CARLETON UNIVERSITY

SCHOOL OF

MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Bayesian Sample Size for M/M/n/n queuing model

AUTHOR: Lixin Zhou

SUPERVISOR: Yiqiang Zhao

DATE:2021/08/23

# Abstract

The wide application of queuing models in the real world fully proves that the in-depth study of queuing models is reasonable, and facts have proved that these models also play a considerable role in fields other than queuing theory. In previous studies, parameters were often estimated based on statistical data, without discussing the planning of the sample size to ensure the accuracy of the estimated parameters. Therefore, the next thing to do in this article is to discuss the determination of the sample size of the M/M/c/c model. In this article, a Bayesian method is described for sample size determination for traffic intensity estimation, one of the most important parameters for performance evaluation of queueing systems.



**Keywords:** *queuing theory; approximate Bayesian computation; parameter estimation; posterior density*


# Introduce

Queuing theory originated from the study of telephone communication queuing wiring. As early as 1909, the Danish mathematician AK Erlang published The Theory of Probabilities and Telephone Conversations, which initially developed the phenomenon of

unstable queues due to the emergence of random demand. Research. In his later work, he found several important conclusions: Automatic telephone communication system can be simulated with two basic probability models: 1. Poisson input, exponentially distributed service time, multiple service flows 2. Poisson input, stable normal state Service time, single service flow. Erlang also proposed the concept of steady-state balance of queues and preliminary optimization methods for queuing systems. After Erlang, many scholars further derived their work: Thornton Fry: Probabilities and Its Engineering Uses, Felix Pollaczek further sorted out the single/multiple service flow model of Poisson input and generalized output. At the same time, Russian mathematicians Kolmogorov, Khintchine Also get involved in this field. Queuing theory originated from the study of actual phenomena, and nearly half a century later, queuing theory was mainly the development of theoretical models (birth and death theory, embedded Markov model). It was not until after World War II that scholars began to give the theory application value, and a large number of studies began to guide how to accurately solve the complex mathematical models left by previous scholars and directly apply them to realistic management decisions. Mainly such as complex queuing model, approximate solution of queuing network and numerical simulation method. Modern queuing theory mainly provides theoretical and simulation support for the development of management decision-making software.

The queuing model has been widely used in the real world. For example, the exit of the supermarket queues up for settlement, the bank waits for the window to handle the business, and the production line is stagnant and waiting due to the uneven tempo.

Among all queuing models, the standard M/M/1 queuing model is the most basic. In view of its characteristics, there are many areas where this model can be used in real life. However, for some special situations, the standard M/M/1 is not well applicable. Therefore, this article attempts to study a variant model of the standard M/M/1 queuing model, the M/M/c/c model. This model means that when c service desks are occupied, customers leave automatically. Therefore, this model is also called the loss-based queuing model. In actual production and construction, we often conduct statistical analysis based on past service conditions to obtain suitable service plans. In our mathematics, this is to estimate the parameters based on the collected samples. And only when we have the value of the parameter, we can calculate the limit probability to achieve higher efficiency. In many previous studies, only the parameters were estimated, and only a few estimated the sample size (such as Quinino, RC, & Cruz, FRB (2016), Dai, Tianyi. (2018), Choudhury, Amit., & Basak , Arpita.(2017).etc). Therefore, in order to ensure the accuracy of parameter estimation, a method for determining the sample size is proposed. Moreover, as we all know, sample size plays an important role in statistical testing. In general, the larger the sample size, the more accurate the parameter value estimation. However, studies have shown that if the sample size is larger than the minimum required sample size, it may cause a waste of scarce resources and expose more resources to any related risks. Therefore, the determination of the appropriate sample size is very important in statistics.

Generally, there are several ways to obtain sample size. In this article, we will study how to use Bayesian methods to determine the sample size for estimating the variable

parameters of the M/M/c/c queuing model. The organization of the article is as follows. Part 2 will explain queuing theory and queuing models. The third part will describe the Bayesian method, how the Bayesian method estimates the sample size of the standard M/M/1 model, and the purpose of this paper to estimate the M/M/c/c sample size. Finally, Part 4 makes a simple analysis and conclusion, the Bayesian method estimates the sample size of the M/M/c/c model

# Queuing theory

In most cases, the following six basic attributes can describe a queue phenomenon. (1) The distribution of the arrival of customers; (2) The service situation of the service desk ;(3) The queuing discipline; (4) The capacity of the system ;(5) The number of service desks ;(6) The number of service processes. Here I would like to make some notes on some of the above terms in order to understand arrival distribution of the customers: Customers here can refer to entity customers, such as those waiting in line at the bank, or machines waiting to be scheduled for maintenance. Arrival distribution refers to how to use probability models to characterize the arrival time of two consecutive customers interval. The service situation of the service desk: the time distribution of the bank window service for different businesses, the distribution of the time spent in each process in the production line, etc. As for the queuing principle, the system capacity, the number of service desks and service processes are all well understood.

**Customer arrival distribution:**

In most queues, the arrival of customers is random (non-random phenomena such as a perfectly balanced production line). Therefore, a probability model is needed to characterize the arrival time interval of two consecutive customers. Common models such as exponential distribution. For why the exponential distribution and Poisson process can satisfaction random phenomena, see section 2.3. At the same time, some special customer performances also need to be referred to. For example: if a customer encounters a long queue, he may refuse to join (balked); existing queue customers leave due to long queue time (impatience, reneged). When multiple service flows occur, customers will flow between different queues, resulting in a theoretically perfect multi-service flow (jockey) of equal length. Finally, if the probability distribution of customer arrival time does not change with time, it is called a stationary arrival distribution (stationary), and vice versa is non-stationary.

**Service status of the service desk:**

The service desk is also called the service organization. The service time of different service desks can be represented by a probability model. For example, the time distribution of each doctor's diagnosis and treatment time in the hospital emergency room, and the processing time distribution of each machine waiting to be repaired in the repair shop. The following special phenomena also need special consideration: The length of service time is related to the length of the queue. This is well understood. When a teller sees more customers, it will naturally speed up business processing speed (state-

dependent). Similarly, service time It can also change over time (maintenance workers gradually accumulate experience), stationary and non-stationary appear.

**Queuing discipline:**

The most common one is first come first service (FCFS). There will also be a last come first service principle (LCFS). For example, when picking up goods from a warehouse, the goods that arrive later are often picked up first because they are stacked on the outermost periphery. Random service principles, such as license plate number lottery, do not follow any first-come-first-served principle.Priority queuing principle: If soldiers arrive, they can be directly ranked first in the queue (priority discipline).

**System capacity, number of service desks, number of service processes:**

These three brothers are relatively easy to understand. The capacity of the system: the train station queuing hall can accommodate up to 1,000 people. The number of service desks: a bank has 12 windows, while the suburban branches have only 4, and the number of service processes (which will form a complex queuing network ): There are 10 items in the order of physical examination items. A production line has 24 procedures. In order to eliminate queues, the production line needs to be balanced.

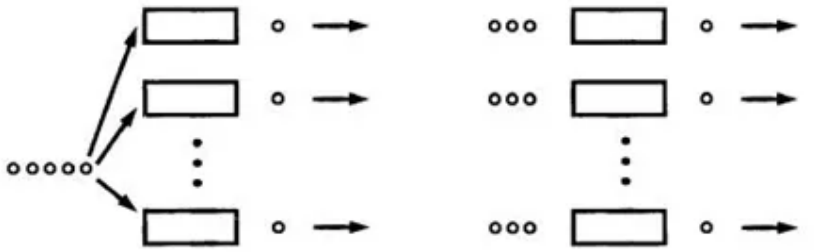

Figure . A multi-service desk, multi-se

**Common Terms and Notation:**

For most common models, we assume the distribution of customer arrival time, and the distribution of service desk service time is independent and identically distributed. The general queue model expression is as follows:

$$A/B/X/Y/Z$$

among which: *A: Probability distribution of customer arrival*

*B: Service time distribution*

*X: Number of service flows*

*Y: System capacity limit*

*Z: Queuing mechanism*

For most common models, we assume the distribution of customer arrival time, and the distribution of service desk service time is independent and identically distributed. The general queue model expression is as follows:

| $Characteristic$ | $Symbol$ | $Explanation$ |
|---|---|---|
| $Interarrival-time\ distribution(A)$ | $M$ | $Exponential$ |
| | $D$ | $Deterministic$ |
| $Service-time\ distribution(B)$ | $E_K$ | $Erlang\ type\ k(k=1,2,\dots)$ |
| | $H_k$ | $Mixture\ of\ k\ exponentials$ |
| | $PH$ | $Phase\ type$ |
| | $G$ | $General$ |
| #of parallel servers (X) | $1,2,\dots,\propto$ | |
| Max.system capacity (Y) | $1,2,\dots,\propto$ | |
| Queue discipline (Z) | $FCFS$ | $First\ come\ ,first\ served$ |
| | $LCFS$ | $Last\ come,first\ served$ |
| | $RSS$ | $Random\ selection\ for\ service$ |
| | $PR$ | $Priority$ |
| | $GD$ | $General\ discipline$ |

Table . General queue model expression

It is worth noting that we will often omit the last two positions, and the default is that the system has no upper limit and the principle of first come, first served. Such as the very classic and basic M/M/1 and M/M/S models (customer arrival and service time are both exponentially distributed, with one or more service desks).

Here are some commonly used variables:

| $\frac{Variables}{termes}$ | $annotation$ |
|---|---|
| $State\ of\ system$ | Number of customers in the queuing system |

| | |
|---|---|
| $length\ of\ queue$ | The number of customers waiting in the queue / the total number of customers in the system minus the customer tree being served |
| $N(t)$ | Number of customers in the system at time t |
| $P_n(t)$ | The probability that there are n customers in the system at time t |
| $S$ | Number of parallel service desks |
| $\lambda_n$ | The average customer arrival rate when there are n customers in the system (the number of customer arrivals per unit time) |
| $\mu_n$ | The average service rate when there are n customers in the system (the number of service completions per unit time) |
| $\lambda, \mu, \rho$ | Notes see below |

Table: Some constant variables and their

$If\ \lambda_n\ does\ not\ change\ with\ n, it\ is\ directly\ represented\ by\ \lambda. \mu_n\ is\ the\ same\ as\ before.$
$About\ the\ definition\ of\ \rho$: $\rho = \frac{\lambda}{\mu}$

Utilization Factor is the ratio of the customer arrival rate of a queuing system to a total of multiple parallel service rates. It is well understood that if this value is greater than 1, the queue will grow indefinitely, and if this value is less than 1, the queue system will gradually become steady-state from the initial transient condition. About this The discussion of state transition and transition is not discussed in this article. The various interesting conclusions presented by the author here are all based on the steady-state queue:

| | |
|---|---|
| $P_n$ | Probability of n customers in the system |
| $L$ | Average number of customers in the system $=\sum_{n=0}^{\infty} n * P_n$ |
| $L_q$ | Queue length $=\sum_{n=s}^{\infty}(n-s) * P_n$ |
| $W$ | The average waiting time of each fast college entrance examination |
| $W_q$ | Average time per customer in the queue |

Table: Some commonly used variables and comments

Little equation:

Assume that $\lambda_n$ is constant for any $n$. John D.C.Little proved to us:

$$L = \lambda W$$

The address of this OR top issue can be found in the last section of this article. Simultaneously:

$$L_q = \lambda W$$

$$W = W_q + \frac{1}{\mu}$$

The above is the general formula for steady state queue.

**2.1 The M/M/1 queue**

M represents the index (Markov), and 1 represents one server. In addition, we also introduce queuing (queuing system) capacity, queuing rules and independence assumptions in the M/M/1 model. M/M/1 queue is an abbreviation of M/M/1/∞/∞/FIFO, where the first "∞" means unlimited queue capacity, the second "∞" means unlimited overall size, and FIFO means first First come first out service. Among them, the queue capacity in the M/M/1 model is the first ∞ ("Introduction to Queuing Theory", n.d., paragraph 2). Taking M/M/3/20/∞/FIFO as an example again, we have 20 queuing capacity, 3 are serving, and 17 are waiting. The queuing rule means that after the server completes the service of the current client, it selects the next client from the queue (if any). In M/M/1, queuing discipline refers to FIFO. The independence assumption refers to the assumption that all customers are independent, customers arrive independently, and customer service hours are independent.

According to Almeida and Cruz (2018), the M/M/1 model assumes that the average customer arrival obeys a Poisson process with parameter $\lambda$, and the service-time process is exponential with parameter $\mu$ (p.2578).

(1) Therefore, the inter-arrival time obeys the exponential distribution:

$$P(x) = \lambda e^{-\lambda x}, x > 0$$

(2) The probability density of service times is:

$$P(x) = \mu e^{-\mu x}, x > 0$$

This is the simplest case.

Next, Almeida and Cruz (2018) mentioned that the queuing system will eventually reach a steady state (equilibrium) after a long period of time, and its traffic intensity $\rho < 1$ (p.2579). Then, the geometric probability distribution of the number of customers N in the departure time system is as follows:

$$P(N = n) = \begin{cases} (1-\rho)\rho^n & for \quad n = 0,1,2 \ldots \\ 0 & for \quad otherwise \end{cases} \tag{1}$$

So, from Eq.(1),$P(N = 0) = 1 - \rho, L = \frac{\rho}{1-\rho}, L_q = \frac{\rho^2}{1-\rho}$.

The above probability distribution shows that the probability of leaving an empty system after the customer leaves is $1 - \rho$. The probability of leaving a non-empty system is $\rho$. Therefore, each leaving customer can be regarded as a Bernoulli variable.

In addition, M/M/1 is one of the simplest and most suitable queuing models in many practical applications and it is continuous.

**2.2 The M/M/c/c queue**

M/M/c/c model comes from M/M/c/K[10],and it is also known as the Erlang–B model.[9]:495 It is a special case of M/M/c/K, where c stands for number of service(lines) and K for system capacity. In an M/M/c/K queue only K customers can queue at any one time (including those in service[8]). Any further arrivals to the queue are considered "lost". Erlang was given the probability that a

new customer is turned away or the probability that all lines are busy in an M/M/c/c system . The Erlang's formula as follow:

$$P_C = \frac{\frac{\rho^c}{c!}}{\sum_{i=0}^{c} \frac{\rho^i}{i!}}, where\ \rho = \frac{\lambda}{\mu} \quad (2)$$

Where $\lambda$ and $\mu$ have the same meaning as M/M/1 model, arrival process is Po ission distribution with $\lambda$ ,service-time process is exponential with parameter $\mu$.

# 3 Bayesian methon

In addition to the Bayesian method mentioned above, Azam et al. (2017) also listed 9 statistical paradigms, such as Bayesian, maximum entropy, non-parametric, etc. (page 6). Bayesian inference is a well-known statistical inference method, which can be applied to many models in different fields. Since some studies have used Bayesian method to find the sample size of M/M/1, this article will use Bayesian to find the sample size of M/M/c/c, which is unresearched.

Next, we need to know how the Bayesian method is used to determine the sample size. For example, according to Sadia and Hossain (2014), if we have an unknown parameter $\theta$ that needs to be estimated, then, assuming that the sample size $n$ needs to be determined, a random sample $X = (X_1, X_2, \dots, X_n)$ will be used Estimate $\theta$. $f(\theta)$is the prior distribution of the parameter $\theta$ and the likelihood function at the sample value $x = (x_1, x_2, \dots, x_n)$ is $L(\theta; x)$, and $L(\theta; x) \propto f(\theta|x)$.

$$L(\theta; x) = \prod_{i=0}^{n} P(x_i; \theta) \tag{3}$$

The posterior marginal distribution of $x$ is thus given by:

$$f(x) = \int_{\Theta} f(x|\theta) f(\theta) d\theta \tag{4}$$

Then the posterior distribution of θ given data $x$ with sample size $c$ is:

$$f(\theta|x, n) = \frac{f(x|\theta) f(\theta)}{\int_{\Theta} f(x|\theta) f(\theta) d\theta} = \frac{f(x|\theta) f(\theta)}{f(x)} \propto L(\theta; x) f(\theta) \tag{5}$$

In addition, if the prior distribution $f(\theta)$ and the posterior distribution $f(\theta|x)$ are the same distribution, they are conjugate each other.

Further, using Highest Posterior Density (**HPD**) Interval approach with given fixed internal, by finding $c$ which gives the maximum coverage of Eq.(5), we can get the sample size $c$ which is most suitable to estimate $\theta$. And the average coverage criterion (ACC) could find the smallest $c$ that meets the following condition:

$$\int_{\chi} \left( \int_{a(x,n)}^{a(x,n)+l} f(\theta|x, n) d\theta \right) f(x) dx \geq 1 - \alpha$$

where f(x) and $f(\theta|x, n)$are given in Eq.(4) and Eq.(5). And $l$ is the HPD credible set of length, $a(x, n)$ is the lower limit of $l$ for posterior density $f(\theta|x, n)$. (pp.421-422). Hence, **ACC** will find the minimum sample size n so that the aveage cover- age probability of the given fixed HPD interval length $l$ is at least $(1 - \alpha)$.

Moreover, as a region to which the parameter of interest belongs with high probability is meaningful, we use average length criteria (**ALC**) first determine the length of the credible interval for the fixed coverage $1-\alpha$ as follow:

$$\int_{a(x,n)}^{a(x,n)+l'(x,n)} f(\theta|x,n)d\theta = 1-\alpha$$

Where $l'(x,n)$is the length of the $(1-\alpha)100\%$ posterior credible interval for $x$.

Second, we can find the smallest $n$ such that:

$$\int_{\chi} l'(x,n)f(x)dx \leq l$$

Where $\chi$ is the data space for $x$, and $l$ is the desired pre-specified average length. This **ALC** is used to find sample size $n$ which would determine the coverage probability $(1-\alpha)$ of $\theta$'s HPD credible set (Sadia and Hossainp, 2014, p.422).

If we prefer a conservative sample size intead of averaging over $\chi$ to guarantee the desired coverage and interval length, we could use worst outcome criterion (**WOC**) to get the minimum sample size $n$ such that:

$$\inf_{x\in\chi}\left[\int_{a(x,n)}^{a(x,n)+l(x,n)} f(\theta|x,n)d\theta\right] \geq 1-\alpha$$

where both $\alpha$ and $l$ are fixed (Sadia and Hossainp, 2014, p.422).

**Exponential family**

We also need to know a little about the exponential family, because we may use it to deduce the conjugate prior more easily. The exponential family is defined as a family of the probability density functions or probability mass functions which can be expressed as:

$$f_\theta(x) = h(x)c(\theta)e^{(\sum_{j=1}^{k} w_j(\theta)t_j(\theta))} \tag{6}$$

in which $h(x) \geq 0, c(\theta) \geq 0$ and $t_1(x) \dots t_k(x)$ are functions of $x$. $f(x|\theta)$ does not depend on $\theta$.

Then according to the propositions the conjugate prior and the posterior are respectively the following

The conjugate prior is given by

$$\pi(\theta) = \varphi(\theta|\mu, \lambda) = K(\mu, \lambda) = e^{\theta\mu - \lambda\psi(\theta)}$$

And the corresponding posterior is given by

$$\pi(\theta|x) = \varphi(\theta|\mu + x, \lambda + 1)$$

Where $K(\mu, \lambda)$ is the normalizing constant of tensity

Example :$x_1, x_2, \ldots, x_n \sim N(\mu, \sigma^2)$,in which $\mu$ is unknown and $\sigma^2$ is known. The density of $X$ is :

$$f_\mu(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Then it can be rewritten as:

$$f_\mu(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{x^2+2x\mu+\mu^2}{2\sigma^2}} = \left(\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{x\mu}{\sigma^2}}$$

Where

$$c(\mu) = \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}, h(x) = e^{-\frac{x^2}{2\sigma^2}}, w(\mu) = -\frac{\mu}{\sigma^2}, t(x) = x$$

It's important to note that both exponential distributions and geometric distributions belong to the exponential family.

## 3.1 Sample size determination for the M/M/1 queue

Now, considering the model M/M/1, by Quinino and Cruz (2016) . And as mentioned above, the number of customers in the system at the departure epoch is given by Eq.(1). Then, we randomly select a sample of the number of customers $n$ left after a departing customer. Therefore, Quinino and Cruz (2016) assume that the number of customers left is given by $x_i$ $a$nd that $x = (x_1, x_2, \ldots, x_n)$ composes our sample size $n$ (p.997).

Next, we can estimate the traffic intensity $\rho$ and get likelihood function from Eq.(1) with parameter $\rho$ as follow:

$$L(x|\rho) = \rho^{y}(1-\rho)^{n} \tag{7}$$

where$y = \sum_{i=1}^{n} x_i$and $n$ is the sample size. Also, Almeida and Cruz(2018) explain that considering the ergodic property of Markov chain , the above data generation process can ensure the independence of sample observations if there is enough space between them (p.2579).

As Eq.(1) belongs to the exponential family, we can infer the prior for $\rho$ directly:

$$f(\rho) \propto \rho^{a-1}(1-\rho)^{b-1}, 0 < \rho < 1, a > 0, b < 0 \tag{8}$$

where $f(\rho)$ gives Beta distribution with parameter $a$ and $b$.

Immediately after, the posterior distribution corresponding to the prior from Eq.(8) would be given by:

$$f(\rho|data) = \begin{cases} \left(\frac{1}{Beta(a+y, n+b)}\right)\rho^{y+a-1}(1-\rho)^{n+b-1} 0 < \rho < 1 \\ 0 \qquad\qquad otherwise \end{cases} \tag{9}$$

in which $Beta(a+y, n+b)$follows a beta function with parameters $y+a$and $n+b$.

Besides,

$$\Pi(\rho|data) = \left(\frac{1}{Beta(a+y, n+b)}\right) exp[(y+a-1)ln\rho][(n+b-1)ln(1-\rho)]$$

Where

$$c(data) = \left(\frac{1}{Beta(a+y, n+b)}\right), h(\rho) = 1$$

$$w_1(data) = y + a - 1\,, w_2(data) = n + b - 1\,, t_1(\rho) = ln\rho\,, t_2(\rho) = \ln(1-\rho)$$

So, we can say that $\Pi(\rho|data)$ also belongs to exponential family.

And then, the Bayesian point estimator, which is also the mean of the posterior beta distribution, for $\bar{\rho}$ would be:

$$\bar{\rho} = \frac{y+a}{y+a+n+b}$$

Second, as mentioned earlier, we retained $n$ customers after a customer left, and Quinino and Cruz (2016) thought it was also possible to estimate the credible area (p.997). In general, the closer the estimated value is to the actual value, the more accurate the probability, but the larger the sample size requirement. In addition, Quinino and Cruz (2016) mentioned that people are looking for the smallest $n$ such that the coverage probability of the credible width area ($w$) is at least $1-\alpha$ (p.997). For example, let $d = w/2$, then it may find the smallest $n$, like this:

$$\int_{\bar{\rho}-d}^{\bar{\rho}+d} \Pi(\rho|data)d\rho \geq 1-\alpha \qquad (10)$$

Where $\bar{\rho}$ is the posterior mean and the credible area is $[\bar{\rho} - d, \bar{\rho} + d]$. And the estimator $\bar{\rho}$ and the coverage probability depend on $x = (x_1, x_2, \dots, x_n)$.

Next, Quinino and Cruz (2016) mention that we can choose n such that the expected posterior coverage probability, $m(x)$, is at least $1 - \alpha$ given by:

$$m(x) = \int_0^1 L(x|\rho)\Pi(\rho)d\rho \tag{11}$$

in which the expected value exceeds the marginal distribution of $x$ induced by the prior distribution (p.997).

Therefore, it is recommended to seek a minimum $n$ by using average coverage criterion (ACC) and satisfies:

$$\sum_{\forall x \in \chi} \left[ \int_{\bar{\rho}-d}^{\bar{\rho}+d} \Pi(\rho|data)d\rho \right] m(x) \geq 1 - \alpha \tag{12}$$

Where $m(x)$ gives weights (p.998).

Besides, if we need a conservative sample size and to satisfy the required coverage and interval length on any possible sample $x$, we could use WOC rather than averaging to find the minimum $n$ as follow:

$$\inf_{x \in \chi} \left[ \int_{\bar{\rho}-d}^{\bar{\rho}+d} \Pi(\rho|data)d\rho \right] \geq 1 - \alpha \tag{13}$$

## 3.2 Sample size determination for the M/M/c/c queue

We know that M/M/ c/c is a special case of M/M/ c /K when the system capacity K is equal to the number of services c. And M/M/ c /K is a case where M/M/ c has a limited capacity. Therefore, we can boldly use the probability calculation formula of sample size obtained by Tianyi Dai (2018)[]:

$$P_n = \begin{cases} \dfrac{\frac{\alpha^n \rho^n}{n!}}{\sum_{m=0}^{c-1} \frac{\alpha^m \rho^m}{m!} + \frac{\alpha^c \rho^c}{c!} \frac{c}{c-\alpha\rho}}, n \le c \\ \dfrac{\frac{\alpha^n \rho^n}{c^{n-c} c!}}{\sum_{m=0}^{c-1} \frac{\alpha^m \rho^m}{m!} + \frac{\alpha^c \rho^c}{c!} \frac{c}{c-\alpha\rho}}, n \ge c+1 \end{cases} \tag{14}$$

Where $\alpha$ is the probability of a customer will join the standard M/M/c queue, so $1-\alpha$ is the same as Eq.(2), represents the probability that a new customer is turned away or the probability that all lines are busy in an M/M/c system .

And then we just have to let $n = c$ ,and that gives us the probability formula for the sample size of M/M/c/c .

$$P_n' = \frac{\frac{\alpha^c \rho^c}{c!}}{\sum_{i=0}^{c-1} \frac{\alpha^i \rho^i}{i!} + \frac{\alpha^c \rho^c}{c!} \frac{c}{c-\alpha\rho}} \tag{15}$$

Then, we take a sample value $x = (x_1 x_2, \dots, x_n)$ from the sample space X to estimate C, and then get the likelihood function. So, the likelihood function for parameter $\rho$ will be:

$$L(x|\rho)=\prod_{j=1}^{n}P'(x_j;\rho)=\prod_{j=1}^{n}\frac{\frac{\alpha^{x_j}\rho^{x_j}}{x_j!}}{\sum_{i=0}^{x_j-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^{x_j}\rho^{x_j}}{x_j!}\frac{x_j}{x_j-\alpha\rho}} \tag{16}$$

Where $y=\sum_{i=1}^{n}x_i$.

Modification of the Eq.(15),we try to deduce a prior distribution for $c$. As the Eq.(15) can be rewritten that follows the from of Eq.(6) as:

$$P'(c|\rho)=\frac{\alpha^c}{c!}\exp\left\{c\ln\rho-\ln\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)\right\} \tag{17}$$

Where $h(c)=1\,,c(\rho)=1,w(\rho)=c\ln\rho,t(c)=\ln\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)$.

Hence, Eq.(15) also belongs to the exponential family, and then we could try a conjugate prior distribution, which is similar to the model M/M/c by Tianyi Dai , for $\rho$ as follow:

$$\pi(\rho)=\varphi(\rho|a,b)=K(a,b)e^{a\ln\rho-b\ln\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)}=K(a,b)\frac{\rho^a}{\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)^b} \tag{18}$$

Where constants $a>0,b>0$ and $K(a,b)$ is the normalizing constant.so,

$$K(a,b)=\frac{1}{\int_0^1\frac{\rho^a}{\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)^b}d\rho}$$

Next, the corresponding posterior distribution for $\rho$ could be obtained

$$\pi(\rho|data) = K(a+y, b+n)\frac{\rho^{a+y}}{\left(\sum_{i=0}^{c-1}\frac{\alpha^i\rho^i}{i!}+\frac{\alpha^c\rho^c}{c!}\frac{c}{c-\alpha\rho}\right)^{b+n}} \quad (19)$$

Where $y = \sum_{i=1}^{n} x_i$.

Hence , the posterior mean for $\bar{\rho}$ would be:

$$\bar{\rho} = E(\pi(\rho|data))$$

Then, the next steps are almost the same as the model M/M/1 in Eq. (10) to Eq. (12). We look for the smallest $n$ such that the coverage probability of the credible width ($w$) area is at least $1-\alpha$. For example, let $d = w/2$, then it may find the smallest $c$, like this:

$$\int_{\bar{\rho}-d}^{\bar{\rho}+d} \pi(\rho|data)d\rho \geq 1-\alpha \quad (20)$$

where$\bar{\rho}$ is the posterior mean and the credible region is [$\bar{\rho}$-d, $\bar{\rho}$+d]. And the esimator $\bar{\rho}$ and the coverage probability depend on $x = (x_1, x_2 \ldots, x_n)$.

Further, $c$ can be selected to make the expected posterior coverage probability

be at least $1-\alpha$.

$$m(x) = \int_0^1 L(x|\rho)\pi(\rho)d\rho \quad (21)$$

in which the expected value exceeds the marginal distribution of $x$ induced by the prior distribution.

Therefore, it is recommended to seek a minimum n by using the average coverage criterion (ACC) and satisfies:

$$\sum_{\forall x \in \chi} \left[ \int_{\bar{\rho}-d}^{\bar{\rho}+d} \pi(\rho|data) d\rho \right] m(x) \geq 1 - \alpha \tag{22}$$

where $m(x)$ gives weights.

And last, if we need a conservative sample size and to satisfy the required coverage and interval length on any possible sample $x$, we could use WOC rather than averaging to find the minimum $c$ as follow:

$$\inf_{x \in \chi} \left[ \int_{\bar{\rho}-d}^{\bar{\rho}+d} \pi(\rho|data) d\rho \right] \geq 1 - \alpha \tag{23}$$

# 4 Conclusion

The estimation of traffic intensity $(\rho)$ in various queueing systems has been considered in many studies involving inventory and maintenance theory. However, in many cases, there is no evaluation of the sample sizes in the planning stage of actually performing the estimate of the parameter which can compromise the analysis developed. In many systems modeled by queues, operational managers would be able to provide tight bounds (both upper and lower) on the traffic intensity such that $c < \rho < d$, in

which $0 < c < d < 1$. Alternatively, they would be able to stipulate percentiles or averages and variances.In this case, under the study in this paper, we can accurately determine the sample size within a certain degree of confidence and accuracy using known criteria, such as the mean coverage criteria (ACC) and the worst outcome criteria (WOC). Among the two standards, WOC generates the largest sample size and is the least sensitive to the different values of parameters a and b of the prior beta. $WOC_{WPD}$ is recommended for cases where there is a serious lack of knowledge about what values should be taken for parameters a and b. However, if these values are not economically or operationally viable, intermediate solutions are used to evaluate the large priors of ACC and adopt the larger sample obtained ($ACC_{WPD}$)

In short, Bayesian methods seem to be very suitable for dealing with inference targets in queuing systems. According to the data output of Tianyi Dai, it can be concluded that using the Bayesian method to estimate the traffic intensity of the M/M/c model has a good predictive effect. Bayesian factors can help test the prior distribution of data, and can infer traffic intensity and calculate the probability of the number of customers in the system. The M/M/c/c model studied in this paper is similar to M/M/c. Its biggest challenge is to calculate the number of posterior distribution in the system to obtain t. For other methods, the process is basically the same as the M/M/c model. It can be seen that the sample size estimation process of M/M/c/c is very similar to model M/M/c. Therefore, the Bayesian method can also be used to estimate the sample size of the M/M/c/c model.

# References

[1] Klugman S A,Panjer H,Willmot G E. John Wiley & Sons .Loss Models:Further Topics.Wiley Series in Probability and Statistics2013

[2] R. C. Quinino,F. R. B. Cruz .Bayesian sample sizes in an M/M/1 queueing systems.The International Journal of Advanced Manufacturing Technology,2017(1-4)

[3] Tianyi Dai .Sample size determination for markovian Queueing modelsCarleton University,2018,URL:https://curve.carleton.ca/6dd320a8-5cfc-4b1b-a884-62097e91502b

[4] Yue Xu .Estimating the sample size of Geo/Geo/1.Carleton Univesity,2020, URL:https://carleton.ca/math/wp-content/uploads/Yue-XuHonours-Project.pdf

[5] F. R. B. Cruz , R. C. Quinino. Bayesian sample sizes in an M/M/1 queuei ng systems. Springer-Verlag London, Int J Adv Manuf Technol (2017) 88:995–1002,DOI 10.1007/s00170-016-8855-2

[6] J.D.C. Little .A Proof for the Queueing Formula: L=W. Operations Research, 9(3):383-387, 1961.

[7] O.J. Boxma , P. R. D. Waal .Multiserver queue with impatient customers. ITC, 1994:743-756

[8] Kleinrock, Leonard (1975). Queueing Systems Volume 1: Theory. pp. 101–103, 404. ISBN 0471491101.

[9] Gautam, Natarajan (2012). Analysis of Queues: Methods and Applications. CRC Press. ISBN 9781439806586.

[10] https://en.wikipedia.org/wiki/M/M/c_queue#cite_note-gautam-1