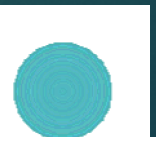
Queueing Theory for Data Science

INTRODUCTION

- we will study a class of models in which
  - Customers arrive in some random manner at a service facility
  - Upon arrival they are made to wait in a queue until it is their turn to be served.
  - Once served they are generally assumed to leave the system.
- For such models we will be interested in determining, among other things, such quantities as
  - Average number of customers in the system (or in the queue).
  - Average time a customer spends in the system (or spends waiting in the queue).

Example

- In general, the queueing system consists of one or more queues and one or more servers, and operates under a set of procedures.
- Let us consider the reservation counter of an airlines where customers from different parts of the world /country arrive and wait at the reservation counter.
- Depending on the served status, the incoming customer either waits at the queue or gets the turn.
- If the system is free at the time of arrival of a customer, the customer can directly enter into the counter for getting service and then leave the system.



Example

- In this process, over a period of time, the system may experience 'customer waiting' and/or 'server idle time'.
- In any service system/manufacturing system involving queueing situation.
- The objective is to design the system in such a manner that the average waiting time of the customers is **minimized** and the percentage **utilization** of the server is maintained above a desired level.

More Examples

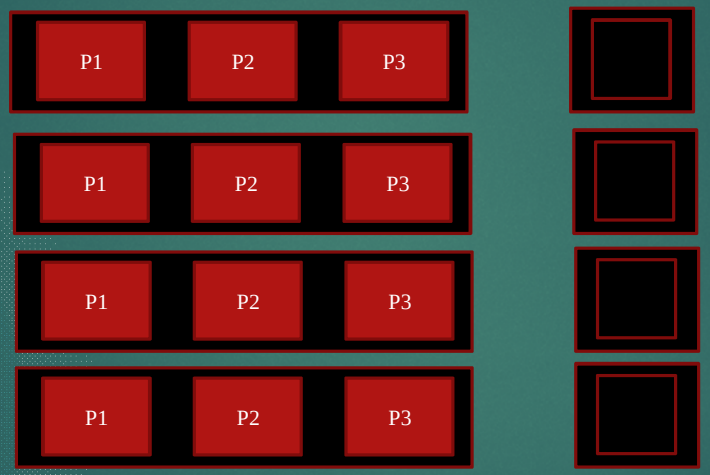
Example	Member of Queue	Server(s)
Bank Counter	Account Holders	Counter Clerk
Toll Gate	Vehicles	Toll Collectors
Library	Students	Counter Clerk
Traffic Signal	Vehicles	Signal Point
Airport runways	Planes	Runways
Maintenance Shop	Breakdown Machines	Mechanics

Diagram: Server Single Server Queueing Model

- We represent a simple queueing system as a diagram.

Server Single Server Queueing Model

Diagram: Multi Server Queueing Model

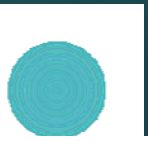


Multi Server Queueing Model

- Customers entering the system should form a queue and if the server is free, then they will get the service. If the server is not free, they wait in the queue and then they get the service. But there are real time situations where there may be variations in the system as follows:
  1. The number of Queues may be more than one.
  2. The number of servers may be more than one. This case is an example of parallel counters for providing service.

CHARACTERISTICS OF QUEUEING PROCESS

- In most cases, six basic characteristics of queueing processes provide an adequate describing of a Queueing system
  1. Arrival pattern of customers.
  2. Service pattern of servers.
  3. Queue discipline.
  4. System capacity
  5. Number of service channels
  6. Number of service stages.



Arrival Pattern of Customers:

- In usual queueing the process of arrivals is stochastic and it is necessary to know the probability distribution describing the times between successive customer arrivals (inter arrival times). It is also necessary to know whether customers can arrive simultaneously (batch or bulk arrivals) and of to the probability distribution describing the size of the batch.
- The assumption regarding the distribution of arrival rate has a great effect upon the Mathematical model. A typical model used is that the arrival rate is randomly distributed according to the Poisson distribution.
- **Mean value of the arrival rate is represented by λ.**

Service (Departure/Patterns):

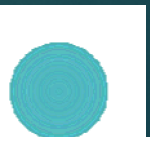
- It represents the pattern in which the number of customers leaves the system.
- Departures may also be represented by the service (inter departure) time. which is the time period between two successive services.
- **The number of customers served per unit of time is called service rate.**
- This rate assumes the service channel to be always busy i.e. no idle time is allowed.
- The assumption regarding the distribution of service is very important in forming a Queueing model. A general assumption used in most of the models is that the service time is randomly distributed according to **exponential distribution**. **Mean value of service rate is denoted by** λ.

Service Channels

- The Queueing System may have single service channel.
- The customers may form a Queue in the clinic.
- Note that the system can also have a number of service channels where the customers may be arranged in parallel or in series or complex combination of both.

Service Discipline:

- Service Discipline is the rule by which the Customers are selected from the queue for service.
- The common discipline is **“first come” “first served”** pattern.
- All The real time situations are coining only in this service discipline.
- The other service disciplines include **“random”** and **“Priority”**.
- In “random” pattern the customers are chosen randomly from the queue.
- “Priority” pattern is a pattern where a customer is chosen head of some other customer in the queue.



Maximum Number of customers allowed in the system

- Maximum Number of customers in the system can be either infinite or finite.
- In some facilities, only a limited number of customers are allowed in the system.
- But in general, number Of customers in the system is infinity.

Calling Source or Population

- The arrival pattern of the customers depends upon the source, which generates them.
- If there are only it few potential customers, the calling source (population) is called finite.
- If there are a large number of potential customers (say over 50 or 60), it is usually said to be infinite.

CLASSIFICATION OF QUEUEING MODELS

- Generally Queueing models may be completely specified in the following symbol form: (a/b/c): (d/e), where
  - a = Probability law for the arrival (or inter arrival) time,
  - b = Probability law according to which the customers are being served.
  - c = Number of Service stations
  - d = The maximum number allowed in the system (in service and waiting)
  - e = Queue Discipline The above notation is called Kendal's Notation.

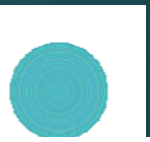
Notation

1	$p_n(t)$	Probability that the size of the population at time 't'
2	P_0	Probability of system being idle or free.
3	ρ	Traffic intensity
4	L_s	Average (or) mean number of customer in the system.
5	L_q	Average (or) mean number of customer in the queue.
6	L_w	Average (or) mean number of customer in the non-empty queues.
7	W_s	Average waiting time of a customer in the system.
8	W_q	Average waiting time of a customer in the queue.

Model- 1: (M/M/1): (∞ /FIFO)

$$P_0 = 1 - \frac{\lambda}{\mu}$$

- $P_0 = 1 - \frac{\lambda}{\mu}$ denotes the probability of system being idel. (i.e.,) the system , is free.
- The quantity $\frac{\lambda}{\mu} = \rho$ is called traffic intensity



Model- 1: (M/M/1): (∞ /FIFO)

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

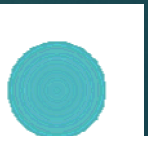
Model- 1: (M/M/1): (∞ /FIFO)

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

- By the average number of customers 1, in the system we mean the number of customers in the queue + the person who is getting serviced.
- To calculate the expected number of customers or average number of customers in the system , we can use the formula given

$$\frac{\rho}{\{1-\rho\}} \text{ or } \frac{\lambda}{\mu - \lambda}$$

- The average number of customers in the system can also be denoted as E (N)



Model- 1: (M/M/1): (∞ /FIFO)

$$L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

- By the average number of customers in the queue we mean the number of customers in the queue excluding the person who is getting serviced.
- To calculate the expected number of customers or average number of customers in the queue we can use the formula given

$$\frac{\rho^2}{(1-\rho)} \quad or \quad \frac{\lambda^2}{\mu(\mu-\lambda)}$$

- The average number if customer in the system can also denoted $E(N_q)$.
- The value of L_q can also be calculated by $L_q = L_s - \frac{\lambda}{\mu}$

Model- 1: (M/M/1): (∞ /FIFO)

L_w	$\frac{\mu}{\mu - \lambda}$
W_s	$\frac{1}{\mu - \lambda}$
W_q	$\frac{\lambda}{\mu(\mu - \lambda)}$
P(N>K)	ρk (For a queue) $\rho k + 1$ (For system)
$P(W_s > t)$	$e^{-(u-\lambda)t}$
ρ	$\frac{\lambda}{\mu}$

Example

- Arrival rate of telephone calls at a telephone booth is according to Poisson Distribution with an average time of 9 minutes between two consecutive arrivals. The length of Telephone call is assumed to be exponentially distributed with mean 3 minutes.
  - Determine the Probability that a person arriving at the booth will have to wait.
  - Find the average Queue length.
  - The telephone company will install a second booth when convinced that an arrival would expect to have to wait at least four minutes for the phone. Find the increase in flow of arrivals which will justify a second booth.
  - What is the Probability that an arrival will have to wo for more than 10 minutes before the Phone is free?
  - What is the Probability that an arrival will have to wait for more than 10 minutes before the phone is available and the call is also complete?
  - Find the fraction of a day that the Phone will be in use.

Solution

- **Step 1:Model Identification**
  - Since there is only one telephone booth, the number of service channels is one. Also, since any number of customers can enter the booth, the capacity of the system is infinity. Hence this problem comes under the mode, (M/M/1):(∞/FCFS).
- **Step 2:Given Data**
  - Arrival rate, $\lambda = 1/9$ per minute
  - Service rate, $\mu = 1/3$ per minute

Solution

- **Step 3: To find the following**
  - The Probability, that a person arriving at the booth will have to wait i.e. ρ
  - The average Queue length i.e., L_s
  - The increase in flow of arrivals which will justify a second booth (We need to apply a different technique)
  - The Probability that an arrival will have to wait for more than 10 minutes before the Phone is free i.e. $P(W_s > 10)$
  - The Probability that an arrival will have to wait fat more than 10 minutes before the phone is available and the call is also complete i.e., Probability [time in system > 10]

Solution

- **Step 4: Required computations**
- Probability that a person will have to wait

$$\rho = \frac{\lambda}{\mu} = \frac{1/9}{1/3} = 0.33$$

Average queue length

$$L_w = \frac{\mu}{\mu - \lambda}$$

$$\frac{1/3}{\frac{1}{3} - \frac{1}{3}} = \frac{\frac{1}{3}}{\frac{2}{9}}$$

Solution

$$\frac{1}{3} \times \frac{9}{2} = \frac{9}{6}$$

$$=1.5$$

Average waiting time in the queue

$$W_q = \frac{\lambda_1}{\mu(\mu - \lambda_1)}$$

[we use λ_1 here as we have used "λ" already]

$$\Rightarrow \qquad 4 = \frac{\lambda_1}{\frac{1}{3}(\frac{1}{3} - \lambda_1)}$$

[The waiting time to install a second booth is 4 minute]

$$\Rightarrow \quad \frac{4}{9} - \frac{4\lambda_1}{3} = \lambda_1$$

$$\Rightarrow \quad \frac{4}{9} = \lambda_1 + \frac{4\lambda_1}{3}$$

$$\Rightarrow \quad \frac{4}{9} = \frac{7\lambda_1}{3}$$

Solution

- $\Rightarrow \quad \frac{4}{3} = 7\,\lambda_1$

$\Rightarrow \quad \lambda_1 = \frac{4}{21}$ arrivals/minute

$\therefore$ Increase in flow of arrivals $\lambda_1 - \lambda = \frac{4}{21} - \frac{1}{9} = 5/63$ per minute

Solution

- Probability of waiting time more than 10 minutes

$$=P(W_s>10)$$

$$= \int_{10}^{\infty} \frac{\lambda}{\mu}(\mu-\lambda)e^{-(\mu-\lambda)t}dt$$

$$(p.d.f)of\ a\ exponentional\ distribution)$$

$$= \frac{\lambda}{\mu}\int_{10}^{\infty} (\mu-\lambda)e^{-(\mu-\lambda)t}dt$$

$$= \frac{\lambda}{\mu}\ \cancel{(\mu-\lambda)}\left[\frac{e^{-(\mu-\lambda)t}dt}{-\cancel{(\mu-\lambda)}}\right]_{10}^{\infty}$$

Solution

$$= \frac{\lambda}{\mu}\left[0 + e^{-10(\mu-\lambda)}\right]$$

$$= \frac{\frac{1}{9}}{\frac{1}{3}} e^{-10(\frac{1}{3}-\frac{1}{9})}$$

$$= \frac{1}{3} e^{-10(\frac{2}{9})}$$

$$= \frac{1}{3} e^{-10(\frac{20}{9})} = \frac{1}{30} \text{ (approximately)}$$

Solution

P[Time in system ≥ 10]

$$= \int_{10}^{\infty} (\mu - \lambda)\, e^{-(\mu-\lambda)t}.\, dt$$

$$(p.d.f) of\ a\ exponentional\ distribution)$$

$$= \frac{\mu}{\lambda}\left\{\frac{\lambda}{\mu}\int_{10}^{\infty} (\mu - \lambda) e^{-(\mu-\lambda)t} dt\right\}$$

$$= \frac{\mu}{\lambda}\left(\frac{1}{30}\right) = \frac{\frac{1}{3}}{\frac{1}{9}}\left(\frac{1}{30}\right) \text{ by(iv)}$$

Solution

- The expected fraction of day that the phone will be in use is equal to

$$= \frac{\lambda}{\mu}$$

$$= \frac{\frac{1}{9}}{\frac{1}{3}} = 0.33$$

Example 2

- Customers arrive at the first class ticket counter of a theatre at a rate of 12 per hour. There is one clerk servicing the customers at the rate of 30 per hour.
  - What is the Probability that there is no customer at the counter?
  - What is the Probability that there are more than 2 customers at the counter?
  - What is the Probability that there is no customer wailing to be served?
  - What is the Probability that It customer is being served and nobody is waiting?

Solution

- **Step 1:Model Identification**
- Since there is only one clerk, the number of service channel is one.
- Also, since any number of persons can come to the counter, the capacity of the system is infinity.
- Hence this problem comes under the model (M/M/1):(∞/FCFS)

Solution

- **Step 2 : Given Data**

  Arrival rate, λ = 12 per hour

  Service rate μ = 30 per hour

solution

- **Step 3: To find the following**
  - Probability that there is no customer at the counter i.e., P_o
  - Probability that there are more than two customers in the counter i.e., P (N > 2)
  - Probability that there is no customer waiting to be served i.e., $P_o + P_1$
  - Probability that a customer is being served and nobody is waiting i.e., P_1

Solution

- **Step 4: Required Computations**
  - Probability that there is no customer at the counter.

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{12}{30} = \frac{18}{30}$$

  - Probability that there are more than two customer in the counter is given by

$$= P(N>2) = 1- \{P(N\leq 2)\}$$

$$= 1-\{P(N=0)+P(N=1)+P(N=2)\}$$

$$= 1 - [P_0 + \rho P_0 + \rho^2 P_0]$$

$$= 1 - (1 + \rho + \rho^2) P_0$$

Solution

$$= 1 - \left[1 + \frac{2}{5} + \left(\frac{2}{5}\right)^2\right] 0.6$$

$$\left[\therefore \rho = \frac{\lambda}{\mu} = \frac{12}{30} = \frac{2}{5}\right]$$

$$= 1 - 0.936 = 0.064$$

Solution

- Probability that there is no customer waiting to be served = Probability that almost one customer at the counter.

$$P(N \leq 1) = P_0 + P_1$$

$$= 0.6 + \rho . P_0 = 0.6 + 0.24$$

$$\mathbf{= 0.84}$$

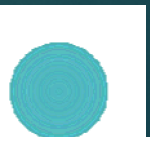
- Probability that a customer is being served and nobody is waiting is given by

$$P_1 = \rho P_0$$

$$= 0.4 \times 0.6 = \mathbf{0.24}$$

Example 3

- **In a super market, the average arrival rate of a customer is 10 in every 30 minutes following a Poisson process. The average time taken by the Manager to list and calculate the Purchase is 2.5 minutes 2.5 minutes, which is exponentially distributed.**
- What is the probability that the queue length exceed 6?
- What is the expected time spent by a customer in the system?



Solution

- **Step 1: Model Identification**
- Since there is only one manager, the number of service channel is one. Also, since any number of person can come to the supermarket, the capacity of the system is infinite. Hence this problem come under the model (M/M/1): (∞/FCFS)
- **Step 2: Given data**

$$\text{Arrival rate, } \boldsymbol{\lambda} = \frac{\mathbf{10}}{\mathbf{30}} Minutes = \frac{1}{3}\ minutes$$

$$\text{Service rate, } \mu = \frac{1}{2.5}\ per\ minute$$

$$\therefore Traffic\ Intensity = \rho = \frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}$$

$$\frac{\frac{1}{3}}{\frac{1}{25}} = 0.8333$$

- **Step 3: To find the following**
  - The probability that the queue size exceed 6 i.e., ρ^6 (or) $\left(\frac{\lambda}{\mu}\right)^6$
  - Waiting time by customer in the system i.e., W_s
- **Step 4 : Required computations**
  - The probability that the queue size exceed *n* is given by ρ^n (or) $\left(\frac{\lambda}{\mu}\right)^n$

$\therefore$ The probability that the queue size exceed 6 = ρ^6

$(0.8333)^6 = 0.3348$

Solution

Waiting time by the customer in the system

W_s = Average number of units in the system = $\frac{1}{\mu-\lambda}$

$$= \frac{1}{\frac{1}{2.5} - \frac{1}{3}}$$

$$= \frac{7.5}{0.5}$$

= 15 minutes