Unit 6: Stochastic process

Simulation of stochastic process:

A number of important characteristics of stochastic processes require lengthy complex computations unless they are estimated by means of Monte Carlo methods. One may be interested to explore the time it takes a process to attain a certain level, the time the process spends above some level or above another process, the probability that one process reaches a certain value ahead of another process, etc. Also, it is often important to predict future behavior of a stochastic process. Simulation is carried out for discrete time process, market chain continuous time process, Poisson process etc.

Queuing Systems:

Definition: A queuing system is a facility consisting of one or several servers designed to perform certain tasks or process certain jobs and a queue of jobs waiting to be processed.

Jobs arrive at the queuing system, wait for an available server, get processed by this server, and leave.

Examples of queuing systems are:

i. a personal or shared computer executing tasks sent by its users;
ii. an internet service provider whose customers connect to the internet, browse, and disconnect;
iii. printer processing jobs sent to it from different computers; – a customer service with one or several representatives on duty answering calls from their customers;
iv. a TV channel viewed by many people at various times;
v. a toll area on a highway, a fast food drive-through lane, or an automated teller machine (ATM) in a bank, where cars arrive, get the required service and depart;
vi. a medical office serving patients; and so on.

Features of queue:

Calling population: It is finite population with independent arrivals and not influenced by queuing system.

Arrival process: Arrival rate follows Poisson distribution with parameter λ.

Queuing configuration: Queue is single waiting line with unlimited space.

Queue discipline: Queue discipline is based upon first come first serve (FCFS).

Service process: Service rate follows exponential distribution with parameter μ.

Main components of a queuing system:

Arrival

Typically, jobs arrive to a queuing system at random times. A counting process A (t) tells the number of arrivals that occurred by the time t. In stationary queuing systems (whose distribution characteristics do not change over time), arrivals occur at arrival rate

$$\lambda_A = \frac{EA(t)}{t}$$

for any $t > 0$, which is the expected number of arrivals per 1 unit of time. Then, the expected time between arrivals is

$$\mu_A = \frac{1}{\lambda_A}$$

Queuing and routing to servers:

Arrived jobs are typically processed according to the order of their arrivals, on a "first come–first serve" basis.

When a new job arrives, it may find the system in different states. If one server is available at that time, it will certainly take the new job. If several servers are available, the job may be randomized to one of them, or the server may be chosen according to some rules.

For example, the fastest server or the least loaded server may be assigned to process the new job. Finally, if all servers are busy working on other jobs, the new job will join the queue, wait until all the previously arrived jobs are completed, and get routed to the next available server.

Service:

Once a server becomes available, it immediately starts processing the next assigned job. In practice, service times are random because they depend on the amount of work required by each task. The average service time is μ_s. It may vary from one server to another as some computers or customer service representatives work faster than others. The service rate is defined as the average number of jobs processed by a continuously working server during one unit of time. It equals

$$\lambda_S = \frac{1}{\mu_s}$$

Departure: When the service is completed, the job leaves the system.

The following parameters and random variables describe performance of a queuing system.

Parameters of a queuing system:

λ_A = arrival rate

λ_S = service rate

$\mu_A = 1/\lambda_A$ = mean interarrival time

$\mu_S = 1/\lambda_S$ = mean service time

$r = \lambda_A/\lambda_S = \mu_S/\mu_A$ = utilization, or arrival-to-service ratio

Random variables of a queuing system:

$X_s(t)$ = number of jobs receiving service at time t

$X_w(t)$ = number of jobs waiting in a queue at time t

$X(t) = X_s(t) + X_w(t)$, the total number of jobs in the system at time t

S_k = service time of the k^{-th} job

W_k = waiting time of the k^{-th} job

$R_k = S_k + W_k$, response time, the total time a job spends in the system from its arrival until the departure

N = Number of customers in the system (waiting and in service)

L_s = Mean (average) number of customers in the system.

L_q = Mean (average) number of customers in the queue.

L_b = Mean (average) length of non empty queue.

W_s = Mean waiting time in the system.

W_q = Mean waiting time in the queue.

P_w = Probability that an arriving customer has to wait.

System with limited capacity

Little's Law:

The Little's Law gives a simple relationship between the expected number of jobs, the expected response time, and the arrival rate. It is valid for any stationary queuing system.

The Little's Law certainly applies to the M/M/1 queuing system and its components, the queue and the server. Assuming the system is functional ($r < 1$), all the jobs go through the entire system, and thus, each component is subject to the same arrival rate λ_A. The Little's Law then guarantees that

$$\lambda_A E(R) = E(X),$$

$$\lambda_A E(S) = E(X_S),$$

$$\lambda_A E(W) = E(X_W).$$

Bernoulli single-server queuing process:

DEFINITION: Bernoulli single-server queuing process is a discrete-time queuing process with the following characteristics: – one server, unlimited capacity, arrivals occur according to a Binomial process, and the probability of a new arrival during each frame is p_A, the probability of a service completion (and a departure) during each frame is p_S provided that there is at least one job in the system at the beginning of the frame, service times and inter arrival times are independent.

Binomial counting processes applies to arrivals of jobs. It also applies to service completions all the time when there is at least one job in the system. We can then deduce that

– there is a Geometric(p_A) number of frames between successive arrivals;

– each service takes a Geometric(p_S) number of frames;

– service of any job takes at least one frame;

– $p_A = \lambda_A \Delta$;

– $p_S = \lambda_S \Delta$.

Moreover, Bernoulli single-server queuing process is a homogeneous Markov chain because probabilities pA and pS never change. The number of jobs in the system increments by 1 with each arrival and decrements by 1 with each departure. Conditions of a Binomial process guarantee that at most one arrival and at most one departure may occur during each frame. Then, we can compute all transition probabilities,

$$P_{00} = P\{\text{ no arrivals }\} = 1-p_A$$

$$P_{01} = P\{\text{ new arrival }\} = p_A$$

and for all $i \geq 1$,

$p_{i,i-1} = P\{\text{ no arrivals} \cap \text{one departure }\} = (1-p_A)p_S$

$p_{i,i}$ = P { no arrivals ∩ no departures } + P { one arrival ∩ one departure } = $(1-P_A)(1-P_S) + P_A P_S$

$P_{i,i+1}$ = P { one arrival ∩no departures } = $P_A(1-P_S)$

Now, the transition probability matrix is

$$P = = \begin{bmatrix} 1-P_A & P_A & 0 & \cdots & \cdots \\ (1-P_A)P_S & (1-P_A)(1-P_S)+P_AP_S & P_A(1-P_S) & \cdots & \cdots \\ 0 & (1-P_A)P_S & (1-P_A)(1-P_S)+P_AP_S & \cdots & \cdots \\ 0 & 0 & (1-P_A)P_S & \vdots & \ddots \\ \vdots & \vdots & \vdots & & \end{bmatrix}$$

<u>Example 13:</u>

Laptop computers arrive at a repair at the rate of four per day. Assume an 8-hour working day. The expected time to complete service on a laptop is 1.25 hours. Model this process as a single-server Bernoulli queuing process with 15-minute frames. (a) Find the service rate. (b) Find the arrival and service probabilities.

<u>Solution:</u> Here;

λ_A = 4 per day = 4 per 8 hour = 0.5 per hour

Δ = 15 minutes = 15/60 hour = 0.25 hrs

Service time of 1 laptop = 1.25 hrs.

∴1 hr = 1/1.25 laptop

λ_S = 0.8 per hour

$P_A = ?$

$P_S = ?$

Now,

$P_A = \lambda_A \Delta = 0.5 \times 0.25 = 0.125$

$P_S = \lambda_S \Delta = 0.8 \times 0.25 = 0.2$

<u>Example 14:</u>

A barbershop has one barber and two chairs for waiting. The expected time for a barber to cut customer's hair is 15 minutes. Customers arrive at the rate of two per hour provided the barbershop is not full. However, if the barbershop is full (three customers), potential customers go elsewhere. Assume that the barbershop can be modeled as single-server Bernoulli queuing process with limited capacity. Use frame size of 3 minutes. (a) Derive the one-step transition probability matrix for this process. (b) Find steady-state probabilities and interpret them.

<u>Solution:</u> Here;

Service time for 1 customer = 15 minutes

Service time for 4 customers = 1 hr

Hence, λ_S =4 per hour

Arrival of customers = 2 per hour

Hence, λ_A =2 per hour

Frame size $\triangle$ = 3 minutes = 3/60 = 1/20 hr = 0.05 hr

Capacity C = 3

$P_A = \lambda_A \triangle = 2 \times 0.05 = 0.1$

$P_S = \lambda_S \triangle = 4 \times 0.05 = 0.2$

$P_{00} = 1 - P_A = 1 - 0.1 = 0.9$

$P_{01} = P_A = 0.1$

For all $i \geq 1$

$P_{i,i-1} = (1-P_A) P_S = (1-0.1) \times 0.2 = 0.18$

$P_{i,i} = (1-p_A)(1-P_S) + P_A P_S$

$= 0.9 \times 0.8 + 0.1 \times 0.2 = 0.72 + 0.02 = 0.74$

$P_{i,i+1} = P_A (1-P_S) = 0.1 (1-0.2) = 0.08$

Now;

Transition probability matrix is

$$P = \begin{bmatrix} 0.9 & 0.1 & 0 & 0 \\ 0.18 & 0.74 & 0.08 & 0 \\ 0 & 0.18 & 0.74 & 0.08 \\ 0 & 0 & 0.18 & 0.82 \end{bmatrix}$$

For steady state distribution,

Now;

$\pi P = \pi$

M/M/1 System:

DEFINITION: An M/M/1 queuing process is a continuous-time Markov queuing process with the following characteristics,

– one server;

– unlimited capacity;

– Exponential inter arrival times with the arrival rate λ_A;

– Exponential service times with the service rate λ_S;

– service times and inter arrival times are independent.

First, let us explain what the notation “M/M/1” actually means.

A queuing system can be denoted as A/S/n/C, Where;

A	denotes the distribution of inter arrival times
S	denotes the distribution of service times
n	is the number of servers
C	is the capacity Default capacity is $C = \infty$ (unlimited capacity)

Letter M denotes Exponential distribution because it is memory less, and the resulting process is Markov.

We study M/M/1 systems by considering a Bernoulli single-server queuing process and letting its frame Δ go to zero. Our goal is to derive the steady-state distribution and other quantities of interest that measure the system's performance. When the frame Δ gets small, its square Δ^2 becomes practically negligible, and transition probabilities for a Bernoulli single-server queuing process can be written as

$$P_{00} = 1 - P_A = 1 - \lambda_A \Delta$$

$$P_{10} = P_A = \lambda_A \Delta$$

For; $i \geq 1$

$$P_{i,\,i-1} = (1 - P_A)\, P_S$$

$$= (1 - \lambda_A \Delta)\, \lambda_S \Delta$$

$$= \lambda_S \Delta - \lambda_A \lambda_S \Delta^2$$

$$P_{i,\,i+1} = P_A (1 - P_S)$$

$$= \lambda_A \Delta\, (1 - \lambda_S \Delta)$$

$$= \lambda_A \Delta - \lambda_A \lambda_S \Delta^2$$

$$= \lambda_A \Delta$$

$$P_{i,\,i} = (1 - P_A)\,(1 - P_S) + P_A P_S$$

$$= (1 - \lambda_A \Delta)\,(1 - \lambda_S \Delta) + \lambda_A \Delta \lambda_S \Delta$$

$$= 1 - \lambda_A \Delta - \lambda_S \Delta + \lambda_A \lambda_S \Delta^2 + \lambda_A \lambda_S \Delta^2$$

$$= 1 - \lambda_A \Delta - \lambda_S \Delta$$

Transition probability matrix is

$$P = \begin{bmatrix} 1-\lambda_A\Delta & \lambda_A\Delta & 0 & 0 & \cdots\cdots \\ \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \lambda_A\Delta & 0 & \cdots\cdots \\ 0 & \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \lambda_A\Delta & \cdots\cdots \\ 0 & 0 & \lambda_S\Delta & 1-\lambda_A\Delta-\lambda_S\Delta & \cdots\cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Steady state distribution of **M/M/1** system.

$$\pi_0 = 1 - P$$

$$\pi_1 = P\,(1-P)$$

$\pi_2 = P^2 (1-P)$; and so on.

Evaluating the system's performance:

Many important system characteristics are as follows:

Utilization rate (r) $= \frac{\lambda_A}{\lambda_S} = P$

Idle rate = 1 – utilization rate = $1 - \frac{\lambda_A}{\lambda_S} = 1-P$

Probability of no customer in queue

$$P_0 = 1 - \frac{\lambda_A}{\lambda_S} = 1-P$$

Probability of one customer in queue

$$P_1 = P\, P_0 = P\,(1-P)$$

Probability of one customers in queue

$$P_2 = P\, P_1 = P^2\,(1-P)$$

Probability of n-customers in queue

$$P_n = P^n\,(1-P);\ P<1,\ n = 0, 1, 2, 3, 4, 5, \dots \dots \dots$$

Probability of server being busy $= 1-P_0 = P$

Expected (average) number of customers in the system

$$L_S = \frac{Utiliation\ rate}{Idle\ rate} = \frac{P}{1-P}$$

Expected queue length (Expected number of customers waiting in the queue)

$$L_q = L_S - \text{Utilization factor}$$

$$= L_S - P = \frac{P}{1-P} - P = = \frac{P-P(1-P)}{1-P} = = \frac{P^2}{1-P}$$

Expected (average) waiting time of a customer in the queue

$$W_q = = \frac{Average\ number\ of\ customer\ in\ queue}{Arrival\ rate} = = \frac{L_q}{\lambda_A}$$

Expected (Average) waiting time of a customer in the system

$$W_b = \frac{1}{\lambda_s - \lambda_A}$$

Probability of k or more customers in the system

$$P\,(n \geq k) = P^K$$

Variance of queue length

$$V\,(n) = \frac{P}{(1-p)^2}$$

Expected number of customers several per busy period

$$L_b = \frac{L_s}{1-P_0} = \frac{1}{1-P}$$

Expected length of non empty queue

$$L_q' == \frac{Expected\ number\ of\ customers\ in\ the\ queue}{P(more\ than\ one\ customer\ in\ queue)} == \frac{L_q}{P(n>1)}$$

Thank you!!!