. *Crisp* can handle over 250 variables. *Crisp*, *Microstat*, *SL-Micro*, and *Systat* allow virtually unlimited numbers of cases. The number of cases permitted by *AbStat* is limited by the RAM available on the system.



Crisp can handle over 250 variables.

If the data you are working with has been created elsewhere, and values that indicate missing data are embedded in it, your statistical package must allow you to define values for missing data. *AbStat*, *Crisp*, and *SL-Micro* provide this capability. *Microstat* and *Systat* provide a special character for missing data; you must recode any applicable values before they can be interpreted as missing.

Finally, you should consider a program's ability to manipulate files. Ideally, a program should allow you to add new cases and variables to a file, merge two existing files, sort files, and select the cases or variables to be output. *Crisp* has an outstanding set of menu-driven utilities for merging and sorting files, while *Systat* has excellent utilities for those who prefer a command-driven system. At the other end of the spectrum, *SL-Micro* does not provide any of the file-manipulating functions mentioned. *AbStat* and *Microstat* offer some of those functions, but with varying degrees of quality.

Accuracy and Speed

Accuracy is a function of the "precision," or number of digits, handled by the program as well as of the algorithms used in calculations. All five packages perform respectably on the Longley data set (*Journal of the Statistical Association of America*, 1967, Vol. 62, 819-831; see Table 2). This data set is commonly used to test the ability of multiple regression programs to deal with extremely high correlations between predictor variables. Most of the regression programs use double precision, allowing a maximum of 16 digits for each number. This avoids the problems associated with rounding error.

If you're tied down to your PC by a strictly interactive package such as *Crisp* or *Microstat*, the program's speed is important. *AbStat* and *Systat* can take advantage of the 8087 coprocessor math chip, which increases the speed of numerical calculations by a factor of ten or more. If you anticipate using large data sets or creating large correlation matrices on a regular basis, you might well consider purchasing the 8087 chip and a package that can use it. On the other hand, you may prefer to use a mainframe for analyzing large data sets, in which case the speed of the program on your PC is less important.

Comparisons of the five packages' speeds in performing sample analyses are presented in Table 3. Elapsed seconds were counted from the time a command was issued until the program was ready for the next task. Any required program overhead is included along with the time required for the procedure itself. For example, *SL-Micro* spends several seconds

Data Manipulation	AbStat	Crisp	Microstat	SL-Micro	Systat
Input from ASCII disk file in specified format	poor	good	fair	exc	exc
Output to ASCII disk file in specified format	good	fair	good	—	good
Recoding, transforming, and computing new variables	good	good	good	fair	exc
Saving transformed and computed variables	good	good	good	poor	good
Specifying missing values for input data	good	good	—	good	—
Assigning and using variable names	poor	good	fair	good	good
Editor to change specific variables and values	good	good	good	—	good
Utilities to sort and merge files and select cases	good	exc	exc	—	good
Ability to accept character as well as numeric data	—	—	—	good	good
Utilities to list, erase, and rename files	—	good	good	—	—
Statistical Procedures Available					
Frequencies	good	exc	exc	exc	fair
Histograms	—	exc	fair	good	exc
Scatterplots	fair	exc	exc	—	exc
Cross-tabulations	fair	exc	poor	good	fair
Specialized exploratory data analysis	—	—	—	—	exc
Analysis of variance (ANOVA)	good	good	good	—	good
Repeated measures analysis of variance	—	good	—	—	good
Pearson's coefficient of correlation	fair	good	good	good	fair
Student's t test on matched and independent groups	good	good	good	—	fair
Ordinary multiple regression	good	exc	good	exc	good
Stepwise multiple regression	—	exc	fair	exc	good
Output of residuals from regression to data file	good	good	good	—	good
Principal components factor analysis	—	—	—	—	good
Canonical correlation and discriminant function	—	—	—	—	good
Multidimensional scaling	—	—	—	—	good
Nonparametric statistics	fair	good	exc	poor	poor
Options for treatment of missing values	good	good	—	—	good
Output from Statistical Procedures					
Readability of standard table format	good	good	good	good	good
Flexibility of table layout, titles, and labels	fair	good	fair	good	poor
Display of significance tests and probability	poor	good	fair	good	poor
Routing output to disk file	good	good	exc	good	exc
Routing output to printer	good	good	exc	good	exc
Ease of Use					
Batch or interactive	both	int	int	batch	both
Menu-driven or command-driven	comm	menu	menu	comm	comm
Ability to exit smoothly at any point	—	good	good	N/A	—
Accuracy (see Table 2)	fair	fair	exc	fair	exc
Speed (see Table 3)	fair	good	good	fair	fair
Documentation					
Overall quality, readability, and organization	good	exc	good	good	exc
Description of computational formulas and methods	good	good	good	fair	fair
On-line help	good	—	—	—	good
Index	—	good	—	good	exc
Listing of error messages	good	—	—	poor	—

Table 1: Statistical software compared

reading the variables from disk, while *Crisp, Microstat,* and *Systat* take as long to write out sorted data as they do to sort it. We did not test the performance of any of the packages using the 8087 chip.

Ease of Use

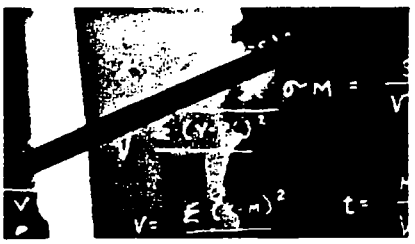
Statistics is a complex field, and there is no reason to expect a statistical package to be easy to use. However, it should be designed so that the user is not led into making errors. Ideally, a program should flag mistakes so they can be corrected and should assign a key to provide easy exit from errors. The listed programs vary greatly in their ease of use (see Table 1).

The ability to exit smoothly from each function at any point is important, particularly for an interactive or menu-driven package. It is frustrating to be forced to begin again at the first menu or, worse, to have to reboot the system entirely after inadvertently issuing the wrong command. *Crisp* and *Microstat* are the only packages that define a special exit key that aborts almost any function.

When judging the ease of use of a package, you should note whether it is command-driven or menu-driven. Novices often prefer menu-driven systems because they do not require knowledge of command syntax. On the other hand, experienced users are often frustrated by having to type responses to each menu item when a one-line command would suffice. *AbStat, SL-Micro,* and *Systat* are command-driven, while *Crisp* and *Microstat* are menu-driven.

Another differentiating feature is whether the system is batch or interactive. Interactive packages require you to issue commands sequentially when prompted, and you must wait until one command is processed before issuing the next. This often causes extensive delays, and all commands must be reentered if a similar analysis is run. The advantage to interactive packages is that if a syntax error is made, it is caught immediately.

With a batch system, an editor or a word processor is typically used to create a file containing all the necessary commands. To execute a function you direct the program to the command file for instructions. With a batch system you are not tied to the console while the batch file is processed, and you can easily alter the command file to run a similar analysis.



Systat offers the widest range of analyses.

SL-Micro is a batch system. *AbStat* and *Systat* are designed as interactive systems but allow the option of specifying a command file for batch processing. *Crisp* and *Microstat* are strictly interactive.

Help and Documentation

Systat and *AbStat* are the only packages reviewed that provide on-line help, a valuable feature, especially for those who like to acquaint themselves with software by booting up the system before unwrapping the manual.

The most comprehensive documentation is provided by *Crisp* and *Systat.* Both have thorough tables of contents and indices and provide satisfactory descriptions of procedures, often backed by references to articles and books from professional statistical literature. The sophisticated statistician will not feel insulted by the level of discussion in *Crisp*'s and *Systat*'s manuals.

Predictors Variables		AbStat	Crisp	[illegible]	SL-Micro	[illegible]	Longley
Deflator	X1	15.0619	15.05666	[illegible]	15.0621		15.06187
GNP	X2	−.03581	−.03582	[illegible]	−.0358		−.03582
Unemployment	X3	−2.02023	−2.02018	[illegible]	−2.0202		−2.02023
Number in Armed Forces	X4	−1.03323	−1.03322	[illegible]	−1.0332		−1.03323
Population	X5	−.05110	−.05113	[illegible]	−.0511		−.05110
Time	X6	1829.15	1829.12744	[illegible]	1829.1596		1829.15146

Table 2: Comparison of accuracy on multiple regression analysis using Longley data set (last column presents calculation from Longley, 1967); shading indicates perfect score on Longley data test

	Sort	Cross-Tabulation	10 by 10 Correlation Matrix
AbStat	10	270	200
[illegible]	[illegible]	[illegible]	[illegible]
SL-Micro	—	230	180
Systat	120	60	240

Table 3: Comparative speed in seconds (all calculations based on 138 cases and 25 variables); shading indicates packages with best scores overall

Statistical Procedures

We found considerable variation in the availability, flexibility, and quality of statistical procedures provided by the packages (see Table 1). All packages reviewed performed basic univariate descriptive statistics, cross-tabulations, and some form of regression. All programs except *SL-Micro* provide Analysis of Variance (ANOVA), although *Crisp* and *Systat* offer repeated measures. *Systat* offers the widest range of analyses, from outstanding exploratory analyses certain to delight statisticians (for instance, stem-and-leaf displays, box plots, and probability plots displaying the shape of a distribution) to the most sophisticated multivariate linear models (canonical correlations, multivariate regression, discriminant function analysis, multidimensional scaling, and principal components factor analysis). *Systat* also offers advanced functions such as log-linear models for analyzing categorical data.

Crisp offers an excellent regression procedure, allowing stepwise and forced inclusion of predictor variables. *SL-Micro*, designed principally for questionnaire data analysis, and *Systat* also allow mixing stepwise and ordinary regression models. *Microstat* provides stepwise regression as an alternative to ordinary regression but does not allow the two models to be mixed. *AbStat* does not provide stepwise regression at all.

Microstat offers by far the widest range of nonparametric tests, being the only package to provide Kolmogornov-Smirnov's goodness of fit test. *Crisp* also offers a range of nonparametrics. *Crisp* and *Microstat* provide Wilcoxon's rank sum, Kruskall-Wallis's one-way Analysis of Variance, and Fisher's exact probability test. *Crisp* and *AbStat* provide the Mann-Whitney U test.

Output from Statistical Procedures

The standard output from all five packages is readable, but quality in other output categories varies. Ideally, you should have the option of creating titles specifying the number of significant digits to be displayed or specifying alternative layouts. In general, *Crisp* and *SL-Micro* provide the most flexibility in layout. For example, standard cross-tabulations print the frequency in each cell with the option of including row, column, and total percentage. *AbStat*, *Crisp*, and *SL-Micro* let you specify which percentages are to be printed in a table. *Systat* allows you to print these percentages only as separate tables, and *Microstat* does not provide any of these options.

Most of the packages are adequate in displaying statistical tests and their significance levels. However, *Systat* does not display the significance level for Pearson's correlation coefficient, and *AbStat* forces you to calculate the probability level separately for all

procedures. *Systat* is the only package that enables you to specify the number of digits to be displayed after the decimal in output.

It is useful to be able to direct output to a disk file for later use in reports prepared with a word processor. Ideally, you should be able to view the results on screen and then decide to route the output to the printer or to a disk. *Microstat, SL-Micro,* and *Systat* provide this capability. With *AbStat* and *Crisp* you must route the output in advance and rerun the analysis if alternative routing is desired.

For some procedures it is important to be able to define the treatment of missing values. For example, when calculating a correlation matrix, you often need to specify whether to exclude all cases with missing variable values (known as listwise deletion) or whether to exclude cases only if they are missing values on one of the two variables (pairwise deletion). *SL-Micro* always uses pairwise deletion in calculating correlations, while *Microstat* uses listwise deletion. The other packages offer both options.

Summary of Specific Packages

AbStat is well documented and easy to learn. It is the only package reviewed that reads and writes *dBASE II* files. It also sorts quickly, in part because the output file need not be saved unless saving is specified.

One minor irritation is *AbStat*'s requirement that all commands be typed in uppercase. *AbStat* provides probability functions, but the probabilities must be calculated separately from significance tests, which is inconvenient but not insurmountable. *AbStat*'s output is unattractive; the variable number is always presented along with its label. Bar graphs, plots, and cross-tabulations are displayed particularly poorly. A major drawback is that every file is limited to 64 variables.

Crisp is generally easy to learn and use, even without its excellent manual. One strength is the program's

ability to transform variables into observations and vice versa. Another unique feature is *Crisp*'s ability to process subgroups, such as male and female, separately without respecifying the analysis. *Crisp* also provides the most flexible specification of regression models of all the packages, allowing both stepwise and ordinary multiple regression to be intermixed easily.

Working with *Crisp* involves several minor frustrations, but none of them poses major difficulty. For example, if you try to print output when the printer is off line, you are bumped all the way back to DOS. Syntax errors in transformations are not immediately checked, and all transformations are lost if you make a syntax error.

Microstat offers the widest range of nonparametric statistics procedures. It also performs time-series analyses and provides a number of probability functions, including binomial, Poisson, normal, F, and Student's t. Other strengths include the ability to display an ASCII file from within the program, the ability to assign a descriptive title to every file, and the ability to direct output to a disk or a printer after viewing it on the screen without respecifying the analysis. A major drawback is that cross-tabulations are limited to 20 rows and 5 columns.

SL-Micro has data definition statements identical to those of *SPSS*. We were able to download a mainframe file created for *SPSS* and run it on *SL-Micro* with only minor editing. If the statistics in *SL-Micro* were sufficient, an *SPSS* user could use this package with minimal learning. Unfortunately, the program has a limited set of statistical procedures, and unless you are certain that the procedures provided will be adequate to meet future needs, you may be better off finding a package that offers more.

Systat provides the widest range of statistical procedures as well as the most flexibility in manipulating data. Fans of the *SAS* mainframe statistical package will be familiar with many of the data and statistical procedures in *Systat*. Although the program is difficult to learn, *Systat* is well documented.

Systat's main drawback is its inflexibility in displaying results. This poses a problem for anyone who wants to present output directly in writing up analyses. But the shortcomings of *Systat*'s output are more than compensated for by an outstanding range of statistical procedures.

Recommendations

Several factors must be considered in deciding what package to buy. Obviously, price is one element. More importantly, you should consider your needs and your level of experience in statistical computing.

Based on tests and the criteria shown in Table 1, *Crisp* and *Systat* are the best systems currently available for the IBM PC. They both offer a wide range of well-documented and easy-to-use statistical procedures. *AbStat* and *Microstat* are recommended for anyone teaching a beginning statistics course; they are both elementary enough that they are easy to learn, and they provide some nice touches such as calculation of probability functions. *SL-Micro* is recommended only for users who are familiar with *SPSS* and do not require a broad range of statistical procedures.

The state of the art of statistical packages currently available for the PC is impressive. During the last few years, many programs have been developed that adequately approximate the luxuries of computing on mainframes. This achievement is magnified when you consider that many of the programs were developed on personal time and funds by talented individuals rather than by the corporate world. The inclination to wait for *SPSS* and other mainframe packages to be released for the PC should be checked by the fact that excellent and affordable packages are already available. ●

Nancy Goodban is a social psychologist and a postdoctoral fellow at Yale University. Kenji Hakuta is an associate professor of psychology at Yale University. Both authors received Ph.D.'s in psychology from Harvard University.

AbStat 3.3
Anderson-Bell
P.O. Box 191
Canon City, CO 81212
303/275-1661
List price: $395
Requirements: 128K, DOS 2.00 with one disk drive, DOS 1.10 with two disk drives

Crisp 83.1
Crunch Software
1541 Ninth Ave.
San Francisco, CA 94122
415/564-7337
List price: $495
Requirements: 128K with DOS 1.10, 152K with DOS 2.00, two disk drives

Microstat Release 4.0
Ecosoft, Inc.
P.O. Box 68602
Indianapolis, IN 46268
317/255-6476
List price: $375
Requirements: 128K, one disk drive

SL-Micro
Questionnaire Service Company
P.O. Box 23056
Lansing, MI 48909
517/641-4428
List price: $250
Requirements: 128K, two disk drives

Systat
Systat, Inc.
1127 Asbury Ave.
Evanston, IL 60202
312/864-5670
List price: $495
Requirements: 256K, two disk drives

A Statistical Glossary

The following list provides definitions of terms frequently used in statistics:

Analysis of Variance (ANOVA). A common method in experimental science for examining the extent to which a set of variables (groupings) predicts the value of another variable.

Box plot. A graph of the values of a variable showing the median and the spread of scores.

Canonical correlation. A method of predicting a set of variables from a different set of variables.

Case. An entity, such as a person, a classroom, an institution, or a machine, on which observations are made.

Categorical variable. A variable whose values are restricted to discrete categories, such as sex.

Chi-square. A commonly used nonparametric statistic used to examine associations between categorical variables.

Continuous variable. A variable that can theoretically take on an infinite range of values: for example, temperature.

Correlation matrix. A table that lists all possible correlation coefficients among a set of variables.

Discriminant function analysis. A method for predicting a categorical variable from a set of predictor variables.

Fisher's exact probability test. A test similar to chi-square.

Format statement. A statement instructing the computer how to read variables in the data file, including information about how many variables are in the file and what columns each variable is in.

Frequency distribution. A list of all values for a variable and the frequency with which each value occurs.

Histogram. A display of the frequency of values for a variable.

Kolmogornov-Smirnov's goodness of fit test. A nonparametric method for estimating the magnitude of association between two categorical variables.

Kruskall-Wallis's one-way analysis of variance. A nonparametric method of Analysis of Variance that is used when values are rank ordered.

Log-linear analysis. A method of analyzing cross-classified categorical data.

Mann-Whitney U. A nonparametric test of differences between groups when the values are rank ordered.

Mean. The arithmetic average of values for a variable.

Median. The midpoint of the values for a variable; half the values fall above the median and the other half below.

Missing value. A value assigned to a variable when no valid value can be collected for the case.

Mode. The most common value for a variable.

Multidimensional scaling. A method for locating a set of variables in multidimensional space according to how closely associated they are.

Multiple linear regression. A method for predicting the value of one variable from the values of a set of different variables.

Multivariate linear models. A set of methods for analyzing multiple sets of variables.

Multivariate regression. A method of multiple regression that allows multiple variables to be predicted.

Nonparametric statistics. A set of methods used when certain properties of the data are unknown or unusual; nonparametric statistics require fewer assumptions about the data than parametric statistics.

Parametric statistics. A set of methods used when samples are drawn from a population that has known statistical properties.

Pearson's correlation coefficient. A commonly used parametric measure of the association between two variables.

Predictor variables. Variables whose values are used to predict the values of other variables.

Principal components factor analysis. A method for determining how a set of variables fit together in smaller subsets, or factors.

Probability functions. Mathematically derived specifications of the likelihood of the occurrence of an event (or a series or group of events) under various conditions. These functions are used to test statistical significance. Common probability functions include normal, Student's t, F, chi-square, binomial, and Poisson distributions.

Range. The difference between the minimum and maximum values of a variable.

Repeated measures Analysis of Variance. An Analysis of Variance in which the same variables are measured for the same cases at more than one time.

Sample. A group of cases, selected from a larger population, on which data has been gathered.

Scatterplot. A method of displaying the relationship between two variables.

Standard deviation. A measure of the spread of values around the mean for a variable.

Stem-and-leaf display. A version of a frequency distribution in which values are grouped into intervals for preliminary exploration of the characteristics of data.

Stepwise multiple linear regression. A version of multiple linear regression in which the predictor variables are entered into calculations one at a time rather than simultaneously.

Sum of squares. A mathematical term commonly used to calculate parametric statistics.

Test of significance. A set of statistical tests, based on probability theory, that determines the likelihood that patterns in sample data have been obtained by chance.

Time-series analysis. A method for analyzing trends in data across time.

Univariate descriptive statistics. A set of statistics used to describe the values of a variable.

Value. The score of a case on a particular variable.

Variable. A characteristic that may take on different values for different cases.

Wilcoxon's rank sum. A nonparametric method of testing differences between two groups.

A Statistical Primer

Statistics are used both to describe data and to make inferences about a population from the description of a sample of that population. All statistical analyses begin with the description of data. Inferences can be made using probability theory, which states the likelihood of an inference being correct by applying a test of significance.

Regardless of the statistical procedure used, data consists of "cases," "variables," and "values." For example, in a market survey each person who responds to the survey represents a case (e.g., Smith or Cohen); each question on the survey is a variable (e.g., How old are you?, How well did you enjoy the program?); and each person's response to each variable is represented by a numerical value (e.g., Smith is 30 years old and liked the program 4 on a scale of 7; Cohen is 23 and liked the program 6).

Descriptive statistics are commonly used to describe the "central tendency" of the values of the cases for a variable (e.g., the average age is 24.6 years old) and the "dispersion" of values around the central tendency (e.g., 50 percent of the respondents are between 22 and 26 years old). Other "univariate" procedures, such as frequency distribution, stem-and-leaf display, and box plots, describe the features of a single variable.

Normally a statistician wants to describe more than one variable at a time. A "scatterplot" is a graph that shows the relationship between two variables, one represented on the vertical axis and the other on the horizontal. Individual cases are plotted at the intersection of the values on the two variables.

The strength of a relationship between two variables can be expressed numerically through "correlation." A statistician can also perform a "regression," based on the correlation, that estimates the most likely value of a case on one variable given a certain value on the other. Regression using more than one variable to estimate a value is called multiple regression. Insurance companies, for example, use multiple regression to predict life expectancy from a large number of predictor variables that, taken together, account for a certain amount of the variation in life expectancy.

In situations in which values are categorical (for example, men and women or Democrat, Republican, and Independent) other kinds of descriptions can be made. Procedures of this kind include cross-tabulations and log-linear models.

When groups of cases (e.g., men versus women) are compared on a continuous variable (e.g., age), a number of statistical procedures are used to assess whether the groups are different. These include ANOVA, Mann-Whitney U, and Kruskall-Wallis. Which test is used depends on the number and kind of groupings to be compared and on statistical assumptions about the nature of the spread of the values. If the spread of scores conforms to a certain set of assumptions, more powerful, "parametric" tests can be used. If it does not, "nonparametric" tests are used.

For further reading in statistics we recommend starting with the following sources, listed in order of increasing complexity: *Modern Elementary Statistics* by J. Freund (Prentice-Hall, Englewood Cliffs, New Jersey, 1979); *Applied Statistics* by J. Neter, W. Wasserman, and G. Whitmore (Allyn & Bacon, Boston, 1982); *Exploratory Data Analysis* by J. Tukey (Addison-Wesley, Reading, Massachusetts, 1977); *Statistical Methods* by G. Snedecor and W. Cochran (Iowa State University Press, Ames, Iowa, 1967); *Statistical Principles in Experimental Design* by B. Winer (McGraw-Hill, New York, 1971).