FIRST PAGE

article title: Queues

authors' names: C. Armero, M.J. Bayarri

full address: Department of Statistics and Operations Research, Universitat de València, Av. Dr. Moliner 50, 46100 Burjassot, Valencia, Spain.

Queues

The intuitive notion that everybody has of the much familiar 'queues' or waiting lines closely parallels its mathematical meaning. Queueing systems are simplified mathematical models to explain congestion. It is a field of impressive activity and growth. It has mainly been concerned with the building of the models and the study of its probability properties. Relatively little effort has been put into their statistical analysis. In this article, the main concepts underlying simple queuing systems are discussed; some Bayesian inferential tools are also given.

1 Queueing Systems.

Broadly speaking, a *queueing system* occurs anytime 'customers' demand 'service' from some facility; usually both, the arrival of the customers and the service times are assumed to be random. If all of the 'servers' are busy when new customers arrive, these will generally wait in line for the next available server. The term 'customer' is very broad and can refer to telephone calls arriving at a switchboard, machine failures, demands of CPU time, patients arriving at emergency services in a hospital, etc. (The same applies to the terms 'service' and 'servers'.) In many interesting applications, customers require more than one service from different servers, and upon completion of service in one node, they have to proceed to another node. If the new server is busy, the customer has to queue up again for service. This is a typical situation modeled by the so called *networks of queues*, which shall not be treated in this article.

Queues (and queueing networks) have become very useful tools to evaluate (at least approximately) the performance of complex interacting stochastic systems in which congestion occurs. Obvious applications occur in airport terminals, healthcare centers, check-out lines at different counters, etc. They have also been traditionally applied to biology (migration, population models, etc.), electrical engineering (flow models) and chemistry (polymerisation, clustering,etc.) among others. However, by far, the most prevalent applications of queueing systems and networks occur in the fields of manufacturing, computer networking, and telecommunications and broadcasting.

Since the economical and social impact of congestion can be considerable, the importance of analyzing queueing systems was clear ever since the pioneering Erlang's work on telecommunications at the beginning of the century. Queues possess for mathematicians an added bonus, namely that the mathematics involved can be hard and intriguing; they also have an enormous potential for varying the conditions and the assumptions of the system. It does not come as a surprise that queues developed quickly into the huge area that it is today. Thousands of papers, and many journals, meetings and books are entirely devoted to the subject. It is impossible to list here even the key references, and only a small relation will be given, from which other references can be obtained. Among the numerous books, it is worth mentioning the general books of Cox and Smith (1961), Takács (1962), Cooper (1981), Gross and Harris (1985), Medhi (1991) and Nelson (1995); Kleinrock (1975, 1976) and Newell (1982) are good sources of applications; Cooper (1990) is a nice survey with an extensive bibliography. Very useful reviews of many different aspects of queueing theory (including history, statistical analysis, simulation methods, applications, etc.) can be found in Dshalalow (1995, 1997).

Simple queueing systems are defined by specifying the following: (a) *The arrival pattern*, which among other things can be deterministic or stochastic, individual or in batches (as travelers arriving to an airport passport control), etc. Also customers can abandon the line before being served, or can interact in some other ways with the rest of the customers. (b) *The service mechanism*, which describes the number of servers, how many customers at the time the server can handle (usually one, but bulk servers such as buses, elevators, etc. can serve more than one), and the distribution of service times (which can be independent among servers, time-dependent, dependent on queue length, etc.). c) *Queue discipline*, usually FIFO (first-in-first-out), but others are possible, such as last-in-first-out or LIFO (as the subroutines in a computer program), priority customers (as express check-out lines, giving priority to shorter service times), customers switching lines, etc. Besides these, other aspects can also be relevant, such as the capacity of the waiting room, the size of the population of customers, etc.

In the early 50's, Kendall introduced a useful shorthand notation for simple queueing systems. It consists on a series of symbols separated by slashes, as $A/B/c$. "A" refers to the distribution of the interarrival times between customers; thus Er stands for Erlang, G for General, M for Exponential (Markovian), etc. "B" characterizes the service distribution for each server, with the same symbols as before, an "c" refers to the number of servers, who are usually assumed to be equally efficient. More symbols are added when needed, to impose restrictions on the system capacity, and/or on the size of the population that feeds the system, and/or to specify the queue discipline if different from FIFO. One of the simplest and most studied queue is the $M/M/c$ queue, specially the $M/M/1$ and $M/M/\infty$, the later specially useful to model self-service facilities or as approximations to systems with a very large number of servers (as emergency services).

2 Equilibrium. Measures of performance.

From the probabilistic point of view, properties of queues are usually derived from the properties of stochastic processes associated with them. Let $N(t)$ denote the number of customers in the system at time $t \geq 0$. The *state probabilities* are defined as $P_j(t|\theta) = Pr(N(t) = j|\theta)$, for $j = 0, 1, 2, \ldots$, where θ denotes all parameters governing the arrival and service processes. The derivation of $P_j(t|\theta)$ is crucial in the probabilistic analysis of queueing systems. For many simple systems, the stochastic process $\{N(t)t \geq 0\}$ is a birth-death process (a 'birth' occurring with every arrival to the system, and a 'death' with every departure), greatly simplifying their study. More complex systems sometimes resort to studying other *embedded* stochastic (hopefully Markovian) processes by contemplating the systems as it would be seen by an arriving customer or by a departing customer.

However, in all but the simplest queues, determination of these probabilities is extremely difficult. Often, however, it is possible to determine their large time limit, the so-called *equilibrium* or *steady-state* distribution

$$P_j(\theta) = \lim_{t \to \infty} P_j(t|\theta)$$

without actually having to compute $P_j(t|\theta)$. The equilibrium distribution does not depend on the initial conditions of the system and it is *stationary* (does not depend on t once the statistical equilibrium is reached). The *ergodic conditions* give the restrictions on θ under which the system will eventually reach the equilibrium (with probability one). Thus, for instance, in an M/M/1 queue, the ergodic condition establishes that the arrival rate (λ) should be smaller than the service rate (μ) (that is, the average number of customers arriving in a unit time is smaller than the average number that the server can handle, a most intuitive condition).

For the most part, queueing theory deals with computations of the steady-state probabilities $P_j(\theta)$, and their use in computing other (steady-state) *measures of performance* of the queue, as the expected number of customers in the system and/or in the waiting line, the distribution of the time spent queueing or in the system or their expected values, the mean (or distribution) of the time the servers are busy (or idle), etc. These steady-state distributions are often very difficult to compute also, and many require solving systems of differential equations, inversion of transforms (generating probability functions, Laplace transfroms, ...), etc. When only the expected values are required, an extremely useful formula for systems in equilibrium is *Little's law* which e stablishes that the expected number of customers in the system (L) is equal to the average time that a costumer spends in the system (W) times the rate (λ) at which costumers enter the system.

3 Statistical Analysis: A Bayesian approach.

Most of the vast effort in queueing theory has been devoted to the probabilistic development of queueing models and to the study of its mathematical properties. That is, the parameters governing the models are, for the most part, assumed given. Statistical analyses, in which uncertainty is introduced, are comparatively very scarce. A good review is Bhat et al. (1997). It could be argued that the general results of the (quite developed) area of inference for stochastic processes (see for instance, Basawa and Prakasa Rao 1980, Lehoczky 1990) could be applied to queues. However, these general results might not take advantage of the special structure of queueing systems nor address its specific questions.

Inference in queueing systems is not easy: Development of the needed sampling distributions can be very involved and often the analysis is restricted to asymptotic results. Also, estimates can be different if the observational process differ, if the initial queue length is or not included in the analysis, etc. The statistical analysis is simpler if approached from the Bayesian perspective. More arguments in favour of Bayesian analyses of queueing systems are given in Armero and Bayarri (1999), where a comprehensive list of references with Bayesian analyses of queues is also given. As an illustration of the main issues, the analysis of the much studied $M/M/1$ queue is outlined here.

According to the assumptions of the $M/M/1$ queue, the interarrival times, Y, are i.i.d. exponential with arrival rate λ, that is

$$f(y \mid \lambda) = \lambda e^{-\lambda y}, \quad y > 0. \tag{1}$$

Similarly, the service times X are also i.i.d., $X \sim \text{Ex}(\mu)$, independently of Y, where μ is the service rate. That is

$$f(x \mid \mu) = \mu e^{-\mu x}, \quad x > 0. \tag{2}$$

Steady-state distributions exist if and only if the *ergodic condition*

$$\rho = \frac{\lambda}{\mu} < 1 \tag{3}$$

holds, where ρ is called the *traffic intensity* and plays an important role in the study of $M/M/1$ queues.

A queue can be observed in a variety of ways and different quantities can be recorded. The most traditional experiments (providing complete information) in queues observe the system over a continuous period of time $(0, T]$. T can either be a value fixed in advance, or be determined by a suitable stopping rule. For instance, the system can be observed until the busy time reaches some preassigned fixed value, until a fixed number of customers have departed from the system, until a fixed number of transitions have been recorded, and so on (references can be found in Armero and Bayarri, 1997). Different quantities can be recorded (instants of arrivals, of departures, of commencing of service, number of arrivals and/or departures, interarrival times, service times, number of transitions, busy times, etc.). If the queue is assumed to be in equilibrium, the observation of the number of customers at time 0 can also be added, providing more informative but rather cumbersome likelihood functions.

Assume that an experiment providing n_a interarrival times, $Y_1, \ldots, Y_{n_a}$, and n_s service completions, $X_1, \ldots, X_{n_s}$ has been performed. Then, from (1) and (2), a likelihood function for (λ, μ) based on the observed $\mathbf{x} = (x_1, x_2, \ldots, x_{n_s})$ and $\mathbf{y} = (y_1, y_2, \ldots, y_{n_a})$ is

$$\ell(\lambda, \mu) \propto f(\mathbf{x}, \mathbf{y} | \lambda, \mu) \propto \lambda^{n_a} e^{-\lambda t_a} \mu^{n_s} e^{-\mu t_s} \tag{4}$$

where $t_a = \sum_{i=1}^{n_a} y_i$, $t_s = \sum_{j=1}^{n_s} x_j$, are the totals of observed interarrival times and service times, respectively. Notice that this likelihood is valid whether n_a and/or n_s are fixed or random (as long as the likelihood function is proportional to (4)). Also, the observation of the arrival and the service processes does not need to be simultaneous, and the $X's$ and $Y's$ does not need to correspond to successive customers. (Of course, many of the observational procedures mentioned above result in likelihood functions proportional to (4).) Notice also that (4) applies whether or not the system is in equilibrium and whether or not equilibrium can eventually be reached.

Maximum likelihood estimators are immediate from (4), and given by

$$\hat{\lambda} = \frac{n_a}{t_a}, \qquad \hat{\mu} = \frac{n_s}{t_s} \qquad \hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}} \tag{5}$$

However, standard errors or confidence intervals do require careful consideration of the specific stopping rule used, and different experimental designs will in general produce different sampling distributions for these estimators, and hence different standard errors and confidence intervals.

Bayesian analyses are insensitive to (non-informative) stopping rules, and all that it is required from the data is the likelihood function (4). A prior distribution has to also be assessed quantifying whatever is known about the system before the data is collected. Usually, there is plenty information a priori about the queue, specially if it is assumed to be in equilibrium (which implies, at the very least, that the system has been working for an extended period of time). A prior distribution incorporating all this information should then be assessed. However, to keep the parallelism with the likelihood analysis and to also avoid the incorporation of further subjective inputs, we shall use a Bayesian analysis usually called 'objective' because the prior distribution used is of the 'non-informative' or 'objective' type. Since both λ and μ are scale parameters, the invariant measures for each will be used, and they will be assumed independent a priori, that is

$$\pi(\lambda,\mu) \propto \frac{1}{\lambda}\,\frac{1}{\mu}. \tag{6}$$

Combining (4) with (6) with Bayes theorem, we get the joint posterior distribution

$$\pi(\lambda,\mu|\mathbf{x},\mathbf{y}) \propto \pi(\lambda,\mu)\; f(\mathbf{x},\mathbf{y}|\lambda,\mu) \propto Ga(\lambda|n_a,t_a)\; Ga(\mu|n_s,t_s), \tag{7}$$

so that λ and μ are independent a posteriori with scaled chi-squared distributions $2t_a\lambda \sim \chi^2_{(2n_a)}$, $2t_s\mu \sim \chi^2_{(2n_s)}$, and since they are independent, $(t_a/t_s)\rho \sim F_{(2n_a,2n_s)}$, an F distribution. These distributions will be *exactly* valid for all experiments producing the likelihood function (4), irrespective of the stopping rule. Usual estimates of λ, μ and ρ are the means (or the modes) of their (marginal) posterior distributions, and 'standard errors' associated to these estimates are most naturally given by the standard deviations of those distributions. Intervals with a given probability content can similarly be given. All these will numerically reproduce their non-Bayesian counterparts when the design has n_a and n_s fixed in advance.

A most important issue when studying queues is how likely it is that steady-state can be reached. From the marginal distribution of ρ, this and other questions of direct interest can be answered. In particular, the probability that the ergodic condition holds is computed from the F distribution above as

$$\pi_e = Pr(\rho < 1|\mathbf{x},\mathbf{y}) = \int_0^1 \pi(\rho|\mathbf{x},\mathbf{y})d\rho = \Psi_{(2n_a,2n_s)}(t_a/t_s) \tag{8}$$

where $\Psi_{(2n_a,2n_s)}$ is the distribution function of an $F_{(2n_a,2n_s)}$ distribution.

Most often, queues are assumed to be in equilibrium. Under this assumption, the restriction $\rho < 1$ imposed by the ergodic condition should be incorporated into the analysis. This is easily done, and in fact amounts to truncating the scaled F posterior distribution for ρ giving

$$\pi(\rho|\mathbf{x},\mathbf{y},\rho<1) = \frac{\pi(\rho|\mathbf{x},\mathbf{y})}{\pi_e} \propto \frac{\rho^{n_a-1}}{(1+(t_a/t_s)\rho)^{n_a+n_s}} \quad \text{for } \rho < 1, \tag{9}$$

and $\pi(\rho|\mathbf{x},\mathbf{y},\rho<1) = 0$ for $\rho \geq 1$. From (9), estimates, standard errors, credible intervals, etc. that take into account the restriction can be produced. (In this simple case, the probability π_e in (8), the proportionality constant in (9) and the moments of the truncated posterior can be given close form expressions in terms of Gauss' hypergeometric functions.) In Armero and Bayarri (1999), an explicit expression for a factor that can be used for adjusting the usual MLE $\hat{\rho}$ is given, as well as (adjusted) 'standard errors'. It can there be seen that incorporating the restriction $\rho < 1$ has the effect of 'shrinking' the usual MLE, the shrinkage being larger the larger $\hat{\rho}$ and the smaller the sample sizes, a most intuitive result.

When statistically analyzing queueing systems, inference about the parameters are rarely (if ever) of direct interest. Usually the main goal is prediction of observable quantities that describe the utilization and degree of congestion of the system. From the Bayesian perspective, it is also very easy to produce predictive distributions incorporating the uncertainty in the estimation of the unknown parameters. The predictive distribution of the steady-state number, N, of customers in the system is computed as

$$p(N = n|\mathbf{x}, \mathbf{y}, \rho < 1) = \int_0^1 Pr(N = n|\rho)\pi(\rho|\mathbf{x}, \mathbf{y}, \rho < 1)d\rho, \tag{10}$$

where $Pr(N = n|\rho)$ is given by the Geometric $Ge(1 - \rho)$ distribution and $\pi(\rho|\mathbf{x}, \mathbf{y}, \rho < 1)$ is given in (9). In this case, again, (10) can be given a close form expression (see Armero and Bayarri, 1994, for details).

Similar arguments apply for prediction of the number of the (steady-state) number of customers in the waiting line, time spent by a costumer in the system (or waiting in line), length of busy (idle) periods, etc. All these predictive distributions are a most natural tool for deciding future interventions in the queue (see Armero and Bayarri, 1997).

Bibliography

Armero C, Bayarri MJ 1994 Bayesian prediction in $M/M/1$ queues. *Queueing Systems* 15:401-417

Armero C, Bayarri MJ 1997 Bayesian questions and answers in queues. In Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds.) *Bayesian Statistics 5.* Oxford University Press, Oxford

Armero C, Bayarri MJ 1999 Dealing with uncertainties in queues and networks of queues: a Bayesian approach. In Ghosh S (ed.) *Multivariate Analysis, Design of Experiments and Survey Sampling.* Marcel Dekker, New York

Basawa IV, Prakasa Rao BLS 1980 *Statistical Inference for Stochastic Processes.* Academic Press, New York

Bhat UN, Miller GK, Rao SS 1997 Statistical analysis of queueing systems. In Dshalalow JH (ed.) *Frontiers in Queueing: Models and Applications in Science and Engineering.* CRC Press, Boca Raton

Cooper RB 1981 *Introduction to Queueing Theory. 2nd ed.* North-Holland, New York

Cooper RB 1990 Queueing Theory. In Heyman DP, Sobel MJ (eds.) *Handbooks in OR & MS , Vol 2.* Elsevier Science Publishers (North-Holland), Amsterdam

Cox DR, Smith WL 1961 *Queues.* Chapmann and Hall, London

Dshalalow JH (ed.) 1995 *Advances in Queueing: Theory, Methods, and Open Problems.* CRC Press, Boca Raton

Dshalalow JH (ed.) 1997 *Frontiers in Queuing. Models and Applications in Science and Engineering.* CRC Press, Boca Raton

Gross D, Harris CM 1985 *Fundamentals of Queueing Theory. 2nd ed.* Wiley, New York

Kleinrock L 1975 *Queueing Systems, Vol. 1: Theory.* Wiley, New York

Kleinrock L 1976 *Queueing Systems, Vol. 2: Computer Applications.* Wiley, New York

Lehoczky J 1990 Statistical Methods. In Heyman DP, Sobel, MJ (ed.) *Handbooks in OR & MS, Vol 2.* Elsevier Science Publishers (North-Holland), Amsterdam

Medhi J 1991 *Stochastic Models in Queueing Theory.* Academic Press, Boston

Nelson R 1995 *Probability, Stochastic Processes and Queueing Theory.* Springer Verlag, New York

Newell GF 1982 *Applications of Queueing Theory. 2nd ed.* Chapman and Hall, London

Takács L 1962 *Introduction to the Theory of Queues.* Oxford University Press, Oxford