The **Queue Inference Engine** and the **Psychology of Queueing**

ESD.86

Spring 2007

Richard C. Larson

Part 1
The Queue Inference Engine

QIE

Queue Inference Engine (QIE)

- Boston area ATMs: reams of data
- Standard approach first
- Then the notion that there may be more information in the transactional data

[Photos of people waiting in line at ATMs removed due to copyright restrictions.]

http://news.bbc.co.uk/olmedia/210000/images/_214989_bank_atm_queue150.jpg
http://europe.cnn.com/SPECIALS/2001/euro/stories/security.fears/story.france.atm.jpg

[Photo of ATM with a visually impaired user removed due to copyright restrictions.]

Source: Ed Roberts Campus

[Photo of a crowd of people waiting outside a bank removed due to copyright restrictions.]

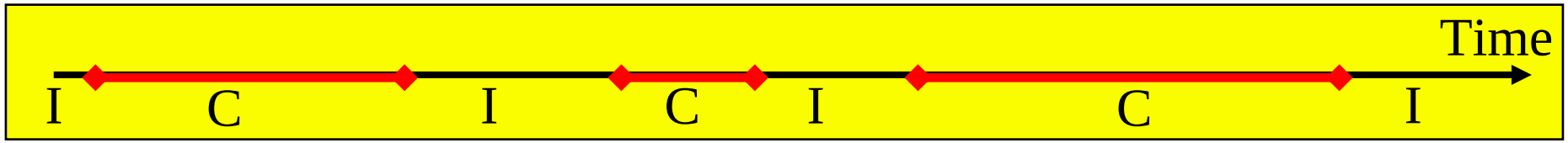
Sometimes the queue is not so orderly!
But the QIE does not care! FCFS is not a requirement for the basic QIE performance measures.

QIE: Assumptions

- (1) A priori, arriving customers are generated by a homogeneous (or slowly time varying) Poisson process;
- (2) The signature of a queue from the transactional data is a service stop time followed very shortly by a service start time;
- (3) Any balking that occurs is dependent only on whether a positive queue delay will be experienced by the prospective customer, not on the line length.

QIE: Assumptions - continued

- Can have single or multiple servers
- Service times need not be i.i.d.
- Unless otherwise specified, queue discipline need not be FCFS
- There is no parameter estimation! The rate parameter λ for the Poisson process does not appear in the analysis.



The Data-set

- The QIE works on one congestion period at a time. C=Congestion period; I=Idle period
- A congestion period commences the instant that all servers become busy, thereby requiring subsequent arrivals to be delayed in queue, and terminates the first moment that one of the servers becomes idle after a service completion because the queue is empty.

The Performance Measures

. For each congestion period the QIE computes, conditioned on the transactional data set, the following quantities:

1. The time-dependent mean number of customers in queue;
2. The mean queue delay experienced by a random customer;
3. The time average number of customers in queue over the duration of the congestion period;
4. The probability that a randomly arriving customer during the congestion period experiences *i* customers ahead of her in line.

The QIE Algorithms

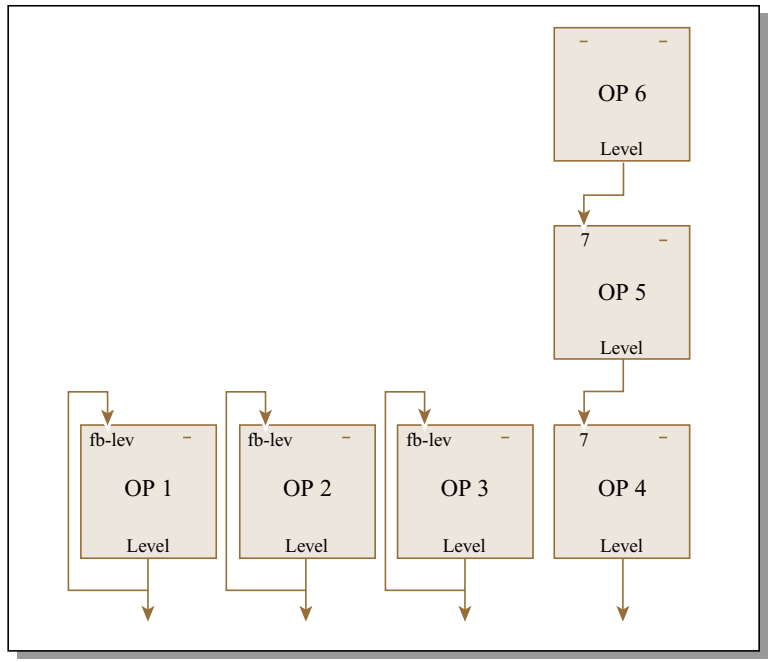


Figure by MIT OCW.

- **Order statistics** (Larson, 1989; Larson & Jones, 1994)
- **Integration** (Bertsimas and Servi, 1990)
- **Markov processes with taboo states** (Daley and Servi, 1991+)
- All O(N^3) in computational complexity

The Applications

- Servi: Air Phone
- Larson: Human server queues @ Logan Airport, Post Offices and Banks

Order Statistics:

The N Unordered Arrivals in [0,T] of a Poisson Process are Mutually Independent and Uniformly Distributed (Urban OR Text, Sec. 2.12.3)

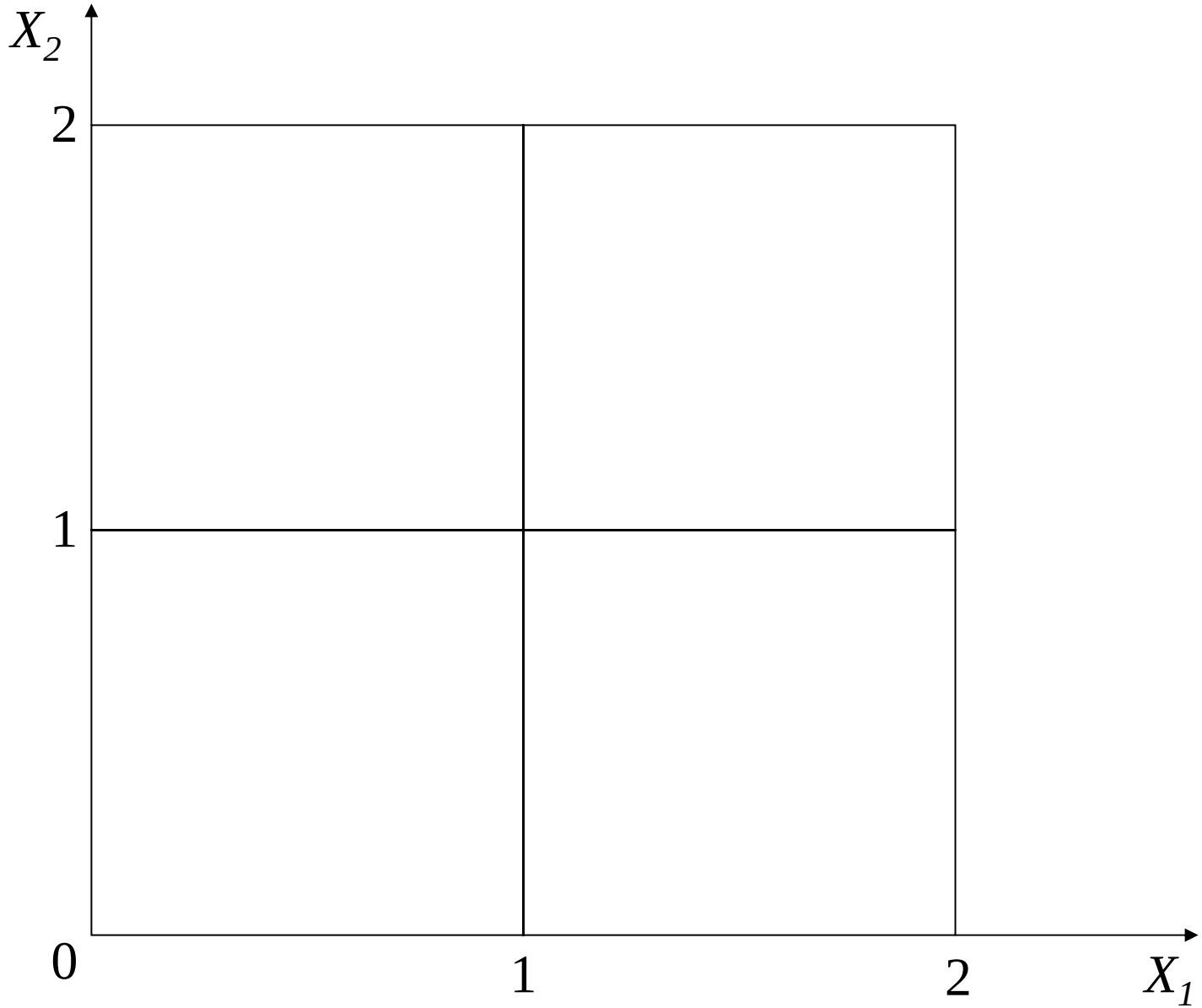
$$E[\, N(t)\,] = (t/T)N$$

$$VAR\,[\, N(t)\,] = N\,(t/T)\,(\,[T - t]/\,T\,)$$

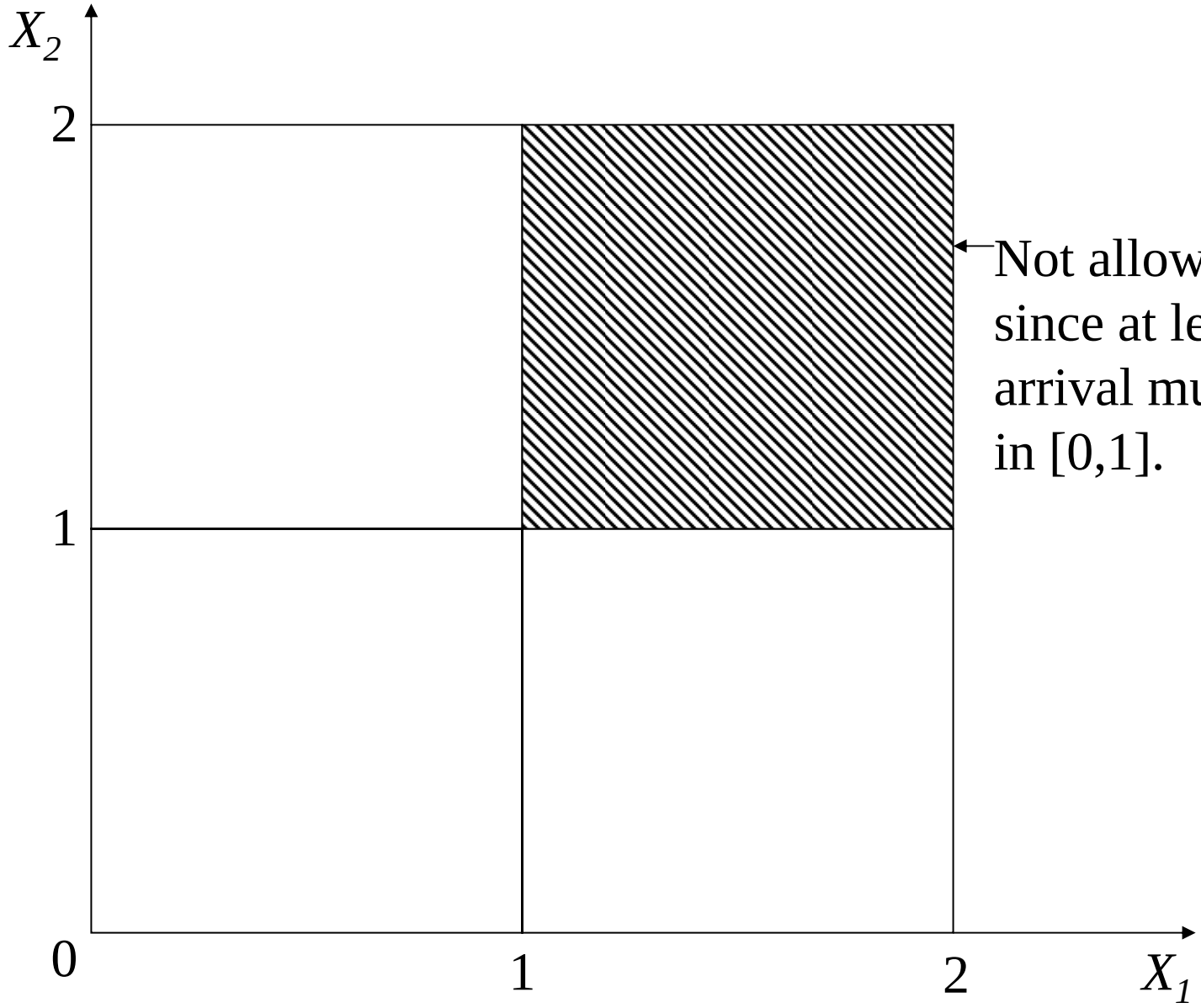
$$\Pr\{N(t) = k\} = \binom{N}{k}\left(\frac{t}{T}\right)^{k}\left(\frac{T-t}{T}\right)^{N-k}$$

Example: *M/D/1* Queue

- Service time = 1.0 minute
- Congestion period contains three customers
- Implies $N = 2$ queued customers
- Busy period = 3 minutes, starting say @ $t=0$
- We know that zero customers arrived during [2,3]. Why?
- So our 2 queued customers arrived during [0,2], with at least one in [0,1]



X_1 and X_2 are our two unordered arrival times



A priori probability of this event = 3/4 = master probability

Arrival time pdf of a randomly queued customer

PDF's for the first and second arrival times of the queued customers

PDF's for the queue waits for the first and second queued customers (FIFO)

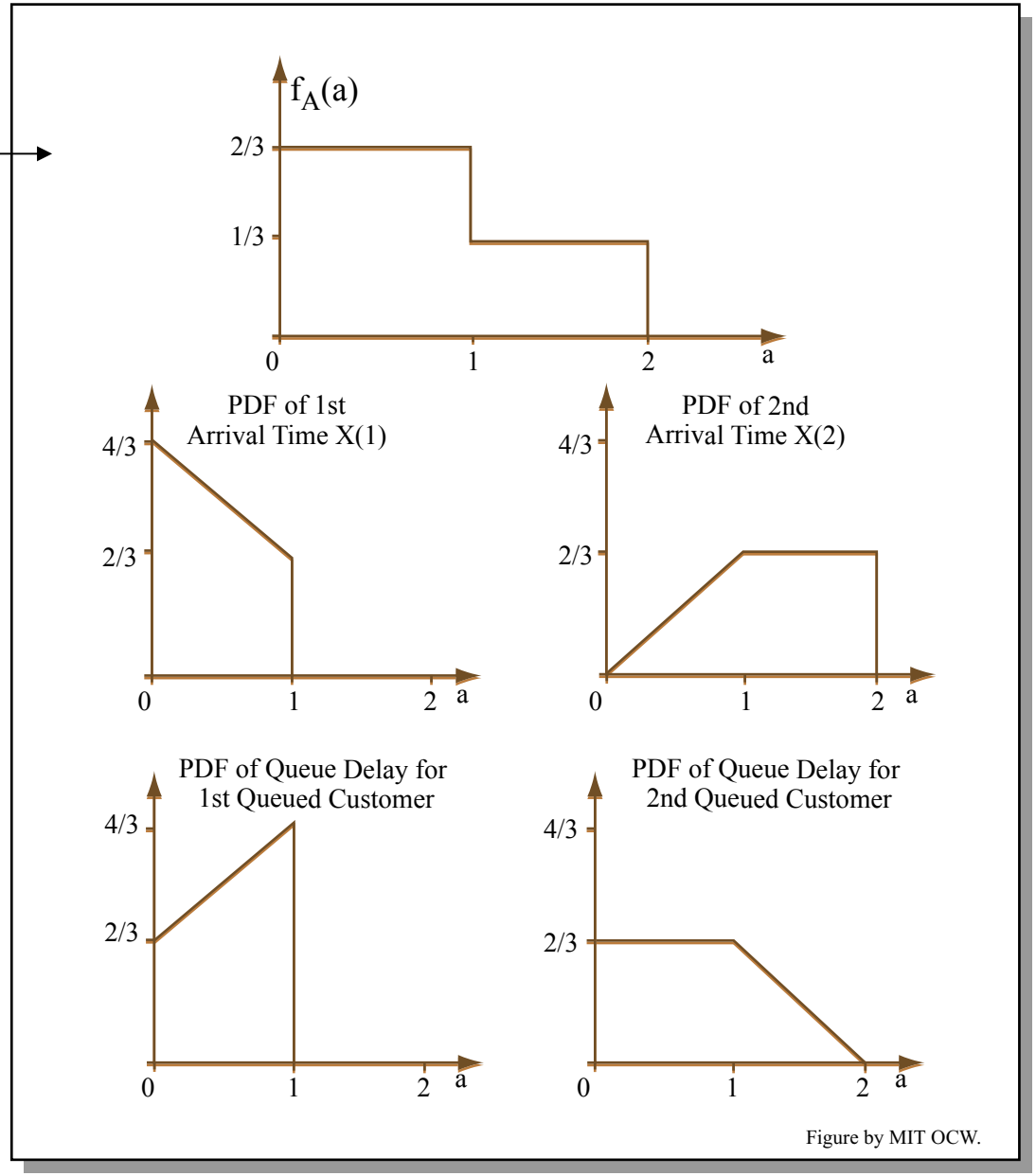


Figure by MIT OCW.

Probability density functions of arrival times and queueing delays of queued customers.

See: Larson, Richard C. "QUEUE INFERENCE ENGINE." In *Encyclopedia of Operations Research and Management Science*, Centennial Edition, Saul I. Gass and Carl M. Harris (eds.). Boston, MA: Kluwer, 2001, pp.674-679

A bank having the QIE for its ATMs could mail you your personal queue delay pdf each month!

We assume a congestion period scaled to [0,1]. We assume that there are N customers queued (and N customers who depart during the congestion period). Here t_i is the departure time for the i^{th} served customer, and it is also the time of initiation of service for the i^{th} customer to leave the queue.

The set of r.v.'s $\{X_1, X_2, ..., X_N\}$ are the i.i.d. uniformly distributed unordered arrival times.

The set of r.v.'s $\{X_{(1)}, X_{(2)}, ..., X_{(N)}\}$ are the corresponding order statistics, e.g., $X_{(2)}$ is the arrival time of the second customer to enter the queue; $X_{(2)}$ is the second smallest from the set $\{X_1, X_2, ..., X_N\}$.

We have the vector $\boldsymbol{t}$ of departure times (and of service initiation times): $\underline{\boldsymbol{t}} = (t_1, t_2, \ldots, t_N)$, where $t_{(i+1)} > t_i$.

Also consider a vector $\underline{\boldsymbol{s}} = (s_1, s_2, \ldots, s_N)$, where $s_{(i+1)} > s_i$, and where $s_i < t_i$.

We have determined how to efficiently and recursively compute the following important probability:

$$\Gamma(\underline{s}, \underline{t}) \equiv \Pr\{s_1 < X_{(1)} \le t_1, s_2 < X_{(2)} \le t_2, \ldots, s_N < X_{(N)} \le t_N | N(1) = N\}, \quad (2)$$

This is the probability that the order statistics fall within a prescribed N-rectangle (in N-dimensional space).
If $\underline{\boldsymbol{s}} = \underline{\boldsymbol{0}}$, then $\Gamma(\underline{\boldsymbol{0}}, \underline{\boldsymbol{t}})$ = 'master probability.'

Illustrative Performance Measures

1. Maximum Experienced Queue Delay (FCFS).

Interested in the cdf for the max of N non-independent r.v.s, the N queue delays.

Define $D(\tau|\underline{\boldsymbol{t}})$ = conditional probability that none of the N queued customers waited in queue τ or more time units, given the observed departure time data.

Set $s_i = s_i^* = \max\{t_i - \tau, 0\}$.

Then

$$D(\tau|\underline{\boldsymbol{t}}) = \Gamma(\underline{\boldsymbol{s}^*},\underline{\boldsymbol{t}})/\Gamma(\underline{\boldsymbol{0}},\underline{\boldsymbol{t}})$$

2. Maximum Queue Length

Set $\underline{\boldsymbol{s}} = \underline{\boldsymbol{s}}^{*K}$ such that

$$s_i^{*K} = t_{(i-K)} \quad \text{for all } i = 1,2,...,N; K = 1,2,...,N,$$

where a nonpositive subscript on t implies a value of zero.

These values for $\underline{\boldsymbol{s}}$ imply that each arriving customer i must arrive *after* the departure time of departing customer i-K during the congestion period. Now we can write

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

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

There are many more quantities we could compute within the QIE framework

- Mean queue delay
- Mean value function of queue length
- CDF of queue delay
- PDF of queue delay for each customer served (FCFS)
- PMF of queue length experienced by a random customer
- And more.....

For additional reading see

- Larson, R.C., "The Queue Inference Engine: Deducing Queue Statistics From Transactional Data." *Management Science* 36(5): 586-601, May 1990.

- Jones, Lee K. and Richard C. Larson, "Efficient Computation of Probabilities of Events Described by Order Statistics and Applications to Queue Inference." *ORSA Journal on Computation*., vol. 7, no. 1, Winter 1995, pp. 89-100.

- **See the following for an overview of the QIE including a summary of published research by Servi, Bertsimas, Daley and others on queue inferencing:** Larson, Richard C., QUEUE INFERENCE ENGINE, chapter in *Encyclopedia of Operations Research and Management Science*, Centennial Edition, Saul I. Gass and Carl M. Harris (eds.), Kluwer, Boston, 2001, pp.674-679.