Brief introduction to the use of the *STATISTICA* software

Adelin Albert
Anne-Françoise Donneau
Lixin Zhang

Biostatistics – School of Public Health
University of Liège

Erasmus – Summer Programme 2008

Foreword

Most datasets that a user encounters in practice can be displayed on a sheet of paper as a table or matrix of observations consisting of rows and columns. The rows correspond to subjects or objects, while the columns represent the variables of the problem. Often, the first column is used to identify the observations (rows) and another one is used as a group-identifier when observations (subjects or objects) belong to different groups or categories. Variables can be of different types: continuous (weight, height), discrete (number of drugs taken), binary (gender), nominal (race, blood group) or ordinal (stage of a cancer). Datasets can be very simple (one variable and a few individuals) or quite complicated (tenths of variables of different types and thousands of subjects or objects).

In the past, the statistician analysed the data, variable by variable, computing mean values or drawing histograms and 2D-plots, a cumbersome and lengthy task done by hand or with a handheld pocket calculator. Today, all this is done electronically by commercial statistical packages in a fast and (too) simple way, provided the dataset has been entered into the computer as an electronic worksheet, in exactly the same way as the ancient sheet of paper. *"If it is a good thing that statistical packages have become cheaper, it is a bad thing that they have become easier to use"*, declared recently a well-known statistician. Indeed, while offering formidable computing power, speed and data analysis capacities, the misuse of statistical software and the lack of adequate statistical knowledge can be scientifically disastrous. By contrast, the wise and well-thought utilisation of automated statistical tools and methods can be a remarkable way of looking at data, a source of richness and scientific progress.

Thus, before starting any data analysis, graphical display or hypothesis testing, it is of prime importance to get the dataset at hand properly entered, stored and saved into a computer file, so that it can be retrieved later on. This process is called "data entry". Next, the stored dataset can be edited, a process called "data management". Indeed, new variables can be created from the old ones, observations can be added or deleted, or sorted out according to several criteria. Even more, the original dataset can be split into several sub-datasets or conversely different stored datasets can be merged together into a single one. The next phase consists in describing and exploring the data. This can be done numerically by computing characteristics like means and standard deviations (summary statistics) or graphically by displaying histograms, X-Y plots or other more advanced graphs (graphical display). The ultimate process is to use the data for inferential purposes: estimating unknown population parameters (with confidence intervals) or testing hypotheses (t-test, chi-squared test). More advanced modelling or multivariate techniques (like multiple regression analysis) are also possible but are not addressed hereafter.

Contents