EFFICIENT COMPUTATION OF PROBABILITIES OF EVENTS DESCRIBED BY ORDER STATISTICS AND APPLICATION TO A PROBLEM OF QUEUES

Lee K. Jones and Richard C. Larson

OR 249-91 May 1991

Efficient Computation of Probabilities of Events Described by Order Statistics and Application to a Problem of Queues

by

Lee K. Jones
Institute for Visualization and Perception Research and
Department of Mathematics
University of Massachusetts-Lowell,
Lowell, Massachusetts 01854

Richard C. Larson
Operations Research Center and
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

May 1991

ABSTRACT

Consider a set of N i.i.d. random variables in [0, 1]. When the experimental values of the random variables are arranged in ascending order from smallest to largest, one has the *order statistics* of the set of random variables. In this note an $O(N^3)$ algorithm is developed for computing the probability that the order statistics vector lies in a given rectangle. The new algorithm is then applied to a problem of statistical inference in queues. Illustrative computational results are included.

Key Words: Order statistics, queues, statistical inference, queue inference engine.

Introduction

Let $X_1, X_2, \ldots, X_{N(1)}$ be an i.i.d. sequence of random variables with values in [0,1] where the sequence length N(1) is an independent random integer. Recently, in an application to queue inference [6,7], an efficient algorithm [$O(N^3)$] has been developed to compute the conditional cumulative probability of the vector of order statistics, $\Pr\{X_{(1)} \le t_1, X_{(2)} \le t_2, \ldots, X_{(N)} \le t_N \mid N(1)=N\}$, for the case of each X_i uniform. (The algorithm presented here will efficiently calculate the latter probabilities for X_i having arbitrary given c.d.f. F(x).) Apparently the question of efficient computation of cumulative probabilities for the order statistics vector has not been previously treated in the literature. (See for example [1], [4].)

It is most natural to ask for an efficient algorithm to calculate the probability of an order statistics vector lying in a given N-rectangle, i.e., to compute $\Gamma(\underline{s}, \underline{t}) \equiv \Pr\{s_1 < X_{(1)} \le t_1, s_2 < X_{(2)} \le t_2, \ldots, s_N < X_{(N)} \le t_N \mid N(1) = N\}$, where $\underline{s} \equiv (s_1, s_2, \ldots, s_N)$, $\underline{t} \equiv (t_1, t_2, \ldots, t_N)$. We note that the method of computing the probability of a rectangle by applying repeated differences to the cumulative will require 2^N evaluations of the cumulative. This is too slow for many applications. In this note we develop an efficient algorithm [$O(N^3)$] to compute these probabilities for arbitrary rectangular regions where the X_i have a given c.d.f. F(x). New applications are given for deducing queue statistics from transactional data.

1. Analysis

Assume $X_i \in (0, 1]$ and $0 \le t_1 \le t_2 \le \ldots \le t_N \le 1$, $0 \le s_1 \le s_2 \le \ldots \le s_N \le 1$ and $s_i \le t_i$ for $i = 1, 2, \ldots, N$. Note that, since $\{t_i\}$ and $\{s_i\}$ are each nondecreasing as sequences,

we may merge the two sequences into $\{v_i\}_{i=1}^{2N}$, ordered according to magnitude, using only O(N) operations. Consider

$$W_{ki} \equiv \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_i\}, s_2 < X_{(2)} \le \min\{t_2, v_i\}, \ldots, \tag{1}$$

$$s_k < X_k \le \min\{t_k, v_i\} \mid N(1) = k\}, \qquad i = 1, 2, \ldots, 2N, \quad k = 1, 2, \ldots, N$$

We want to compute $W_{N,2N} \equiv \Gamma(\underline{s}, \underline{t})$. This will require recursive computation of entries of the matrix $\mathbf{W} = (W_{ki})$, starting with k = 1. By noting impossible events we see that for all k,i such that $s_k \ge v_i$, $W_{ki} = 0$, implying that $W_{k1} = 0$ for k = 1,2, . . . , N. We also require as a boundary condition

$$W_{0i} = 1 \qquad \text{for } i = 1,2, \ldots, 2N - 1, \tag{2}$$

which can be interpreted to be the probability that the event inequalities will be satisfied, given no random variables (hence no inequalities) in [0, 1].

Theorem. W_{ki} can be computed using the following recursion:

$$W_{ki} = \sum_{\substack{j=0,1,2,\ldots \\ \text{s.t. } v_i \le t_{k-j+1}}} \binom{k}{i} W_{k-j,\,i-1}\left[F(v_i) - F(v_{i-1})\right]^j$$

Proof.

$$W_{ki} \equiv \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_i\}, s_2 < X_{(2)} \le \min\{t_2, v_i\}, \ldots,$$

$$s_k < X_k \le \min\{t_k, v_i\} \mid N(1) = k\}$$

$$= \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \le \min\{t_2, v_{i-1}\}, \ldots,$$

$$s_k < X_k \le \min\{t_k, v_{i-1}\} \mid N(1) = k\}$$

$$+ \quad \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \le \min\{t_2, v_{i-1}\}, \ldots,$$

$$s_{k-1} < X_{(k-1)} \le \min\{t_{k-1}, v_{i-1}\}, v_{i-1} < X_{(k)} \le \min\{t_k, v_i\} \mid N(1) = k\}$$

$$+ \quad \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \le \min\{t_2, v_{i-1}\}, \ldots,$$

$$s_{k-2} < X_{(k-2)} \le \min\{t_{k-2}, v_{i-1}\}, v_{i-1} < X_{(k-1)} \le \min\{t_{k-1}, v_i\},$$

$$v_{i-1} < X_{(k)} \le \min\{t_k, v_i\} \mid N(1) = k\}$$

$$+ \ldots + \quad \Pr\{s_1 < X_{(1)} \le \min\{t_1, v_{i-1}\}, s_2 < X_{(2)} \le \min\{t_2, v_{i-1}\}, \ldots,$$

$$s_{k-j} < X_{(k-j)} \le \min\{t_{k-j}, v_{i-1}\}, v_{i-1} < X_{(k-j+1)} \le \min\{t_{k-j+1}, v_i\}, \ldots,$$

$$v_{i-1} < X_{(k)} \le \min\{t_k, v_i\} \mid N(1) = k\} + \ldots$$

The term explicitly displaying $X_{(k-j+1)}$ on the RHS can be nonzero only if $\min\{t_{k-j+1}, v_i\} = v_i$. Hence we can write

$$W_{ki} = W_{k,i-1} + \binom{k}{1} W_{k-1,i-1}\left(F(v_i) - F(v_{i-1})\right)$$

$$+ \binom{k}{2} W_{k-2,i-1}\left(F(v_i) - F(v_{i-1})\right)^2$$

$$+ \ldots + \binom{k}{j} W_{k-j,i-1}\left(F(v_i) - F(v_{i-1})\right)^j + \ldots + W_{0,i-1}\left(F(v_i) - F(v_{i-1})\right)^k$$

for all j satisfying $v_i \le t_{k-j+1}$ and where the last term on the RHS utilizing the boundary condition Eq. (2) is included only if $v_i \le t_1$. ■

As a verification of the recursion we obtain as expected at the first iteration

$$W_{1i} = F(\min\{t_1, v_i\}) - F(s_1) \qquad i = 2, 3, \ldots, 2N.$$

The matrix $\mathbf{W} = (W_{ki})$ can be partitioned into three regions:

(1) $W_{ki} = 0$ for $k \geq i$;

(2) $W_{ki} \geq 0$ for $i - N < k < i$;

(3) $W_{ki} > 0$ for $k \leq i - N$;

Hence the maximum possible number of nonzero terms in row k is $2N - k$, and the minimum number is $N - k + 1$. The recursion to obtain W_{ki} requires computation and addition of up to $k + 1$ terms. Thus, row k of (W_{ki}) requires computation of up to $(2N - k)(k + 1)$ terms. The total number of terms required to compute (W_{ki}) is

$$\sum_{k=1}^{N} (2N-k)(k+1) = \frac{2}{3}N^3 + 2N^2 - \frac{2}{3}N$$

yielding an $0(N^3)$ procedure. For the special case $s_i = 0$, $i = 1, 2, \ldots, N$, all W_{ki} in region (2) of **W** are zero and we have the problem of Refs. [6,7].

2. Applications to the *Queue Inference Engine*

Ref. [6] uses events of order statistics to derive an algorithm, the "Queue Inference Engine," to compute various performance measures of Poisson arrival queues. In particular, N is the total number of customers to arrive to the queueing system during a *congestion period*, a continuous time interval during when all servers are busy and all arriving customers must queue for service. And t_i is the observed time of departure of the i^{th} customer to leave the system during the congestion period. Using the fact that the N unordered arrival times during any fixed time interval (0,T] are i.i.d. uniform and scaling the congestion period to (0, 1], then in our notation here $\Gamma(0, t)$ is the *apriori* probability that the (unobserved)

arrival times $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$, obey the inequalities $X_{(i)} \leq t_i$ for all $i = 1, 2, \ldots, N$, a condition that must hold for the congestion period to persist. (Additional new work on the Queue Inference Engine is reported in [2], [3] and [5].)

2.1 The Maximum Experienced Queue Delay

Assume we have a first-come, first-served (FCFS) queue. Suppose we set $\underline{s} = \underline{t} - \tau$, i.e., $s_i = \text{Max}\{t_i - \tau, 0\}$ for all $i = 1, 2, \ldots, N$. Then $\Gamma(\underline{t} - \tau, \underline{t})$ is the *a priori* probability that the observed departure time inequalities will be obeyed *and* that no arrival waits more than τ time units in queue. Define

$D(\tau \mid \underline{t}) \equiv$ conditional probability that none of the N customers waited more than τ time units, given the observed departure time data.

Clearly,

$$D(\tau \mid \underline{t}) = \Gamma(\underline{t} - \tau, \underline{t}) / \Gamma(0, \underline{t}). \tag{3}$$

2.2 The Cumulative Distribution of Queue Delay

Again assume we have a FCFS queue. Suppose we set $\underline{s} = \underline{s}^j$, defined so that

$$s_i^j = 0 \qquad i = 1, 2, \ldots, j-1$$
$$s_i^j = \text{Max}\{t_j - \tau, 0\} \qquad i = j, j+1, \ldots, N.$$

Then if we define

$\beta_j(\tau \mid \underline{t}) \equiv \Pr\{$ j^{th} customer to arrive during the congestion period waited less than τ time units $\mid$ observed departure time data$\}$,

we can write

$$\beta_j(\tau \mid \underline{t}) = \Gamma(\underline{s}^j, \underline{t}) / \Gamma(0, \underline{t}). \tag{4}$$

This result allows us to determine for any congestion period the probability that a *random* customer waited more than τ time units, given the observed departure data. We simply compute Eq. (4) once for each value of j and average the results. Or, if a less accurate computation is permitted, just select the customer j at random from the N available and apply Eq. (4) to the selected customer. By applying Eq. (4) for varying values of τ, we can determine the c.d.f. of queue delay, conditioned on the observed departure time data.

2.3 Maximum Queue Length

Finally, without any assumption regarding queue discipline, suppose we define $\underline{s} = \underline{s}^{*M}$ such that

$$s_i^{*M} = t_{(i-M)} \qquad \text{for all } i = 1, 2, \ldots, N,$$

where a negative subscript implies a value of zero. These values for $\underline{s}$ imply that each arriving customer i is to arrive no earlier than the departure time of departing customer i – M during the congestion period. Now we can compute the conditional probability that the queue length did not exceed M during the congestion period:

$$P(Q \le M \mid \underline{t}) = \Pr\{\text{queue length did not exceed M during the congestion period} \mid \text{observed departure time data}\}$$

$$= \Gamma(\underline{s}^{*M}, \underline{t})/\Gamma(0, \underline{t}). \qquad (5)$$

2.4 Probability Distribution of Queue Length

Following the same arguments as in [4], we can utilize the $O(N^3)$ computational algorithm to determine for any queue discipline the probability distribution of queue length at departure epochs, and by a balance of flow argument, this distribution is also the queue length distribution experienced by arriving customers.

2.5 Behavior of Priority Queues

As a final example, consider a multiserver queue with L priority classes of customers. At any given time during a congestion period there exists up to L distinct queues, indexed $l = 1, \ldots, L$. A customer from class i is said to be higher priority than one from class j if $i < j$. Upon completion of service of a customer, the newly available server will select a customer from the highest priority nonempty queue. There must be at least one, otherwise the congestion period would be over. We define

l_i = priority of the customer whose service commences at time t_i.

We assume l_i is known for each customer, thus $\underline{l} = (l_i)$ is an additional vector in the transactional data set.

Analysis of the priority sequences in $\underline{l}$ uncovers *subcongestion periods* within the universal congestion period. As an example, suppose for a three priority system we have a single congestion period commencing at time t = 0, with $\underline{l} = (3, 2, 1, 1, 1, 2, 3)$, $t=(t_1, t_2, t_3, t_4, t_5, t_6, t_7)$. Here, for example, the first priority 3 customer commences service at time t = 0, the first priority 1 customer to enter service does so at time $t = t_2$, and the entire congestion period terminates at time $t = t_7$. The time interval $[t_1, t_4]$ is a continuous period of time during which the three consecutive customers to

enter service are priority 1; this period of time is a subcongestion period for priority 1 customers and a component of longer subcongestion periods for priority 2 and 3 customers. For instance, the subcongestion period for the single queued priority three customer is $[0, t_6]$. Each such subcongestion period may be analyzed separately using the ideas above to compute queue performance by priority class. If queue discipline is first-come, first-served within each priority class, then one can use Eq.(4) to determine the probability distribution of queue delay for each priority class.

2.6 Illustrative Computational Results

Perhaps the most important queue inference application of the algorithm is use of Eq.(4) in computing points on the c.d.f. of the in-queue waiting time for a random customer, assuming a FCFS queue. We have done this for several different queues for which limiting or equilibrium results are known.

One set of Monte Carlo simulation runs modeled the well known M/M/1 (Poisson customer arrivals, i.i.d. negative exponential service times, single server) queue under alternative load factors (ratio of customer arrival rate to available customer service rate). As one illustrative example an M/M/1 queue was simulated with an average of 10 customers arriving per hour, available service rate of 20 customers per hour (i.e., mean service time of 1/20 hour or 3 minutes) for a total of 1000+ simulated hours. The average load factor was 0.5. The transactional data of each of the 4961 observed congestion periods were analyzed with Eq.(4) to estimate points on the in-queue waiting time c.d.f. If W_q is the random variable of interest, then Eq.(4) yields the following c.d.f. estimates: $P\{W_q = 0\} = 0.4922$, $P\{W_q \leq 1 \text{ min.}\} = 0.5913$, $P\{W_q \leq 2 \text{ min.}\} = 0.6476$, $P\{W_q \leq 3 \text{ min.}\} = 0.6972$. From the theory of M/M/1 queues, the analytically obtained limiting results are $P\{W_q = 0\} = 0.5000$, $P\{W_q \leq 1 \text{ min.}\} = 0.5768$, $P\{W_q \leq 2 \text{ min.}\} = 0.6417$, $P\{W_q \leq 3 \text{ min.}\} = 0.6967$.

Acknowledgements

The work of the second author was supported by the National Science Foundation. Both authors thank Sue Hall for programming the algorithm and obtaining the computational results.

References

1. Barlow, R. E., D. J. Bartholomew, J. M. Bremmer and H. D. Brunk, *Statistical Inference Under Order Restriction*, John Wiley and Sons, New York, 1972.

2. Bertsimas, D. J. and L.D. Servi, "Deducing Queueing from Transactional Data: The Queue Inference Engine Revisited," Technical Report OR 212-90, Operations Research Center, Massachusetts Institute of Technology, Cambridge, Mass. 1991.

3. Daley, D. J. and L. D. Servi, "Exploiting Markov Chains to Infer Queue-Length From Transactional Data," submitted to *Journal of Applied Probability*, 1991.

4. David, H. A., *Order Statistics*, John Wiley and Sons, New York, 1981.

5. Hall, S. A. and R. C. Larson, "The Queue Inference Engine (QIE) with Partial Queue Length Information," paper presented at TIMS/ORSA Nashville National Meeting, May 15, 1991.

6 Larson, Richard C., "The Queue Inference Engine: Deducing Queue Statistics from Transactional Data," *Management Science*, Vol. 36, No. 5, 1990, pp.586-601.

7. Larson, Richard C. "The Queue Inference Engine: Addendum," to appear in *Management Science*, 1991.