

Quantiles

Marks 22 of the NAG Fortran Library and the NAG Toolbox for MATLAB include a new routine for fast extraction of quantile values from unordered sets of real numbers. Put simply, a quantile can be viewed as a pointer into a data set such that a certain proportion of values lie below the quantile and the remainder lie above. Some quantiles have more familiar names; the well-known term median is the name given to the middle value of a data set when the data are arranged into ascending order. If there happen to be an even number of points, of course there is no middle value, and by definition the average of the two middle values is taken. In the terminology of the new NAG routine, the median is the 0.5 quantile.

Percentile values are given by the quantiles 0.01, 0.02, 0.03 etc. The first percentile value is the data point one hundredth of the way up the series when the values are ordered. The requirement for the software came from a major financial institution, which needed a faster way of extracting this information. The obvious way to compute quantile values is to rearrange the data into ascending order, and then pick out the value of interest. However, for large data sets this is not efficient - the time needed to sort the data is significant.

In response to this requirement, NAG collaborator Professor Mike Giles, of Oxford University, developed a partial Quicksort algorithm, using the Quicksort method already in the NAG Libraries as a starting point. As hoped, it turned out that the new routine performed much more efficiently because it is possible to extract the quantile value without sorting the entire data set.

The graph shown below shows just how fast the new routine is; it compares how much time it takes to get the median of a data set using the new routine G01AMF as opposed to sorting using routine M01CAF.

