Algorithm 727
Quantile Estimation Using Overlapping Batch Statistics

SHERIF HASHEM and BRUCE SCHMEISER
Purdue University

An efficient algorithm for estimating the qth quantile of a set of n data points is introduced. The standard error of the quantile estimate is estimated using overlapping batch statistics. A driver program is included.

Categories and Subject Descriptors: E.1 [**Data**]: Data Structures—*trees*; G.3 [**Mathematics of Computing**]: Probability and Statistics—*statistical software*

General Terms: Algorithms

Additional Key Words and Phrases: Aggregation, autocorrelation, Monte Carlo, stochastic simulation, time series

1. DESCRIPTION

The function **obq** is a **C** function that estimates the qth quantile for a given set of data as well as the standard error of that estimate.

2. METHOD

The qth quantile is estimated by the jth order statistic, where $j = \max(1, \lceil n * q \rceil)$ and n is the number of observations.

The standard error of the point estimator of the qth quantile is estimated using Overlapping Batch Statistics (OBS), as described in Schmeiser et al. [1990]. Carlstein [1986] discusses the statistical properties of similar standard-error estimators for general statistics from covariance-stationary time-series data.

Through the use of an efficient ranking and selection method and a 2-3 tree data structure (see Aho et al. [1974]), the run time of the OBS algorithm was reduced from $O(n \log(n))$ to $O(n \log(m))$, where m is the batch size.

This work was supported by PRF Research Grant 6901627 from Purdue University and by NSF Grant DMS-8717799.
Authors' address: The School of Industrial Engineering, Purdue University, 1287 Grissom Hall, W. Lafayette, IN 47907-1287.

The batch size $m \in \{2, 3, \ldots, n-1\}$ is an algorithm parameter. Its value does not affect the value of the point estimate, but it does affect the value of the standard-error estimate. A quantile estimate is computed for each successive batch of m observations, so choosing m to satisfy $m \geq 1/\min\{q, 1-q\}$ is necessary for reasonable performance. Choosing m too small leads to large absolute bias because the batch quantiles are biased; choosing m too large leads to large variance because there are few batch quantiles. We have found that using $m = n/20$, when consistent with the value of q, is a reasonable heuristic.

2.1 Arguments List

—Input:

a: array of size n containing input data points (*float*).

n: size of the input array a (*unsigned long int*).

m: overlapping batch size (*between 2 and (n − 1) inclusive, unsigned long int*).

q: quantile to be estimated (*between 0. and 1. inclusive, float*).

—Output:

The output of **OBQ** is returned to the calling program using *float* pointers to the following variables:

qq: point estimate of the qth quantile (pointer: *pqq*).

stdv: estimate of the standard deviation of qq (pointer: *pstdv*).

3. MODULES

The following is a description of the basic modules.

DRIVER. Example of a main program. Artificial input data is generated in reverse sequence between 0 and 99, passed to the module **OBQ**, and the results are printed. (The batch size, m, is set to 4 and the quantile $q = 0.4$.)

OBQ. Computes the point estimate of the qth quantile and the estimate of its standard deviation. It employs the two modules SELCT and XTREE described below.

SELCT. An $O(n)$-time function used for estimating the qth quantile using *divide-and-conquer* strategy.

XTREE. An $O(n \log(m))$-time function used for estimating the variance of the qth quantile's estimate using OBS with batch size m. A 2-3 tree is constructed with the elements of the first batch in $O(m \log(m))$ time. Using a simple search-delete-insert scheme, the tree is updated to reflect the data in the subsequent overlapping batches, and hence batch estimates are obtained in $O(\log(m))$ time per batch.

REFERENCES

AHO, A., HOPCROFT, J., AND ULLMAN, J. 1974. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Mass., 146–152.

CARLSTEIN, E. 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Stat.*, *14*, 3, 1171–1179.

SCHMEISER, B., AVRAMIDIS, T., AND HASHEM, S. 1990 Overlapping batch statistics. In *Proceedings of the 1990 Winter Simulation Conference*. IEEE, New York, 395–398.

Received September 1991; revised February 1993 and April 1993; accepted April 1993