MANAGEMENT SCIENCE
Vol. 36, No. 5, May 1990
Printed in U.S.A.

THE QUEUE INFERENCE ENGINE: DEDUCING QUEUE STATISTICS FROM TRANSACTIONAL DATA*

RICHARD C. LARSON

Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

The transactional data of a queueing system are the recorded times of service commencement and service completion for each customer served. With increasing use of computers to aid or even perform service one often has machine readable transactional data, but virtually no information about the queue itself. In this paper we propose a way to deduce the queueing behavior of Poisson arrival queueing systems from only the transactional data and the Poisson assumption. For each congestion period in which queues may form (in front of a single or multiple servers), the key quantities obtained are mean wait in queue, time-dependent mean number in queue, and probability distribution of the number in queue observed by a randomly arriving customer. The methodology builds on arguments of order statistics and usually requires a computer to evaluate a recursive function. The results are exact for a homogeneous Poisson arrival process (with unknown parameter) and approximately correct for a slowly time varying Poisson process.
(QUEUES; INFERENCE; DATA ANALYSIS; POISSON)

1. Introduction

Consider a Poisson arrival queueing system for which we have transactional data. That is, we know the time of service commencement and time of service completion for each customer who has been served by the system. Whenever there is a queue of customers waiting for service, we assume that following a customer's departure from service the next customer to enter service from the queue does so virtually immediately following said departure. Thus, the "signature" of a queue existing is a service completion time followed virtually immediately by a service initiation time. The transactional data, when rank ordered, provide such signatures and allow us to identify "congestion periods" during which arriving customers must wait in queue prior to service.

Our objective is to derive the queue statistics, including mean time spent waiting in queue, and the time-dependent mean number in queue from the transactional data. In other words, *we wish to deduce queue behavior without observing the queue* but by drawing inferences from the transactional data and from the Poisson arrival assumption. There are many potential applications, including analysis of customers queueing at automatic teller machines (ATM's), automobile traffic delayed at signalized intersections, and individuals queued awaiting access to a limited number of communications channels.

Our approach focuses on a single congestion period. Since the completion (or commencement) of a congestion period constitutes a renewal point in any Poisson arrival queue, once we have obtained the results for one congestion period we have in essence solved the entire problem. As will become clear, our approach exploits arguments drawn from the field of "order statistics" (cf. Barlow et al. 1972 and David 1981). We will find that we do not need to know the arrival rate parameter of the Poisson process. In all of

* Accepted by Linda Green; received August 1987. This paper has been with the author 3 months for 3 revisions.

0025-1909/90/3605/0586$01.25

our work, *the server or servers can be completely general*; for instance, successive service times need not be i.i.d. We call our derived results the "queue inference engine" or QIE.

2. Examples

EXAMPLE 1. *Automatic Teller Machines*. Consider a facility housing k automatic teller machines (ATM's) fed by a single queue. The system is said to be operating within a congestion period whenever all k ATM's are simultaneously busy, requiring any new arrivals to wait in queue. A congestion period commences (terminates) whenever the number of busy ATM's jumps from $k-1$ to k (k to $k-1$, respectively). A customer service time is the time (s)he "occupies" the space directly in front of the ATM. For many ATM systems this time is closely approximated by the magnitude of the difference in times between the customer's ATM card insertion and the machine's card ejection. These transaction times may be routinely recorded in a master data file. When the data for all k ATM's are merged and time-ordered, they constitute (to close approximation) the customer transaction times required to determine queue statistics developed herein. The queue statistics derived from QIE may be used by bank managers to monitor the use of ATM sites, providing an accurate means to identify those sites requiring additional (fewer) ATM's.

EXAMPLE 2. "*Invisible*" *Queues in Communications Systems*. Many finite capacity communications systems have during periods of congestion invisible queues of customers outside the system, continuously trying to gain access to it.

One example is a k-channel land mobile radio system. Whenever all k channels are simultaneously in use, potential users in the field (in vehicles) having a message to transmit continuously monitor channel use and attempt to acquire a channel as soon as any one of the current k communications is completed. If at a given time t there are $n(t)$ such potential users awaiting a channel, they constitute a spatially dispersed invisible queue, a queue in which one of the waiting customers enters service very shortly after another customer completes service. Of course, this queue can grow in size due to (Poisson) arrivals of new users desiring channel access. Service discipline is not necessarily first-come, first-served. The user entering service next is the one who successfully "locks in" the channel very shortly after termination of a previous message. We assume the radio technology is designed to avoid deadlock or paralysis due to simultaneous channel demands by multiple queued users. Within the context of this paper the customer transaction times are the moments of gaining channel access (service initiation) and message termination (service completion). These times can be routinely monitored and recorded by electronic devices measuring energy in the various broadcast channels, and thus QIE can be used to deduce queueing behavior.

Another communication system example is a telephone system having system capacity j, capacity measured by the maximum number of customers allowed in service and in queue. This system is "congested" whenever j customers are in the system and subsequent potential customers ("callers") are lost (they get a "busy" signal). If all such lost customers continuously and repeatedly call back until they successfully enter the system, then the real time population $m(t)$ of such lost customers constitutes an invisible queue. Within the context of this paper, initiation of "service" occurs the moment a caller successfully enters the system and "termination" of service occurs the moment the telephone conversation is completed; hence the "service time" of this paper represents the sum of queueing delay and telephone conversation time in the telephone system.

EXAMPLE 3. *Traffic Queued at Intersections*. Imagine a street intersection in which one of the streets entering the intersection is equipped with a pressure-sensitive cable placed across the street. Whenever a vehicle passes over the cable, its presence is detected and recorded. Suppose that vehicles traveling along that street toward the intersection

arrive in the vicinity of the intersection according to a Poisson process. As the vehicles stop at the intersection, perhaps due to a stop sign or a traffic light, a queue may form. This queue is depleted as vehicles pass over the cable and enter the intersection.

Within the context of this paper, the service initiation time for each vehicle is the time that the vehicle's front axle passes over the cable. The service completion time is the time the rear axle passes over the cable plus some reasonable constant perhaps dependent on vehicular speed (the calibration details requiring additional research) to allow for space between vehicles. A congestion period exists whenever the cable is registering vehicular movement and, if the intersection is signalized, whenever the light is "red" for vehicles attempting to pass over the cable and enter the intersection. Note that with a signalized intersection (1) successive moveups in vehicular queue position are not i.i.d., and (2) congestion periods can be caused by exogenous events (a "red light") as well as by simple queueing congestion.

QIE allows a traffic engineer to deduce the queueing behavior of vehicles at the intersection simply from the cable-recorded information, without ever observing the queue.

EXAMPLE 4. *Queueing Networks*. A not so obvious application is in communication networks. At any given node of a communications network one has in general a complex queueing system in which arrivals are not Poisson (and not even regenerative) and the service process is complicated, typically not following i.i.d. or other "nice" assumptions. However, the cause of analytical tractability would be served if the (complex) arrival process could be approximated to be Poisson. Using transactional data (from the real system), one could estimate queue behavior at the node using the methods herein and compare to observed queue behavior; if the two are "similar," then the Poisson arrival assumption is probably a reasonable approximation for modeling purposes.

3. Preliminaries

Suppose we consider a homogeneous Poisson process with rate parameter $\lambda > 0$. Over a fixed time interval $[0, T]$ we are told that precisely N Poisson events occur. The N *ordered* arrival times are $0 \leqq X_1 \leqq X_2 \leqq \cdots \leqq X_N \leqq T$ (by implication $X_{N+1} > T$). The N unordered arrival times are $U_1, U_2, \ldots, U_N$, $0 \leqq U_i \leqq T$ $(i = 1, 2, \ldots, N)$. From the theory of order statistics, it is well known that the U_i's are independent, uniformly distributed over $[0, T]$. If we now let $N(t)$ be the number of arrivals over $[0, t]$, $0 \leqq t \leqq T$, without further conditioning information the following are well known for $N(t)$:

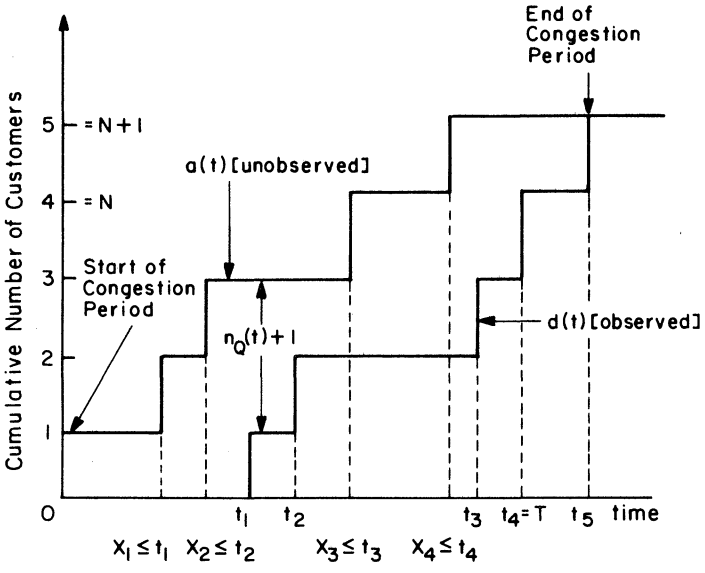
$$E[N(t)] = (t/T)N, \tag{1a}$$

$$\text{VAR}\,[N(t)] \equiv \sigma^2_{N(t)} = N\left(\frac{t}{T}\right)\left(\frac{T-t}{T}\right), \tag{1b}$$

$$\Pr\{N(t) = k\} = \binom{N}{k}\left(\frac{t}{T}\right)^k\left(\frac{T-t}{T}\right)^{N-k}. \tag{1c}$$

In a queueing environment, $N(t)$ could represent the number of customers in queue at time t, assuming *bulk service* of all N waiting customers at time T, such as occurs at signalized pedestrian crosswalks.

In more general queueing environments, customers usually leave one-at-a-time. Their service completion times within a congestion period impose a set of inequalities on the arrival times of other customers who waited in queue. It is this set of inequalities that produces precise conditioning information within the general context of order statistics, conditioning information that we use to deduce queue behavior.



$a(t)$ = cumulative number of arrivals from commencement of congestion period through time t,
$d(t)$ = cumulative number of departures from commencement of congestion period through time t,
$n_Q(t)$ = number of customers in queue at time t,
t_i = departure time of ith customer served,
X_i = arrival time of ith customer to enter queue during the congestion period ($i = 1, 2, 3, 4$).

FIGURE 1. Illustrative Sample Function for a Three-Server Queue.

To illustrate key ideas and introduce notation, consider the sample function for a three-server queue shown in Figure 1. In the example the congestion period commences at $t = 0$ upon arrival of a customer who changes the remaining idle server's status from idle to busy. From transactional data the queue exhibits both service departures and service commencements at times t_1, t_2, t_3 and t_4, indicating that (1) all three servers were continuously busy during this time; (2) a queue existed at least at times t_1-, t_2-, t_3-, and t_4-; and (3) that the total number of customers delayed in queue during the congestion period was $N = 4$. At time t_5 the transactional data indicate a service completion but no service commencement, thus ending the congestion period and thereby creating an idle server. From the transactional data, the cumulative number of departures through time t, $d(t)$, is an observed function whereas the cumulative number of arrivals $a(t)$ is not. From the conditioning information we know that the first arrival during the congestion period occurred prior to the first departure, i.e., $X_1 \leqq t_1$, and that subsequent arrivals obey the departure time inequalities $X_2 \leqq t_2$, $X_3 \leqq t_3$, $X_4 \leqq t_4 = T$. (Note that the end point of the conditional arrival interval for queued customers is $T = t_4$, not t_5.) During the congestion period the number of customers in queue is $n_Q(t) = a(t) - d(t) - 1$. (For values of t equal to service completion times, i.e., $t = t_j$, one must be careful whether one is considering t_j+ or t_j-, as the former subtracts from the queue the customer who enters service at time t, whereas the latter does not.) The total number of customers in the system (in service and in queue) at time t is $n(t) = n_Q(t) + 3$.

The same concepts apply in more general queueing systems, including those with state-dependent service rates, shortest-job-first queue discipline, etc. The key idea is to locate those service completion times which are accompanied by (nearly) simultaneous service commencement times.

4. Main Results

In this section we show how to deduce from transactional data (1) mean number of customers in queue t time units after commencement of a congestion period; (2) time average queue length; (3) mean delay in queue; and (4) incidence probabilities. All of the results follow simply once we can determine, using order statistics, the *a priori* probability that the arrivals during a congestion period obey the time orderings imposed by the observed departure times.

1. *Computing the Fundamental A Priori Conditional Probability*

For a given congestion period commencing at time $t = 0$ and terminating at time $t = t_{N+1} > t_N$, we wish to find the a priori probability of the event that our transactional data indicate has occurred. Define

$\mathbf{t} \equiv (t_1, t_2, \ldots, t_N)$,

$\mathbf{t} - y \equiv (t_1 - y, t_2 - y, \ldots, t_N - y)$,

$O(\mathbf{t}) \equiv \text{Event}\{X_1 \leqq t_1, X_2 \leqq t_2, \ldots, X_N \leqq t_N\}$,

$N(t) \equiv$ cumulative number of arrivals to the system over the interval $(0, t]$, $t \leq t_N$; i.e., the cumulative number of customers delayed in queue during $(0, t]$.

The conditional probability we wish to compute is $P\{O(\mathbf{t})|N(t_N) = N\}$.

We derive two recursive procedures for computing $P\{O(\mathbf{t})|N(t_N) = N\}$, the first putting tagged customers in a "left-hand" interval $[0, t_1]$ and the second putting them in a "right-hand" interval $[t_{i-1}, t_i]$. Define two conditional probabilities necessary for the recursions:

$$\Psi_k(\mathbf{t}, T) \equiv \Pr\{X_1 \leqq t_1, X_2 \leqq t_2, \ldots, X_k \leqq t_k | N(T) = k\}$$

where $k \leqq N$, $T \geqq t_k$ and $\Psi_0 \equiv 1$, and

$$\alpha_{ki}(\mathbf{t}) = \Pr\{X_1 \leqq t_1, X_2 \leqq t_2, \ldots, X_i \leqq t_i, \ldots, X_k \leqq t_i | N(t_N) = k\},$$

where $k \geqq i$. $\Psi_k(\mathbf{t}, T)$ is the a priori conditional probability that the arrival times in $[0, T]$ satisfy the first k departure time inequalities, given exactly k arrivals in $[0, T]$. $\alpha_{ki}(\mathbf{t})$ is the a priori conditional probability that all arrivals in $[0, t_N]$ occur before the ith departure and obey the first $(i - 1)$ other departure time inequalities, given exactly k arrivals in $[0, t_N]$. The desired fundamental a priori conditional probability is

$$P\{O(\mathbf{t})|N(t_N) = N\} = \Psi_N(\mathbf{t}, t_N) = \alpha_{NN}(\mathbf{t}). \tag{2}$$

First consider $\Psi_k(\mathbf{t}, T)$. For $N = 1$ we have

$$\begin{aligned}\Psi_1(\mathbf{t}, T) &= \Pr\{X_1 \leqq t_1 | \text{ precisely one Poisson arrival in } [0, T]\} \\ &= t_1/T, \qquad T \geqq t_1.\end{aligned} \tag{3}$$

We now find that $\Psi_k(\cdot)$ can be computed from $\Psi_0(\cdot), \Psi_1(\cdot), \ldots, \Psi_{k-1}(\cdot)$ by the recursion in

LEMMA 1.

$$\Psi_k(\mathbf{t}, T) = \sum_{j=1}^{k} \binom{k}{k-j+1} \left(\frac{t_1}{T}\right)^{k-j+1} \left(\frac{T - t_1}{T}\right)^{j-1} \Psi_{j-1}(\mathbf{t} - t_1, T - t_1). \tag{4}$$

PROOF. (Induction) Equation (3) demonstrates that (4) holds for $k = 1$. Suppose (4) holds for k; we prove it holds for $k + 1$.

The argument proceeds as follows:

$$\begin{aligned}
\Psi_{k+1}(\mathbf{t}, T) &= \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_k \le t_k, X_{k+1} \le t_{k+1} \,|\, N(T) = k+1\} \\
&= \Pr\{X_{k+1} \le t_1\} + \Pr\{X_k \le t_1 \text{ and } t_1 < X_{k+1} \le t_{k+1}\} \\
&\quad + \Pr\{X_{k-1} \le t_1 \text{ and } t_1 < X_k \le t_k,\ t_1 < X_{k+1} \le t_{k+1}\} + \cdots \\
&\quad + \Pr\{X_1 \le t_1, \text{ and } t_1 < X_2 \le t_2, \ldots, t_1 < X_k \le t_k,\ t_1 < X_{k+1} \le t_{k+1}\} \\
&= \left(\frac{t_1}{T}\right)^{k+1} + (k+1)\left(\frac{t_1}{T}\right)^{k}\left(\frac{T-t_1}{T}\right)\Psi_1(\mathbf{t}-t_1, T-t_1) \\
&\quad + \binom{k+1}{2}\left(\frac{t_1}{T}\right)^{k-1}\left(\frac{T-t_1}{T}\right)^{2}\Psi_2(\mathbf{t}-t_1, T-t_1) + \cdots \\
&\quad + \binom{k+1}{k}\left(\frac{t_1}{T}\right)\left(\frac{T-t_1}{T}\right)^{k}\Psi_k(\mathbf{t}-t_1, T-t_1). \qquad \blacksquare
\end{aligned}$$

Now consider $\alpha_{ki}(\mathbf{t})$. To calculate $\alpha_{ki}(\mathbf{t})$ first note that

$$\alpha_{k1}(\mathbf{t}) = (t_1/t_N)^k, \qquad k = 1, 2, \ldots, N. \tag{5}$$

The fundamental recursion is given by

LEMMA 2.

$$\alpha_{ki}(\mathbf{t}) = \sum_{j=0}^{k-i+1}\binom{k}{j}\alpha_{(k-j)(i-1)}(\mathbf{t})\left(\frac{t_i - t_{i-1}}{t_N}\right)^j, \qquad k \ge i. \tag{6}$$

PROOF. (Induction) Equation (5) demonstrates that (6) holds for $i = 1$. Suppose (6) holds for i; we prove it holds for $i+1$ ($i \le k$).

$$\begin{aligned}
\alpha_{k(i+1)}(\mathbf{t}) &= \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_{i+1} \le t_{i+1}, \ldots, X_k \le t_{i+1} \,|\, N(t_N) = k\} \\
&= \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_{i+1} \le t_i, \ldots, X_k \le t_i | N(t_N) = k\} \\
&\quad + \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_{i+1} \le t_i, \ldots, t_i < X_k \le t_{i+1} \,|\, N(t_N) = k\} \\
&\quad + \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_{i+1} \le t_i, \ldots, t_i < X_{k-1} \le t_{i+1}, \\
&\qquad t_i < X_k \le t_{i+1} \,|\, N(t_N) = k\} \\
&\quad + \cdots + \Pr\{X_1 \le t_1, X_2 \le t_2, \ldots, X_i \le t_i, t_i < X_{i+1} \le t_{i+1}, \ldots, \\
&\qquad t_i < X_k \le t_{i+1} \,|\, N(t_N) = k\} \\
&= \alpha_{ki}(\mathbf{t}) + \binom{k}{1}\alpha_{(k-1)i}(\mathbf{t})\left(\frac{t_{i+1}-t_i}{t_N}\right) + \binom{k}{2}\alpha_{(k-2)i}(\mathbf{t})\left(\frac{t_{i+1}-t_i}{t_N}\right)^2 + \cdots \\
&\quad + \binom{k}{k-i}\alpha_{ii}(\mathbf{t})\left(\frac{t_{i+1}-t_i}{t_N}\right)^{k-i}. \qquad \blacksquare
\end{aligned}$$

To compute equation (6) iteratively one is filling out a lower triangular matrix $\mathbf{A}(\mathbf{t}) \equiv (\alpha_{ki}(\mathbf{t}))$, including terms on the diagonal. One first uses equation (5) to compute all N entries of the first column of $\mathbf{A}(\mathbf{t})$. Then to compute the kth entry ($k \ge 2$) in the second column, one adds k terms, the jth involving a multiplication with entry $(k-j+1)$ in the first column. In this way, one sweeps through the matrix column by

column, starting in column one. The number of separate terms that have to be computed to complete the matrix is equal to

$$\sum_{i=1}^{N} \frac{i(i+1)}{2} = \frac{1}{6}N^3 + \frac{1}{2}N^2 + \frac{1}{3}N,$$

yielding an $O(N^3)$ procedure for evaluating $P\{O(\mathbf{t})|N(t_N) = N\}$.

Regarding the computational work associated with equation (4), one can show that 2^N separate terms must be computed. The two recursions are comparable in computational effort for $N \leqq 5$, but the $O(N^3)$ procedure of Lemma 2 is clearly superior for larger N.

2. *Computing Arrival Time Cumulative Probabilities*

We now concentrate on the event $X_k \leqq t_i$, that is, the event that the kth arrival precedes the ith departure. For this reason define

$$\beta_{ki}(\mathbf{t}) \equiv \Pr\{X_k \leq t_i | O(\mathbf{t}), N(t_N) = N\}.$$

Since the kth arrival must precede the kth departure, one clearly has $\beta_{ki}(\mathbf{t}) = 1$ for all $k = 1, 2, \ldots, i$.

There are two alternative methods for computing the matrix $\beta(\mathbf{t}) \equiv (\beta_{ki}(\mathbf{t}))$, depending on whether one uses Lemma 1 or Lemma 2 for the fundamental recursions. In the context of Lemma 1, for $k > i$, we compute the arrival time cumulative probabilities as follows:

$$\begin{aligned}\beta_{ki}(\mathbf{t}) &= \Pr\{X_k \leq t_i | O(t), N(t_N) = N\} \\ &= \frac{\Pr\{X_1 \leq t_1, \ldots, X_i \leq t_i, X_{i+1} \leq t_i, \ldots, X_k \leq t_i, X_{k+1} \leq t_{k+1}, \ldots, | N(t_N) = N\}}{P\{O(\mathbf{t})|N(t_N) = N\}}\end{aligned} \tag{7}$$

or

$$\beta_{ki}(\mathbf{t}) = \frac{\Psi_N((t_1, t_2, \ldots, t_i, t_i, \ldots, t_i, t_{k+1}, \ldots, t_N), t_N)}{P\{O(\mathbf{t})|N(t_N) = N\}}. \tag{8}$$

With Lemma 2 the notation for determining $\beta(\mathbf{t}) = (\beta_{ki}(\mathbf{t}))$ is somewhat more complex, but the computational effort for large N is considerably less. First, it should be clear that the bottom row of $\beta(\mathbf{t})$ is obtained by a simple division,

$$\beta_{Ni}(\mathbf{t}) = \frac{\alpha_{Ni}(\mathbf{t})}{\alpha_{NN}(\mathbf{t})} = \frac{\alpha_{Ni}(\mathbf{t})}{P\{O(\mathbf{t})|N(t_N) = N\}}. \tag{9}$$

For the general term, rewrite equation (7) as

$$\begin{aligned}\beta_{ki}(\mathbf{t}) = \frac{1}{\alpha_{NN}(\mathbf{t})} \{ & P\{X_1 \leq t_1, \ldots, X_i \leq t_i, \ldots, X_k \leq t_i, X_{k+1} \leq t_i, X_{k+2} \leq t_{k+2}, \ldots, \\ & X_N \leq t_N | N(t_N) = N\} \\ & + P\{X_1 \leq t_1, \ldots, X_i \leq t_i, \ldots, X_k \leq t_i, t_i < X_{k+1} \leq t_{k+1}, \\ & t_i < X_{k+2} \leq t_{k+2}, \ldots, t_i < X_N \leq t_N | N(t_N) = N\}\}.\end{aligned}$$

The first probability in the brackets, when divided by $\alpha_{NN}(\mathbf{t})$, is seen to be $\beta_{(k+1)i}(\mathbf{t})$, thereby giving rise to a recursion. To compute the second term, consider the $\binom{N}{k}$ ways of selecting k of the N unordered arrival times to obey inequalities in the "left-hand" interval $[0, t_i]$ and the remaining $(N-k)$ to obey inequalities in the "right-hand" interval $[t_i, t_N]$. Those selected for the left would have to obey the first k inequalities in the second

probability term above, while those selected for the right would have to obey the final $N - k$ inequalities. Invoking independence of the unordered arrival times, we can now write the recursion

$$\beta_{ki}(\mathbf{t}) = \beta_{(k+1)i}(\mathbf{t}) + \binom{N}{k}\alpha_{ki}(\mathbf{t})\eta_{ki}(\mathbf{t})/\alpha_{NN}(\mathbf{t}) \qquad \text{where} \tag{10}$$

$$\eta_{ki}(\mathbf{t}) \equiv P\{t_i < X'_1 \leq t_{k+1}, \ldots, t_i < X'_{N-k} \leq t_N | N - k \text{ arrivals in } [0, t_N]\} \tag{11}$$

and X'_j is the jth smallest arrival time of the $N - k$ selected for the right interval ($0 \leqq X'_j \leqq t_N$).

If we define the time-shifted vector $\mathbf{t}' \equiv (t'_j)$,

$$t'_j = \begin{cases} t_{k+j} - t_i & \text{for} \quad j = 1, 2, \ldots, N - k, \\ t_N & \qquad\quad j = N - k + 1, \ldots, N, \end{cases}$$

and invoke uniformity of the probability measure over the joint sample space of the $N - k$ unordered arrival times, we find an event having identical probability to that indicated in equation (11), i.e.,

$$\eta_{ki}(\mathbf{t}) \equiv P\{0 < X'_1 \leq t_{k+1} - t_i, 0 \leq X'_2 \leq t_{k+2} - t_i, \ldots,$$
$$0 \leq X'_{N-k} \leq t_N - t_i | N - k \text{ arrivals in } [0, t_N]\}.$$

Thus, (11) can be computed using the algorithm already developed for computing $\alpha_{ki}(\ )$,

$$\eta_{ki}(\mathbf{t}) = \alpha_{(N-k)(N-k)}(\mathbf{t}'). \tag{12}$$

Computation of $\beta_{ki}(\mathbf{t})$ using (12) requires $O((N - k)^3)$ new computations (i.e., using the algorithm of Lemma 2). Summing all the $O([N - k]^3)$ terms below the diagonal of $\beta(\mathbf{t})$, one finds that the number of separate terms required to evaluate the entire matrix $\beta(\mathbf{t})$ is $O(N^5)$. In practice, for large problems terms far from the diagonal are often small enough to be approximated as zero, so the computational work tends to grow more slowly than $O(N^5)$.

3. *The Mean Cumulative Number of Arrivals at Time t*

We now wish to compute

$\bar{N}_a(t) \equiv$ the expected cumulative number of arrivals to the system up to and including time t, given $O(\mathbf{t})$ and $N(t_N) = N$.

This is the quantity analogous to $E[N(t)]$ displayed in equation (1) (a) for unconditioned order statistics. To avoid counting ambiguities we assume in Lemma 3 a strict ordering of the t_i's: $0 < t_1 < t_2 < \cdots < t_N$. The generalization to nonstrict inequalities is straightforward and will not be stated here.

LEMMA 3.

(i)

$$\bar{N}_a(t_j) = \sum_{k=1}^{N} \beta_{kj}(\mathbf{t}) \qquad \textit{for all} \qquad j = 1, 2, \ldots, N. \tag{13}$$

(ii) *Define* $t_0 \equiv 0$. *For* $t_{j-1} < t \leq t_j$, $j = 1, 2, \ldots, N$,

$$\bar{N}_a(t) = \frac{t_j - t}{t_j - t_{j-1}}\bar{N}_a(t_{j-1}) + \frac{t - t_{j-1}}{t_j - t_{j-1}}\bar{N}_a(t_j). \tag{14}$$

REMARK. (i) states that the expected cumulative number of arrivals up to and including time t_j is equal to a simple sum of arrival time cumulative probabilities. (ii) states

that $\bar{N}_a(t)$ grows linearly during any time interval between two successive departure times t_{j-1} and t_j.

PROOF. (i) Write

$$N_a(t_j) = \sum_{k=1}^{N} Y_k(t_j), \qquad \text{where}$$

$$Y_k(t_j) = \begin{cases} 1 & \text{if at least } k \text{ arrivals in } (0, t_j], \text{ given } O(\mathbf{t}) \text{ and } N(t_N) = N, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\bar{N}_a(t_j) = \sum_{k=1}^{N} \bar{Y}_k(t_j) = \sum_{k=1}^{N} \beta_{kj}(\mathbf{t}). \qquad \blacksquare$$

(ii) Suppose $N_a(t_{j-1}) = l$ and $N_a(t_j) = l + m$, $m \geqq 0$. Then over $(t_{j-1}, t_j]$ we have m random variables that are conditionally independent, uniformly distributed, the m "unordered arrival times" over $(t_{j-1}, t_j]$, where the expected value of the cumulative number of arrivals through time t, $t_{j-1} < t \leqq t_j$, grows linearly with t (with zero growth, of course, for the case $m = 0$). Thus,

$$\bar{N}_a(t | N_a(t_{j-1}) = l \text{ and } N_a(t_j) = l + m) = l + \frac{m}{t_j - t_{j-1}}(t - t_{j-1}).$$

Unconditioning first on $N_a(t_{j-1})$,

$$\bar{N}_a(t | N_a(t_j) - N_a(t_{j-1}) = m) = \bar{N}_a(t_{j-1}) + \frac{m}{t_j - t_{j-1}}(t - t_{j-1}).$$

Then unconditioning on $N_a(t_j) - N_a(t_{j-1})$,

$$\bar{N}_a(t) = \bar{N}_a(t_{j-1}) + \frac{[\bar{N}_a(t_j) - \bar{N}_a(t_{j-1})](t - t_{j-1})}{t_j - t_{j-1}}$$

which simplifies to equation (14).

As a final interesting property regarding $\bar{N}_a(t)$, we have

LEMMA 4. *For $t \geqq 0$, $\bar{N}_a(t)$ is a concave function of t.*

PROOF. See Appendix I.

REMARK. Lemma 4 is useful in developing bounds and approximations for large N.

Note that none of the results of this section depend on the value of the Poisson rate parameter λ, nor do they depend on the number or type of servers. They only require the assumptions of (1) simple homogeneous Poisson arrivals and (2) queue "signatures" from the transactional data, i.e., a service initiation following virtually immediately after a service completion whenever there is a queue.

4. *Numerical Example*

To illustrate the mechanics, we solve using both Lemmas 1 and 2 a simple $N = 3$ example with $t_1 = \frac{1}{3}$, $t_2 = \frac{2}{3}$ and $t_3 = T = 1$. These data correspond to a queueing system for which (1) a congestion period commences at time $t = 0$; (2) departures followed immediately by service initiations occur at t_1, t_2, and t_3; and (3) the departure occurring sometime later at time t_4 is not followed immediately by a service initiation, thereby signaling the end of the congestion period. Hence, a queue existed at least at times t_1-, t_2-, and t_3-. Here $\mathbf{t} = (\frac{1}{3}, \frac{2}{3}, 1)$.

Using Lemma 1, we compute from (3) and (4)

$$P\{O(\mathbf{t})|N(1)=3\} = \Psi_3(\mathbf{t}, 1)$$

$$= \left(\frac{1}{3}\right)^3 + \binom{3}{2}\left(\frac{1}{3}\right)^2 \frac{2}{3}\Psi_1\left(\mathbf{t}-\frac{1}{3}, 1-\frac{1}{3}\right) + \binom{3}{1}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^2 \Psi_2\left(\mathbf{t}-\frac{1}{3}, 1-\frac{1}{3}\right).$$

Clearly

$$\Psi_1\left(\mathbf{t}-\frac{1}{3}, 1-\frac{1}{3}\right) = 1,$$

$$\Psi_2\left(\mathbf{t}-\frac{1}{3}, 1-\frac{1}{3}\right) = \left(\frac{1/3}{2/3}\right)^2 + 2\left(\frac{1/3}{2/3}\right)\frac{1/3}{2/3}\frac{1/3}{2/3}.$$

Combining results, we obtain $\Psi_3(\mathbf{t}, 1) = \frac{16}{27}$. This is the *a priori* probability that the arrival times, given 3 arrivals over [0, 1], obey the departure time inequalities.

Using Lemma 2, we find the matrix

$$\mathbf{A}(\mathbf{t}) = \begin{bmatrix} 1/3 & — & — \\ (1/3)^2 & 1/3 & — \\ (1/3)^3 & 7/27 & 16/27 \end{bmatrix}$$

where $\alpha_{33}(\mathbf{t}) = P\{O(\mathbf{t})|N(1)=3\} = \frac{16}{27}$, as just computed using Lemma 1. To illustrate computation of one of the terms in $\mathbf{A}(\mathbf{t})$,

$$\alpha_{32}(\mathbf{t}) = \alpha_{31}(\mathbf{t}) + \binom{3}{1}\alpha_{21}(\mathbf{t})\cdot\left(\frac{1}{3}\right) + \binom{3}{2}\alpha_{11}(\mathbf{t})\cdot\left(\frac{1}{3}\right)^2 = \frac{7}{27}.$$

We now wish to obtain the matrix of arrival time cumulative probabilities,

$$\beta(\mathbf{t}) = \begin{bmatrix} 1 & 1 & 1 \\ \beta_{21}(\mathbf{t}) & 1 & 1 \\ \beta_{31}(\mathbf{t}) & \beta_{32}(\mathbf{t}) & 1 \end{bmatrix}.$$

We illustrate use of Lemma 1 by computing the most complicated entry,

$$\beta_{32}(\mathbf{t}) = \Pr\{X_3 \le t_2|O(\mathbf{t}), N(1)=3\} = \frac{\Psi_3((1/3, 2/3, 2/3), 2/3)}{16/27}$$

$$= \frac{27}{16}\left[\left(\frac{1}{3}\right)^3 + \binom{3}{2}\left(\frac{1}{3}\right)^2\frac{2}{3}\cdot\frac{1}{2} + \binom{3}{1}\frac{1}{3}\left(\frac{2}{3}\right)^2\Psi_2\left(\left(\frac{1}{3}, \frac{1}{3}\right), \frac{2}{3}\right)\right].$$

But $\Psi_2 = ((\frac{1}{3}, \frac{1}{3}), \frac{2}{3}) = (\frac{1}{2})^2$, thus $\beta_{32}(\mathbf{t}) = \frac{7}{16}$.

However, computation of $\beta(\mathbf{t})$ is much simpler using Lemma 2 with equations (8), (10) and (12). Using (8) we immediately determine the bottom row of $\beta(\mathbf{t})$,

$$\bar{\beta}_{3\cdot}(t) = \left(\frac{(1/3)^3}{16/27}, \frac{7/27}{16/27}, \frac{16/27}{16/27}\right) = (1/16, 7/16, 1).$$

Using (10) and (12) we find that

$$\beta_{21}(\mathbf{t}) = \beta_{31}(\mathbf{t}) + \binom{3}{2}\alpha_{21}(\mathbf{t})\eta_{21}(\mathbf{t})/\alpha_{33}(\mathbf{t}) = \frac{1}{16} + \frac{9}{16}\eta_{21}(\mathbf{t}).$$

But $\eta_{21}(\mathbf{t}) = \alpha_{(3-2)(3-2)}(\mathbf{t}') = \alpha_{11}(\mathbf{t}') = \frac{2}{3}$, where $\mathbf{t}' = (\frac{2}{3}, 1, 1)$, yielding $\beta_{21}(\mathbf{t}) = \frac{7}{16}$. The complete matrix, together with the column sums representing mean cumulative number of arrivals, is given by

$$\beta(\mathbf{t}) = \begin{bmatrix} 1 & 1 & 1 \\ 7/16 & 1 & 1 \\ 1/16 & 7/16 & 1 \end{bmatrix}.$$

$$\bar{N}_a: \quad 1.5 \qquad 39/16 \quad 3$$

Finally, using (13) and (14) the mean queue length as a function of time is displayed in Figure 2.

5. *Expected Queue Length*

Letting $\bar{N}_Q$ represent the time average queue length over a congestion period of length T, we have

$$\bar{N}_Q = \frac{1}{T} E\left[\int_0^T N_Q(t)\,dt\right] = \frac{1}{T}\int_0^T \bar{N}_Q(t)\,dt.$$

Since $\bar{N}_Q(t)$ is piecewise linear, with drops of magnitude one at t_i ($i = 1, 2, \ldots, N$), we can easily evaluate $\bar{N}_Q$ as follows (defining $t_0 \equiv 0$):

$$\bar{N}_Q = \frac{1}{2T}\sum_{i=1}^{N} (t_i - t_{i-1})[\bar{N}_Q(t_i-) + \bar{N}_Q(t_{i-1}+)]. \tag{15}$$

EXAMPLE. Drawing from our continuing $N = 3$ example,

$$\bar{N}_Q = \tfrac{1}{2}(\tfrac{1}{3})[1.5 + (1.4375 + 0.5) + (1.0 + 0.4375)] = 0.8125.$$

Note that $\bar{N}_Q$ is the time average queue length during the congestion period for which the departure instants are known; $\bar{N}_Q$ is not the average queue length observed by a random customer arriving during the congestion period, because the conditioning information removes the Poisson arrival assumption (!).

To find the time average queue length over larger time intervals, including multiple congestion and uncongestion periods, one simply computes appropriate (time) weighted averages.

It is well known that Poisson arrivals see time averages (Wolff 1981). Assuming that the queueing system is ergodic (which would be true for instance if each congestion period is governed by the same probability laws) our computations for $\bar{N}_Q$ and incidence probabilities (see §4.7) when averaged over many congestion periods would approach time averages.

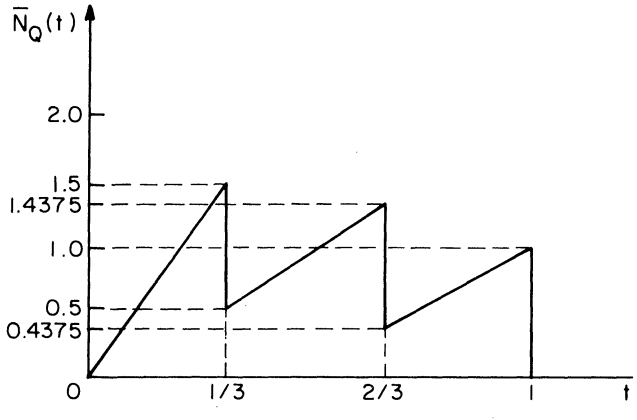


FIGURE 2. Mean Queue Length as a Function of Time for $N = 3$ Numerical Example.

6. *Mean Delay in Queue*

The expected total number of minutes spent in queue by customers during a congestion period is

$$E\left[\int_0^T N_Q(t)dt\right] = \int_0^T \bar{N}_Q(t)dt = T\bar{N}_Q.$$

Since there are N customers arriving during the congestion period, the average amount of time spent in queue per customer is

$$E[W_Q] \equiv \bar{W}_Q = \frac{1}{N}\int_0^T \bar{N}_Q(t)dt = \left(\frac{T}{N}\right)\bar{N}_Q. \tag{16}$$

Since N customers arrive (depart) during the period $(0, T)$, the quantity (N/T) is the average arrival (departure) rate of customers during the congestion period. Equation (16), when rewritten

$$\bar{N}_Q = \left(\frac{N}{T}\right)\bar{W}_Q$$

is equivalent to Little's formula $L_Q = \lambda W_Q$ (Little 1961). In our running numerical example, $\bar{W}_Q = 0.2708$.

7. *Incidence Probabilities*

In this section we wish to compute the probability distribution of the queue length upon arrival of a random customer during a congestion period. Since the congestion period commences and terminates with zero customers in queue, we use the observation that for each queue length transition from i to $i + 1$ during the congestion period there must be a transition from $i + 1$ to i ($i = 0, 1, 2, \ldots$). If we define

$$\Pi_k \equiv \text{Prob}\{\text{a randomly arriving customer finds } k \text{ customers in queue}\},$$
$$k = 0, 1, 2, \ldots,$$

then Π_k can be found by computing the probability that a randomly *departing* customer leaves behind k customers in queue. (This is a familiar argument found in the analysis of many different queues, including $M/G/1$ queues [cf. Kleinrock 1975]).

We can write

$$\begin{aligned}\Pi_k &= \frac{1}{N}\sum_{j=1}^{N} \text{Prob}\{j\text{th departing customer leaves behind } k \text{ in queue}\} \\ &= \frac{1}{N}\sum_{j=1}^{N} \text{Prob}\{\text{exactly } j + k \text{ arrivals in } (0, t_j]\} \\ &= \frac{1}{N}\sum_{j=1}^{N} [\text{Prob}\{\text{at least } j + k \text{ arrivals in } (0, t_j]\} - \text{Prob}\{\text{at least } j + k + 1 \\ &\qquad \text{arrivals in } (0, t_j]\}] \qquad \text{or}\end{aligned}$$

$$\Pi_k = \frac{1}{N}\sum_{j=1}^{N} (\beta_{(j+k)j}(\mathbf{t}) - \beta_{(j+k+1)j}(\mathbf{t})), \tag{17}$$

where $\beta_{j+k,k} = 0$ for $j + k > N$.

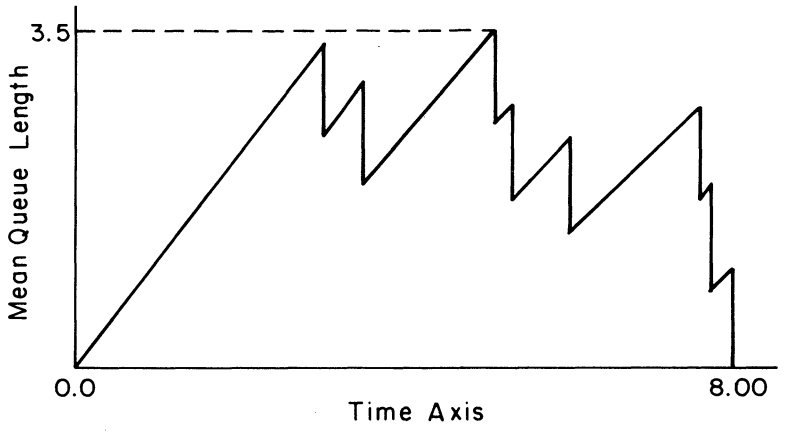


FIGURE 3. $\bar{N}_Q(t)$ for $N = 8$ Example.

For our continuing numerical example, we find the following:

$$\Pi_1 = \tfrac{1}{3}(\tfrac{7}{16} - \tfrac{1}{16} + \tfrac{7}{16}) = \tfrac{13}{48} \approx 0.271,$$

$$\Pi_2 = \tfrac{1}{3}\,\tfrac{1}{16} = \tfrac{1}{48} \approx 0.021,$$

$$\Pi_0 = 1 - (\Pi_1 + \Pi_2) = \tfrac{34}{48} \approx 0.708 \qquad \text{or}$$

$$\boldsymbol{\Pi} = (\tfrac{34}{48}, \tfrac{13}{48}, \tfrac{1}{48}) \approx (0.708, 0.271, 0.021).$$

The average queue length experienced by an arriving customer, call it $\bar{l}_Q$, is

$$\bar{l}_Q = 0 \cdot \tfrac{34}{48} + 1 \cdot \tfrac{13}{48} + 2 \cdot \tfrac{1}{48} = \tfrac{5}{16} \approx 0.3125,$$

in this case considerably less than the time average queue length $\bar{N}_Q = 0.8125$.

We have developed a computer program that carries out all of the computations of this paper, including plotting $\bar{N}_Q(t)$. We show in Figure 3 $\bar{N}_Q(t)$ for a congestion period having $N = 8$ simultaneous departures and service initiations as follows: Congestion

TABLE 1

Detailed Statistics for N = 8 Example

Matrix of the Betas

1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.9485	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.7299	0.8647	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.4220	0.5940	0.9816	1.0000	1.0000	1.0000	1.0000	1.0000
0.1696	0.2923	0.8143	0.8718	1.0000	1.0000	1.0000	1.0000
0.0446	0.0957	0.4868	0.5569	0.7984	1.0000	1.0000	1.0000
0.0070	0.0188	0.1870	0.2321	0.4445	0.9789	1.0000	1.0000
0.0005	0.0017	0.0338	0.0459	0.1227	0.7184	0.8209	1.0000

Cumulative Expected Number of Customers

3.3222	3.8673	5.5035	5.7067	6.3656	7.6972	7.8209	8.0000

Incidence Probabilities

Π_0	Π_1	Π_2	Π_3	Π_4	Π_5	Π_6	Π_7
0.2169	0.3009	0.2877	0.1322	0.0501	0.0111	0.0010	0.0001

- Average Number of Customers in the Queue as seen by a randomly arriving customer = 1.5354.
- Time Average Number of Customers in the Queue = 2.0332.
- Average Waiting Time for Customers in the Queue = 2.0332.

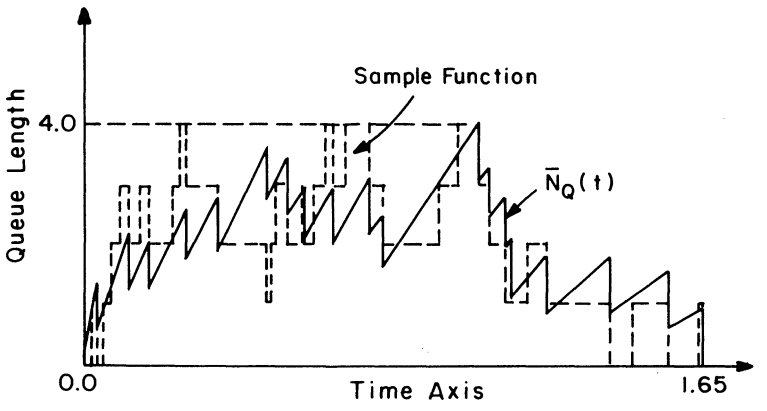


FIGURE 4. Mean Value Function $\bar{N}_Q(t)$ and Sample Function for an $N = 19$ Congestion Period of a Simulated $M/M/1$ Queue.

period starts at $t = 0$. The vector of departure times is $\mathbf{t} = (3.0, 3.5, 5.1, 5.3, 6.0, 7.6, 7.75, 8.0)$. The key statistics for this example are displayed in Table 1.

In Figure 4 we display $\bar{N}_Q(t)$ for an $N = 19$ congestion period of a simulated $M/M/1$ queue, together with the underlying sample function of the simulation. The close "tracking" of the sample function demonstrated in Figure 4 is typical of other empirical results we have obtained.

5. Summary and Conclusions

In this paper we have shown how to apply ideas of order statistics to deduce the behavior of Poisson-arrival queues without observing them. We simply use transactional data (i.e., times of service commencement and service completion) for each customer served together with the Poisson assumption to derive time-dependent mean number in queue, mean wait in queue and the probability distribution of the number of customers in queue upon arrival of a random customer. Using the same ideas, additional performance measures could be devised if desired.

None of our formulas contains the rate parameter λ of the Poisson process. This is because the total number of (Poisson) arrivals over a congestion period is given as part of the conditioning information. Thus our results could be averaged over congestion periods occurring during times of different Poisson rate intensities. In fact, λ could be a slowly varying function of time, $\lambda(t)$, and our results would be approximately correct, as long as $\lambda(t)$ does not "change very much" over any congestion period.

A limitation in implementing the methods proposed herein is that evaluation of the matrix $\beta(\mathbf{t})$ requires up to $O(N^5)$ computations for a congestion period having N arrivals. Clearly this is not practical for very large N. However, with today's computers, such calculations are feasible certainly for $N \leqq 50$ and probably for $N \leqq 100$. As a benchmark, the average number of customers who queue in an $M/G/1$ system during a period of congestion is $\rho/(1-\rho)$ (where $\rho = \lambda E$ [service time]), which is less than 10 for $\rho < 0.9$ (Kleinrock 1975, p. 217). So for many important applications the fifth order growth in computational work with N should not be an impediment to implementation. For more saturated systems, we require additional research aimed at developing faster exact algorithms or bounds and approximations.[1]

[1] This research was supported by the National Science Foundation under Grant #SES 8709811. I thank Christopher Athaide for excellent research assistance, both in computer programming and in suggesting Lemma 2. I also thank for comments on an earlier draft A. Barnett, S. Graves, D. Gross, A. Odoni, and two anonymous referees.

Appendix I

LEMMA 4. *For $t \geqq 0$, $\bar{N}_a(t)$ is a concave function of t.*

PROOF. From Lemma 1 we know that $\bar{N}_a(t)$ is piecewise linear, continuous, monotone nondecreasing. We first prove the theorem for $N = 2$, then for general N. For $x_2 > x_1$ define the "truncated ramp function"

$$l(t; x_1, x_2) \equiv \begin{cases} 0 & \text{for} \quad t \leq x_1, \\ (t - x_1)/(x_2 - x_1) & \text{for} \quad x_1 < t \leq x_2, \\ 1 & \text{for} \quad t > x_2. \end{cases}$$

Without loss of generality we can assume that the N time-conditioned arrivals occur in $[0, 1]$. Define

$\bar{N}_a(t|\Gamma) \equiv$ mean number of arrivals in $[0, t]$, given event Γ.

$N = 2$. Let the unordered arrival times be U_1, U_2. The time conditioning information is MIN $[U_1, U_2] \leqq t_1$ where $0 < t_1 < 1$, and MAX $[U_1, U_2] \leqq 1$. Without time conditioning, call that event A, U_1 and U_2 are i.i.d., uniformly over $[0, 1]$ and $\bar{N}_a(t|A) = 2t$, $0 \leqq t \leqq 1$. Hence, given A, one can write

$$2t = p_1 2l(t; 0, t_1) + p_2 2l(t; t_1, 1) + p_3[l(t; 0, t_1) + l(t; t_1, 1)],$$

where $p_1 > 0$, $p_2 > 0$, $p_3 > 0$ represent probabilities that the two unordered (unconditioned) arrival times are (1) both in $[0, t_1]$; (2) both in $(t_1, 1]$; and (3) such that one is in $[0, t_1]$ and the other is in $(t_1, 1]$. But the time conditioning information excludes possibility (2), implying that

$$\bar{N}_a(t) = \frac{p_1}{1 - p_2} 2l(t; 0, t_1) + \frac{p_3}{1 - p_2}[l(t; 0, t_1) + l(t; t_1, 1)].$$

Since $p_2 > 0$, we must have at $t = t_1$,

$$\bar{N}_a(t_1) > p_1 2l(t_1; 0, t_1) + p_3 l(t_1; 0, t_1) = 2t_1,$$

implying $\bar{N}_a(t)$ is concave.

Arbitrary N. (contradiction) If $\bar{N}_a(t)$ is not concave then there must exist at least one k for which

$$\bar{N}_a(t_k) < \bar{N}_a(t_{k-1}) + [\bar{N}_a(t_{k+1}) - \bar{N}_a(t_{k-1})][t_k - \bar{N}_a(t_{k-1})]/[t_{k+1} - t_{k-1}],$$

where $\mathbf{t} \equiv (t_i)$ is the vector of conditioning times such that the ith smallest U_j must be less than or equal to t_i, where we assume $0 \equiv t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N \equiv 1$.

Expanding the logic shown for $N = 2$, we can write

$$\bar{N}_a(t) = \sum_{j=0}^{N-1} \sum_{i=1}^{N-j} il(t; t_j, t_{j+1}) p_{ij} \tag{A1}$$

where we define the probability

$p_{ij} \equiv \Pr\{\text{exactly } i \text{ unordered arrival times fall in the interval } (t_j, t_{j+1}]\}$.

But (A1) can be written

$$\bar{N}_a(t) = \sum_{j=0}^{k-2} \sum_{i=1}^{N-j} il(t; t_j, t_{j+1}) p_{ij} + \sum_{j=k-1}^{k} \sum_{i=1}^{N-j} il(t; t_j, t_{j+1}) p_{ij} + \sum_{j=k+1}^{N-1} \sum_{i=1}^{N-j} il(t; t_j, t_{j+1}) p_{ij}. \tag{A2}$$

For $t_{k-1} < t \leqq t_{k+1}$, the first term in (A2) contributes a positive constant to $\bar{N}_a(t)$ and the third term contributes zero. Hence to determine concavity we focus on the second term and on the intervals $(t_{k-1}, t_k]$, $(t_k, t_{k+1}]$.

Suppose in any given realization of the process, we are given additional conditioning information that the random variable $N_a(t_{k-1}) = j$ for some $j \geqq k$. Then of the remaining $N - j$ time-conditioned arrivals, we may have any positive number (up to $N - j$) of them uniformly (conditionally) independently distributed over the joint interval $(t_{k-1}, t_{k+1}]$, with the remainder distributed appropriately (given the time conditioning information) over $(t_{k+1}, 1]$. For each such possibility, for $t_{k-1} < t \leqq t_{k+1}$, the conditional contribution to $\bar{N}_a(t)$ is a positively sloped straight line; probabilistically weighting each possibility, the corresponding weighted sum of straight lines is a positively sloped straight line, a property that does not violate concavity.

Now focus on the conditioning information $N_a(t_{k-1}) = k - 1$. Assume further (for the moment) that $N_a(t_{k+1}) = k - 1 + m$, i.e., m arrivals occur in $(t_{k-1}, t_{k+1}]$, for $m = 2, 3, \ldots, N - k + 1$. If the m arrivals were uniformly independently distributed over $(t_{k-1}, t_{k+1}]$, then we could write for $t_{k-1} < t \leqq t_{k+1}$,

$$\bar{N}_a(t|\beta_m) = k - 1 + [m/(t_{k+1} - t_{k-1})](t - t_{k-1})$$

$$= k - 1 + \sum_{i=0}^{m} p_i\{il(t; t_{k-1}, t_k) + (m - i)l(t; t_k, t_{k+1})\}$$

for appropriate conditional probabilities $p_i > 0$ ($i = 0, 1, \ldots, m$) and where $\beta_m \equiv$ event that $k - 1$ time-conditioned arrivals are in $(0, t_{k-1}]$ and m arrivals are uniformly independently distributed in $(t_{k-1}, t_{k+1}]$. But considering $N_a(t)$, time-conditioning prohibits the event whose probability is p_0, i.e., the event having zero of the m arrivals in $(t_{k-1}, t_k]$. Let $\beta'_m = \beta_m - \{\text{event that all } m \text{ arrivals are in } (t_k, t_{k+1}]\}$.

Then

$$\bar{N}_a(t|\beta'_m) = k - 1 + \sum_{i=0}^{m} \frac{p_i}{1 - p_0} \{il(t; t_{k-1}, t_k) + (m - i)l(t; t_k, t_{k+1})\}$$

and at the "breakpoint" t_k we have

$$\bar{N}_a(t_k|\beta'_m) = k - 1 + \sum_{i=0}^{m} \frac{ip_i}{1 - p_0} > k - 1 + [m/(t_{k+1} - t_{k-1})](t_k - t_{k-1}).$$

Hence for any m ($m = 2, 3, \ldots, N - k + 1$) we have shown that $\bar{N}_a(t|\beta'_m)$ is concave over $[t_{k-1}, t_{k+1}]$. To complete the proof we multiply each $\bar{N}_a(t|\beta'_m)$ by the appropriate conditional probability, sum to obtain $\bar{N}_a(t)$ over $[t_{k-1}, t_{k+1}]$, and use the fact that a sum of concave functions is concave.

References

BARLOW, R. E., D. J. BARTHOLOMEW, J. M. BREMNER AND H. D. BRUNK, *Statistical Inference Under Order Restriction*, John Wiley and Sons, New York, 1972.

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

KLEINROCK, L., *Queueing Systems*, Vols. 1 and 2, John Wiley and Sons, New York, 1975.

LITTLE, J. D. C., "A Proof of the Queueing Formula $L = \lambda W$," *Oper. Res.*, 9 (1961), 383–387.

WOLFF, R. W., "Poisson Arrivals See Time Averages," *Oper. Res.*, 30 (1987), 223–231.