Contributed paper

A SUBJECTIVE BAYESIAN APPROACH TO THE THEORY OF QUEUES I – MODELING

Michael F. MCGRATH

Office of the Secretary of Defense
Washington, DC 20301, U.S.A.

Donald GROSS
Nozer D. SINGPURWALLA*

Department of Operations Research
*and Department of Statistics**
The George Washington University
Washington, DC 20052, U.S.A.

Received 21 November 1985
(Revised 25 August 1986)

Abstract

This is the first of an expository two-part paper which outlines a point of view different from that currently used in queueing theory. In both parts, the focus is on concepts. Here we adopt a personal probability point of view to all sources of uncertainty in the theory of queues and explore the consequences of our approach by comparing our results to those that are currently available. For ease of exposition, we confine attention to the $M/M/1/\infty$ and the $M/M/1/K$ queues. In Part I we outline the general strategy and focus on model development. In Part II we address the problem of inference in queues within the subjective Bayesian paradigm and introduce a use of Shannon's measure of information for assessing the amount of information conveyed by the various types of data from queues.

Keywords

Queueing theory, Bayesian analysis of queues, Shannon information, subjective probability, discrete DFR distributions.

1. Introduction

The theory of queues deals with the problem of making statements of uncertainty about future events in waiting lines. The models used in the theory consist of deterministic and stochastic components. It appears that, from the inception of the theory and to date, the stochastic components have been viewed as quantifying the variability of observed interarrival and service times rather than as quantifying the uncertain nature of the times yet to be observed. That is, the distributions of the times between arrivals and the times to complete services have been viewed as being *objective* with a fixed but unknown value for the underlying parameters. This means that the distributions in question can be realized via an indefinite repetition of the interarrival and service processes. Such a view has influenced the direction in which the subject has evolved, with respect to both the development of models and the attitude towards inference. The latter has subscribed to the sample-theoretic paradigm and its accompanying paraphernalia of point estimates, confidence intervals, tests by hypothesis, etc., for the unknown vector of parameters, say θ. Other issues of general interest, such as the informative or non-informative nature of various types of data, the role of stopping rules, the amount of information provided by the data, etc., appear not to have been carefully addressed.

Philosophical objections to, and practical difficulties with, adopting an objective view of probability have been well voiced in the literature. A seminal source of this is the two-volume work on probability theory by de Finetti [9]. These objections and criticisms apply equally to the existing results in queueing theory.

In this paper, we adopt a *subjective* point of view of the theory of queues according to which we treat all the stochastic elements in the theory of queues as expressions of the analyst's betting behavior. That is, any probability statement corresponds to the odds that one is willing to place on the outcome of an unknown quantity (event or parameter). That this point of view can lead to results that are quite different from those currently available, will be apparent from a reading of the subsequent text.

Familiarity with the terminology and results of queueing theory will be assumed; for convenience, we shall refer to these results as *classical queueing theory*. Since the main tenets of the subjective view towards the quantification of uncertainty may not be familiar to the intended audience of this paper, a quick review is given in the section below, along with some notation and a criticism of some current approaches used in queueing theory. In sect. 3 we shall give the specific objectives and an overview of this paper.

2. The subjective probability paradigm

There are three main tenets of the subjective approach towards the treatment of uncertainty. The first of these is that "probability is the only satisfactory way of quantifying uncertainty" (Lindley [14]). This means that notions such as point estimates, confidence limits, tests of hypotheses, etc., which are used in connection with inference in classical queueing theory are, according to this paradigm, untenable. The second tenet states that all probability statements must *always be conditional*, conditioned on $\mathcal{H}$, the background information available to the assessor of uncertainty, at the time the probability statement is made. The role of data, or new information, is to expand the knowledge base $\mathcal{H}$, and thus data should cause us to revise (update) our previous statement of uncertainty. The third tenet of subjective probability states that the revision of uncertainty statements should be done using the calculus of probability alone, of which Bayes' Rule is the appropriate vehicle. It is because of this tenet that we have used the term "subjective Bayesian" in the title of this paper.

Typically, the dimensions of $\mathcal{H}$ are very large because $\mathcal{H}$ could conceivably include everything that we know. This could make the initial specification and the subsequent updating of our probability statements a difficult proposition. To simplify this task, abstract quantities called *parameters* are introduced in our specification of the probabilities. This introduction of parameters is facilitated by the calculus of probability, via what has sometimes been referred to as "the law of the extension of conversation" (cf. Lindley [15]), which is mathematically equivalent to the law of total probability. The parameters have dimensions much smaller than $\mathcal{H}$, and a simplification of the probability statement is achieved via an assumption of independence, which defacto justifies the replacing of $\mathcal{H}$ by the parameter(s) $\boldsymbol{\theta}$. Since a specification of probability models is subjective, the models may vary from individual to individual, and thus θ need not have a unique fixed value (as is assumed in classical queueing theory). Furthermore, since $\boldsymbol{\theta}$ is an unknown quantity, its uncertainty must be described via probability. Contrast this to the strategy taken in classical queueing theory in which uncertainty about $\boldsymbol{\theta}$ is expressed via point estimates and confidence limits, typically obtained by the method of maximum likelihood and often involving an appeal to a large sample theory, and the restrictive assumption that the queueing system is in a "steady state"; an exception is Basawa and Prahbu [6]. Note that for some queueing models (e.g. the $M/M/1/\infty$) inference about θ under steady state is a circular proposition, because such an assertion can only be made when we have precise knowledge about $\boldsymbol{\theta}$. From a subjective Bayesian point of view, the assumption of a steady state violates Cromwell's Rule (Lindley [15]) which states that no uncertain quantity shall be assigned, a priori, a probability of 0 or 1. Furthermore, no amount of data can lead us to claim a steady state with probability 0 or 1. Finally, inference under steady state, even under the sample theoretic paradigm implies an approximation which becomes exact only when the data used for inference is obtained from a queueing system which has been in operation for an infinite length of time.

3. Objectives and overview

Notwithstanding the philosophical differences in basic attitudes between classical queueing theory and the point of view proposed here, a subjectivist may view the available results of queueing theory as being incomplete, they being conditional on θ, with the next step of averaging out with respect to the uncertainty of θ being omitted. Thus a *first objective* of this paper is to complete the specification of uncertainty about the various measures of performance of queueing systems by averaging out the uncertainty of θ via its distribution. This is done in sect. 5, where for simplicity only the $M/M/1/\infty$ and the $M/M/1/K$ models of classical queueing theory are considered. It may be useful to remark that a lack of appreciation of the basic differences in philosophical attitudes between objective and subjective probabilities may cause one to view the process of averaging out the uncertainty of θ as one which is just probabilistic mixing which merely leads to an enrichment of the existing models of queueing theory. Such a view is limited, and perhaps even incorrect, even though the net results would be the same.

We have stated before, that observations of the queueing process lead us to revise our previous assessments of uncertainty via Bayes' Rule. An intermediate, though fundamental, step in this process is also the revision of our statement of uncertainty about the abstract quantity θ – this process will be referred to as *Bayesian inference for* θ. The *second objective* of this paper is to develop an approach to inference about θ which does not depend on the assumption of steady state, which does not appeal to a large sample theory, and whose ultimate goal is to make revised statements of uncertainty about the measures of performance of queues in a logically consistent* (coherent) manner. In undertaking this objective, we will also be able to revise our statements of uncertainty about the attainment of a steady state in the light of new information (data). All of the above will be undertaken in Part II, sect. 6, where again, only the $M/M/1/\infty$ and the $M/M/1/K$ models will be considered. In sect. 6 we will also introduce and discuss the informative or non-informative nature of the various types of data, and a use of the "Shannon Information Measure" for assessing the role and importance of certain data for inference about θ. This latter feature is motivated by a desire to determine the relative loss in information caused by ignoring certain data which are computationally burdensome to deal with in our inference regarding θ.

Finally, we would like to state at the outset, that the examples presented in this paper are admittedly narrow – they involve computer intensive numerical solutions, are restricted to a small class of models of queueing theory, and focus on limited issues within this class. However, since the goal of this paper is to advocate and

*It is by now well known, that interval estimates based on the notion of confidence limits and other such sample theoretic arguments may lead to logical inconsistencies – see for example Cornfield [7] or Robinson [24].

explore the ramifications of a point of view different from that of classical queueing theory, the contributions of this paper should be judged in the light of its expository merit. It is possible (and quite likely) that the numerical work can be undertaken more efficiently and that it can be carried out to a higher level of completeness than we have attempted. The next steps in this research would be to undertake the above and also to extend the methodology we have outlined to other queueing models.

4. Terminology, notation and statement of some known results

To establish terminology and notation, and also to facilitate the development of the subsequent material, we present here some well known results from classical queueing theory pertaining to the $M/M/1/\infty$ and the $M/M/1/K$ queues.

The notation $M/M/1/\infty$ [$M/M/1/K$] denotes a single channel queueing system with interarrival and service times that are conditionally (given the parameters λ and μ, respectively) independent and exponentially distributed, and with unlimited [truncated at system size K] waiting room capacity. The parameters λ and μ are often referred to as the *arrival* and the *service* rates, respectively, and λ^{-1} and μ^{-1} are the means of the interarrival and service time distributions. Arguments which motivate the choice of an exponential distribution based on subjective considerations are given in Singpurwalla [26]. Let $N(t)$ denote the number of customers in the queueing system (including the one being served) at time t; $N(t)$ is referred to as the *system size*. Suppose that the limit probability

$$(p_n|\lambda, \mu) = \lim_{t \to \infty} P\{N(t) = n|\lambda, \mu\}$$

exists. Note that $(p_n|\lambda, \mu)$ is merely *conceptual* since it is defined as a limit. The queueing system is said to be in a *steady state* at any finite time t, if

$$P\{N(t) = n|\lambda, \mu\} = (p_n|\lambda, \mu);$$

when this happens $N(t)$ is said to be in statistical equilibrium. The aforementioned limit always exists when $K < \infty$, but for the $M/M/1/\infty$ model, it exists

$$\text{iff } \rho = (\lambda/\mu) < 1.$$

The parameter ρ is referred to as the *traffic intensity*, and the condition $\rho < 1$ is known as the *ergodic condition* for the $M/M/1/\infty$ queue. The following well known results (see for example Gross and Harris [11], p. 137) will be needed.

FOR THE M/M/1/∞ MODEL:

$$(p_{in}(t)|\lambda,\mu) = P\{N(t) = n|N(0) = i,\lambda,\mu\}$$

$$= e^{-(1+\rho)\mu t}\Big[\rho^{-(i-n)/2} I_{n-i} + \rho^{-(i-n+1)/2} I_{n+i+1}$$

$$+ (1-\rho)\rho^{n} \sum_{\ell=n+i+2}^{\infty} \rho^{-\ell/2} I_{\ell}\Big], \tag{4.1}$$

where

$$I_n = \sum_{j=0}^{\infty} \frac{(\mu t\sqrt{\rho})^{n+2j}}{j!(n+j)!},$$

is a modified Bessel function of the first kind and $\rho = \lambda/\mu$.

$$(p_n|\rho) = \begin{cases} \rho^n(1-\rho), & \text{for all } n, \quad \text{if } \rho < 1 \\ 0, & \text{for all } n < \infty, \quad \text{if } \rho \geq 1. \end{cases} \tag{4.2}$$

That

$$\lim_{t\to\infty} (p_{in}(t)|\lambda,\mu) = (p_n|\rho) \quad [0]$$

when $\rho <$ [=] 1, has been verified by Gross and Harris [11], p. 137.

FOR THE M/M/1/K MODEL (Morse [22], p. 66):

$$(p_{in}(t)|\lambda,\mu) = (p_n|\rho) + \frac{2\rho^{(n-i)/2}}{K+1} \sum_{s=1}^{K} \frac{1}{\gamma_s}\left[\sin\frac{si\pi}{K+1} - \sqrt{\rho}\,\sin\frac{s(i+1)\pi}{K+1}\right]. \tag{4.1a}$$

$$\left[\sin\frac{sn\pi}{K+1} - \sqrt{\rho}\,\sin\frac{s(n+1)\pi}{K+1}\right] e^{-\gamma_s \mu t},$$

for $n = 0, 1, \ldots, K$, where $\gamma_s = 1 + \rho - 2\sqrt{\rho}\cos(s\pi/(K+1))$ and

$$(p_n|\rho) = \begin{cases} \dfrac{\rho^n(1-\rho)}{1-\rho^{k+1}}, & \text{for all } n \leq K, \quad \text{if } \rho \neq 1 \\ \\ \dfrac{1}{K+1}, & \text{for all } n \leq K, \quad \text{if } \rho = 1. \end{cases} \tag{4.2a}$$

The quantities $(p_{in}(t)|\lambda, \mu)$ and $(p_n|\rho)$ are often referred to as the *transient distribution* and the *steady state distribution* of system size, respectively.

5. Specification of uncertainty in the parameters and its consequences

With the notation of sect. 4, suppose that $\theta = (\lambda, \mu)$ and that our uncertainty about θ is described by $\pi(\theta|w)$ a prior distribution for θ given a vector of "hyperparameters", say w. We shall consider two possibilities, the first involving independence of λ and μ, given w, and the second involving their dependence. We are also able to model the case $\lambda < \mu$ (for the M/M/1/-queue, where ergodicity is an issue) with probability 1, but refrain from doing so because there is no physical nor intuitive basis for certainty regarding the relationship between these two abstract quantities in question. Accordingly, let $w = (\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma)$ where $\alpha_i, \beta_i > 0$, $(i = 1, 2)$, and the value of γ depends on the nature of dependence between λ and μ.

(a) THE INDEPENDENCE CASE

Suppose that $\gamma = 0$ and α_i $(i = 1, 2)$, is a positive integer. Then a reasonable strategy for describing the independence between λ and μ is via a product of two *Erlang distributions* with means α_i/β_i and variances α_i/β_i^2 $(i = 1, 2)$, respectively. The Erlang density is flexible enough to capture many subjective feelings about λ and μ that the analyst has, and α_i and β_i can be chosen to correspond to our best guesses about λ and μ, and our measure of uncertainty associated with these guesses. Thus the joint density at λ and μ may be written

$$\pi(\theta|w) = \exp\,(-\lambda\beta_1 - \mu\beta_2)\ \lambda^{\alpha_1 - 1}\ \mu^{\alpha_2 - 1} \prod_{i=1}^{2} \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}. \tag{5.1}$$

Since the ratio of two independent Erlang distributions can be transformed to have an F distribution, it follows that the density at ρ induced by (5.1) is of the form

$$\pi(\rho \mid w) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)} \left(\frac{\beta_1}{\beta_2}\right)^{\alpha_1} \frac{\rho^{\alpha_1 - 1}}{\left(1 + \left(\frac{\beta_1}{\beta_2}\right)\rho\right)^{\alpha_1 + \alpha_2}} \quad (\rho \geq 0). \tag{5.2}$$

Thus, with the assignment of (5.1) as a prior distribution for θ, the induced distribution for ρ would admit values of $\rho > 1$, even if $(\alpha_1/\beta_1) < (\alpha_2/\beta_2)$. This implies, a priori, a non-zero probability that the ergodic condition will *not* be met for the M/M/1/∞ queue. The a priori probability that the ergodic condition will be met is given by the F-distribution function based on the density (5.2), evaluated at 1. As pointed out by Armero [4], the probability that $\rho < 1$ is itself a very useful quantity that has no parallel in classical queueing theory.

(b) THE DEPENDENCE CASE

It may happen that in the subjective opinion of the analyst, the arrival and the service processes bear a qualitative relationship to each other. Such an opinion might be prompted by the physics of the queueing process, wherein an increase in the arrival pattern, as measured by some criteria, may tend to cause an increase (or a decrease) in the service process. A reasonable strategy for modeling dependencies is via a Bivariate Normal Distribution (BVN) with a positive (negative) value for the coefficient of correlation; independence can be modeled as a special case by setting the correlation coefficient equal to 0. An approach for the subjective assessment of the correlation coefficients in the multivariate normal case has been given by Gokhale and Press [10]. Since λ and μ are positive quantities, a suitable approach for describing the stochastic relationship between them is via a Bivariate Lognormal Distribution (BVL). That is, $\log \lambda$ and $\log \mu$ are BVN with a parameter vector w whose elements α_i and β_i $(i = 1, 2)$ represent the location and scale parameters, respectively, and γ the correlation coefficient. Since the functional form of the BVL is well known (see for example Aitchison and Brown [3], p. 11), we may denote the prior density at θ by

$$\pi(\theta \mid w) = \mathrm{BVL}(\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma). \tag{5.3}$$

In addition to the flexibility in modeling any subjective opinions about λ and μ and their dependence, the BVL model for θ has a technical advantage, in the sense that the induced distribution for ρ is also a lognormal with location $\alpha = \alpha_1 - \alpha_2$ and scale $\beta = \beta_1^2 + \beta_2^2 - \gamma\beta_1\beta_2$ (see, for example, Anderson [2], p. 19). That is, the prior density at ρ can be written

$$\pi(\rho \mid w) = \frac{1}{\sqrt{2\pi}\;\beta\rho} \exp\left\{-\frac{1}{2}\left(\frac{\log\,(\rho - \alpha)}{\beta}\right)^2\right\} \quad (\rho \geq 0). \tag{5.4}$$

The discussion following eq. (5.2) applies here also, with appropriate modifications.

An approach for the subjective elicitation of the parameters α_i and β_i $(i = 1, 2)$ based on expert opinions and a modulation of such opinions based on the perceived expertise of the experts is given by Lindley and Singpurwalla [16] and also by Singpurwalla and Song [17]. Such an approach can be easily implemented for the scenario considered here.

(c) EFFECTS OF UNCERTAINTY IN PARAMETERS ON THE ASSUMED DISTRIBUTIONS

Recall that for the M/M/1 queue, *given* $\lambda > 0$ and $\mu > 0$, the interarrival and service times are independent and identically exponentially distributed with densities denoted by $a(t|\lambda) = \lambda\, e^{-\lambda t}$ and $s(t|\mu) = \mu e^{-\mu t}$, respectively. The assignment of (5.1) for $\pi(\theta|w)$ implies that, given w, the marginal distributions of λ and μ are Erlang, so that for $(\alpha_1, \beta_1) \in w$, the density at t for any *single* interarrival time is

$$a(t|\alpha_1, \beta_1) = \int_0^\infty a(t|\lambda) \frac{\beta_1^{\alpha_1}\ \lambda^{\alpha_1 - 1}\ e^{-\beta_1 \lambda}}{\Gamma(\alpha_1)}\, d\lambda$$

$$= \alpha_1\ \beta_1^{\alpha_1}/(t + \beta_1)^{\alpha_1 + 1}\,,$$

a Pareto (α_1, β_1) density, with mean $\beta_1/(\alpha_1 - 1)$. Similarly, $s(t|\alpha_2, \beta_2)$ the density at t for any single service time, given (α_2, β_2), is also a Pareto (α_2, β_2).

It is important to note that the above observation does *not* imply that what we have here is a set-up for a "Pareto/Pareto/1" queueing process. Such an assertion would be true if, given w, for any $n > 0$, the sequences $\{T_i\}$ and $\{S_i\}$ $(i = 1, \ldots, n)$ of interarrival and service times, respectively, are independent and identically distributed with Pareto marginals; in our case, as we shall soon see, the $\{T_i\}$ and $\{S_i\}$ constitute an *exchangeable sequence* (Kingman [12]). If desired, a Pareto/Pareto/1 queueing process can be constructed via what is known as an *Empirical Bayes* type construction (cf. Morris [21]). That is, we first generate $\{\lambda_i\}$ $(i = 1, \ldots, n)$ a sequence of independent variables from an Erlang distribution with parameters (α_1, β_1), and then for *each* λ_i we generate a variable A_i from an exponential distribution with parameter λ_i. Under the above construction, the independence of the λ_i's, given (α_1, β_1), ensures that the $\{A_i|\lambda_i\}$ $(i = 1, \ldots, n)$ is independent and non-identically exponentially distributed, whereas unconditionally the sequence $\{A_i\}$ $(i = 1, \ldots, n)$ is independent and identically distributed with a Pareto (α_1, β_1) distribution. A similar construction would enable us to generate the independent and identically Pareto (α_2, β_2) distributed sequence $\{S_i\}$ $(i = 1, \ldots, n)$. In the above construction, if a generation of the sequences $\{\lambda_i\}$ and $\{\mu_i\}$ $(i = 1, \ldots, n)$ can be physically motivated (as has been done by Harris (1974) for the μ_i's), then a proper Bayesian attitude to this situation would be to assign a prior distribution to w, and then operate within Deely and Lindley's [8] Bayes Empirical Bayes paradigm. We shall not further pursue this line of argument here.

Returning to our particular case, we note that the sequence $\{A_i|\lambda\}$ $(i = 1, \ldots, n)$ is independent and identically distributed with λ having an Erlang distribution with parameters (α_1, β_1). This makes the sequence $\{A_i\}$ $(i = 1, \ldots, n)$ exchangeable, with a joint density at $t_1, \ldots, t_n$ given by

$$a(t_1, \ldots, t_n|\alpha_1, \beta_1) = \int_0^\infty \prod_{i=1}^n \lambda e^{-\lambda t_i} \frac{\beta_1^{\alpha_1} \lambda^{\alpha_1 - 1} e^{-\beta_1 \lambda}}{\Gamma(\alpha_1)} d\lambda$$

$$= \frac{\beta_1^{\alpha_1}}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_1 + n)}{\sum_1^n t_i + \beta_1},$$

a *multivariate Lomax* distribution (Nayak [23]). If we let $A_i^* = A_i + \beta_1$, then the joint distribution of $A_1^*, \ldots, A_n^*$ would be the *multivariate Pareto* (Type I) of Mardia [18]. A similar argument leads us to an exchangeable distribution for the service times. Since exchangeable sequences are not necessarily independent – they constituting a weak form of dependence (Shaked [25]), we conclude that our set-up involving uncertainty in the parameters has led us to a consideration of *queues with exchangeable interarrival and service times*. The same conclusion would also be true under the assignment of (5.3) for $\pi(\theta|w)$ except that the marginal densities would no longer be Paretos.

(d) EFFECTS OF UNCERTAINTY IN PARAMETERS ON MEASURES OF PERFORMANCE

Our uncertainty about θ, as expressed via $\pi(\theta|w)$, influences our previous statements of uncertainty about system size at time t (recall that these were conditional on θ being known). If we average out the effect of θ, we have

$$(p_{in}(t)|w) = \int_\theta (p_{in}(t)|\theta)\ \pi(\theta|w)\ d\theta, \tag{5.5}$$

where $(p_{in}(t)|\theta)$ is given by (4.1) or (4.1a), depending on the queueing model considered. In either case, the evaluation of (5.5) will have to be undertaken numerically. A computer code which undertakes the numerical evaluation of (5.5) when $(p_{in}(t)|\theta)$ is given by (4.1a) is given in McGrath [20]. A similar exercise when $(p_{in}(t)|\theta)$ is given by (4.1) proves to be involved (from a computational point of view) and is therefore not given here.

If we let $t \to \infty$, extend the conversation to include λ and μ, we may consider,

$$\lim_{t \to \infty} (p_{in}(t)|w) = \lim_{t \to \infty} \int_{\theta} [(p_{in}(t) \mid (\lambda, \mu, \lambda < \mu, w))\ \pi(\lambda, \mu, \lambda < \mu|w)$$

$$+ (p_{in}(t) \mid (\lambda, \mu, \lambda \geq \mu, w))\ \pi(\lambda, \mu, \lambda \geq \mu|w)]\ d\theta \qquad (5.6)$$

$$= \int_0^{\infty} \frac{\rho^n(1-\rho)}{1-\rho^{K+1}}\ \pi(\rho|w)\ d\rho, \text{ for the M/M/1/}K \text{ case} \quad (0 \leq n \leq K). \qquad (5.6a)$$

For the M/M/1/∞ case, the last term in (5.6) drops off giving us the result

$$\lim_{t \to \infty} (p_{in}(t|w)) = \int_0^1 \rho^n(1-\rho)\ \pi^*(\rho|w)\ d\rho \qquad (0 \leq n < \infty), \qquad (5.6b)$$

where $\pi^*(\rho|w)$ is a truncated prior density on [0, 1).

(e) COMPARISON OF RESULTS WITH THOSE OF CLASSICAL QUEUEING THEORY

The M/M/1/K case. In fig. 5.1, we plot $(p_{in}(t)|\lambda, \mu)$, the transient distribution of system size for the *M/M/1/8* queue, using eq. (4.1a), with $i = 0$, $t = 2$, $\lambda = 9$ and $\mu = 10$, for $n = 0, 1, \ldots, 8$. An inspection of fig. 5.1 suggests that $(p_{in}(t)|\lambda, \mu)$ decays geometrically in n, like $(p_n|\rho)$, the steady state distribution of system size for the M/M/1/K queue – see eq. (4.2a), which for $\rho > (<)\ 1$ places most of its mass at $n = O(K)$. Thus, as is expected, the transient distribution of system size has a tendency to approach the steady state distribution.

By way of a comparison, in fig. 5.2, we show a plot of $(p_{in}(t)|w)$, as given by eq. (5.5) with $\pi(\theta|w)$ having the parameters $\alpha_1 = 9$, $\beta_1 = 1$, $\alpha_2 = 5$ and $\beta_2 = .5$, for $n = 0, 1, \ldots, 8$. This assignment makes $E(\lambda) = 9$ and $E(\mu) = 10$, a situation compatible with that of fig. 5.1. The computation of $(p_{in}(t)\,|\,\lambda, \mu)$ has to be undertaken numerically. An inspection of fig. 5.2 reveals its *U-shapeness* with an accumulation of probability mass towards both ends at $n = 0$ and $n = 8$. This can be explained in the light of the behavior of $(p_n\,|\,\rho)$ for the M/M/1/K queue, since the specified uncertainty about θ admits the possibility that ρ can be less than or greater than 1. The difference in transient distributions manifests itself in terms of differences in $E[N(t)]$, the expected system size, which, in the case of fig. 5.1(5.2), is 3.04(3.39).

The M/M/1/ case. For the M/M/1∞ queue, the computation of the transient distributions of system size via eq. (4.1) and eq. (5.5) with $(p_{in}(t)\,|\,\theta)$ replaced by (4.1), poses numerical difficulties. The steady state expressions are more tractable and so in this section we compare the corresponding steady state distributions under the assumption of an ergodic condition. That is, in the case of uncertain parameters we assign a probability zero to the event $\lambda > \mu$. Thus we seek to compare the quantities

$$(p_n\,|\,\rho, \rho < 1) = \rho^n(1 - \rho), \text{ and} \tag{5.7}$$

$$(p_n^*\,|\,w) = \int_0^1 (p_n\,|\,\rho, \rho < 1)\ \pi^*(\rho\,|\,w)\ \mathrm{d}\rho, \tag{5.8}$$

where $\pi^*(\rho\,|\,w)$ is $\pi(\rho\,|\,w)$ *truncated* at 1 – see eqs. (5.2) and (5.4).

A qualitative comparison of eqs. (5.7) and (5.8) is possible via the following arguments which are common in reliability theory. First, we note that (5.7) is a geometric distribution, a member of the class of distributions having a decreasing failure rate (DFR) – see Barlow and Proschan [5], p. 55. Next, we prove that (5.8), being a continuous mixture of discrete DFR distributions, is also DFR. Finally, in theorem 1 below, we state that the bound on the survival function of our discrete DFR distribution is given by the survival function of a geometric distribution.

Let

$$\bar{F}(n\,|\,w) = \sum_{n=m}^{\infty} p_n^*(w)$$

be the survival function of $p_n^*(w)$, and let

$$\bar{G}(m\,|\,\rho^*) = \sum_{n=m}^{\infty} (\rho^*)^n\ (1 - \rho^*)$$

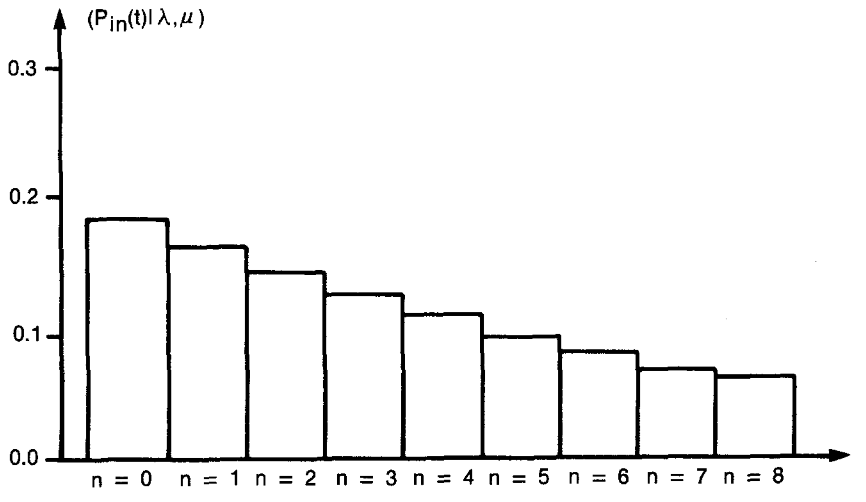

Fig. 5.1. A plot of the transient distribution of system size for the M/M/1/K queue with specified parameters

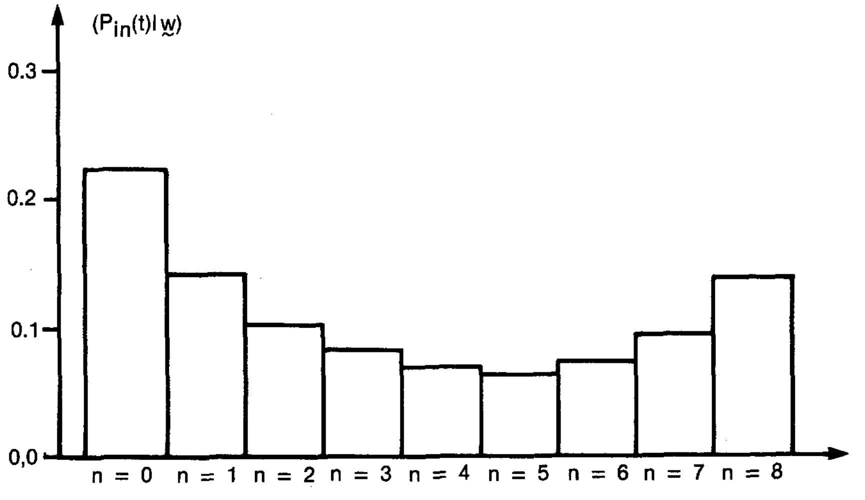

Fig. 5.2. A plot of the transient distribution of system size for the M/M/1/K queue with uncertain parameters

be the survival function of the geometric distribution $(p_n | \rho^*, \rho^* < 1)$, where $\rho^* = E(\rho)$ with respect to the distribution $\pi^*(\rho | w)$. Then

THEOREM 1

a) $\bar{F}(n|w) = \bar{G}(n|\rho^*) \quad (n = 0, 1)$,

b) $\bar{F}(n|w) \geq \bar{G}(n|\rho^*) \quad (n = 2, 3, \ldots)$.

To prove this theorem, we first need to state and prove the lemmas given below.

Let ζ_{D} be the class of discrete distributions that are DFR, and let ζ_{DE} be the set of all the extreme points of ζ_{D}. Then Langberg, Leon, Lynch and Proschan [13] show that ζ_{D} is a *closed convex set*.

DEFINITION 1

Let $F(n|\theta)$ the distribution function at n, of a discrete random variable N, be indexed by θ, and let $\pi(\theta)$ be the distribution function at θ of a continuous random variable Θ, $\Theta \in [a, b]$. Then $\int_a^b F(n|\theta)\, d\pi(\theta)$ is said to be a *continuous mixture* of $F(n)$.

LEMMA 1

A continuous mixture of discrete DFR distributions is DFR.

Proof of lemma 1

Let $a = x_0 < x_1 < \ldots x_r = b$ be a partition of $[a, b]$ by points of subdivision $x_0 < x_1 < \ldots < x_r$, and let ξ_r be an arbitrary point in $[x_{k-1}, x_k)$, $k = 1, \ldots, r$. Consider the convex combination of distribution functions

$$G_r(n) = \sum_{k=1}^{r} F(n|\xi_k)\, [\pi(x_k) - \pi(x_{k-1})], \qquad \text{and let}$$

$$\Delta = \max\left\{[x_1 - x_0), [x_2 - x_1), \ldots, [x_r - x_{r-1})\right\}.$$

For any r and Δ, $G_r(n)$ is a convex combination of the members of ζ_D and hence $G_r(n) \in \zeta_D$. By definition of the Lebesque-Stieltjes integral

$$\lim_{\Delta \to 0} G_r(n) = G(n) = \int_a^b F(n|\theta)\, d\pi(\theta).$$

Since ζ_D is a closed convex set $G(n) \in \zeta_D$.

DEFINITION 2 (Langberg et al. [13])

A discrete distribution function F is DFR if $\bar{F}^2(n+1) \le \bar{F}(n)\,\bar{F}(n+2)$, for all n; $\bar{F} = 1 - F$.

DEFINITION 3 (Marlow [19], p. 363)

A sequence $f(k)$ is a *convex sequence* if $f(k+1) \le (f(k) + f(k+2))/2$.

LEMMA 2

If F is a discrete DFR distribution, then $\log \bar{F}(n)$ is a convex sequence.

Proof of lemma 2

Follows from definitions 2 and 3.

Proof of theorem 1

a) By definition $\bar{F}(m|w) = 1 - F(m-1|w) = \sum_{n=0}^{m-1} p_n^*(w)$,

$$\text{and so } \bar{F}(1|w) = 1 - \int_0^1 (1-\rho)\ \pi^*(\rho|w),$$

$$= \rho^* = \bar{G}(1|\rho^*).$$

b) Suppress w and let $\bar{F}(n) = e^{-R(n)}$ so that $R(n) = -\log \bar{F}(n)$, which by lemma A5 is concave in n. The proof follows.

The implication of theorem 1 is that the steady-state result of the classical $M/M/1\infty$ queue underestimates the expected system size, in the face of uncertainty in the parameters; this matches our intuition.

For a discussion on possible extensions of the work presented here and some concluding comments, we refer the reader to sect. 7 of Part II.

References

[1] M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1970).
[2] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. (Wiley, New York, 1984).
[3] J. Aitchison and J.A.C. Brown, *The Lognormal Distribution* (Cambridge University Press, Cambridge, 1963).
[4] C. Armero, Bayesian Analysis of $M/M/1/\infty$ FIFO Queues, in: *Bayesian Statistics, 2*, Eds. J.M. Bernardo et al. (North-Holland, Amsterdam, 1985)613.
[5] R.E. Barlow and F. Proschan, *Mathematical Theory of Reliability* (Wiley, New York, 1965).
[6] I.V. Basawa and N.U. Prabhu, Estimation in single server queues, Naval Research Logistics Quarterly 28(1981)475.
[7] J. Cornfield, The Bayesian outlook and its application, Biometrics 25(1969)617.
[8] J.J. Deely and D.V. Lindley, Bayes empirical Bayes, Journal of the American Statistical Association 76(1981)833.
[9] B. de Finetti, *Theory of Probability*, Vols. 1 and 2 (Wiley, London, 1974).
[10] D.V. Gokhale and S.J. Press, Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution, Journal of the Royal Statistical Society, A, 145, Part 2(1982)237.
[11] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*, 2nd ed. (Wiley, New York, 1985).
[12] J.F.C. Kingman, Uses of exchangeability, The Annals of Probability 6(1978)183.
[13] N.A. Langberg, R.V. Leon, J. Lynch and F. Proschan, Extreme points of the class of discrete decreasing failure rate distributions, Mathematics of Operations Research, 5(1980)35.
[14] D.V. Lindley, Scoring rules and the inevitability of probability, International Statistical Review 50(1982)1.
[15] D.V. Lindley, The Bayesian approach to statistics, in: *Some Recent Advances in Statistics*, Eds. J. Tiago de Oliviera and B. Epstein (Academic Press, London, 1982).
[16] D.V. Lindley and N.D. Singpurwalla, Reliability (and fault tree) analysis using expert opinion, Journal of the American Statistical Association 81, 393(1986)87.
[17] N.D. Singpurwalla and M.S. Song, The analysis of Weibull Lifetime Data Incorporating Expert Opinion, 1986, Proc. Symp. on Probability and Bayesian Statistics in Honor of de Finetti, ed. R. Viertl (Plenum, New York) to appear.
[18] K.V. Mardia, Multivariate Pareto distributions, Annals of Mathematical Statistics 33(1962) 1008.

[19] W.H. Marlow, *Mathematics for Operations Research* (Wiley, New York, 1978).
[20] M.F. McGrath, A subjective Bayesian approach to queueing theory, Doctoral Dissertation, The George Washington University (1985).
[21] C. Morris, Parametric empirical Bayes inference: Theory and applications, Journal of the American Statistical Association 78(1983)47.
[22] P.M. Morse, *Queues, Inventories and Maintenance* (Wiley, New York, 1958).
[23] T.K. Nayak, Multivariate Lomax distributions: Properties and usefulness in reliability, The George Washington University, Institute for Reliability and Risk Analysis, Technical Report GWU/IRRA/Serial TR-85/7 (1985).
[24] G.K. Robinson, Some counterexamples to the theory of confidence intervals, Biometrika 62, 1(1975)155.
[25] M. Shaked, A concept of positive dependence for exchangeable random variables, Annals of Statistics 5(1977)505.
[26] N.D. Singpurwalla, Foundational issues in reliability and risk analysis, SIAM Review (1987), to appear.