International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

On Markovian Queuing Models

Tonui Benard C.[1], Langat Reuben C.[2], Gichengo Joel M.[3]

[1, 2., 3]University of Kabianga, Mathematics and Computer Science Department, P.O Box 2030-20200 Kericho, Kenya

Abstract: ***The ultimate objective of the analysis of queuing systems is to understand the behaviour of their underlying process so that informed and intelligent decisions can be made by the management. The application of queuing concepts is an attempt to minimize cost through minimization of inefficiency and delays in a system. Various methods of solving queuing problems have been proposed. In this study we have explored single –server Markovian queuing model with both interarrival and service times following exponential distribution with parameters λ and μ, respectively, and unlimited queue size with FIFO queuing discipline and unlimited customer population. We apply this model to catering data and estimate parameters for the same. A sensitivity analysis is the carried out to evaluate stability of the system.***

Keywords: Queuing models, Markovian models, Single-server, Interarrival times, Service times**,** Sensitivity analysis.

1. Introduction

Queuing theory is a branch of applied probability theory used to describe the more specialized mathematical models for waiting lines or queues. It uses Queuing models to represent the various types of Queuing systems that arise in practice. The models enable finding an appropriate balance between the cost of service and the amount of waiting. The concept of Queuing theory has been developed largely in the context of telephone traffic engineering originated by A. K. Erlang in 1909. Queuing models find applications in a wide variety of situations that may be encountered in health care, engineering, and operations research (Gross and Harris, 1998). Queuing systems are comprised of customer(s) waiting for service and server(s) who serve the customer. They are frequently observed in some areas of day-to-day life, for example:

1) People waiting at the check-in counter of an airport
2) Aeroplanes arriving in an airport for landing
3) Online train ticket reservation system
4) People waiting to be served at a buffet
5) Customers waiting at a barber shop for a hair cut
6) Sequence of emails awaiting processing in a mail server

Queues are usually characterized by the *arrival pattern* (Poisson, deterministic or a general distribution), *Service pattern* (constant, exponential, hyper exponential, hypo-exponential or general distribution*), number of servers* (single server or multiple servers), *maximum system capacity* (number of customers in a system can range from one to infinity), *population size* (queue can have infinite or finite length) and *queue discipline* (order of service delivery can be First In First Out (FIFO), random order, Last In First Out (LIFO) or priorities), see Zukerman (2013) and Adan Resing (2002).

To incorporate these features, Kendall (1953) introduced a Queuing notation *A/B/C/X/Y/Z* in where: *A* is the interarrival time distribution, *B* is the service time distribution, *C* is the number of servers, *X* is the system capacity, *Y* is the population size and *Z* is the queue discipline.

In this study, an infinite customer population and service in the order of arrival (*FIFO*) are default assumptions. There is also an additional default assumption: inter-arrival and service times are independent.

2. Basic Markovian Queuing Models

In Markovian models, the analysis is conducted using the memoryless property of exponential distribution

2.1 Markovian Single-Server Models

1) $M/M/1/\infty$ **Queuing System:** *M* stands for Markovian or memoryless. The first *M* denotes arrivals following a Poisson process, the second *M* denotes service time following exponential distribution, 1 refers to a single server and refers to infinite system capacity.
2) $M/M/1/N$ **Queuing System:** This system is a type of $M/M/1/\infty$ queue with at most *N* customers allowed in the system.

2.2 Markovian Multiserver Models

1) $M/M/c/\infty$ **Queuing system:** This is a Markovian Queuing model with *C* number of servers
2) $M/M/c/c$ **Loss system:** This is also known as the Erlang loss system and its system state follows a truncated Poisson distribution.
3) $M/M/c/K$ **Finite –Capacity Queuing system:** In this Queuing model, the system has a finite capacity of size *K* and we assume that $c < K$.
4) $M/M/\alpha$ **Queuing system:** This is a Markovian Queuing model without any queue. There are infinitely many servers such that every incoming customer finds an idle server immediately.

Queuing models play an essential role for business process re-engineering purposes in administrative tasks. "Queuing models provide the analyst with a powerful tool for designing and evaluating the performance of Queuing systems." (Banks, Carson, Nelson & Nicol, 2001). Sometimes, inefficiencies in services also occur due to an undue wait in service may be because of new employee. Delays in service jobs beyond their due time may result in losing future business opportunities.

Detailed discussion on Markovian models are found in Castaneda et al. (2012). In this study we concentrate on the $M/M/1/\infty$ Queuing System.

3. $M/M/1/\infty$ Queuing System

The $M/M/1/\infty$ or simply *M/M/1* Queuing system describes a Queuing system with both interarrival and service times following exponential distribution with parameters λ and μ, respectively, one server, unlimited queue size with *FIFO* Queuing discipline and unlimited customer population. The *M/M/1* is one of the earliest systems to be analyzed. As it is neatly described by Chee-Hock and Boon-Hee (2008), the single-server queue is a place where customers arrive individually to obtain service from a service facility. The service facility contains one server that can serve one customer at a time. If the server is idle, the customer is served immediately. Otherwise, the arriving customer joins a waiting queue. This customer will receive his service later, either when he reaches the head of the waiting queue or according to some service discipline. When the server has completed serving a customer, the customer departs.

Theorem 1: Let X_t be a random variable denoting the number of customers in the $M/M/1$ Queuing system at any time t.
Define:

$$P_n(t) = P(X_t = n), n = 0,1,2,\cdots, t > 0$$

Let $\lambda/\mu = \rho$. *When* $\rho < 1$, *the steady state probabilities are given by:*

$$P_n = \lim_{t\to\infty} P(X_t = n) = (1-\rho)\rho^n, n = 0,1,2,\ldots$$

Proof: The stochastic process $\{X_t \geq 0\}$ in $M/M/1$ Queuing system can be modeled by a birth-and-death process with birth states $\lambda_n = \lambda,\ n = 0,1,\ldots$ and death states $\mu_n = \mu,\ n = 1,2,\ldots$
The steady state balance equations can be obtained are given by:

$$-\lambda P_0 + \mu P_1, \qquad n = 0$$
$$\lambda P_{n-1} - (\lambda+\mu)P_n + \mu P_{n+1}, \qquad n = 1,2,\ldots$$

Solving the above equations, we get:

$$P_1 = \frac{\lambda}{\mu} P_0, \qquad n = 0$$
$$P_{n+1} = \frac{\lambda}{\mu} P_n, \qquad n = 1,2,\ldots$$

So that: $P_{n+1} = \left(\frac{\lambda}{\mu}\right)^{n+1} P_0,\ n = 1,2,\ldots$

To obtain the value of P_0, we use the fact that $\sum_{n=0}^{\infty} P_n = 1$
Hence, when $\rho < 1$,

$$P_0 = \frac{1}{1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2 + \cdots} = 1 - \rho$$

The steady state probabilities are given by:

$$P_n = (1-\rho)\rho^n,\ n = 0,1,2,\cdots \qquad (1)$$

Note 1: *Equation (1) represents the probability mass function of a discrete random variable denoting the number of customers in the system in the long run. Clearly this distribution follows a geometric distribution with parameter* $1-\rho$.

Theorem 2: *Let* L_s *be the average number of customers in the M/M/l Queuing system and* L_q *be the average number of customers in the queue. Then:*

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho};\ L_q = \frac{\rho^2}{1-\rho}$$

Proof: Using (1), the mean number of customers can be found. Note that the number of customers in the system is the sum of the number of customers in the queue and the number of customers in service.

Hence:

$$L_s = \sum_{n=1}^{\infty} nP_n = \sum_{n=1}^{\infty} n(1-\rho)\rho^n \qquad (2)$$

This yields;

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$$

As expected, these equations show that with increasing load, i.e., as $\rho \longrightarrow 1$, the mean number of customers in the system grows and the probability of an idle system decreases.

Similarly, the average number of customers in the queue can be computed as:

$$L_q = \sum_{n=1}^{\infty} (n-1)P_n$$
$$= L_s - (1-P_0)$$
$$= \frac{\rho^2}{1-\rho}$$

Note 2: *The probability that the server is busy is another performance measure of the Queuing system. The probability that the server is busy when the system is in equilibrium is known as the utilization factor (traffic intensity) and is denoted by* ρ.
For the $M/M/1$ *queue,*

$$\rho = 1 - P_0 \qquad (3)$$

The number of customers in the system is of importance from the management's perspective and interest. Besides, the average queue size and average system size are also important parameters that represent the quality of service.

Two more measures important from the customer's point of view are the average time spent in the system (T_s) and the average time spent in the queue (T_q).

Little (1961) derived the following formula, which gives the relation between the average number of customers in the system (L_s) and the average time spent in the system (T_s) and also between the average number of customers in the queue (L_q) and Tq.

$$L_s = \lambda T_s, \qquad L_q = \lambda T_q \tag{4}$$

It is justified that, using the average time spent in the system, T_s, the average number of the customers during this time is λT_s, where λ is the average number of arrivals per unit time. It is very important to note that no assumption is made on the interarrival distribution, the service time distribution and the queue discipline.
From (2) we have

$$T_s = E(T) = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda} \tag{5}$$

for the $M/M/1$ Queuing system.

It can be deduced that:

$$T_s = T_q + \frac{1}{\mu}$$

Rust (2008) said that the Little's theorem can be useful in quantifying the maximum achievable operational improvements and also to estimate the performance change when the system is modified.

4. Research Design

The study was conducted in the University of Kabianga campus mess. Two days data was collected using observation method because the objective of the study didn't rely on opinion of the customers. Two variables of interest namely inter-arrival times and service times were recorded. The population studied included all the arrivals between Noon and 1.00pm of the study period. Data emerging from balking and/or reneging was disregarded. A pilot study was done to test the reliability of research instrument and it was validated. Two-sample *t*-test was used to test of equality of means of inter-arrival times and service times. Windows-based Quantitative Software for Business (WINQSB) suite and STATA software were used for computation of the above parameters, testing of hypothesis and performing sensitivity analysis.

5. Results

5.1 Testing of Hypotheses.

For the service time, the hypothesis tested was:
H_0: service time day 1 = service time day 2
V/s H_1: service time day 1 $\neq$ service time day 2
The results are shown below.

Two-sample t test with unequal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
servicel	249	37.10843	1.492468	23.55074	34.16891	40.04796
service2	125	38.312	1.425275	15.93506	35.49098	41.13302
combined	374	37.5107	1.101173	21.29566	35.34541	39.67598
diff		-1.203566	2.063702		-5.262782	2.855649

diff = mean(servicel) - mean(service2) t = -0.5832

Ho: diff = 0 Satterthwaite's degrees of freedom = 340.392

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.2801 Pr(|T| > |t|) = 0.5601 Pr(T > t) = 0.7199

The two-tailed *t*-test yielded a *P*-value of *0.5601* indicating that there is no significant difference in the two means.

Similarly, for the inter-arrival time, the following hypothesis was tested:
H_0: Inter-arrival time day 1 = Inter-arrival time day 2 V/s
H_1: Inter-arrival time day 1 $\neq$ Inter-arrival time day 2

The results are shown below.

Two-sample t test with unequal variances

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
inter_~1	260	13.23462	1.219914	19.67052	10.8324	15.63683
inter_~2	258	12.9186	1.0302	16.54745	10.8899	14.94731
combined	518	13.07722	.798138	18.16531	11.50923	14.64521
diff		.3160107	1.596716		-2.821049	3.453071

diff = mean(inter_arr1) - mean(inter_arr2) t = 0.1979

Ho: diff = 0 Satterthwaite's degrees of freedom = 502.555

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.5784 Pr(|T| > |t|) = 0.8432 Pr(T > t) = 0.4216

Since the *P*-value for the two-tailed *t*-test is far much beyond the 0.05 index, we concluded that there was no significant difference between inter-arrival times at 95% level of confidence. Therefore, we decided to pick on 38.312 seconds and 12.9186 seconds as the effective service time and the effective inter-arrival time respectively.

5.2 Estimation of the Queue Parameters

Applying the data analysis tools as mentioned in section 3 above, we obtain:

$$\mu = \frac{1}{38.312} \times 60 = 1.5660$$

$$\lambda = \frac{1}{12.9186} \times 60 = 4.6445$$

Theoretically, the service rate was 1.5660 customers per minute while the inter-arrival time was 4.6445 customers per minute. This therefore implied that the system utilization was:

$$\rho = \frac{4.6445}{1.5660} = 2.966 \ (3\ d.p)$$

Since the system utilization factor is greater than one, the system is unstable for analysis. Hence the main task is to

perform sensitivity analysis by making reasonable changes to the queue parameters so as make it reach steady state suitable for analysis.

The system utilization factor ideally indicates the percentage of time the servers are busy. Thus if it exceeds 100%, the queue grows indefinitely. This causes dissatisfaction among customers and may lead to possible losses due to inefficiency.

5.3 Sensitivity Analysis

Two major possible parameter changes were made in an attempt to stabilize the system. Changes were made on the;

1) Number of servers holding all other factors constant.
2) Service rate holding all other initial factors constant

This yielded the following results:

Table 1: Sensitivity analysis for number of servers

No. of servers	*Arrival rate* λ	*Service rate* μ	ρ	L_s	L_q	T_s	T_q	P_0	P_w
1	4.645	1.566	-	-	-	-	-	-	-
2	4.645	1.566	-	-	-	-	-	-	-
3	4.645	1.566	0.989	87.92	84.95	18.93	18.29	0.002	0.98
4	4.645	1.566	0.742	4.39	1.42	0.94	0.31	0.04	0.49
5	4.645	1.566	0.593	3.30	0.33	0.71	0.07	0.05	0.23

The above findings indicate that at the current rates of arrival and service, the Queuing system would stabilize for analysis when there are at least three servers (or service points), with the corresponding system utilization factors. This means that probability of a customer waiting in the queue (P_w) when there are three servers is 0.98, while the probability of the system having no customer (P_0) would be 0.002. A customer would spend about 18.93 minutes in the entire system (T_S) while they have to wait to get to the server for an average of 18.29 minutes (T_q). Still at three servers, the service points would be busy 98.86% of the time with the queue length being an average of 88 customers including those being served (L_s). This means that those who would be waiting for service would be about 85.

Table 2: Sensitivity Analysis for the Service Rate

Arrival rate λ	*Service rate* μ	ρ	L_s	L_q	T_s	T_q	P_0	P_w
4.4665	1.5660	-	-	-	-	-	-	-
4.4665	2.5660	-	-	-	-	-	-	-
4.4665	3.5660	-	-	-	-	-	-	-
4.4665	4.5660	0.9782	44.89	43.91	10.05	9.83	0.02	0.98
4.4665	5.5660	0.8025	4.06	3.26	0.91	0.73	0.20	0.80
4.4665	6.5660	0.6802	2.13	1.45	0.48	0.32	0.32	0.68

The above results were obtained after making an arbitrary assumption of a 1 unit shift for the service rate. The system stabilizes for analysis at a service rate of at least 4.5660 customers per minute. This implies that the service point would be busy 97.82% of the time with about 44 customers awaiting service. The entire system would have about 45 people including the one being served. A customer would have to wait for service for an average of 9.83 minutes with a probability of 0.98 and would end up spending an average of 10.05 minutes in the entire system. The probability of the queue having no customer would be 0.02.

6. Conclusion

From the research findings, it is evident that the mess at the University of Kabianga main campus experiences an unstable Queuing system (not statistically analyzable) leading to numerous inefficiencies like infinite growth of waiting line. This therefore implies that customers may be dissatisfied and the probability of them pulling away from the premises is high. This could see the enterprise plunge into consequent losses.

However, we developed a sample sensitivity analysis scheme that could assist the management to make necessary changes so that the queue conforms to statistical standards. The adoption of our findings could be beneficial to both the customers (by reducing time spent in the system) and the enterprise (by optimizing operations and task force). The number of servers could be increased to at least three to enhance an effective queue. Alternatively, the current service rate can be increased to not less than 4.5660≈5 customers per minute.

References

[1] Adan I. and Resing J., (2002). Queuing Theory .Eindhoven University of Technology, Eindhoven.
[2] Allen O. (1990), Probability, Statistics and Queuing Theory with Computer Science Applications, 2nd Edition, academic press, Boston
[3] Banks J; Carson J; Nelson B. L. and Nicole D. M. (2001). Discrete Event System Simulation. Prentice Hall International series, 3rd Edition.
[4] Castaneda, L. B., Arunachalam, V. and Dharmaraja, D (2012). Introduction to Probability and Stochastic Processes with Applications. John Wiley & Sons, Inc.
[5] Chee-Hock N. and Boon-Hee, (2008). Queuing Modeling Fundamentals with Applications. 2nd Edition, John Wiley and sons, England
[6] Jain R (2008). Lecture notes on Introduction to Queuing Theory. Washington University in Saint Louis Saint Louis.
[7] Little J. D. C (1961), a proof for Queuing formula, operations research volume 9(3).
[8] Rust K.,(2008) "Using Little's Law to estimate cycle time and cost", proceedings of the 2008 winter simulation conference. IEEE press.
[9] Zukerman M. (2013), Introduction to Queuing Theory and Stochastic Teletraffic Models. arXiv:1307.2968v3. Reprint.