THE GENERATION OF ORDER STATISTICS IN DIGITAL COMPUTER SIMULATION: A SURVEY

Bruce W. Schmeiser

ABSTRACT

Order statistics are often needed in computer simulation. Common examples are quantile estimation and censored data test statistics. Methods for generating order statistics in various contexts are surveyed. Sorting and the use of histograms, the most general methods, are first discussed, followed by a method for non-identically distributed samples. Finally, the very powerful methods applicable to iid random variables are surveyed.

I. INTRODUCTION

Let $x_1, x_2, \ldots, x_n$ be a set of random observations. The associated order statistics are $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ where $x_{(i)}$ is the i^{th} smallest value. Although the assumption of independent and identically distributed values is commonly made, (David [7], Gibbons [10]), in simulation studies the independence assumption need not hold and the identically distributed assumption often does not hold.

Order statistics arise in simulation models in a number of ways. $X_{(i)}$ is the time of failure for a system requiring i of n components to operate, where the distribution of the times to failure for each component are not necessarily independent or identically distributed. In PERT simulation the value of $x_{(n)}$, the maximum of n observations, is the time at which a future node is realized, where the observations may not be independent and usually are not identically distributed. In next-event timekeeping, the time of the next event can be viewed as the first order statistic of the times in the future events calander.

In other cases, order statistics may be the desired output rather than an input of only intermediate interest. Queuing simulations often estimate the p^{th} percentile of the waiting time distribution by the appropriate order statistics, as do Monte Carlo studies of distributions of test statistics.

This paper considers order statistics in the simulation context. In Section II, the various ways order statistics arise in simulation are categorized. Later sections then discuss by category the methods available for generating order statistics.

II. CATEGORIZATION OF ORDER STATISTICS

Order statistics in simulation can be categorized several ways. In this section a categorization is given yielding twelve separate situations. While each is different, in later sections it will become clear that similar methods sometimes apply to more than one situation.

Letting F_i^{-1} denote the inverse cummulative distribution function for the i^{th} random variable, the categorization is based upon

1. Distribution of the x_i's is
   a. known, but F_i^{-1} is not available for some $i = 1, 2, \ldots, n$
   b. known, and F_i^{-1} is available for all $i = 1, 2, \ldots, n$
   c. unknown.
2. The x_i's are
   a. independent
   b. dependent
3. The x_i's are
   a. identically distributed
   b. not identically distributed

which results in 3 x 2 x 2 = 12 separate categories of order statistics in simulation.

The first partitioning is based on the manner in which the order statistics arise. If the distribution of the x_i's is known, as in the PERT and reliability examples above, the problem is to generate order statistics for the known distributions. On the other hand, the simulation may be used to estimate quantiles or entire distributions, which corresponds to the x_i's being the simulation output having unknown distributions. In the latter case some sort of selection or sorting must be performed on the x_i's to obtain the order statistics.

In the former case of the distribution of the x_i's being known, two situations arise. If the inverse cumulative distribution functions, F_i^{-1} $i = 1, 2, \ldots, n$, are available, efficient methods become available for more direct generation of order sta-

tistics than selection or sorting.

A second partition arises as to whether the x_i's are independent. As might be expected, independence sometimes leads to efficient methods.

The third partition is whether or not the x_i's are identically distributed. Identically distributed values lead to better methods than non-identically distributed values.

III. SORTING, SELECTION, AND HISTOGRAMS

A general method applicable to all twelve cases, but not the best for several cases, is to generate $x_1, x_2, \ldots, x_n$ and to sort the n values to obtain $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Inefficient methods require time proportional to n^2, while efficient methods require time proportional to n ln n (see Knuth [14]). There is no option to sorting if 1) all order statistics are needed and 2) no amount of approximation is allowed, except as noted in Section IV.

If only some order statistics are needed, improvement over n ln n sorting can be made. Chambers [3] gives algorithm PSORT for partially sorting $x_1, x_2, \ldots, x_n$ to obtain only specified order statistics. For a small fixed number of specified order statistics, the time for PSORT is nearly proportional to n. An 80 line FORTRAN implementation is given in [3]. (See also Chambers [4].) An example when only a fraction of the order statistics are needed would be in plotting the distribution function from a sample of many observations. Determining every tenth, say, order statistic may be satisfactory.

Floyd and Rivest [8] give an algorithm for selecting the i^{th} order statistic from n values. In [9] they show the number of comparisons is n + min (k, n-k) + o(n).

If some approximation is allowed, a histogram can be used. A histogram having $m << n$ cells can be used to sort the observations in one pass (as they are generated). For discrete distributions, histograms can often be used with no loss of information. For continuous values, however, information is lost in that the c_i values in cell i can not be recovered completely. The cell containing the i^{th} order statistic may be determined with certainty, but how to best determine the point in the cell to represent the i^{th} order statistic is not obvious. The cell mid-point is the most straight forward choice. David and Mishriky [6], who considered only $n \leq 100$, state that "the effect of grouping... is... of minor importance for h = 0.6 (standard deviations) which represents quite coarse grouping." Schmeiser and Deutsch illustrate, via deterministic calculations, three potentially serious idiosyncrasies which occur with the large sample sizes common in quantile estimation via simulation. Two are potential problems concerning order statistics directly:

1. The expected value of the midpoint estimator can depend heavily upon both the histogram cell width and the placement of the histogram.
2. Larger cells can result in smaller variance, thus incorrectly indicating an answer more accurate than actually obtained.

Schmeiser [22] suggests estimating the r^{th} order statistic as $a+b[q-(\sum_{i=1}^{q} c_i - r+1)/(c_q+1)]$ where q is the smallest integer such that $\sum_{i=1}^{q} c_i \geq r$. This estimator behaves considerably more like the true order statistic than does the crude cell midpoint estimator.

IV. METHODS WHEN F_i^{-1} ARE AVAILABLE

When F_i^{-1} is available, values of x_i can be generated using $F_i^{-1}(u)$ where $u \sim U(0,1)$. This inverse transformation method leads to efficient methods for order statistics when F_i^{-1} is available. Often F_i^{-1} is available as a closed-form formula, as is the case for the Weibull, exponential, Cauchy, double exponential, uniform, triangular, and power series distributions. The general families of distributions of Burr [12], Ramberg and Schmeiser [19,20] and Schmeiser and Deutsch [24] also have closed-form F_i^{-1}. It has sometimes been overlooked that numerical solution of F_i^{-1} is equally valid, even if not as easy to implement. For example, most computer packages have routines to provide the u^{th} quantile of the normal and gamma distributions.

Schmeiser [23] gives a method for reducing the effort to generate the maximum, minimum and range of a sample arising from non-identically distributed random variables having easy-to-evaluate F_i^{-1}. Based on a preliminary check of u_i to a partition p_i, the calculation of $x_i = F_i^{-1}(u_i)$ is sometimes avoided. Since the evaluation of F_i^{-1} is often time consuming, substantial reduction in computation time can result in simulations where generation times play a significant role. While conceptually valid for dependent random variables, the conditional distribution function $F_i^{-1}(u|x_1, x_2, \ldots, x_{i-1})$ must be available. Other than for the multivariate normal, this is rare.

The most impressive gains in efficiency, ease of implementation, and memory requirements are obtained for independent, identically distributed values generated via F_i^{-1}. Schucany [27] gives the following method for generating $x_{(i)}$:

Let $v_1, v_2, \ldots, v_n$ be independent U(1,0) values.

Set $u_{(n)} = v_1^{1/n}$, $u_{(n-1)} = u_{(n)} v_2^{1/(n-1)}$ and in general $u_{(n-i)} = u_{(n+1-i)} v_{i+1}^{1/(n-i)}$

Set $x_{(i)} = F^{-1}(u_{(i)})$ for all desired order statistics. That the i^{th} uniform order statistic is tranformed directly to the i^{th} order statistic of the distribution of interest follows from F^{-1} being a

monotonically nondecreasing function. This recursive algorithm requires effort linear in n and is therefore preferable to complete sorting for large n. Even for small n this approach is efficient if only a few of the extreme order statistics are required.

If the extreme order statistics are not needed, this procedure can be improved by noting that

$$u_{(i)} = 1 - \exp\left[\sum_{j=1}^{i} (\ln v_j)/(n+1-j)\right]$$

for i = 1, 2, ..., n,

since v^m is calculated as exp(m ln v). Lewis [15] carries this a step further by pointing out that it is faster to obtain ln v_j as the negative of a standard exponential variate than to take the logarithm of a randomly generated U(0,1) value v. Either Marsalia's method (see Knuth [13]) or Lewis and Learmouth [16] would be appropriate to generate the exponential values.

Lurie and Hartley [17] published a method analagous to Schucany's at about the same time, but began their recursion at $x_{(1)}$ and proceeded toward $x_{(n)}$. In addition to the recursive algorithm, under the heading "Simultaneous generation of order statistics for a multiplicity of sample sizes," they note that the i^{th} order statistic $u_{(i)}$ may be generated as a ratio of gamma random variables. (Note the erroneous substitution of "1" for "i" in their equation 10.) Ramberg and Tadikamalla [21]noted more directly that $u_{(i)}$ has a beta distribution with parameters i and n+1. Given the recent advances in beta variate generation (Cheng[5] and Schmeiser and Shalaby [26]) this method is quite efficient when only one statistic from the center of the sample is needed. If more than one is needed from the center of the sample, the recursion algorithms can be applied beginning at the value generated as a beta variate.

Rabinowitz and Berenson [18] compare the "grouping method" to the other methods for independent identically distributed random variables when all n order statistics are needed. The grouping method consists of partitioning the unit interval into several subintervals, generating n U(0,1) values, performing a permutation sort on each group of observations for each subinterval, and using F^{-1} to generate the $x_{(i)}$'s from the sorted $u_{(i)}$'s. This grouping method was fastest, but obviously was somewhat more difficult to implement.

V. MISCELLANEOUS RESULTS

1. Sillitto [31] gives some relationships between expectations of order statistics which may be useful when various sample sizes are of interest.
2. Some recent work on quantile estimation not referenced above included Goodman, Lewis and Robbins [11], Iglehart [12], and Seila[28,29,30].

REFERENCES

1. Burr, I.W., "Cumulative frequency functions," Annals of Mathematical Statistics, 13(1942), 215-232.
2. Burr, I.W., "Parameters for a general system of distributions to match a grid of α_3 and α_4," Communications in Statistics, 2(1973), 1-21.
3. Chambers, J. "Algorithm 410. Partial Sorting [M1]," Comm. ACM, 14(1971), 357-8.
4. Chambers, John M., Order Statistics: Sorting and Partial Sorting (Computational Methods for Data Analysis), Wiley, N.Y. (1977).
5. Cheng, R.C.H., "Generating Beta Variates with Nonintegral Shape Parameters," Communications of the ACM, 21 (1978), 317-322.
6. David, H.A. and Mishriky, R.S., "Order Statistics for Discrete Populations and for Grouped Samples," Journal of the American Statistical Association, 63 (1968), 1390-1398.
7. David, H.A., Order Statistics, John Wiley and Sons, New York, 1971.
8. Floyd, Robert W. and Rivest, Ronald W., "Expected Time Bounds for Selection," Communications of the ACM, 18 (1975), 165-172.
9. Floyd, Robert W. and Rivest, Ronald W., "Algorithm 489. The Algorithm SELECT for Finding the i^{th} Smallest of n Elements," Communications of the ACM, 18 (1975), 173.
10. Gibbons, J.D., Nonparametric Statistical Inference, McGraw-Hill Book Co., New York, 1971.
11. Goodman, A.S., Lewis, P.A.W. and Robbins, H.E., "Simultaneous Estimation of Large Numbers of Extreme Quantiles in Simulation Experiments," October, 1972, unpublished.
12. Iglehart, D.L., "Simulating Stable Stochastic Systems, VI: Quantile Estimation," Journal of the Association for Computing Machinery, 23 (1976), 347-360.
13. Knuth, D.E., The Art of Computer Programming, Volume 2/Seminumerical Algorithms, Addison-Wesley Publishing Company, Reading, Massachusetts, 1969.
14. Knuth, D.E., The Art of Computer Programming, Volume 3/Sorting and Searching, Addison-Wesley Publishing Company, Reading, Massachusetts, 1973.
15. Lewis, P.A.W., "Large-Scale Computer-aided Statistical Mathematics," Proceedings of Computer Science and Statistics: Sixth Annual Symposium on the Interface, (1972), 1-15.
16. Lewis, P.A.W., and Learmonth, G., "Naval Postgraduate School Random Number Generator Package LLRANDOM," Naval Postgraduate School Report NPS55Lw73061A, Monterey, CA, 1973.
17. Lurie, D. and Hartley, H.O., "Machine-generation of Order Statistics for Monte Carlo Computations," The American Statistician, 26 (1972), 26-27.
18. Rabinowitz, M. and Berenson, M.L., "A Comparison of Various Methods for Obtaining Random Order Statistics for Monte Carlo Computations," The American Statistician, 28 (1974), 27-29.
19. Ramberg, John S. and Schmeiser, Bruce W., "An Approximate Method for Generating Symmetric Random Variables," Communications of the ACM, 15 (1972) 987-990.
20. Ramberg, John S. and Schmeiser, Bruce W., "An Approximate Method for Generating Asymmetric Random Variables," Communications of the ACM, 17 (1974), 78-82.
21. Ramberg, J.S. and Tadikamalla, P.R., "On the Generation of Subsets of Order Statistics," Journal of Statistical Computation and Simulation, 6 (1978), 239-41.
22. Schmeiser, B.W., On Monte Carlo Distribution Sampling, with Application to the Component Randomization Test, Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Ga., 1975.
23. Schmeiser, B.W., "Generation of the Maximum (Minimum) Value in Digital Computer Simulation," Journal of Statistical Computation and Simulation,

to appear.
24. Schmeiser, Bruce W. and Deutsch, S.J., " A Versatile Four Parameter Family of Probability Distributions, Suitable for Simulation," AIIE Transactions, 9 (1977), 176-182.
25. Schmeiser, B.W. and Deutsch, S.J., "Quantile Estimation from Grouped Data: The Cell Midpoint," Communications in Statistics: Simulation and Computation, B6(3) (1977), 221-234.
26. Schmeiser, Bruce W. and Shalaby, M.A., "Acceptance/Rejection Methods for Beta Variate Generation," Technical Report IEOR 77014, Southern Methodist University, Dallas, Texas 75275.
27. Schucany, W.R., "Order Statistics in Simulation," Journal of Statistical Computation and Simulation, 1 (1972), 281-286.
28. Seila, A.F., "On the Performance of Two Methods for Quantile Estimation in Regenerative Processes," Bell Telephone Laboratories, Holmdel, New Jersey, 1978.
29. Seila, A.F., Quantile Estimation Methods for Discrete Event Simulations of Regenerative Systems, Ph.D. Thesis, University of North Carolina, Chapel Hill, 1976.
30. Seila, A.F., "Quantile Estimation Methods in Discrete Event Simulations of Regenerative Systems," Technical Report 76-12 (1976), Operations Research and Systems Analysis, University of North Carolina at Chapel Hill.
31. Sillitto, G.P., "Some Relations between Expectations of Order Statistics in Samples of Different Sizes," Biometrika, 51 (1964), 259-262.