Abstract

Queueing theory is pivotal in understanding waiting lines and their pervasive role in everyday life. Its profound application spans across diverse sectors, including computer programming networks, healthcare, transportation, and so on. Researchers have applied many statistical distributions in analyzing queueing data. This article focuses on the analysis of a single server queueing system with Poisson input and service times distributed according to a two-parameter gamma distribution. The study elaborates on the construction and estimation of key queueing properties such as steady state equations, average number of customers in the system, average queue length, expected size of the non-empty queue, distribution of waiting time, average waiting time in the system, and average waiting time in the queue of the \( M/Gamma/1 \) queueing model using the two-parameter gamma distribution.

Keywords: Gamma distribution; Poisson; service; single server; queueing model.
1. Introduction

Statistical inference has emerged as an important tool in the field of queueing theory, capturing the attention of many researchers in the past few decades [1]. The significance of statistical methods in queueing analysis lies in their ability to estimate crucial parameters like arrival rates (λ) and service rates (µ), along with various performance measures such as traffic intensity (ρ), system size (L), and the expected waiting time in the system (W). Originally, the queueing model foundation was rooted in Markovian models, where arrival times followed Poisson distribution and service time followed exponential distribution. Historically, these models were also used as a decision-making aid in the early stages of queueing theory. Bhat and Rao [2] estimated the arrival rate in an M/D/1 queue using single-packet probing. Cox and Smith [3] and Medhi [4] derived the steady-state equations and performance of various queueing models. Ikeda [5] presented a method to compute the mean waiting times of Gamma/Gamma/1, Gamma/D/1, and D/Gamma/1 queues, and to estimate their errors. Abate et al., [6] showed the calculation of the GI/G/1 waiting-time distribution and its cumulants from Pollaczek’s formulas. Kimura [7] studied diffusion approximation for an M/G/m queue.

The busy period in relation to the single-server queueing system with general independent arrivals and Erlangian service time was studied by Conolly [8]. Legros [9] studied the M/G/1 queue with event-dependent arrival rates. Moreover, Insua et al., [10] studied Bayesian analysis of M/Er/1 and M/Hk/1 queues, later Levy and Yechiali [11] showed the utilization of idle time in an M/G/1 queueing system. The waiting time distribution of a Pareto service self-similar queueing model for wireless network nodes was derived by Xu et al., [12]. Ramirez Cobo et al., [13] obtained inference for double Pareto Log-Normal queues with applications. Courtois and Georges [14] studied a single-server queue where the arrival rate is a decreasing function of the system length. Insua et al., [10] showed extensions to Er/M/1 and Er/M/s queues. Recently, this approach has been significantly developed in several publications by [15,16,17,18]. Further Swamy et al., [19] studied a single-server Poisson queueing model with additive Exponential service time distribution. The Markov process representing the number of customers in such systems is called the birth-death process, which is widely used in population models. Birth-death terminology is used to indicate increases and decreases in population size. In queueing systems, the corresponding events are arrivals and departures.

In real-world scenarios where the observed data may deviate from the idealized expected exponential distribution, the gamma distribution offers a more flexible modeling approach as it can better represent the real world. The utility of the Gamma distribution is further enhanced by its connection to the exponential distribution and its adherence to the Markov property. The single server queueing system with Poisson input and exponential service time has been extensively analyzed in the literature.

The unique properties of exponentially distributed inter-arrival times and service times give rise to a queueing system on which the number of customers in the system, at any sequence of times, constitutes an embedded Markov chain. Therefore, knowledge of the number of customers in the system at a given time is equivalent to complete knowledge of the state of the system, and steady-state equations involving the probabilities of the number of customers in the system are readily obtained.

However, when the service times not exponentially distributed, the analysis of the queueing system is more becomes more challenging. The Gamma service time distribution is often employed in queueing systems to model variability in customer service times. This variability is crucial because
real-world service times rarely remain constant; they can exhibit randomness due to factors such as processing variation, resource availability, and system interruptions. By considering a Gamma distribution for service time, queueing models can better capture the inherent variability in service processes, providing more accurate predictions and insight into system performance [20], [21]. It is commonly used as a model for financial losses or insurance claim sizes in finance and insurance sectors due to its flexibility, positive-skewed nature, and convenient mathematical and statistical properties. Additionally, it finds its application in modeling waiting time distribution.

1.1. Purpose of study

This article considers a queueing model where arrivals follow a Poisson distribution, service times follow a two-parameter gamma distribution with an average service time.

2. Methods and materials

\( \alpha \mu \) per unit time, there is a single service channel, the system capacity is unlimited and the service discipline is FIFO. The order governing steady-state equations and performance measures for the model are obtained in the pdf of the two-parameter Gamma distribution with parameter \((\mu, \alpha)\) is

\[
f(t) = \frac{\mu^{\alpha} \cdot t^{\alpha-1} \cdot e^{-\mu t}}{\Gamma(\alpha)}, \quad t > 0, \quad (\mu, \alpha) > 0
\]

with mean service rate

\[
E(x) = \int_0^\infty tf(t)dt = \frac{\alpha}{\mu}
\]

The M/Gamma/1 queues are particularly relevant when trying to analyze and optimize a system where service time variability significantly affects performance measures like waiting times, queue length, and overall system. By using this type of queueing model, practitioners and researchers can gain insight into how different levels of service times variability impact the system’s behavior and performance, ultimately adding to decision-making and resource allocation. This work endeavors to formulate and estimate the steady state equations, average number of customers in the system, average queue length, the expected size of the nonempty queue, waiting time distribution, mean waiting time in the system and mean waiting time in the queue.

2.1. Steady-state equations

We have considered, \( P_n(t) \) to be the probability that there are \( n \) customers in the system at time \( t \) and \( P_n (t + \Delta t) \) to be the probability that the system has \( n \) customers at a time \( (t + \Delta t) \). This event occurs in the following mutually exclusive and exhaustive ways

(i) There are \( n \) units in the system at time \( t \) and none have arrived or been served in the period \( \Delta t \).
(ii) At time \( t \) there are \( (n-1) \) units in the system and at a small time \( \Delta t \) there is 1 arrival but no service has been performed.
(iii) The system has \((n + 1)\) units at time \( t \) and at small time \( \Delta t \) there is no arrival time, but one service completion.
(iv) At time $t$ there are $n$ units in the system, and at time $\Delta t$ the service has not arrived or has finished.

Hence, the steady-state equation for the model is

$$P_n'(t) = -\frac{\lambda}{\mu} P_n(t) + \lambda P_{n-1}(t) + \frac{\alpha}{\mu} P_{n+1}(t), \quad n \geq 1$$

$$P_0'(t) = -\lambda P_0(t) + \frac{\alpha}{\mu} P_1(t)$$

$$P_{n+1}(t) = \frac{\lambda}{\mu} P_n(t) - \lambda P_{n-1}$$

and

$$P_0(t) = \frac{\lambda \mu}{\alpha} P_0$$

Since,

$$\sum$$

2.2. **Performance measures**

The steady-state probability distribution of the system size allows us to calculate the system’s measures of effectiveness [22], [23]. The effectiveness of any queueing system is characterized by the distribution of some output indicators, such as the number of customers in the system or the queue, customer waiting time in the system or the queue, etc. These distributions are conveniently summarized in terms of their expected values, called performance measures that are the function of both $\lambda$ and $\mu$. Evaluation of these measures is necessary to answer questions such as whether there is a shortage of servers, or whether holding capacity should be limited so that expected customer wait time and expected server idle time, or even more generally, some associated probability of overrun, are minimized. The performance measures of the single server Gamma model are given below:

2.2.1. **Expected number of customers in the system**

Let $N$ represent the number of customers in the system, then
\[ L = E(N) \]
\[ = \sum_{n=0}^{\infty} n P_n \]
\[ = \sum_{n=0}^{\infty} n \frac{\lambda \mu}{\alpha} n^{\alpha - 1} \left(1 - \frac{\lambda \mu}{\alpha}\right)^{n-1} \]
\[ = 1 - \frac{\lambda \mu}{\alpha} \sum_{n=0}^{\infty} \left(1 - \frac{\lambda \mu}{\alpha}\right)^n \]
\[ = \alpha - \lambda \mu \]

2.2.2. **Expected number of customers in the queue**

The number of customers in the queue under steady state is denoted by \( N_Q \). Then

\[ L_q = E[N_q] = \sum_{n=1}^{\infty} (n - 1) P_n \]
\[ = \sum_{n=1}^{\infty} n P_n - \sum_{n=1}^{\infty} P_n \]
\[ = \sum_{n=1}^{\infty} n \frac{\lambda \mu}{\alpha} n^{\alpha - 1} \left(1 - \frac{\lambda \mu}{\alpha}\right)^{n-1} - \sum_{n=0}^{\infty} P_n \]
\[ = \lambda \mu \frac{\lambda \mu}{\alpha} \frac{1}{\alpha} \]
\[ = \frac{\lambda \mu}{\alpha} \]
\[ \approx 1 - \frac{\lambda \mu}{\alpha} \]

2.2.3. **The expected size of the nonempty queue**

It is denoted by \( L'_q \)

Thus,
\[ L_q = E[N_q | N_q \neq 0] \]
\[ = \sum_{n=2}^{\infty} (n-1)p_n \]
\[ p_n = \frac{P(n \text{ in the system and } n \geq 2)}{p(n \geq 2)} \]
\[ = \frac{\sum_{n=2}^{\infty} p_n}{\sum_{n=2}^{\infty} p_n} \]
\[ = \frac{1 - p_0 - p_1}{\alpha - \lambda \mu} \]

2.2.4. **Waiting time distribution**

Let \( \phi \, dt \) be the probability that the customer is served in the interval \((t, t + \Delta t)\), where \( t \) is measured from the time of the customer’s arrival. Let the customer arrive at time \( t = 0 \) and the service starts in the interval \((t, t + \Delta t)\), then

(i) If the system is empty, the waiting time is \( P_0 \) with zero probability. Then \( p_0 = (1 - \frac{\lambda \mu}{\alpha}) \)

(ii) If there are already \( n \) customers in the system when \((n + 1)\)th customers arrive, \( n \) customers must leave the system before the service of \((n + 1)\)th customer starts or \((n + 1)\) customers must be served in the time period \((t, t + \Delta t)\)

The mean service rate is \( \frac{\alpha}{\mu} \) unit of time or \( (\frac{\alpha}{\mu})t \) in time \( t \) and probability of \((n - 1)\) departures in time \( t \) is given by Poisson Distribution:

\[ \frac{\alpha}{\mu} \, \frac{1 - \frac{\alpha}{\mu}t}{(n - 1)!} \]  

(7)

If there are \( n \) customer in the system, then

\[ \phi_n(t) = P[(n-1) \text{ customer are served by time } t] \cdot P[n^{th} \text{ customer is served during } \Delta t] \]

\[ \phi_n(t) = \left( \frac{\alpha}{\mu}t \right)^{n-1} \frac{e^{-\frac{\alpha}{\mu}t}}{(n-1)!} \alpha \Delta t \]  

(8)

\[ \frac{1}{\mu} \]
The PDF of the waiting time of a customer is given by:

\[
\sum_{n=1}^{\infty} \phi_n(t) P_n \\
\sum_{n=1}^{\infty} \alpha = \\
= \sum_{n=1}^{\infty} \frac{\lambda \mu}{\alpha} \left( 1 - \frac{\lambda \mu}{\alpha} \right) \frac{(\frac{\alpha}{\mu})^{n-1} e^{-\frac{\mu}{\alpha}}}{(n-1)!} \frac{\alpha}{\mu} \Delta t \\
= \lambda \left( 1 - \frac{\lambda \mu}{\alpha} \right) e^{-\left(\frac{\alpha}{\mu}\right)t} \Delta t \\
= \lambda \left( 1 - \frac{\lambda \mu}{\alpha} \right) e^{-\lambda t} \Delta t \\
\]

\[
\text{if system is empty} \\
= \lambda \left( 1 - \frac{\lambda \mu}{\alpha} \right) e^{\lambda t} \Delta t \\
\text{otherwise}
\]

2.2.5. The mean waiting time in the system

\[
W = \frac{\lambda \mu}{\alpha - \lambda \mu} \frac{1}{\lambda} \\
= \frac{\mu}{\alpha - \lambda \mu}
\]

2.2.6. Mean waiting time in the queue

\[
W_q = \frac{\lambda \mu}{\alpha} \frac{1}{\lambda \\
= \mu \left( 1 - \frac{\lambda \mu}{\alpha} \right) \\
\sim \frac{\lambda \mu}{\alpha}
\]

3. Results

3.1. Numerical illustration

In this section, the performance of the model is discussed through the different values of the parameters considered for arrival rate \( \lambda \) and service time parameter \( \mu, \alpha \) for the given value of \( \lambda = 0.5, 0.6, 0.7, 0.8, 0.9, \mu = 0.5, 0.6, 0.7, 0.8, 0.9, \alpha = 1, 2, 3, 4, 5 \) and \( n=1,2,3,4,5 \). The probability that the system is empty and the probability of a busy server is computed and presented in Table 1. The relationship between the parameter and the probability of idleness is represented in Figure 1.

Table 1 shows how different combinations of \( \alpha, \lambda, \) and \( \mu \) impact the probability of the queue being empty. In the first section, \( \mu \) varies from 0.5 to 0.9, while \( \alpha \) and \( \lambda \) are fixed. As the service
rate $\mu$ increases, the probability of the queue being empty ($p_0$) decreases, indicating that as the system becomes more efficient at serving customers, it’s less likely to be empty.

### Table I

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$p_0$</th>
<th>$1 - p_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.750</td>
<td>0.250</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.700</td>
<td>0.300</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.7</td>
<td>0.650</td>
<td>0.350</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.8</td>
<td>0.600</td>
<td>0.400</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.9</td>
<td>0.550</td>
<td>0.450</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.875</td>
<td>0.125</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.850</td>
<td>0.150</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.5</td>
<td>0.825</td>
<td>0.175</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.5</td>
<td>0.800</td>
<td>0.200</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>0.5</td>
<td>0.775</td>
<td>0.225</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.750</td>
<td>0.250</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.875</td>
<td>0.125</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.916</td>
<td>0.083</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.937</td>
<td>0.062</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.950</td>
<td>0.050</td>
</tr>
</tbody>
</table>

In the second section, with fixed $\alpha$, and $\mu$, as the arrival rate $\lambda$ increases, the probability of the queue being empty ($p_0$) decreases, suggesting that a higher arrival rate results in a lower likelihood of an empty queue. In the third section, $\alpha$ varies from 1 to 5, while $\lambda$ and $\mu$ remain constant. As $\alpha$ increases, the probability of the queue being empty ($p_0$) decreases, indicating that a higher $\alpha$ value corresponds to a lower likelihood of an empty queue. This implies that an increasing $\alpha$ corresponds to a positively skewed gamma distribution, which in turn means a lower probability of an empty queue ($p_0$). For different value parameters, the average number of customers in the system and the average number of customers in the queue is computed and presented in 2.

![Fig 1](image1.png)  
(a) $\lambda$ vs $P_0$  
(b) $\mu$ vs $P_0$  
(c) $\alpha$ vs $P_0$  

**Fig 1.** Relationship between parameter ($\lambda$, $\mu$) and Probability that there is no customer in the system ($p_0$)
It is observed from Table 2 that the performance measures of the proposed model are significantly affected by the model parameters for different combinations of $\lambda$ and $\mu$. In the first
section, $\mu$ varies from 0.5 to 0.9, while $\alpha$ and $\lambda$ are kept fixed. As the service rate increases, the average number of customers in the system increases ($L$), indicating that it leads to increased congestion and longer waiting times. Again, with fixed $\alpha$ and $\mu$, as the arrival rate $\lambda$ increases, the average number of customers in the queue ($L_q$) increases, suggesting that there will be increased variability in waiting times and it may be necessary to allocate additional resources, such as more service agents or servers, to handle the increased demand. However, if $\alpha$ varies from 1 to 5, while $\lambda$ and $\mu$ remain constant. As $\alpha$ increases, the average number of customers in the system ($L$) and as well as in the queue ($L_q$) decreases, indicating shorter waiting times, reduced congestion, improved service quality, and potential cost savings. The relationship between the parameters ($\lambda$, $\mu$), and average number of customers in the system ($L$), and the average number of customers in the queue ($L_q$) is shown in Figure 2. Table 3 shows the average waiting time of the customer in the system and the queue for different values of the parameter.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\lambda$</th>
<th>$\mu$</th>
<th>$W$</th>
<th>$W_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.666</td>
<td>0.166</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.857</td>
<td>0.257</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.7</td>
<td>1.076</td>
<td>0.376</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.8</td>
<td>1.333</td>
<td>0.533</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.9</td>
<td>1.636</td>
<td>0.736</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.285</td>
<td>0.142</td>
</tr>
<tr>
<td>2</td>
<td>0.6</td>
<td>0.5</td>
<td>0.294</td>
<td>0.176</td>
</tr>
<tr>
<td>2</td>
<td>0.7</td>
<td>0.5</td>
<td>0.303</td>
<td>0.212</td>
</tr>
<tr>
<td>2</td>
<td>0.8</td>
<td>0.5</td>
<td>0.312</td>
<td>0.250</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>0.5</td>
<td>0.322</td>
<td>0.290</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.666</td>
<td>0.166</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.5</td>
<td>0.285</td>
<td>0.142</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.5</td>
<td>0.181</td>
<td>0.136</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.133</td>
<td>0.133</td>
</tr>
<tr>
<td>5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.105</td>
<td>0.131</td>
</tr>
</tbody>
</table>

(a) $\lambda$ vs $W$  
(b) $\mu$ vs $W$  
(c) $\alpha$ vs $W$

Fig 3. Relationship between parameters ($\lambda$, $\mu$) and $W$ and $W_q$

Table 3 shows how different combinations of $\alpha$, $\lambda$, and $\mu$ impact the probability of the queue being empty. In the first section, $\mu$ varies from 0.5 to 0.9, while $\alpha$ and $\lambda$ are fixed. As the service rate $\mu$ increases, the average waiting time of customers in the system and the average waiting time of customers in the queue increases. An increase in the average waiting time of customers in both the queue and the system is a sign that more customers may choose to abandon the queue or system without receiving service. Again, with fixed $\alpha$, and $\mu$, as the arrival rate $\lambda$ increases, both the average waiting time of customers in the system and in the queue increases. However, if, $\alpha$ varies from 1 to 5, while $\lambda$ and $\mu$ remain constant, the average waiting time of customers in the system and average waiting time of customers in the queue decreases, indicating that the queuing system is performing well and meeting customer needs efficiently. This can have a range of positive effects on customer satisfaction, operational efficiency, and overall business performance. In Figure 3, $W$ and $W_q$ are represented graphically.

4. Conclusion

This article sought to derive steady-state equations, performance measures, and waiting time distributions while conducting a numerical analysis of the impact of gamma distributions on service times within single-server queueing systems. The selection of gamma distributions was motivated by the anticipation that changes in skewness would have notable effects. However, our findings revealed that the effects were relatively modest. Remarkably, they were more pronounced in scenarios with less variable service times, resulting in slight increases in the average queue length, average system size, the probability of an empty system, the probability of the server being busy, and mean waiting time.

Therefore, it is worth reflecting that even a slight change in service time skewness contributes to increased mean waiting times. Moreover, the shape parameter $\alpha$ has a significant impact on the performance measures. More specifically, as $\alpha$ increases the service time required by the customer in the system and average waiting is increased. With the aforementioned explanation, it is evident that the model developed in this study has greater practical utility in closer approximation to real-world phenomena. For future research, the above model can be extended to a multi-server queueing model. Other popular distributions such as the Weibull distribution, Pereto distribution, and beta distribution can also be considered as service time distribution of a queueing.
Conflict of Interest:
The authors declare that they have no conflict of interest.

References


