Statistical Software Smorgasbord

A number-crunching program for every taste and need

by Bob Trotter

I bought my Kaypro to save my neck. College professors live in a publish-or-perish environment, and word processing is the heart of publishing. I am strongly opposed to perishing, so I bought a machine with a great reputation as a word processor. It came as a pleasant surprise to discover that the friendly machine living in my office (and occasionally being lugged home) can replace the university mainframe for another important research task: number crunching.

Scientific publications do not live by words alone; they also take numbers—not just numbers you can add up, numbers that have to be *crunched* (unmasticated they are too raw for publication). The traditional way to crunch numbers is to use a statistical package on a mainframe. The only problem is that these packages have a tendency to gobble CPU time like old-fashioned cars guzzle gas. They can rapidly deplete a departmental budget to the point where hard choices have to be made.

The solution? The discovery that, using canned statistical packages that work on a micro, Kaypro's data manipulation capacity is sufficient to keep me from having to make that hard choice.

Finding the right package

Statistical packages are for folks who are blind to the beauty of formulas. They are for those of us who must present our data in organized ways, backed up by numbers. They are not for people who can achieve nirvana by deriving formulas. They are for the rest of us, who come to statistics not out of love, but out of need.

A good statistical package is not just a library of programs. The best ones allow you to enter, check, edit, and manipulate data easily, run statistical analyses, then label and print the results, so all you have to do is to stick it in a report and say, "Ah-*ha*! This proves I was right all along!"

Looking for the right package, I went to the *Kaypro Software Directory*, scanned the ads in *PROFILES*, and talked to friends. Then I sent off to companies for their literature. Statistical packages cost anywhere from $250 to $600. No one can afford to get a program that does not work well.

I discovered that each package differs and none has all of the procedures that mainframes offer. However, the good packages have the most commonly used procedures, plus a few extras. Some of them have good data entry and editing systems. Some do not.

From the many statistical packages available (the April 1984 issue of *Byte* researched no fewer than 24) I selected five according to the following criteria. Each had to: (1) run on CP/M Kaypros; (2) be integrated in the sense that procedures were well linked and easy to move among—not just separate programs; (3) offer at least the usual procedures commonly required by researchers; and (4) offer clear and otherwise professional documentation.

In short, I tried to narrow a very wide field of complicated packages for the specific benefit of Kaypro users. If other Kaypro-compatible packages exist that rank with those reviewed here, I have not found them.

I tested each package using my own data—data I had previously run on the university mainframe. After putting each one through its paces, here's what I found.

ABstat (Anderson-Bell, Inc.)

ABstat is friendly, easy-to-learn and built like a Dagwood sandwich—with a little bit of this and a little bit of that. It is probably best used by people

who are in a hurry or who have simple tastes. It takes virtually no time to get the package up and running.

ABstat is driven by a set of easy-to-learn commands. One particularly nice feature, for those of us who are occasionally fumble-fingered, is the "!" command. It allows you to retreat from any command without penalty. And if you forget a command, or are not sure which one to use, the "?" command gets you all the help you need in virtually any situation. Systat is the only other package reviewed here that has a help function. The other packages could benefit considerably from these two examples.

Data entry is easy to handle with ABstat. Hand entry is prompted by case number and variable. In addition, the package has a utility that allows you to read data from ASCII or dBASE II files. Equally nice, it also allows you to convert ABstat files to ASCII or dBASE II files, so you can use them with other programs you might have. I tried it both ways and it was essentially hassle free. Data editing is also easy to accomplish with ABstat. This is an important feature, especially if several people are entering data into your data files. ABstat also has a sophisticated set of features for handling missing data. This is a must for survey research, where there is always data missing because someone refuses to answer a question or two.

With ABstat's 30-day trial system, the package comes initially configured to process only very small data sets. You can get used to ABstat's command structure and decide if you want it. If you decide to keep the package, you open a sealed envelope and type a magic phrase at the proper prompt. This transforms ABstat into the full-capacity package and, simultaneously, makes your money-back guarantee disappear. This trial system is a useful feature for people who have no experience with statistical packages and those who need a few practice sessions to decide if they want to buy one. For more sophisticated users, this feature is simply an extra step in the installation process, taking perhaps an extra minute or two.

The most serious weakness of ABstat is its limited capacity to handle large data sets. It will handle a maximum of 64 variables. This is one-fourth the capacity of most of the other packages. This limitation severely restricts its use in survey research, where questionnaires often have more than 64 variables imbedded in the questions.

If I could make three wishes for each of the packages I'm discussing, my wishes for ABstat would be for lots more procedures, a much higher variable capacity, and the ability to run in batch mode.

StatPac (Walonick Associates, Inc.)

If you do social science survey research and you can only buy one package, this is probably the package for you. It was obviously designed with survey research in mind. It creates a code book, then allows you to save, modify, and print it. StatPac offers three different types of data entry; two give you protection against the problems inherent in using research assistants to code data. StatPac allows both variable labels (40 characters) and value labels (16 characters). And StatPac has two methods for handling missing data.

The procedures in StatPac are of the solid meat-and-potatoes variety. They provide a decent meal—nothing particularly fancy, just solid stuff. The package contains basics, such as frequencies, crosstabs, regression, and analysis of variance (ANOVA), but lacks a number of procedures that are nice to have for those rare but inevitable times when you need them. The software is well designed. For example, with the Crosstabs program you can generate tables with a maximum size of 50 by 50 cells, which is adequate for virtually any project.

The procedures of StatPac provide a solid meal.

StatPac does offer three procedures missing from the other packages but valuable to survey researchers. "Multiple Variable Response" allows you to print summary reports for a group of variables when those variables have the same response categories. And Breakdown allows you to do things such as break out income statistics by race, sex and age. The other packages can deliver the same information, but only through a much more time-consuming process. Finally, StatPac has a utility that calculates ideal sample sizes, allowing one to do a survey from start to finish.

Statpac makes it very easy to hand-enter data. I had a lot of fun trying out the line-prompted and the template-directed entry, and they worked very well. The keypunch emulator worked equally well, but I have never considered keypunching fun. It was also easy to edit the data once it was entered. What I found much more difficult was handling data that I already had sitting around in other files, either in the mainframe, on spreadsheets, or in my dBASE II

COMPARISON OF STATISTICAL PACKAGES

PROCEDURES	ABstat	StatPac	Microstat	NWA Statpak	Systat
Elementary					
1. frequencies	Y	Y	Y	Y	Y
2. crosstabs	Y	Y	Y	Y	Y
3. arithmetic mean	Y	Y	Y	Y	Y
4. median	Y	Y			Y
5. mode	Y	Y			
6. variance	Y	Y	Y	Y	
7. minimum	Y	Y	Y	Y	Y
8. maximum	Y	Y	Y	Y	Y
9. range	Y	Y			Y
10. standard deviation	Y	Y	Y	Y	Y
11. standard error	Y	Y	Y	Y	Y
12. skewness	Y	Y	Y	Y	Y
13. kurtosis	Y	Y	Y	Y	Y
14. moments about mean			Y	Y	
15. coef. of variation	Y		Y	Y	Y
16. sum of squares		Y	Y	Y	Y
17. mean square	Y	Y	Y	Y	Y
18. chi squared	Y	Y	Y	Y	Y
19. correlation analysis	Y	Y	Y	Y	Y
20. linear regression	Y	Y	Y	Y	Y
21. multiple regression	Y	Y	Y	Y	Y
22. polynomial regression				Y	Y
23. stepwise regression			Y	Y	Y
24. fourier analysis				Y	
25. analysis of variance	Y	Y	Y	Y	Y
26. two-way ANOVA	Y	Y	Y	Y	Y
27. three-way ANOVA				Y	Y
28. multiple analysis of var.					Y
29. analysis of co-variance					Y
Non-Parametric					
30. Kruskal-Wallis			Y	Y	
31. Mann-Whitney	Y			Y	
32. Spearman			Y	Y	Y
33. Wilcoxon signed ranks			Y	Y	
34. Kendal's tau			Y	Y	
35. Wald-Wolfowitz			Y		
36. Wilcox rank-sum			Y		
37. Kolmovorov-Smirov			Y	Y	
Miscellaneous					
38. goodness of fit	Y		Y	Y	Y
39. log linear models					Y
40. Poisson	Y		Y	Y	
41. F	Y	Y	Y	Y	Y
42. time series analysis			Y	Y	Y
43. factorial calculation			Y	Y	Y
44. combinations			Y	Y	
45. permutations				Y	
46. prob. of no repetition				Y	
47. Bayes formula				Y	
48. geometric mean				Y	
49. harmonic mean				Y	
50. quadratic mean				Y	
51. trimmed means					Y
52. moving average			Y	Y	
53. standardized scores				Y	Y
54. residual analysis		Y	Y	Y	Y
55. multi-way correlation				Y	Y
56. binomial distributions	Y		Y	Y	Y
57. bivariate normal				Y	
58. gamma function				Y	Y
59. exponential functions			Y	Y	Y
60. hypergeometric functions			Y	Y	
61. inverse normal functions			Y	Y	
62. Khrigian-Mazin				Y	
63. logarithmic functions				Y	
64. negative binomial				Y	
65. normal functions				Y	
66. Pearson product moment				Y	
67. t statistic	Y	Y	Y	Y	Y
68. Fisher's exact test			Y	Y	Y
69. contingency tables			Y	Y	

	ABstat	StatPac	Microstat	NWA Statpak	Systat
70. Durbin-Watson					
71. Freidman	Y	Y	Y	Y	Y
72. hypothesis tests (means)			Y	Y	
73. hypothesis tests (prop.)			Y		
74. save residuals			Y		Y
75. correlation matrix	Y	Y	Y		Y
76. Z-value transformation		Y	Y	Y	Y
77. partial r	Y	Y	Y	Y	Y
78. Cramer's V		Y	Y	Y	Y
79. log-linear models		Y			Y
80. covariance matrix					
81. Pearson correlation matrix					Y
82. gamma matrix					Y
83. Guttman mu2 coefficients					Y
84. Wilk's lambda					Y
85. theta					Y
86. analysis of repeated measures					Y
87. Pillai trace					Y
88. Hotelling-Lawley trace					Y
89. nested factors					Y
90. trend surface analysis					Y
91. principle components analysis					Y
92. multi-dimensional scaling					Y
93. Kruskal stress formula 1					Y
94. Guttman coef. of alienation					Y
95. discriminant analysis					Y
96. factor analysis					Y
97. canonical correlation					Y
98. Monte Carlo simulation					Y
99. Spearman rank corr. matrix					Y
100. frequencies histogram	Y	Y			Y
101. scatterplots	Y			Y	Y
102. stem-leaf plots			Y	Y	Y
103. box plots					Y
104. split plot analysis					Y
105. probability plots					Y
106. break down					Y
107. multiple var. response		Y			
108. random number generator		Y			
109. sample size generator	Y	Y	Y	Y	Y
		Y			
total	35	35	53	69	75

DATA MANAGEMENT and DATA ENTRY

	ABstat	StatPac	Microstat	NWA Statpak	Systat
1. create code book					
2. label reports		Y			
3. variable labels	Y	Y	Y		
4. value labels	Y	Y	Y	Y	Y
5. keypunch emulator entry		Y			Y
6. line entry		Y			
7. template defined entry	Y	Y	Y	Y	Y
8. editing possible		Y			
9. merge files	Y	Y	Y	Y	
10. delete cases	Y	Y	Y	Y	
11. create data subsets	Y	Y	Y	Y	Y
12. sort files	Y	Y	Y	Y	Y
13. data transformations	Y	Y	Y	Y	
14. recode	Y	Y	Y	Y	Y
15. recode if	Y	Y	Y	Y	Y
16. select		Y			
17. select if	Y	Y	Y		
18. compute		Y			
19. lag transformations	Y	Y	Y	Y	Y
20. curve smoothing	Y		Y	Y	Y
21. deseasonalization			Y	Y	
22. rank order			Y	Y	
	Y	Y	Y	Y	Y
total	14	19	16	13	10

files. The package has a downloading utility, so it was possible to use general ASCII files from the mainframe, but that is about it. The other packages offer much more flexibility in this area.

My three wishes for this package would be for it to have an index in the manual (an irritating oversight), to include a lot more procedures, and to really beef up the package's ability to use externally generated data files.

Microstat (Ecosoft)

Microstat is a good, solid statistical meal, with a bit of dessert. The statistical programs can be accessed through both interactive and batch mode, which is not possible with the other packages. The printouts include variable but not value labels. Data entry is prompted by case number and the variable name and number. Editing data is relatively easy. The package has a good description of missing data handling, as well as in-depth information on how missing data affects each of the major operations.

The accompanying tables demonstrate that Microstat falls between the other packages in procedures calculated and in the sophistication of data labeling and entry. For people who can only buy one package, it is a decent compromise between the best data entry and the most statistical procedures available. Apparently no one has yet decided to do both extremely well.

I really appreciated Microstat's ability to use externally generated data files with its programs. Microstat will accept comma-delimited ASCII files, fixed-field ASCII files, .DIF files, and dBASE II (.DBF) files. This greatly multiplies one's access to a wide variety of data sources. I did, however, have trouble using Microstat's batch mode or command file format. It takes at least one dry run through the program to set up the file. This feature is useful only when you want to run a lot of statistics and don't want to babysit the computer.

You will need considerably more time to put these files together than you would if you were using the batch processing system of StatPac, but it can be done. Researchers running tests on a single variable, or even a small number of variables, will find Microstat's data entry structure less time-consuming than StatPac's, since no code book has to be created just to get a single statistic run.

If I had three wishes for Microstat's next version, I would wish for a code book utility like StatPac's, the inclusion of value labels, and more procedures. You can never have too many procedures. They are of great psychological value. Even if you never use a particular procedure, you usually feel more comfortable knowing it is there, just in case.

NWA Statpak has to be called a statistical smorgasbord.

NWA Statpak (Northwest Analytical, Inc.)

If the first three packages provide basic sustenance, then NWA Statpak has to be called a statistical smorgasbord. There is so much variety in its statistical tables that you are not sure where to start, but you can keep sampling until you are stuffed. It has four different calculations of means (arithmetic, geometric, harmonic, and quadratic), eight different analyses of variance (two one-way ANOVAs, three two-way ANOVAs, three three-way ANOVAs), and 14 different distribution functions (see chart).

NWA Statpak is different from the other packages in that it can be both command or menu driven—though it's primarily oriented toward command use. It is strongly geared for running single procedures. You are essentially tied to your computer by having to ask for a separate, sequential analysis for each procedure and each variable you want to analyze. Its structure makes it flexible, but considerably more difficult to learn and use.

The impression I get from using this package is that it was designed primarily for interface with already existing data files. It is possible to convert files either to or from the Statpak file format to interface with files created by editors (ED, EDLIN, EDITOR), word processors (WordStar, Final Word, Writeone, and Image), spreadsheets (Lotus 1-2-3, SuperCalc), and data bases (dBASE II and R:BASE). If you have these programs you can greatly increase the sophistication of your data analysis with NWA Statpak. The package seems better designed for this purpose than for original data entry. More space is

given in the manual to describing the data translation process than is given to data entry, which should tell you something. At the same time, I had no difficulty using the Speed program to enter my own data and run analyses of it, and I used the Convert program to fool with some data from one of my spreadsheets. Both worked quite well.

One problem I encountered was that Statpak lacks both variable and value labels. This is not terribly serious when you only want a single statistic or a series of runs on a single variable. You are probably going to add the information directly into a report. But for many of us, there is a necessary "shelf life" for data. It sits until we can write an article, then sits again until parts of it can be used in another article. Unless you hand-label the NWA Statpak data, you will run the risk of having to guess what a particular printout means six months later when you want to expand on what you have already done.

Systat is like a 17-course meal with subtle sauces and a wine list.

My three wishes for NWA Statpak would be for variable and value labels, optional batch processing, and perhaps just a few more procedures.

Systat (Systat, Inc.)

Systat is the most powerful package I reviewed. It is complex, even intricate. To finish my gustatory analogy, it is a 17-course meal with subtle sauces and an extensive wine list for every course. It contains virtually everything from frequencies to canonical correlation. If your research is of the "one-man band" variety, you can probably get along without Systat by using a mainframe for the procedures the other packages lack. But if you use a team that includes a statistician, make sure the statistician has Systat, even if you don't. Systat should also be high on the "must have" software list of research institutions.

The four packages discussed above can be used

by those with minimal to moderate statistical skills. The full power of Systat is open only to pros. The fact that you could probably become a pro by spending a lot of time with Systat does not mitigate its complexity.

Overall, Systat compares favorably not just with other micro packages, but with many mainframe packages as well. It is command driven and has a very large set of commands. It has extremely good documentation on the algorithms it uses and their precision, and it provides you with resources for studying them further. It has more plotting facilities than any of the other four packages. Since it is written primarily in FORTRAN, rather than BASIC, it runs faster than the other packages. It is not a package you will be able to use five minutes after opening the cover, but it is one you will still be using five years later.

The most troublesome thing I found with Systat was data editing. You have to know something about programming in BASIC (the more the better) in order to effectively edit your data, even to change a single variable. Data entry also could be far better designed. It is possible to hand-enter data, and, with some experimentation, you should be able to convert some of your files from other sources to use in Systat. The manual recommends using a mainframe, a database, or a spreadsheet to create ASCII data files that can be read by Systat. As long as the files are in the format recognized by Systat, this will work quite well. But the other packages have better solutions to this problem.

I have three wishes for the Systat package, as well. First, I would like a user-friendly data entry system, to go along with the acceptable cordiality of the rest of the package. As a user I am very concerned that data be placed in my files easily and without errors. Second, the package needs a better internal data editor. Finally, and most surprisingly, the package lacks non-parametric statistics. This seems odd in a package that has virtually any other procedure you might ever want or need.

Comparing the packages

The results of my comparisons are summarized in the charts. I set up these charts from the documentation that accompanied the packages, from my use of the packages, and from phone conversations with the companies.

Overall

The programs described in this article warrant serious consideration by professionals. Their documentation ranges from good to excellent. The procedures offer high levels of precision, and once learned, are at least as easy to use as those for mainframe packages. Each package should handle the data sets for most research projects, though each requires some prior knowledge of statistics. Then again, they are not intended to *teach* statistics; they're designed to let you *do* statistical work conveniently and quickly.

MS-DOS versions

Systat Inc. released a new MS-DOS version of Systat in January 1985 that adds a full screen editor, non-parametric statistics, cluster analysis, and time series analysis, plus four new parametric statistics programs, to the package, without changing the price. If you have a Kaypro 16 or a Kaypro with the "Plus 88" configuration it would be worth researching this new version, since it is designed to overcome criticisms of the package reported in this article. However, the CP/M version reviewed remains unchanged and the criticisms of it are accurate.

Anderson-Bell released a revised MS-DOS version of ABstat in September 1984 that includes the addition of several non-parametric statistics. Otherwise the package remains the same. □

Robert T. Trotter, II, is an anthropologist and Director for Research Development at Pan American University in Edinburg, Texas.

Quick Reference Summary

All of the statistical packages reviewed will run on Kaypro 2s, 4s, or 10s. These companies also have MS-DOS versions.

PACKAGE	COST	COMPANY	ADDRESS
ABstat	$395	Anderson-Bell	P.O. Box 191 Canon City, CO 81212
StatPac	$400	Walonick Associates, Inc.	5624 Girard Av. Minneapolis, MN 55419
Microstat	$375	Ecosoft, Inc.	6413 N. College Ave. Indianapolis, IN 46220
NWA Statpak	$495*	Northwest Analytical, Inc.	520 N.W. Davis St. Portland, OR 97209
Systat	$495	Systat, Inc.	603 Main St. Evanston, IL 60202

*NWA Statpak offers an educational institution discount.