



StatNews #23: Desktop Analysis of Large Public-Use Data Files

September 24, 1997

Until recently, extracting subsets of large public-use data files was tedious and time-consuming, and required knowledge of mainframe computing systems. Government agencies that distribute public use data have simplified this process by distributing data on CD-ROM and diskette rather than on reel tapes, and by distributing a software package for data access along with many datasets. Now it is possible to perform secondary analyses of large data files entirely on your personal computer.

Many government datasets are packaged with a data extraction software package called SETS (Statistical Export and Tabulation System). SETS is a DOS-based package that allows users to browse data and documentation, search and print documentation, develop subsets of the data using logical expressions, perform simple tabulations of variables, and export data to common spreadsheet, database, or statistical software packages.

The program is menu-driven and offers on-line help as well as a manual that may be printed. The Browse menu allows users to read the documentation on-screen, search for key words, and print the documentation. You can also scroll through the data in a spreadsheet layout that provides variable names and explanations. SETS can be used to create simple tables, count records with specified characteristics, and perform basic mathematical functions on the data. If you want to perform further analyses on the data, SETS lets you select the records and variables you need and export them in ASCII or dBase format. SETS will also create SAS, SPSS, BMDP or Epi Info programs that will read the ASCII file and preserve variable names, labels, formats and value explanations from the original documentation.

While experienced programmers may still prefer to write their own programs to extract data from large public use data files, the SETS program makes these files accessible to all researchers. The forthcoming Windows 95 version of SETS, currently being tested, will make using data on CD-ROM even easier.

The Office of Statistical Consulting has several health and nutrition data sets on CD-ROM with SETS for use by Cornell researchers, including the 1994 Continuing Survey of Food Intakes by Individuals and the Third National Health and Nutrition Examination Survey. Contact Cara Olsen if you have any questions. For data on other topics, contact the Cornell Institute for Social and Economic Research (CISER) data archive.

Author: Cara Olsen

This newsletter was distributed to faculty and graduate students in the Division of Nutritional Sciences and the College of Human Ecology, and faculty in the College of Agriculture and Life Sciences, by the Office of Statistical Consulting. Please forward it to any interested colleagues and research staff. Anyone not receiving this newsletter who would like to be added to the mailing list for future newsletters should e-mail statcons@cornell.edu. Information about the Office of Statistical Consulting can be obtained at World Wide Web address <http://www.human.cornell.edu/admin/statcons/>.