GENERATING AUTOMATED DATASET SUMMARIES

Agustin Calatroni and Herman Mitchell
Rho, Inc. Chapel Hill, NC

Every time a dataset is created, either for data management proposes or for statistical analyses, it is imperative that the each variable be reviewed. Not only should the evaluation provide summary statistics and graphical displays to detect data errors, it should also present the results in a thorough, but succinct manner. To accomplish this goal, descriptive summaries for each variable should be created according to their characteristics. As simple as this task sounds, no major commercial statistical software package has a shell procedure to do it.

The best available option for generating descriptive data set summaries is found in the Hmisc (Harrell Miscellaneous) package for the R statistical programming environment. However these freely available tools have two main draw backs. First, the user must invest a substantial amount of time in learning a new programming language. Second, most statistical and clinical coordinating center data sets are stored as SAS files which are not straight forward analyzable using R.

The following presentation introduces a SAS macro that automatically creates a thorough and succinct data book where the variable type (dichotomous, ordinal, nominal, continuous, etc) dictates which summary statistics are displayed. The SAS macro combines Hmisc functions, Sweave and Latex to create an innovative databook. All the processes are run behind the scenes; as such anyone with access to SAS can create these displays with no need to learn a new programming language.