

Micah Altman
Associate Director, Harvard-MIT Data Center
Harvard University
Cambridge, MA 02138
617-496-3847 (phone)
425-740-0715 (fax)
Micah_Altman@harvard.edu

Word Count: 993

Statistical Packages are collections of software designed to aid in statistical analysis and data exploration. The vast majority of quantitative and statistical analysis relies upon statistical packages for their execution. An understanding of statistical packages is essential to correct and efficient application of many quantitative and statistical methods.

Although researchers can, and sometimes do, implement statistical analyses in generalized programming languages like FORTRAN, C++, and Java, statistical packages offer a number of potential advantages over such 'hand-coding'. First, statistical packages save the researcher time and effort: Statistical packages provide software for a range of analyses that would take years for an individual programmer to re-implement. Second, statistical packages can provide a unified operating framework, and common interface, for data manipulation, visualization, and statistical analysis. At their best, tools within a package can easily be combined and extended, enabling the researcher to easily build upon previous work. Furthermore, even where a statistical procedure is simple enough or

specialized enough to be programmed by hand, statistical packages offer user interfaces that make it far easier for others to use new procedures. Third, statistical packages are more likely to produce correct results than hand-coded routines. Even simple tasks, like computing the standard deviation, can be tricky to implement in a way that yields reliably accurate results. Statistical packages are likely to be more extensively tested for bugs and numerical problems, and the implementations of these packages are usually more familiar with modern numerical methods than is the typical researcher. Fourth, statistical packages are likely to be faster than average implementations: A considerable amount of effort and refinement goes into the choice of algorithm for a particular analysis or statistical tool, and into tuning the software to work well on a specific computer architecture. (For an introduction to statistical algorithms see Gentle 2002.)

The chief problems researchers face when using statistical packages are choosing a package from the hundreds now available, and ensuring that the results obtained from it are correct and replicable. Obviously, each package will differ in terms of the features that it support, although there is considerable overlap among many large packages. Commercial giants like SAS[®] and SPSS[®], and large open-source packages like 'R' offer thousands of data manipulation, visualization and statistical features. The interfaces of the packages vary considerably as well: from the point-and-click style of SPSS to the command-line based approach of

SAS and R. Unless a particular package offers a crucial feature, the researcher may want to consider a modern statistical language like ‘S’ which has broad functionality and both commercial (‘S+’) and open-source implementations (‘R’). (see Venables and Ripley 2002 for an introduction)

Less obviously, packages differ significantly in terms of speed, accuracy and extensibility. When approaching a statistical package, researchers may want to ask: Has this package been tested for accuracy? Does it document the algorithms it uses to compute statistical analyses and other quantities of interest? Was it designed to efficiently process data of the size and structure to be used in the proposed research project? What programming facilities are available? Does the programming language support good programming practices, like object-oriented design? Are the internal procedures of the package available for inspection and extension? How easily is it to use this package with an external library, or as part of a larger programming and data analysis environment?

While statistical packages are viewed as essential tools for constructing an analysis, they are often considered highly interchangeable. Thus, once results are obtained, researchers often fail to document the package used in subsequent publications. In fact, however, statistical packages vary greatly with respect to accuracy and reliability, and reported results may be dependent on the specific

package and version, and the computational options used during the analysis.

(McCullough & Vinod 1997)

There are several approaches to ensuring accurate results. One may want to replicate the results across different packages and using different algorithms, to check that different methods of computing the results yield the same answer. In packages that support extended precision, increasing the numerical precision of computations during the analysis can be useful in avoiding certain types of numerical error. Another approach is sensitivity analysis – where small amounts of noise are introduced into calculations, data, and/or starting values, in order to verify that the solution found is numerically robust. (Altman, Gill & McDonald 2003)

Finally, three steps should be taken to increase the replicability of the results. Since complex analysis may depend on the particular computational strategy used both in researchers analysis and within the statistical package: reporting (or archiving) the code used to run the analysis, the package and version used, and any computational options used during the runs is essential to ensuring replicable results. Archiving a copy of the statistical package itself may be warranted in some cases where exact replication is vital – although a few packages have the ability to replicate the syntax rules and computational options used in older

versions, most packages change both the programming syntax and implementation over time. Second, many social science data sets are difficult or impossible to re-create -- publicly archiving any unique data used for an analysis is often a necessity for successful replications. Third, the formats that statistical packages use are often both proprietary and subject to change over time, posing grave obstacles to later use. Researchers can ensure that the data remains accessible even as proprietary formats change by exporting data using a plain-text format, along with adequate documentation of missing values, weights, and other special coding of the data. (Altman, Gill & McDonald 2003)

MICAH ALTMAN

References

Altman, M., J. Gill, and M.P. McDonald, 2004. *Numerical Methods in Statistical Computing for the Social Sciences*, John Wiley and Sons: New York.

B. D. McCullough, H. D. Vinod, 1997. "The Numerical Reliability of Econometric Software", *Journal of Economic Literature* 37 (2): 633-665

Gentle, James, 2003. *Computational Statistics*, Springer-Verlag: New York.

W. N. Venables and B. D. Ripley, 2002. Modern Applied Statistics with S.
(Fourth Edition), Springer-Verlag: New York.