Invited paper

STATISTICAL ANALYSIS OF QUEUEING SYSTEMS

U. NARAYAN BHAT

Department of Statistics, Southern Methodist University, Dallas, Texas 75275, USA

and

S. SUBBA RAO*

Department of Management and Systems, Washington State University, Pullman, Washington 99164, USA

Received 2 May 1986
(Revised 2 June 1986)

Abstract

This paper provides an overview of the literature on statistical analysis of queueing systems. Topics discussed include: model identification, estimation, hypothesis testing and other related aspects. Not all of these statistical problems are covered in books on queueing theory or stochastic processes. The bibliography is not exhaustive, but comprehensive enough to provide sources from the literature.

Keywords and phrases

Statistical analysis of queueing systems, estimation problems in queueing theory, hypothesis testing in queueing theory, inference in queueing theory.

0. Introduction

Statistical analysis is an integral part of formulating a mathematical model for a real system. A model is not of much use unless it is related with the system through empirical data analyses, parameter estimation and tests of relevant hypotheses. However, in queueing theory statistical analyses have taken a backseat due to two major

*Current address: Department of Information Systems and Operations Management, University of Toledo, Toledo, Ohio 43606, USA

© J.C. Baltzer A.G., Scientific Publishing Company

reasons. Unlike time series, the input process of queues is fully described and statistical analyses of various elements of the input process can be carried out through established procedures, thus making the inference studies of the underlying process less urgent. From a global view, stochastic processes underlying queueing systems are special cases of more general stochastic processes, and the general theory of inference on stochastic processes has made major strides over the last two decades (see Basawa and Prakasa Rao [4], and Jacobsen [33]).

Despite these considerations, queueing systems present special problems not usually confronted in general statistical investigations and provide much more specific structure so as to be able to go beyond the general theory on stochastic processes. In this survey, we provide an overview of this subject area, with particular emphasis on topics not covered by books and earlier survey articles.

The paper is in six sections. Section 1 deals with statistical problems arising from model identification. In sects. 2, 3, 4 we survey estimation problems encountered in queueing systems and in sect. 5, hypothesis testing and sequential analysis. The final section 6 identifies some related topics and prospects for further work. We use the well-known Kendall [35] notation GI/G/s in representing different systems. Any departure from the usual description will be specified whenever necessary.

Earlier articles giving overviews of statistical analysis in queueing theory emphasizing various aspects are the following: Cox [19], Harris [32], Reynolds [48] and Gross and Harris [30], sect. 6.6. Some of the techniques discussed in Cox and Lewis [20] and Lewis [38] are very pertinent to the statistical analysis of queueing systems.

1. Identification of models

In the formulation of a queueing model, one starts with the identification of its elements and their properties. The system structure is easily determined. What remains is the determination of the form and properties of the input and service processes. Four major steps are essential in this analysis: (i) collection of data, (ii) tests for stationarity in time, (iii) tests for independence, and (iv) distribution selection.

(i) COLLECTION OF DATA

The required data form depends largely on the proposed model and the nature of results sought. For instance, in an M/M/1 queue (Poisson arrivals, exponential service and single server), traffic intensity can be estimated as the ratio of the estimates of arrival and service rates. Alternatively, noting that the traffic intensity provides the utilization factor for the system, we may use the empirical utilization factor as its estimate. Some of the pitfalls of this approach are indicated by Cox [19], who notes that if ρ is the traffic intensity of the system, the efficiency of this approach

is given by $1 - \rho$. Also see the discussion by Burke on Cox's article regarding the bias resulting from estimating the load factor in an M/M/s loss system as (average number of customers in the system)/(1 − probability of loss) and Descloux [24].

An additional problem relates to the sampling plan of the investigation. How long should the system be observed – for a specified length of time or until a specified number of events has occurred? If the arrival process is Poisson, Birnbaum [11] has shown that the second alternative is better in determining the sample size than the first one. But when nothing is known regarding the processes, no such statements can be made and the efficiency of different schemes should be considered in individual cases. Another aspect of the sampling plan is the mode of observation; for discussions of what is known as the snap reading method and systematic sampling, the reader is referred to Cox [19] and Cox [18], p. 86, respectively.

(ii) TESTS FOR STATIONARITY

A comprehensive treatment of tests for stationarity has beeen given by Cox and Lewis [20]. In addition to the treatment of data on the occurrence of events as a time series and the determination of second-order properties of the counting process, they consider statistical problems related to renewal processes and provide tests of significance in some general as well as special cases. Lewis [38] updates this study and considers topics such as trend analysis of non-homogeneous Poisson processes.

In many queueing systems (such as airport traffic and telephone traffic), the non-stationarity of the arrival process leads to a periodic behavior. Furthermore, even though the process is non-stationary when the entire period is considered, it might be possible to consider it as a piecewise stationary process in which stationary periods can be identified (e.g. a rush hour). Under such circumstances, a procedure that can be used to test the stationarity of the process as well as to identify stationary periods is the Mann–Whitney–Wilcoxon test (see, for example, Bradley [15], Conover [17], Randles and Wolfe [45]), appropriately modified to handle ties in ranks (Putter [44], Mielke [41]). The data for the test can be obtained by considering two adjacent time intervals $(0, t_1]$ and $(t_1, t_2]$ and observing the number of arrivals during such intervals for several time periods. Let $X_1, X_2, \ldots, X_n$ be the number of arrivals during the first interval for n periods, and $Y_1, Y_2, \ldots, Y_m$ be the number of arrivals during the second interval for m periods (usually $n = m$). If F and G represent the distribution of X's and Y's, respectively, then the hypothesis to be tested is $F = G$ against the alternative $F \neq G$ for which the Mann–Whitney–Wilcoxon statistic can be used. Using this test, successive stationary periods can be delineated and the system can be studied in detail within such periods (see Moore [42], who gives an algorithm for the procedure).

To analyze cyclic trends of the type discussed above, we may also use the periodogram method described by Lewis [38] for the specific case of a non-homogeneous Poisson process.

(iii) TESTS FOR INDEPENDENCE

While formulating queueing models, several assumptions of independence are made regarding its elements. Thus, most of the models assume that inter-arrival times and service times are independent sequences of independent and identically distributed random variables. If there are reasons to suspect such assumptions, statistical tests can be used for verification. Some of the tests that can be used to verify independence of a sequence of observations are: tests for serial independence in point processes, described by Lewis [38], and various tests for trend analysis and renewal processes given by Cox and Lewis [20]. To verify the assumption of independence between inter-arrival times and service times, non-parametric tests seem appropriate. Tests such as Spearman's rho and Kendall's tau (Bradley [15], Conover [17], and Randles and Wolfe [45]) test for the correlation between two sequences of random variables, whereas Cramér-von Mises type statistics (see Koziol and Nemec [37] and other references cited therein) test for bivariate independence directly using the definition of independence of random variables.

Tests for the dependence structure can also be carried out on the process output such as the number of customers in the system. Then a check for Markovian dependence can be made using well-known tests for Markov chains (see Bhat [7], ch. V and references cited therein).

(iv) DISTRIBUTION SELECTION

The next step in the model identification process is the determination of the best model for arrival and service processes. The distribution selection problem is a standard one, and based on the nature of data and the availability of analytical models approximate distributions can be chosen. For a comprehensive discussion of this problem, readers are referred to Gross and Harris [30], pp. 389–397. It should be noted however, that it is advisable to start with simple distributions such as the Poisson, since analysis under such assumptions is considerably simpler. After all, a mathematical model is essentially an approximation to the real process. The simpler the model, the easier it is to analyze it and extract information from it. Thus, the selection of the distribution should be made with due regard to the tradeoff between the advantages of the sophistication of the model and our ability to derive information from it.

2. Parameter estimation: the maximum likelihood method

Estimation problems in queueing theory are of three types: (i) parameter estimation based on the maximum likelihood method, (ii) the method of moments and non-parametric methods, and (iii) process mean value estimation based on autocorrelation and second-order properties of an underlying stationary process. Clarke [16], Beneš [5], Wolff [53] and Cox [19] have explored the parameter estimation

problem via the maximum likelihood method for the M/M/s type queues. Harris [32] has extended this method to $M/E_2/1$ queues. More recently, this method has been used by Harishchandra and Rao [31] for parameter estimation in $M/E_k/1$ queues. Basawa and Prabhu [3] derive moment estimates as well as maximum likelihood estimates for general queues over random time horizon. A good review of auto-correlation functions in queues is made by Reynolds [48]. Notable papers using this approach for process mean value estimation are by Daley [22,23], Blomqvist [12,13], Gafarian and Ancker [26], Reynolds [47] and Aigner [1]. In the remainder of this section we review maximum likelihood estimates (m.l.e.). In the following two sections we describe the methods (ii) and (iii) mentioned above.

(a) PARAMETER ESTIMATION IN MARKOVIAN SYSTEMS

A landmark paper in parameter estimation is by Clarke [16], who assumes that the queue M/M/1 is fully observed over a period of time and complete information is available in the form of arrival epochs, and the points of beginning and end of service on each customer. Let n_a, n_s, t_e, t_b represent the number of arrivals, number of service completions, the time spent in the empty state, and the time spent in the busy states, respectively, in the observation interval $[0, t]$. Further, let n_0 be the initial number in the queue. Denote by λ and μ the arrival and service rates of the system, which we assume to be in equilibrium. The likelihood function can be written as

$$L(\lambda, \mu) = \left(\frac{\lambda}{\mu}\right)^{n_0} \left(1 - \frac{\lambda}{\mu}\right) \lambda^{n_a} \mu^{n_s} e^{-(\lambda+\mu)t_b} e^{-\lambda t_e}, \tag{1}$$

the m.l.e.'s of λ and μ are found to be

$$\hat{\lambda} = (\hat{\mu} - \hat{\lambda})(n_a + n_0 - \hat{\lambda}t) \quad \text{and} \quad \hat{\lambda} = (\hat{\lambda} - \hat{\mu})(n_s - n_0 - \hat{\mu}t_b). \tag{2}$$

Estimating $\hat{\mu}$ from the second equation gives a quadratic in $\hat{\lambda}$. Of the two solutions, any negative solution is rejected and for the remaining values of $\hat{\lambda}$, corresponding $\hat{\mu}$ is obtained. Further, any pair $(\hat{\lambda}, \hat{\mu})$ would be rejected for which $\hat{\mu} \leqslant 0$ or $\hat{\lambda}/\hat{\mu} \geqslant 1$. If both solutions are valid, then the solution which maximizes the likelihood function is chosen.

If $n_s - n_0$ is large, Clarke gives a simple approximation for $\hat{\lambda}$ and $\hat{\mu}$ as

$$\hat{\lambda} \cong (n_a + n_0)/t, \quad \hat{\mu} \cong (n_s - n_0)/t_b . \tag{3}$$

If we ignore the initial queue size, the estimates of λ and μ are, respectively, n_a/t and n_s/t_b. Whether this can be done depends upon whether we have observations from a very long realization or we observe a number of independent, fairly short realizations making up the sample. In the latter case, as Cox [19] points out using the Fisher

information measure, the information provided by the initial state could be remarkably high. A practical consequence is that in a given situation, it may often be advantageous to split the observations into a number of independent periods while observing the initial state in each section.

The above analysis can be extended to most Markovian queueing systems. In particular, for those queues satisfying the generalized birth-death process, the conditional likelihood is of the form

$$e^{-\Sigma(\lambda_i+\mu_i)t_i}\prod\lambda_i^{n_{a_i}}\mu^{n_{s_i}}, \tag{4}$$

where λ_i, μ_i are the rates of arrival and service completions in state i, n_{a_i} and n_{s_i} are the numbers of arrivals and serive completions in state i, and t_i is the total time spent in state i during the observation interval $(0, t]$. For the finite state birth-death queue, ignoring the contribution of the initial queue size, the m.l.e.'s of λ_i and μ_i are given by

$$\hat{\lambda}_i = n_{a_i}/t_i \quad (0 \leqslant i \leqslant M-1), \qquad \hat{\mu}_i = n_{s_i}/t_i \quad (1 \leqslant i \leqslant M). \tag{5}$$

The above results and similar estimates for parameters in M/M/s, M/M/∞ and machine interference problems given below are due to Wolff [53].

If we assume $\lambda_i = ah(i)$ and $\mu_i = bg(i)$, where a and b are the parameters to be estimated and $h(i)$ and $g(i)$ are known functions of i, the log-likelihood function becomes

$$\sum_{i=0}^{\infty} n_{a_i}\log ah(i) + \sum_{i=1}^{\infty} n_{s_i}\log bg(i) - \sum_{i=0}^{\infty} t_i\{ah(i) + bg(i)\}. \tag{6}$$

This yields the estimates

$$\hat{a} = \sum_{i=0}^{\infty} n_{a_i} \Big/ \sum_{i=0}^{\infty} t_i h(i), \quad \hat{b} = \sum_{i=1}^{\infty} n_{s_i} \Big/ \sum_{i=1}^{\infty} t_i g(i). \tag{7}$$

For ergodic queues, note that

$$\lim_{t\to\infty} \sum_{i=0}^{\infty} n_{a_i} \Big/ \sum_{i=1}^{\infty} n_{s_i} = 1$$

with probability one. Therefore, for large samples, Σn_{a_i} and Σn_{s_i} can be replaced with $n/2$, n being the total number of transitions (total of all arrivals and service completions) in $(0, t]$. For large samples this yields the estimates

$$\hat{a} \cong n \Big/ \left[2 \sum_{i=0}^{\infty} t_i h(i)\right], \quad \hat{b} \cong n \Big/ \left[2 \sum_{i=1}^{\infty} t_i g(i)\right]. \tag{8}$$

The estimates for M/M/∞, M/M/s and machine-interference models are obtained from the above as follows:

M/M/∞ : $a = \lambda,\ h(i) = 1,\ b = \mu,\ g(i) = i.$

M/M/s : $a = \lambda,\ h(i) = 1,\ b = \mu,\ g(i) = i,\ i \leqslant s-1,\ g(i) = s,\ i \geqslant s.$

Machine-interference : $a = \lambda,\ h(i) = i,\ i \leqslant M,\ b = \mu,\ g(i) = 1,\ i \leqslant M-1,$ $g(i) = 0,\ i \geqslant M.$

The effect of neglecting the initial queue length on the estimates is seen by comparing (5) and (3). Estimates can also be different if the observational procedures are different. For example, for the M/M/∞ model let us assume that during an interval $[0, t]$, the following are observed:

n_0 : number of calls at the start of the period,

n_a : number of calls arriving during the period,

n_d : number of calls terminating during the period,

$\bar{n}$: average number of calls during the period.

Beneš [5] has shown that n_0, n_a, n_d and $\bar{n}$ are sufficient statistics to estimate the arrival rate λ and service rate μ. The estimates are

$$\hat{\lambda} = \frac{1}{t}(n_a + n_d) - \hat{\mu}\bar{n}$$

$$\hat{\mu} = \frac{1}{\bar{n}t}\left\{n_d - n_0 - \bar{n} + [(n_d - n_0 - \bar{n})^2 + 4\bar{n}(n_a + n_d)]^{1/2}\right\}. \tag{9}$$

For the same model, Wolff gives the estimates as in (7) with $h(i) = 1$ and $g(i) = i$. The differences are due to the fact that the estimates are based on different sets of observations and statistics. While Wolff uses the counts up and down out of a state, Beneš uses the number of arrivals and services in the observation period.

The means, variances and correlation coefficients of these estimators, namely $\hat{\lambda}$ and $\hat{\mu}$ are also given by Beneš [5].

As pointed out by Cox [19], confidence intervals for λ, μ and ρ in an M/M/1 system can be obtained by observing that $2\lambda(t_e + t_b)$ can be treated as a chi-square

variate with $2n_a$ degrees of freedom and $2\hat{\mu}t_b$ as a chi-square variate with $2n_s$ degrees of freedom and noting that the ratio of two chi-square variates leads to an F-distribution. For details, readers are referred to Lilliefors [39] and Gross and Harris [30], p. 383.

(b) ESTIMATION IN NON-MARKOVIAN SYSTEMS

Regarding the general problem in non-Markovian systems, Cox [19] observes that the maximum likelihood estimates of the arrival and service distributions can be determined for the following more general class of queueing systems: (i) arrivals occur as a point process with a specified probabilistic structure except for unknown parameters; (ii) with each customer is associated a service time which is a random variable independent of the arrival pattern; (iii) given the arrival epochs and service times, the entire process is either uniquely determined or has a distribution independent of the unknown parameters.

Under these conditions, the likelihood function will be the product of those for arrival patterns and for the service times, the time spent in service, say x_Q, by the very last customer, and the probability for the initial number of customers. The parameters can be estimated, at least numerically, by maximizing the likelihood function once a plausible functional form has been chosen for the inter-arrival time and service time densities. For an illustration of this approach, see Gross and Harris [30], p. 386, where m.l.e.'s for parameters in the $M/E_2/1$ queue are considered.

(c) THE GI/G/1 SYSTEM – ESTIMATION OF ARRIVAL AND SERVICE TIME PARAMETERS

Basawa and Prabhu [3] obtain the m.l.e.'s of parameters of the arrival and service time distributions with continuous densities $f(u; \theta)$ and $g(v; \phi)$, respectively. The sampling scheme is to observe the queue until the first n customers have departed from the system and note the service times of these n customers, say $(v_1, v_2, \ldots, v_n)$. Let the nth departure epoch be D_n and observe the inter-arrival times of all customers who arrive during $(0, D_n]$, obtaining the inter-arrival sequence $(u_1, u_2, \ldots, u_{N_A})$, where $N_A = N_A(D_n) = \max\, k: u_1 + \ldots + u_k \leqslant D_n$. Under this sampling scheme, the likelihood function is

$$L_n(f, g) = \left\{ \prod_{i=1}^{N_A} f(u_j; \theta) \right\} \left\{ \prod_{i=1}^{n} g(v_j; \phi) \right\} \cdot [1 - F(x_n; \theta)], \tag{10}$$

where

$$x_n = x_n(D_n) = D_n - \sum_{1}^{N_A} u_j .$$

Since the factor $[1 - F(x_n; \theta)]$ causes difficulty in obtaining simple estimates, consider the alternative approximate likelihood function

$$L_n^a(f, g) = \left\{ \prod_{i=1}^{N_A} f(u_j; \theta) \right\} \left\{ \prod_{i=1}^{n} g(v_j; \phi) \right\} . \tag{11}$$

If $\hat{\theta}_n^a$, $\hat{\phi}_n^a$ are the m.l.e.'s of θ and ϕ based on $L_n^a(f, g)$, they are the solutions of equations

$$\sum_{1}^{N_A} \frac{\partial}{\partial \theta} \log f(u_j; \theta) = 0, \quad \sum_{1}^{n} \frac{\partial}{\partial \phi} \log g(v_j; \phi) = 0. \tag{12}$$

They prove that $\hat{\theta}_n^a$ and $\hat{\phi}_n^a$ are consistent estimators of θ and ϕ and

$$\begin{bmatrix} \sqrt{n}\,(\hat{\theta}_n^a - \theta) \\ \sqrt{n}\,(\hat{\phi}_n^a - \phi) \end{bmatrix} \Rightarrow N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2/\eta, & 0 \\ 0, & \sigma_\phi^2 \end{pmatrix} \right\} , \tag{13}$$

where N_2 represents a bivariate normal density,

$$\sigma_\theta^2 = \left[E\left(\frac{\partial}{\partial \theta} \log f \right)^2 \right]^{-1}, \quad \sigma_\phi^2 = \left[E\left(\frac{\partial}{\partial \phi} \log g \right)^2 \right]^{-1} \tag{14}$$

and $\eta = \max(1, \rho)$, ρ being the traffic intensity.

Let $\hat{\theta}_n$ and $\hat{\phi}_n$ be the estimators based on the full likelihood function (10). It is seen that $\hat{\phi}_n = \hat{\phi}_n^a$, and $\hat{\theta}_n$ differs from $\hat{\theta}_n^a$, but it can be shown that $\hat{\theta}_n$ and $\hat{\theta}_n^a$ have the same limit distributions whenever

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \log (1 - F(x_n; \theta)) \to 0 \tag{15}$$

in probability. This condition is satisfied for Erlangian arrivals. For large samples, estimates of θ and ϕ can be determined from (12) at least numerically, if not in a closed form. Using (13), confidence intervals for θ and ϕ can also be constructed. From a practical point of view, it is significant to note that the limit properties of statistics are obtained without the imposition of steady state assumptions on the system. Further, in their paper Basawa and Prabhu also consider m.l.e.'s for arrival and service rates in the M/M/1 queue based on a sample function observed over a fixed

interval $(0, t]$, as done by Wolff [53], and obtain limit distributions of the m.l.e.'s without any restrictions on ρ.

(d) ESTIMATION USING IMBEDDED MARKOV CHAINS

By observing a process only at the imbedded Markov points, a certain amount of information is lost. However, the gain is the analytical tractability when inter-arrival or times are not exponential. Suppose we are interested only in the estimation of the traffic intensity instead of the individual rates of arrival and service. Then we may use observations on the imbedded Markov chain in some queueing systems of the class M/G/1 or GI/M/1. In particular, when the arrivals are Poisson and the service times are Erlangian (E_k) the number of arrivals during service intervals form a sequence of i.i.d. random variables with a probability distribution consisting of only two parameters, the shape parameter k and the traffic intensity ρ. Thus, if X_n denotes the number of arrivals during the service of the $(n+1)$st customer in the queue $M/E_k/1$, then X_n has the negative binomial distribution given by

$$\Pr(X_n = x) = f(x, \rho) = \binom{x+k-1}{x} \left(\frac{\rho}{\rho+k}\right)^x \left(\frac{k}{\rho+k}\right)^k, \quad (x = 0, 1, 2, \ldots) \tag{16}$$

Suppose the system is observed only at departure epochs, and let $x_1, x_2, \ldots, x_n$ be the number of arrivals during the first n service times, respectively. The likelihood function for this sample is then

$$L(x_1, x_2, \ldots, x_n; \rho) = \prod_{i=1}^{n} \binom{x_i+k-1}{x_i} \left(\frac{\rho}{\rho+k}\right)^{x_i} \left(\frac{k}{\rho+k}\right)^k. \tag{17}$$

The m.l.e. of ρ is found to be $\hat{\rho} = \Sigma x_i/n$. This estimator is unbiased and consistent, since $E(\hat{\rho}) = \rho$ and $\mathrm{Var}(\hat{\rho}) = \rho(\rho+k)/(kn)$. Further, it turns out that $\hat{\rho}$ is also the minimum variance bound (MVB) estimator and therefore the uniformly minimum variance unbiased estimator (UMVUE) of ρ. It can be shown that the probability distribution of X belongs to the one-parameter exponential family and hence $T = \Sigma x_i$ is a sufficient statistic for ρ. Finally, for large values of n,

$$\frac{1}{\sigma}\sqrt{n}(\hat{\rho} - \rho) \to N(0, 1), \tag{18}$$

where

$$\sigma^2 = \left[E\left(\frac{\partial}{\partial \rho} \log f(x, \rho)\right)^2\right]^{-1} = \frac{\rho(\rho + k)}{k} . \tag{19}$$

Using this result, large sample confidence intervals for ρ can be computed. These results are due to Harishchandra and Rao [31].

Unfortunately, a similar approach for the $E_k/M/1$ system does not work since, to obtain an imbedded Markov chain, the random variables $\{X_n\}$ are defined as potential services during an inter-arrival period and they are not observable during idle periods.

3. Parameter estimation; method of moments

When it is not possible to observe the system completely, several interesting estimation problems arise. One such situation occurs when we observe the output process of the M/G/1 queue and we wish to estimate the mean inter-arrival and mean service times. If the process is in equilibrium then, as pointed out by Cox [19], the arrival rate can be estimated with full asymptotic efficiency since it should nearly equal the departure rate over a long time period. If the service time is also exponential, no inference is possible about the mean service time since in that case the limiting distribution of the output is the same as that of the input, i.e. Poisson (Burke [55] and Reich [56]). On the other hand, if the service times are other than exponentially distributed, estimation of the mean service time is feasible. This is facilitated by the relation (Gross and Harris [30], p. 387),

$$C(t) = \frac{\lambda}{\mu} B(t) + \left(1 - \frac{\lambda}{\mu}\right) \int_0^t B(t - x) \lambda e^{-\lambda x} dx , \tag{20}$$

where $C(t)$ and $B(t)$ are the distributions of inter-departure and service times, respectively. In particular, if the service time is a constant ($= \mu^{-1}$), it can be shown that its estimate is given by the minimum observed inter-departure time.

When the service time distribution is other than exponential or deterministic, the method of moments can be used. From (20) the Laplace–Stieltjes transforms (LST) are found to be

$$C^*(s) = \frac{\{1 + (s/\mu)\} B^*(s)}{(1 + s/\lambda)} , \tag{21}$$

where $C^*(s)$ and $B^*(s)$ are, respectively, the LST of the inter-departure and service times. If β_r and γ_r are the cumulants of service and inter-departure times, expanding (21) in powers of s and taking logs, we obtain

$$\gamma_1 = \frac{1}{\lambda}, \quad \gamma_2 = \beta_2 - \beta_1^2 + \frac{1}{\lambda^2}, \quad \gamma_3 = \beta_3 - 2\beta_1^3 + \frac{2}{\lambda^3}, \text{etc.} \tag{22}$$

If we assume a particular form for $B(x)$, we can have as many equations (22) as there are parameters in $B(x)$ and estimate these parameters by equating them to the observed moments of the inter-departure times. However, there are some problems in using this method. This has to do with the fact that the dependence of successive inter-departure times and auto-correlation need to be taken into consideration while calculating moments of the inter-departure times from observed data. We may either test for the absence of correlation by computing the auto-correlation coefficient or see that data are spread sufficiently far apart to ensure an approximate random sample (see Harris [32], p. 355, and Cox [19], p. 301).

When observations on the waiting time are available, the estimation of the arrival rate and the service time parameters for the M/G/1 queue can be made by using the Pollaczek–Khintchine formula. For further details, see Cox [19] and Gross and Harris [30], p. 388.

For the GI/G/1 queue, Basawa and Prabhu [3] propose the following moment estimates for the means a and b of the arrival and service time distributions:

$$\hat{a}_n = \frac{1}{N_A} \sum_{1}^{N_A} u_j, \quad \hat{b}_n = \frac{1}{n} \sum_{1}^{n} v_j. \tag{23}$$

The sampling scheme and the quantities u_j, v_j, N_A and n are the same as described earlier in connection with the maximum likelihood method. It should be observed that while $\hat{b}_n$ is the usual sample mean, $\hat{a}_n$ is based on a random number of observations. They show that $\hat{a}_n$ and $\hat{b}_n$ are consistent estimators for a and b and further that

$$\begin{bmatrix} \sqrt{n}\,(\hat{a}_n - a) \\ \sqrt{n}\,(\hat{b}_n - b) \end{bmatrix} \Rightarrow N_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2/\eta, & 0 \\ 0, & \sigma_2^2 \end{pmatrix} \right\}, \tag{24}$$

where N_2 is the bivariate normal density, σ_1^2 and σ_2^2 are, respectively, variances of arrival and service times, and $\eta = \max\{\rho, 1\}$. As observed earlier, these properties of the estimates do not require the imposition of steady-state conditions. Further, the estimates are "natural" estimates and simple. However, if it is required to find estimates of other parameters of either the arrival or service times, it is not clear whether this simplicity can still be maintained.

4. Covariance structure, auto-correlation and process mean value estimation

In the previous sections, we have identified procedures to estimate various parameters of the arrival and service processes based on random samples of observations on a queueing system. These estimates in turn provide estimates of the queue characteristics such as the mean queue length $E[X(t)]$ and mean waiting time $E(W_n)$. However, it is also possible to estimate these quantities directly from sample observations. Thus, let us consider the $X(t)$ process, which we assume to be stationary in the wide sense. To estimate the process mean value $\mu = E[X(t)]$, we observe $X(t)$ over some interval $(0, T]$ and construct a suitable sample mean. Two obvious candidates for estimating μ are (Reynolds [48]):

$$\hat{\mu}_1 = \sum_{r=1}^{n} X(rh)/n, \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{T}\int_0^T X(t)\,dt, \tag{25}$$

where $nh = T$.

It can be seen that $\hat{\mu}_2$ is the limiting form of $\hat{\mu}_1$ as $n \to \infty$ and $h \to 0$. Both $\hat{\mu}_1$ and $\hat{\mu}_2$ are unbiased estimators for μ. To assess their precision as well as to compare them with other estimators for μ, we need to evaluate their standard errors. These are given by:

$$\text{Var}(\hat{\mu}_1) = \frac{\sigma^2}{n}\left[1 + \frac{2}{n}\sum_{j=1}^{n-1}(n-j)\,\rho(jh)\right] \tag{26}$$

$$\text{Var}(\hat{\mu}_2) = \frac{2\sigma^2}{T}\int_0^T \left(1 - \frac{x}{T}\right)\rho(x)\,dx, \tag{27}$$

where $\sigma^2 = \text{Var}\{X(t)\}$, $\rho(x)$ is the auto-correlation function of the $X(t)$ process defined by $\rho(x) = \text{cov}[X(t), X(t+h)]/\sigma^2$, and $\rho(jh)$ are similarly defined for the discrete process $X(rh)$ sampled from the $X(t)$ process. From (26) and (27) it is obvious that the sampling errors of $\hat{\mu}_1$ and $\hat{\mu}_2$ (and other such estimators) can not be evaluated without the knowledge of the covariance structure of $X(t)$. Further, it may be noted that for large n

$$\text{Var}(\hat{\mu}_1) \cong \frac{\sigma^2}{n}\left[1 + 2\sum_{j=1}^{\infty}\rho(jh)\right]. \tag{28}$$

These results have motivated the study of the covariance structure in queues and of the asymptotic behavior of $\Sigma_{j=1}^{\infty} \rho(jh)$. For a good review of the work in the area, see Reynolds [48].

To obtain confidence intervals for the process mean in the system M/M/1, Gebhard [28] uses (28) along with the result

$$\text{Var}(\hat{\mu}_1) \cong A(\hat{\mu}_1) \frac{\sigma^2}{n}, \tag{29}$$

where $A(\hat{\mu}_1) = 1 + 2\rho(1 + \rho)/(1 - \rho)^2$. By applying the central limit theorem for dependent variables, Gebhard shows that the distribution of $\hat{\mu}_1$ is asymptotically normal with mean $E[X(t)]$ and variance given by (29). These results enable the construction of confidence intervals.

We proceed to discuss some ramifications of the above results.

(a) SAMPLE SIZE DETERMINATION

An obvious use of the above results is in determining the minimum sample size in simulation runs consistent with some required precision. It is clear that a larger sample will be required to estimate the process mean when the observations are dependent. In particular, when the observations are independent, the required sample size n_I is proportional to σ^2, while when the observations are serially correlated, the required sample size n_c is proportional to $\sigma^2 (1 + 2 \Sigma \rho(jh))$ so that

$$n_c/n_I = 1 + 2\sum \rho(jh). \tag{30}$$

The key to sample size determination is the evaluation of $\Sigma\rho(jh)$. For simple queues this may not be too much of a problem. If we consider the GI/M/1 queue, the equilibrium distribution is geometric. Using this fact and the property of stochastic monotonicity in Markov chains, for this system Daley [22] shows that as $\rho \to 1$, $(1-\rho)^2$ $\Sigma\rho(jh) \to 1 + \alpha_2/2\alpha_1^2$, where α_1 and α_2 are the first and second moments of the arrival time distribution. Combining this result with (28), it can be concluded that the coefficient of variation of $\hat{\mu}_1$ is constant for ρ near 1 when $n \cong k(1 - \rho)^{-2}$, for some constant k. A similar result should hold for the GI/M/s queue as well.

(b) TIME SLICING AND EVENT SEQUENCING

Up to this point we have considered the estimator $\hat{\mu}_1$ when the $X(t)$ process is observed only at times rh $(r = 1, 2, \ldots, n; nh = T)$. This sampling scheme is referred to as "time-slicing" by Gafarian and Ancker [26]. In order to use $\hat{\mu}_2$, we have to observe the process continuously during $(0, T]$ and the sampling is then called "event sequencing". In this latter case, $\hat{\mu}_2$ can be approximated by

$$\hat{m}_2 = \frac{1}{T} \sum_{i=1}^{N} S_i,$$

where N is the number of customers arriving in $(0, T]$ and S_i is the service time of the ith customer. For small values of ρ in the M/M/1 case, it can be shown that for large T,

$$\text{Var}(\hat{m}_2) \cong 2\rho^2/(\lambda T), \tag{31}$$

a result due to Beneš [6].

Comparisons of the efficiencies of $\hat{\mu}_1$ and $\hat{\mu}_2$ require a little more detailed investigation, which has been done by Gafarian and Ancker and later by Reynolds [47] for the M/M/s type of queues. Recall that the efficiency of $\hat{\mu}_1$ relative to $\hat{\mu}_2$ is defined as

$$E_h = \text{Var}(\hat{\mu}_2)/\text{Var}(\hat{\mu}_1). \tag{32}$$

For a class of processes for which $\rho(h) = e^{-ch}$ $(c > 0)$, E_h is obtained as

$$E_h = \frac{2(1 - e^{-ch})^2 \, (nch - 1 + e^{-nch})}{(ch)^2 \, [n(1 - e^{-2ch}) - 2e^{-ch}(1 - e^{-nch})]} . \tag{33}$$

The constant c is a measure of how rapidly the correlation between two samples, h units of time apart, decreases as h increases. By graphing E_h against ch and n, Gafarian and Ancker conclude that for any reasonable efficiency $E_h > 0.90$, the sampling interval h must be less than the constant $1/c$, preferably by a substantial margin. The M/M/1, M/M/∞ and the loss system M/M/s satisfy the condition $\rho(h) = e^{-ch}$. Qualitatively similar results have been obtained by Reynolds [47] for finite Markov queues, wherein the eigenvalue structure is exploited in exhibiting $\rho(h)$ and E_h. He concludes that $E_h < 1$ for all $h > 0$ and therefore event sequencing is always asymptotically more efficient than time slicing. However, from a practical point of view it is easier to handle time slicing than event sequencing. Since $E_h \to 1$ as $h \to 0$, a value of h can be determined to ensure that E_h is as close to 1 as required, in which case time slicing can be used with only this acceptable loss of efficiency.

(c) REPLICATION VERSUS EXTENSION

A final question in a simulation experiment to estimate the process mean is whether one should extend the observation interval from $[0, T]$ to $[0, mT]$ or carry out m independent simulations of the system on $[0, T]$, the objective being the

reduction of the standard error of the estimate (say) $\hat{\mu}_2$. Gafarian and Ancker [26] show that replication is preferable to extension. In this connection, we should also note the observation made by Cox [19] based on Fisher's information measure for the two estimates.

(d) ESTIMATION OF WAITING TIME

We now consider the sequence $\{W_r\}$, where W_r is the waiting time of the rth customer. We shall assume the queue discipline to be first-come, first-served. The mean waiting time $\mu_W = E\{W_r\}$ is estimated by $\hat{\mu}_W$ (Blomqvist [12]), where

$$\hat{\mu}_W = \frac{1}{n} \sum_{r=1}^{n} W_r . \tag{34}$$

We have

$$E(\hat{\mu}_W) = \mu_W, \quad \text{Var}\,(\hat{\mu}_W) = \frac{\sigma_W^2}{n} \left[1 + \frac{2}{n} \sum_{1}^{n-1} (n-j)\,\rho_j \right], \tag{35}$$

where

$$\rho_j = \gamma_j/\gamma_0, \quad \gamma_j = \text{Cov}\,(W_r, W_{r+j}), \quad \text{and} \quad \sigma_W^2 = \text{Var}\{W_n\} .$$

In the M/G/1 case, Blomqvist has shown that for large n,

$$\text{Var}\,(\hat{\mu}_W) \cong \frac{1}{n} \left(2 \sum_{r=0}^{\infty} \gamma_r - \gamma_0 \right) . \tag{36}$$

The sum $\Sigma\gamma_r$ is evaluated in terms of the derivatives of the LST of the waiting time distribution. Blomqvist has demonstrated numerically the effect of auto-correlation in the case of the M/E_k/1 queue by tabulating n Var $(\hat{\mu}_W)$/Var (W_n), i.e. the factor by which the variance of the sample mean of uncorrelated observations should be multiplied to allow for the effect of auto-correlation.

A later paper by Blomqvist [14] gives heavy traffic results for the covariance functions in the GI/G/1 queue and considers the problem of estimating $P(W > w)$ under steady-state conditions. Further details of the behavior of the correlation structures for the waiting time process can be found in Daley [23] and Craven [21].

(e) DIRECT ESTIMATOR VERSUS MLE

Let $W_n^S = W_n + S_n$ be the time spent by the customer in the system. A direct estimator for $\mu_W^S = E(W)$, based on the $X(t)$ process, was proposed by Jenkins [34] as

$$\hat{\mu}_W^S = \frac{\int_0^T X(t)\,dt}{A(T)} = \frac{I}{A}, \tag{37}$$

where $A(T)$ is the number of arrivals in $(0, T]$. An approximation to Var $(\hat{\mu}_W^S)$ is given by Jenkins in the form

$$\text{Var}\left(\frac{I}{A}\right) \cong \left[\frac{E(I)}{E(A)}\right]^2 \left[\frac{\text{Var}\,(I)}{E^2(I)} + \frac{\text{Var}\,(A)}{E^2(A)} - \frac{2\,\text{Cov}\,(I, A)}{E(I)E(A)}\right]. \tag{38}$$

Evaluation of these relevant quantities in the M/M/1 case leads, for large T, to

$$\text{Var}\,(\hat{\mu}_W^S) \cong \rho^2\,(1+\rho)^2/\{\lambda^3\,(1-\rho)^4\,T\}. \tag{39}$$

On the other hand, the m.l.e. of μ_W^S is $\hat{\mu}_W^{\star S} = (\hat{\mu} - \hat{\lambda})^{-1}$, where $\hat{\mu}$, $\hat{\lambda}$ are the m.l.e.'s of μ and λ. This has the variance

$$\text{Var}\,(\hat{\mu}_W^{\star S}) \cong \rho^2\,(1+\rho^2)/\{\lambda^3\,(1-\rho)^4\,T\}. \tag{40}$$

Comparing (39) and (40), we find that the efficiency of the direct estimator $\hat{\mu}_W^S$ relative to $\hat{\mu}_W^{\star S}$ is $E = (1+\rho^2)/(1+\rho)^2$, which varies between 1 and 1/2 as ρ varies between 0 and 1. This points to the fact that except for cases of low traffic intensity, there is a marked loss of efficiency in estimation using only the waiting time of individual customers.

(f) ESTIMATION USING CROSS-SECTIONAL DATA

The previous sections examined parameter estimation from single sample realizations or time series data. Sometimes it may be possible to obtain cross-sectional data from a number of identical queueing systems. Because there exist proportionality relations such as $L = \lambda W$, one can use ratio and least-squares estimation methods with cross-sectional data that may provide statistics with useful properties. An interesting paper on parameter estimation using this approach in an M/M/1 queue is by Aigner [1]. Suppose the following observations are made:

u = time between successive arrivals (inter-arrival times),

v = service times,

n = number of units (customers) in the system,

q = number of units (customers) in the queue,

z = waiting time in the system, and

x = waiting time in the queue.

Then for the queue M/M/1, the following relations hold:

$$E(n) = \lambda E(z),\quad E(x) = \rho E(z),\quad E(q) = \lambda E(x),\quad E(q) = \rho E(n),\quad E(n) = \mu E(x).$$

A "direct" ratio (R) estimator for λ based on a sample of size N is given by

$$\hat{\lambda}^{R}_{nz} = \bar{n}/\bar{z}, \tag{41}$$

where $\bar{n}$ and $\bar{z}$ are, respectively, the sample means of observations on n and z. We find that $\hat{\lambda}^{R}_{nz}$ is asymptotically unbiased under random sampling assumptions and its asymptotic variance is

$$V(\hat{\lambda}^{R}_{nz}) = \frac{\lambda^2}{N}\,\frac{1-\rho}{\rho}. \tag{42}$$

Moreover, in this case $\hat{\lambda}^{R}_{nz}$ is also the maximum likelihood estimator $\hat{\lambda}^{ML}_{nz}$.

The least-squares (LS) estimator for λ proposed by Aigner is generated by a homogeneous regression of n on z. This estimator is then given by

$$\hat{\lambda}^{LS}_{nz} = \sum_{i=1}^{N} n_i z_i \Big/ \sum_{i=1}^{N} z_i^2, \tag{43}$$

where $(n_i, z_i), i = 1, 2, \ldots, N$ denote the paired observations on n and z. In general, even though the LS estimator could be biased and inconsistent, for the M/M/1 queue it is found that LS estimators are consistent. The large sample variance of $\hat{\lambda}^{LS}_{nz}$ is obtained as

$$V(\hat{\lambda}^{LS}_{nz}) = \frac{3}{2}\left[\frac{\lambda^2}{N}\cdot\frac{1-\rho}{\rho}\right]. \tag{44}$$

Another estimate of λ with the same asymptotic variance as $\hat{\lambda}^{R}_{nz}$ is $\hat{\lambda}^{R}_{qx} = \bar{q}/\bar{x}$. Comparing the estimators $\hat{\lambda}^{R}_{nz}$, $\hat{\lambda}^{LS}_{nz}$, $\hat{\lambda}^{R}_{qx}$ for λ with the MLE $\hat{\lambda}^{ML}_{u} = 1/u$, Aigner observes that when the traffic intensity is low, $\hat{\lambda}^{ML}_{u}$ is the most efficient, if $\rho > 0.5$, $\hat{\lambda}^{ML,R}_{nz}$ and $\hat{\lambda}^{R}_{qx}$ are generally more efficient than $\hat{\lambda}^{ML}_{u}$, and the least-squares estimator $\hat{\lambda}^{LS}_{nz}$ is always less efficient than the alternatives.

Similar arguments on estimators of μ lead to the following observation: The best estimator (with the least asymptotic variance for all values of ρ) of μ is given by:

$$\hat{\mu}_{nz}^{\mathrm{ML}} = \frac{1+\bar{n}}{z} \tag{45}$$

with asymptotic variance $(\mu^2/N)(1-\rho)$. The major drawback of this estimator is that it requires observations on waiting times which are generally expensive to collect.

Finally, based on asymptotic variance, except where ρ is small (< 0.4), the best estimator of ρ is the ML and the LS estimator.

$$\hat{\rho}_{n}^{\mathrm{ML,LS}} = \frac{\bar{n}}{1+\bar{n}}, \tag{46}$$

which uses only the number of customers in the system. This estimator has the asymptotic variance $\rho(1-\rho)^2/N$. When ρ is small, the estimator

$$\hat{\rho}_{q}^{\mathrm{ML}} \cong \frac{1+\bar{q}}{2+\bar{q}} \tag{47}$$

has a smaller asymptotic variance given by

$$\frac{\rho^2}{N}\left[\frac{\rho+1-\rho^2}{1+(1-\rho)^2}\right]. \tag{48}$$

(g) OPTIMAL PREDICTION

A natural extension of the analysis of the covariance structure of the output process is the use of their properties in their prediction. For the queue GI/M/1, Stanford et al. [50] provide an algorithm to obtain the optimal mean square predictor $E[Y_n|\underset{\sim}{Q}_k]$, where Y_n can be Q_n (the imbedded queue length at the nth arrival epoch), W_n (the waiting time of the nth customer) or S_n (the system time for the nth customer), and $\underset{\sim}{Q}_k = (Q_0, Q_1, \ldots, Q_n)$. The mean squared errors of the predictors are obtained through a bounding approach. These results have been extended to the GI/M/s system by Woodside et al. [54].

5. Hypothesis testing

Hypothesis testing in queueing systems still remains a vast unexplored area. A hypothesis testing problem arises when we are required to make inferences about parameters of arrival and service time distributions or measures such as traffic intensity, as well as the form of distributions, based upon sampled data from the system. Inferences may lead to control procedures. The sampled data, as we have seen in the previous section, can be from queueing processes such as queue length, waiting time or the output. In this section, these aspects of parameter testing will be discussed in some detail.

(a) SIGNIFICANCE TESTS FOR ARRIVAL AND SERVICE PARAMETERS IN M/M/1 SYSTEMS

For an M/M/1 queue, significance tests for λ and μ can be based on a chi-square distribution. As noted in subsect. 2(a), $2\hat{\lambda}t = 2\hat{\lambda}(t_e + t_b)$ and $2\hat{\mu}t_b$ are chi-square variates with $2n_a$ and $2n_s$ degrees of freedom, respectively, and therefore a test for $\rho = \lambda/\mu$ can be based on the F-distribution. This procedure assumes that the state of the system (empty or busy) is under observation throughout the interval $(0, t)$.

As proposed by Cox [19], a simple test for the proportion of idle times in an M/M/1 system, which is also the probability that the waiting time is zero, is obtained by observing that under the null hypothesis that the system is M/M/1, the number n_{ae} of arrivals to the empty queue has a binomial distribution with index n_a and parameter t_e/t.

(b) A TEST FOR EXPONENTIAL SERVICE USING WAITING TIME DATA

At times, full observation of a system may not be possible. Suppose we are able to observe only $W_1, W_2, \ldots, W_n$, the waiting times of the first n successive customers. We assume that the inter-arrival times have an exponential density, and we wish to test the hypothesis that the service times are exponential which can be stated as H_0: G = M in an M/G/1 queue (Thiagarajan and Harris [52]).

The main difficulty arises from the fact that $\{W_n\}$ are serially correlated. Let us assume that none of the W_n are zero. Then we have the well-known relationship $W_{n+1} = W_n + Y_n$, where $Y_n = v_n - u_n$, with u_n = inter-arrival time and v_n = service time. Here, $\{v_n\}$ and $\{u_n\}$ are assumed to be i.i.d. random variables, and therefore $\{Y_n\}$ are i.i.d. as well. Under H_0: G = M, the conditional densities of Y, given $Y > 0$ and $Y < 0$ are, respectively,

$$g(y \mid Y > 0) = \mu e^{-\mu y} \ (y > 0), \quad \text{and} \quad g(y \mid Y < 0) = \lambda e^{\lambda y}. \tag{49}$$

Equation (49) suggests that the test for G = M can be stated as follows: Split the data Y_n into two groups, one consisting of positive numbers and another consisting of negative numbers. Test for exponentiality separately, using the test proposed by Gnedenko (see Gnedenko et al. [29]). For details and tests for cases when there are zero waiting times, see Thiagarajan and Harris [52].

(c) A UMP TEST FOR ρ IN AN M/E_k/1 SYSTEM

Using the property (16), namely that the number of customers $(X_1, X_2, \ldots, X_n)$ arriving during service periods in an M/E_k/1 queue are i.i.d. random variables with a negative binomial distribution, Harishchandra and Rao [31] have developed a likelihood ratio test for ρ based on a sample $\underset{\sim}{x} = (x_1, x_2, \ldots, x_n)$. By the Neyman–Pearson lemma, a uniformly most powerful test of size α for $H_0: \rho = \rho_0$ against $H_1: \rho > \rho_0$ is given by

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \text{if} \quad \sum x_i > c \\ \gamma(\underset{\sim}{x}) & \text{if} \quad \sum x_i = c \\ 0 & \text{if} \quad \sum x_i < c \end{cases}, \tag{50}$$

where c and γ are determined such that

$$\alpha = P\left[\sum x_i > c \,|\rho_0\right] + \gamma P\left[\sum x_i = c \,|\rho_0\right], \tag{51}$$

and $\phi(\underset{\sim}{x})$ is the probability of rejecting H_0.

The procedure is to reject H_0 with probability 1 whenever $\Sigma x_i > c$, reject H_0 with probability γ whenever $\Sigma x_i = c$, and accept H_0 otherwise. Note that this is a randomized test. The power function of the test is given by $\beta(\rho) = P(\Sigma x_i > c \,|\, \rho) + \gamma P(\Sigma x_i = c \,|\, \rho)$. An example is provided by the following:

EXAMPLE

Consider the M/M/1 queue and let $H_0 : \rho = 0.8$ against $H_1 : \rho > 0.8$, and set $\alpha = 0.05$, $n = 10$. In this case

$$f(x, \rho) = \left(\frac{\rho}{1+\rho}\right)^x \frac{1}{1+\rho}, \qquad (x = 0, 1, 2, \ldots),$$

so $Y = \Sigma x_i$ has the distribution

$$f(y, \rho) = \binom{y+n-1}{y} \left(\frac{\rho}{1+\rho}\right)^y \left(\frac{1}{1+\rho}\right)^n.$$

The quantities c and γ for the test are determined from the relation

$$0.05 = \sum_{y=c+1}^{\infty} \binom{y+10-1}{y} \left(\frac{0.8}{1.8}\right)^y \left(\frac{1}{1.8}\right)^{10}$$

$$+ \gamma \binom{c+10-1}{c} \left(\frac{0.8}{1.8}\right)^c \left(\frac{1}{1.8}\right)^{10}.$$

which yields $c = 15$ and $\gamma = 0.56995$.

The test procedure is then:

reject H_0 with probability 1 whenever $\Sigma x_i > 15$;

reject H_0 with probability 0.56995 whenever $\Sigma x_i = 15$;

accept H_0 otherwise.

(d) SEQUENTIAL PARAMETER CONTROL

In operating a queueing system, monitoring and control of the parameters of the system are essential to ensure that the system performance is up to design standards, or exigencies of the environment call for changing parameters to keep the system under stable conditions. The parameter control problem in effect involves a problem of testing hypothesis $H_0 : \theta = \theta^0$, where θ^0 is the desired vector of parameters, against a suitable alternative. If the hypothesis is not rejected at a suitably chosen level of significance, we conclude that the system parameters have not changed, while rejection of the hypothesis is indicative of the fact that the system has changed from the desired state. Once detection of change is achieved, appropriate control action can be taken.

Bhat and Rao [9] considered the problem of controlling the traffic intensity paramter ρ in M/G/1 and GI/M/1 queues. In an analogy with statistical quality control, the technique used is such that the queueing system is left undisturbed as long as it satisfies certain conditions, but when the conditions are violated, it is readjusted so as to make it consistent with the original objectives. For an ideal control technique, the type I and type II errors should be under control.

Let $t_0, t_1, t_2, \ldots$ be a discrete set of epochs at which the system is observed and Q_n the number of customers observed at epoch t_n, $(n = 0, 1, 2, \ldots)$. The technique is based on the sample function $\{Q_n\}$. In general, the parameter control technique consists of two phases and two sets of control limits for $\{Q_n\}$. The first phase (a warning phase) indicates the time at which the sample function gets out of the region covered by upper and lower control limits, say c_u and c_ℓ; the second phase (the testing phase) is intended to see whether the process returns to the control region within a specified amount of time and involves two limits, say d_u and d_ℓ. For M/G/1 and GI/M/1 queues, if we select $\{t_n\}$ to be the points of regeneration so that $\{Q_n\}$ is the imbedded Markov chain, the first set of limits are approximately determined using the equilibrium distribution of $\{Q_n\}$. Let $Q^\star = Q_\infty$ and α_u and α_ℓ two specified probabilities. Then c_u and c_ℓ are integers such that

$$c_u = \min\{k \mid P(Q^\star \geqslant k) \leqslant \alpha_u\}, \quad c_\ell = \max\{k \mid P(Q^\star \leqslant k) \leqslant \alpha_\ell\}. \qquad (52)$$

It is somewhat harder to determine the second set of limits (d_u and d_ℓ) as this involves first passage distributions. Bhat and Rao have shown that in the M/G/1 case, if β_u and β_ℓ are two specified probabilities and $c_\ell = 0$, then

$$d_u = \min\{n \mid P(T_1 > n) \leqslant \beta_u\}, \quad d_\ell = \min\{n \mid k_0^n \leqslant \beta_\ell\}, \tag{53}$$

where T_1 is the length of a busy period initiated by a single customer and k_0 is the probability that no customers arrive during a service period. Note that c_u, c_ℓ, d_u, d_ℓ are determined under $H_0 : \rho = \rho_0$.

Once these limits are obtained, the control technique may be described for given values of α_u, α_ℓ, β_u, β_ℓ as follows.

(i) Starting with an initial queue length i and traffic intensity ρ_0, leave the system alone as long as Q_n lies between c_u and c_ℓ, or when it goes out of these limits if it returns within bounds before d_u and d_ℓ transitions, respectively.

(ii) If the system does not return within bounds before d_u or d_ℓ consecutive transitions as the case may be, conclude that the traffic intensity has changed from ρ_0 and reset the system to bring traffic intensity back to the level ρ_0.

(iii) Repeat the procedure in (i) and (ii) above using the last state of the system as the initial state.

Tables for c_u, c_ℓ, d_u and d_ℓ are provided by Bhat and Rao [9] for the $M/E_k/1$ queue for some selected values of k. The second phase (testing phase) of the sequential parameter control technique requires the knowledge of the first passage distributions and a great amount of numerical work. An attempt to alleviate this problem in the system M/G/1 has been made by Bhat [8] using a modified distribution free procedure based on a censored sample of number of customers arriving during consecutive service periods.

(e) SEQUENTIAL PROBABILITY RATIO TESTS

When the difference between parameter values under null and alternative hypotheses is large, a sequential test has the advantage of using a considerably smaller sample size. With this objective, Rao et al. [46] have developed a procedure for testing the hypothesis $H_0 : \rho = \rho_0$ against an alternative $H_1 : \rho \neq \rho_1$ using Wald's Sequential Probability Ratio Test (SPRT) for the systems M/G/1 and GI/M/s, in which queue length processes have imbeded Markov chains $\{Q_n\}$. Let the transition probability of the chain be $p_{ij}(\rho)$ and let n_{ij} be the number of transitions $i \to j$ in $\{Q_n\}$ up to and including the nth transition. Then the likelihood ratio for the SPRT is

$$L_n = \prod_{i,j} p_{ij}^{n_{ij}}(\rho_1) \Big/ \prod_{i,j} p_{ij}^{n_{ij}}(\rho_0). \tag{54}$$

Let $A = (1-\beta)/\alpha$, and $B = \beta/(1-\alpha)$, where α and β are the probabilities of the errors of the first and second type. The SPRT procedure is as follows:

(a) if $L_n \geqslant A$, accept H_1;

(b) if $L_n \leqslant B$, accept H_0, and

(c) if $B < L_n < A$, observe the next queue length Q_{n+1}, compute L_{n+1} and repeat steps (a), (b) and (c).

The mechanics of applying the test are easier if logarithms are used. For the case of systems M/M/1, M/E_k/1, E_k/M/s, M/M/s/s and the machine-interference problem, the logarithm of (54) takes the form

$$\log L_n = an + \sum_{i,j} n_{ij} c_{ij}, \tag{55}$$

where a and c_{ij} are constants depending upon ρ_0, ρ_1, and the transition probabilities of the imbedded Markov chain.

When the state space of $\{Q_n\}$ is finite, the operating characteristic (OC) function for the SPRT can be obtained as

$$\begin{aligned} L(\rho) &\cong \frac{A^{t_0(\rho)} - 1}{A^{t_0(\rho)} - B^{t_0(\rho)}} \qquad \text{if } t_0(\rho) \neq 0 \\ &\cong \frac{\log A}{\log A - \log B} \qquad \text{if } t_0(\rho) = 0, \end{aligned} \tag{56}$$

where $t_0(\rho)$ is the non-zero real root of the equation $\lambda_0(t) = 1$. Here, $\lambda_0(t)$ is the largest real positive latent root of the matrix

$$P(t) = \left\{ p_{ij}(\rho) \left[\frac{p_{ij}(\rho_1)}{p_{ij}(\rho_0)} \right]^t \right\}. \tag{57}$$

Note that $\lambda_0(t)$ is a function of ρ. The average sample number (ASN) can then be obtained as

$$\begin{aligned} E(n;\rho) &\cong \frac{L(\rho)\log B + \{1 - L(\rho)\}\log A}{\lambda'(0)}, \qquad \text{if } \lambda'(0) \neq 0, \\ &\cong \frac{L(\rho)(\log B)^2 + \{1 - L(\rho)\}(\log A)^2}{\lambda''(0)}, \quad \text{if } \lambda''(0) = 0. \end{aligned} \tag{58}$$

The ASN and OC functions have been numerically evaluated for the M/M/1/10 case, as shown in table 1.

Table 1

ASN and OC functions: M/M/1/10 queue $H_0 : \rho = 0.8, H_1 : \rho = 2.0$; $\alpha = 0.05, \beta = 0.10$

ρ	ASN $E(n; \rho)$	OC $L(\rho)$
0.8	10.950	0.95000
1.0	14.449	0.83138
1.2	17.929	0.63049
1.4	18.539	0.41720
1.6	17.862	0.26374
1.8	16.991	0.16065
2.0	15.840	0.10000

A limitation of the procedure adopted in the determination of OC and ASN functions is that the state space must be restricted to a finite number for the associated imbedded Markov chain. Further, the computation of ASN and OC functions for the SPRT is quite tedious. Simplifications, expecially by way of making use of the i.i.d. sequence $\{X_n\}$ of the number of customers arriving during successive service periods in the system M/G/1 as in Bhat [8], seem desirable.

(f) LARGE SAMPLE TESTS

Drawing upon the asymptotic tests for Markov processes developed by Billingsley [10], Wolff [53] gives a number of tests for the parameters in the M/M/1, M/M/∞, M/M/s loss and M/M/s systems, where the arrival and service rates are λ_n and μ_n when the number of customers in the system is n. For the M/M/s loss system, his approach is illustrated as follows. Suppose the queue is fully observed and let u_j be the number of upward transitions and d_j the number of downward transitions from state j, γ_j the total time spent in state j, and n the total number of observed transitions during $(0, T]$. The null hypothesis is stated as

$$H_0 : \theta = \theta^0, \quad \text{where} \quad \theta^0 = (\lambda_0^0, \lambda_1^0, \ldots, \lambda_{s-1}^0, \mu_1^0, \ldots, \mu_s^0). \tag{59}$$

Then the likelihood ratio statistics can be obtained as

$$\chi^2 = \sum_{j=0}^{s-1} u_j \log (u_j / \lambda_j^0 \gamma_j) + \sum_{j=1}^{s} d_j \log (d_j / \mu_j^0 \gamma_j)$$

$$+ \sum_{j=0}^{s} \gamma_j (\lambda_j^0 + \mu_j^0) - n. \tag{60}$$

Under the null hypothesis, (60) is asymptotically distributed as χ^2 with $2s$ degrees of freedom. The power of the test can be determined using the non-central χ^2 distribution.

When the system is not Markovian but if Markov chains can be identified in them, large sample tests can also be constructed using i.i.d. sequences of random variables responsible for the Markovian structure. Thus, in the imbedded Markov chain characterization of an M/G/1 queue, if X_n is the number of customers arriving during the nth service period, then $\{X_n\}$ is an i.i.d. sequence with $E(X_n) = \rho$, the traffic intensity. For large n, therefore, we may consider

$$Z_n = \frac{\bar{X} - \rho_0}{s/\sqrt{n}} \tag{61}$$

as being approximately standard normal for testing the hypothesis $\rho = \rho_0$. Note that s is the standard deviation of the sample. In general, this is a distribution free test. When the form of the service time distribution is known, better tests can be constructed. In particular, for the $M/E_k/1$ queue, Harishchandra and Rao [31] use the statistic

$$Z_n = \frac{\bar{X} - \rho_0}{\sqrt{\dfrac{\rho_0(\rho_0 + k)}{kn}}} \tag{62}$$

and compare the power of this test with the power of the exact UMP test given by eq. (50) for $n = 30$. An illustration is provided by table 2. Even for moderate sample sizes, the approximation by the large sample test appears to be good.

Table 2

Power of the UMP test and the large sample test for ρ. Queue M/M/1, $H_0: \rho = 0.8, \alpha = 0.05, n = 30$

ρ	0.80	0.85	0.90	1.00	1.20	1.40	1.60	1.80	2.00
UMP test	0.05	0.08076	0.12122	0.22880	0.49735	0.72998	0.87402	0.94605	0.97859
Large sample test	0.05	0.08693	0.13795	0.26764	0.55332	0.76296	0.88104	0.94065	0.96993

For testing $\rho = \rho_0$ in an M/G/1 queue using the i.i.d. sequence $\{X_n\}$, additional large sample tests have been suggested by Basawa [2].

In GI/G/1 queues, Basawa and Prabhu [3] have shown that the m.l.e.'s of the arrival and service time parameters given by eq. (12) have an asymptotic normal distribution [eq. (13)]. This property can be used to construct large same tests for appropriate parameters. Further work comparing these tests with other exact tests may also throw some light on where and when to use approximate tests.

6. Other related topics and future prospects

In this section we briefly mention some related topics which in our opinion indicate the type of problems that are significant in the future development of the subject area.

(1) In the use of queueing models, convenience of observation plays a major role. For example, as observed by Neal and Kuczura [43], traffic parameters in the Bell System are generally determined by obtaining the following three measurements during a time period $(0, t)$: (i) $A(t)$, the number of calls, (ii) $O(t)$, the number of unsuccessful calls (overflow), and (iii) $L_d(t)$, an estimate of usage based on discrete samples ($n = 36$). Using these functions, one can determine:

$$\text{call congestion} = O(t)/A(t); \quad \text{load } \hat{\alpha} = \frac{L_d(t)/n}{1 - O(t)/A(t)} .$$

Intuitively, we may identify call congestion as representing the probability of loss and load $\hat{\alpha}$ as representing the effective traffic intensity. These measures can be obtained from analytical models by determining arrival and service rates. However, it is convenient to get these quantities directly as shown above, instead of using the detailed information on the arrival and service processes. When such measurements are made, accuracy of the result as measured by the standard error of the sample function is of interest. This aspect has received considerable attention from several investigators; see Kosten et al. [36], Syski [51], p. 654, Beneš [6], Riordan [49], Descloux [24], and Neal and Kuczura [43]. In particular, Descloux considers the joint distribution of three processes in an M/M/s loss system: $N(t)$ = number of busy channels at time t and $A(t)$ and $O(t)$ as defined above, and derives cov$[A(t), O(t)]$ and the variance of the call congestion $O(t)/A(t)$. Comparing call congestion with the proportion of time when all servers are busy (time congestion), Descloux concludes that the standard deviation of call congestion is larger than that of the time congestion when the offered load is approximately less than the number of channels and the inequality is reversed otherwise.

(2) A problem closely related to measurements is that of sampling. Extensive results on sampling as applicable to simulation of stochastic systems are available; see Fishman [25] and comments made earlier in sect. 1.

(3) There are times when only the output process of the queue can be observed. A question that naturally arises is to what extent the output process provides information on the input process. This characterization problem is also important in the study of queueing networks where departures from one queueing node form arrivals into another. One approach to output analysis is provided by time series analysis techniques, but the input-output transfer functions used in time series are purely theoretical models and they may not have any resemblence to the actual process. In queueing theory, much more information is available on the input-output transfer structure and therefore, for the best use of time series analysis techniques, considerable work needs to be done in their adaption to queueing systems analysis.

(4) Applied queueing theory relies heavily on results from Markovian systems. The robustness of such results is therefore a significant factor. Queueing theorists have tackled this problem in an indirect manner by using approximating systems. When simpler systems are used to approximate more complex ones, validation is essential. At the present time, simulation and checking through some simple cases seems to be the more prevalent methods available for this purpose. In this case of experimentation, more sensitive analysis is needed. Wider use needs to be made of statistical techniques related to point and interval estimation. A study of system robustness using approximation systems does not tackle the problem at its roots. What is needed is the information on the impact of changing distribution assumptions for system elements.

(5) One of the major problems in gathering necessary data from queueing systems for statistical analysis is the inability to obtain complete information, either due to system structure or due to prohibitive costs. An example is the test for the exponentiality of the service time in the M/G/1 queue using waiting time data, as given by Thiagarajan and Harris [52]. The sequential parameter control technique discussed in subsect. 5(d) is aimed at using observations only on the queue length process. It is our belief that if analysis techniques are to be useful for the practitioner, they should make use of only observations that are easy to collect.

(6) In queueing theory, most of the resulting random variables are non-normal, while most of the standard tests are based on normal variates. Therefore, an area that needs exploration is the use of non-parametric tests. We may use simple tests such as the Kolmogorov–Smirnov test to see whether a queueing situation can in fact be represented by a specific queueing model with known properties. When steady-state distrbutions are characterized, this procedure can be easily applied. However, for more complex systems and hypotheses, we need to develop simple tests.

As queueing theory finds new application areas, new problems emerge. For example, during recent years queueing problems in computer communication systems have been a major area of queueing theory research. These are mainly network-related systems and consequently statistical analyses of queueing networks have become necessary (see Gaver and Lehoczky [27]).

Finally, we may also mention that new perspectives on statistical inference are also influencing research on inference on queueing systems. The subjective Bayesian

approach to the theory of queues initiated by McGrath and Singpurwalla [40] is in this spirit. In any case, statistical analysis of queueing systems is an area that brings together the practitioner and the theoretician with a common purpose.

Acknowledgement

The authors wish to thank the Editor, Professor N.U. Prabhu, for carefully going through the original version of the paper and making suggestions for its improvement.

References

[1] D.J. Aigner, Parameter estimation from cross-sectional observations on an elementary queueing system, Oper. Res. 22, 2(1974)422.
[2] I.V. Basawa, private communication (1985).
[3] I.V. Basawa and N.U. Prabhu, Estimation in single server queues, Naval Res. Log. Quart. 28(1981)475.
[4] I.V. Basawa and B.L.S. Prakasa Rao, *Statistical Inference for Stochastic Processes* (Academic Press, New York, 1980).
[5] V.E. Beneš, A sufficient set of statistics for a simple telephone exchange model, Bell Syst. Tech. J. 36(1957)939.
[6] V.E. Beneš, The covariance function of a simple trunk group with applications to traffic measurement, Bell. Syst. Tech. J. 40(1961)117.
[7] U.N. Bhat, *Elements of Applied Stochastic Processes*, 2nd ed. (Wiley, New York, 1984).
[8] U.N. Bhat, A sequential technique for the control of traffic intensity in Markovian queues, Ann. Oper. Res. 8(1987)151.
[9] U.N. Bhat and S.S. Rao, A statistical technique for the control of traffic intensity in queueing systems M/G/1 and GI/M/1, Oper. Res. 20(1972)955.
[10] P. Billingsley, *Statistical Inference for Markov Chains* (University of Chicago Press, Chicago, 1961).
[11] A. Birnbaum, Statistical methods for Poisson processes and exponential populations, J. Amer. Statis. Soc. 49(1954)254.
[12] N. Blomqvist, The covariance function of the M/G/1 queueing system, Skand. Aktuar. 50 (1967)157.
[13] N. Blomqvist, Estimation of waiting time parameters in the GI/G/1 queueing systems, Part I: General results, Skand. Aktuar. 51(1968)178.
[14] N. Blomqvist, Estimation of waiting time parameters in the GI/G/1 queueing systems, Part II: Heavy traffic approximations, Skand. Aktuar. 52(1969)125.
[15] J.V. Bradley, *Distribution-Free Statistical Tests* (Prentice Hall, Englewood Cliffs, N.J., 1968).
[16] A.B. Clarke, Maximum likelihood estimates in a simple queue, Ann. Math. Stat. 28(1957) 1036.
[17] W.J. Conover, *Practical Nonparametric Statistics* (Wiley, New York, 1971).
[18] D.R. Cox, *Renewal Theory* (Methuen and Co., London, 1962).
[19] D.R. Cox, Some problems of statistical analysis connected with congestion, *Proc. Symp. on Congestion Theory*, ed. W.L. Smith and W.B. Wilkinson, University of North Carolina, Chapel Hill, N.C. (1965).

[20] D.R. Cox and P.A.W. Lewis, *The Statistical Analysis of Series of Events* (Meuthen and Co., London, 1966).
[21] B.D. Craven, Asymptotic correlation in a queue, J. Appl. Prob. 6(1969)573.
[22] D.J. Daley, Monte Carlo estimation of mean queue size in a stationary GI/M/1 queue, Oper. Res. 16(1968)1002.
[23] D.J. Daley, The serial correlation coefficients of waiting times in a stationary single server queue, J. Austr. Math. Soc. 8(1968)683.
[24] A. Descloux, On the accuracy of loss estimates, Bell Syst. Tech. J. 44, 6(1965)1139.
[25] G.S. Fishman, *Principles of Discrete Event Simulation* (Wiley, New York, 1978).
[26] A.V. Gafarian and C.J. Ancker, Mean value estimation from digital computer simulations, Oper. Res. 14(1966)25.
[27] D.P. Gaver and J.P. Lehoczky, Random parameter Markov population process models and their likelihood, Bayes, and empirical Bayes analysis, NPS55-85-020, Monterey, California: Naval Postgraduate School (1985).
[28] R.F. Gebhard, A limiting distribution of an estimator of mean queue length, Oper. Res. 11 (1963)1000.
[29] B.V. Gnedenko, Y.K. Belyayev and A.D. Solovyev, *Mathematical Methods of Reliability Theory* (Academic Press, New York, 1969).
[30] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory,* 2nd ed. (Wiley, New York, 1985).
[31] K. Harishchandra and S.S. Rao, Statistical inference about the traffic intensity parameter of $M/E_k/1$ and $E_k/M/1$ queues, Report, Indian Institute of Management, Bangalore, India (1984).
[32] C.M. Harris, Some new results in the statistical analysis of queues, *Proc. Conf. on Math. Methods in Queueing Theory,* ed. A.B. Clarke, Vol. 98, Lecture Notes in Economics and Mathematical Systems (Springer-Verlag, New York, 1974) p. 157.
[33] M. Jacobsen, *Statistical Analysis of Counting Processes* (Springer-Verlag, New York, 1982).
[34] J.H. Jenkins, The relative efficiency of direct and maximum likelihood estimates of mean waiting time in the simple queue M/M/1, J. Appl. Prob. 9(1972)396.
[35] D.G. Kendall, Stochastic processes occurring in the theory of queues and their analysis by the method of imbedded Markov chains, Ann. Math. Stat. 24(1953)338.
[36] L. Kosten, J.R. Manning and F. Garwood, On the accuracy of measurements of probabilities of loss in telephone systems, J. Roy. Statis. Soc. B11(1949)54.
[37] J.A. Koziol and A.F. Nemec, On a Cramér-von Mises type statistic for testing bivariate independence, Can. J. Stat. 7(1979)43.
[38] P.A.W. Lewis, Recent results in the statistical analysis of univariate point processes, in: *Stochastic Point Processes,* ed. P.A.W. Lewis (Wiley, New York, 1972) p. 1.
[39] H.W. Lilliefors, Some confidence intervals for queues, Oper. Res. 14(1966)723.
[40] M.F. McGrath and N.D. Singpurwalla, A subjective Bayesian approach to the theory of queues: Part I – Modeling, Part II – Inference and information, GWU/IRRA/Serial TR-85/14, 15, The George Washington University, Washington D.C. (1985).
[41] P.W. Mielke, Jr., A note on some squared rank tests with existing ties, Technometrics 9 (1967)312.
[42] S.C. Moore, Approximate techniques for non-stationary queues, Ph.D. Dissertation, Computer Science/Operations Research Center, Southern Methodist University, Dallas, Texas (1972).
[43] S.R. Neal and A. Kuczura, A theory of traffic-mesurement errors for loss systems with renewal input, *Proc. Conf. on Math. Methods in Queueing Theory,* ed. A.B. Clarke, Vol. 98, Lecture Notes in Economics and Mathematical Systems (Springer-Verlag, New York, 1974) p. 199.

[44] J. Putter, The treatment of ties in some nonparametric tests, Ann. Math. Stat. 26(1955)268.
[45] R.H. Randles and D.A. Wolffe, *Introduction to the Theory of Non-parametric Statistics* (Wiley, New York, 1979).
[46] S.S. Rao, U.N. Bhat and K. Harishchandra, Control of traffic intensity in a queue – A method based on SPRT, Opsearch 21(1984)63.
[47] J.F. Reynolds, Asymptotic properties of mean length estimators for finite Markov queue, Oper. Res. 20, 1(1972)52.
[48] J.F. Reynolds, The covariance structure of queues and related processes: A survey of recent work, Adv. Appl. Prob. 7(1975)383.
[49] J. Riordan, *Stochastic Service Systems* (Wiley, New York, 1962).
[50] D.A. Stanford, B. Pagurek and C.W. Woodside, Optimal prediction of times and queue lengths in the GI/M/1 queue, Oper. Res. 31(1983)322.
[51] R. Syski, *Introduction to Congenstion Theory in Telephone Systems* (Oliver and Boyd, London, 1960).
[52] T.R. Thiagarajan and C.M. Harris, Statistical tests for exponential service from M/G/1 waiting-time data, Naval Res. Log. Quart. 26, 3(1979)511.
[53] R.W. Wolff, Problems of statistical inference for birth-and-death queueing models, Oper. Res. 13(1965)343.
[54] C.M. Woodside, D.A. Stanford and B. Pagurek, Optimal prediction of queue lengths and delays in GI/M/m multiserver queues, Oper. Res. 32(1984)809.
[55] P.J. Burke, The output of a queueing system, Oper. Res. 4(1956)699.
[56] E. Reich, Waiting times when queues are in tandem, Ann. Math. Stat. 28(1957)768.