MICT-5101: Probability and Stochastic Process[1]

Dr. Md. Rezaul Karim
PhD(KULeuven & UHasselt), MS(Biostatistics), MS(Statistics)
Professor, Department of Statistics and Data Science
Jahangirnagar University (JU), Savar, Dhaka - 1342, Bangladesh

MS-2024

[1]These course slides should not be reproduced nor used by others (without permission).

Lecture Outline I

Introduction

1 Introduction

1.1 Text & Reference Book List

Text Book

1. Ross, S. (2010): *Introduction to Probability Models*, 10th edition, Pearson, Prentice Hall.
2. Anthony J. Hayter (2012): *Probability and Statistics for Engineers and Scientists* 4th Edition, Duxbury Press.

Reference Book List

1. Mehdhi, J. (2009): *Stochastic Processes* , 3rd Revised Edition, New Age Science.
2. Beichelt F. (2016): *Applied probability and stochastic processes*, 7th edition, CRC Press.
3. Ross, S. (2020): *Introduction to Probability and Statistics for Engineers and Scientists*, 6th Edition, Pearson Education Inc.

Fundamentals of Probability Models

1. **Part I:** Probability Theory
   - Basic Concepts of Probability
   - Random Variable
   - Expectation
   - Some Probability Distributions
     - Bernoulli
     - Binomial
     - Poisson
     - Uniform and
     - Normal
     - exponential
     - ...
2. **Part II:** Stochastic Processes
   - Basics of Stochastic Processes
   - Random Point Processes
   - Discrete-Time Markov Chains
   - ...

Chapter 4: Queuing Process

2 Chapter 4: Queuing Process

What is Queuing Process?

A **queuing process** is a mathematical model used to describe systems in which entities (such as customers, tasks, or data packets) wait in line for service or processing. It typically involves:

- A **queue** (waiting line),
- A **server** (the resource providing service),
- A set of rules governing the arrival of entities, their waiting times, and the service mechanism.

Queuing processes are often characterized by parameters such as:

- The arrival rate of entities,
- The service rate,
- The number of servers,
- The number of entities allowed in the system.

These processes are widely used in fields like operations research, telecommunications, computer networks, and traffic flow to analyze performance, optimize resources, and predict system behavior.

Examples of Queuing Processes

Here are a few common examples of queuing systems:

- **Bank Teller Queues:** Customers wait in line to be served by one or more tellers. The arrival rate of customers and service time influence the overall waiting time.
- **Call Centers:** Calls arrive at a call center where agents serve customers. The queue length, number of agents, and service times impact customer satisfaction.
- **Internet Data Traffic:** Packets of data wait in line to be transmitted through routers or switches. The network bandwidth and packet arrival rate are key factors.
- **Supermarket Checkout Lines:** Shoppers wait in line at checkout counters. The number of counters and the time it takes to scan and bag items determine the queue dynamics.

Components of a Queuing Process

A queuing process typically involves several key components that determine its behavior and performance. The main components are as follows:

1. **Arrival Process:**
   The arrival process defines how customers or tasks arrive at the system. This is often modeled as a stochastic (random) process. Common models include:
   - Poisson Process: A common model where arrivals occur randomly and independently over time with a constant average arrival rate, denoted λ (customers per unit time).
2. **Service Process:**
   The service process describes how the server(s) handle the entities in the system. It typically involves the following:
   - μ: The service rate (the rate at which customers are served, typically in customers per unit time).
   - Exponential Distribution: The service times are often assumed to follow an exponential distribution in the M/M/1 queue.

3. **Queue Discipline:**
   The queue discipline refers to the rules that determine the order in which entities are served. Common types include:
   - First-Come, First-Served (FCFS): Entities are served in the order they arrive.
   - Priority Queueing: Entities with higher priority are served first.
   - Last-Come, First-Served (LCFS): The last arriving entity is served first (less common).
4. **Number of Servers:**
   The number of servers in the system can affect the queue behavior:
   - Single Server: A single server handles all arriving entities.
   - Multiple Servers: Multiple servers handle the entities simultaneously (e.g., M/M/c queue).
5. **Queue Capacity:**
   The queue capacity defines how many entities can be in the system (including both in service and in the queue). This could be:
   - ∞: An infinite queue capacity (most common assumption).
   - K: A finite queue capacity, where the system can hold at most K entities (including those in service).

6. **System Capacity:**
   The total capacity of the system may also be limited. For example:
   - ∞: Unlimited capacity (no limit on the number of entities that can be in the system).
   - K: A system with a finite capacity, meaning the system can hold at most K entities in total.
7. **Service Time Distribution:**
   The service times are often modeled using various probability distributions, including:
   - Exponential Distribution: Common for service time in M/M/1 systems.
   - General Distribution: More complex systems may use arbitrary service time distributions.

Common Queuing Models (Kendall Notation)

Queuing systems are commonly described using Kendall's notation: A/B/C

- A: Arrival process (e.g., M for Markovian, D for deterministic)
- B: Service time distribution (e.g., M for exponential, D for deterministic)
- C: Number of servers (e.g., 1 for single-server, c for multi-server)

Examples:

- M/M/1: A simple model with exponential arrival times, exponential service times, and a single server.
- M/M/c: Similar to M/M/1, but with multiple servers available to handle customers.
- M/G/1 Queuing Model: Involves a single server, exponential inter-arrival times, and general service times (not necessarily exponential).
- G/G/1 Queuing Model: A more generalized model, where both arrival and service processes follow arbitrary distributions.

M/M/1 Queuing System

The M/M/1 queue is one of the simplest and most widely used queuing models. It assumes:

- M: Markovian (Poisson) arrival process with rate λ.
- M: Markovian (Exponential) service time with rate μ.
- 1: A single server.

The system is characterized by:

- Arrival rate λ, service rate μ, and server utilization $\rho = \frac{\lambda}{\mu}$.
- The queue discipline is typically FIFO (First-In-First-Out).

In the M/M/1 queuing model, the number of entities in the system (queue + service), n, follows a Geometric Distribution with the following probability mass function (PMF):

$$P_n = (1-\rho)\cdot\rho^n, \quad n = 0, 1, 2, \ldots$$

where:

- $\rho = \frac{\lambda}{\mu}$ is the utilization factor. ρ must satisfy $\rho < 1$ for stability (i.e., $\lambda < \mu$).
- P_n is the probability that there are exactly n customers in the system at any given time.

Interpretation:

- $P_0 = (1-\rho)$: The probability that there are no customers in the system (i.e., the system is empty).
- $P_1 = (1-\rho)\cdot\rho$: The probability that there is exactly 1 customer in the system.
- $P_2 = (1-\rho)\cdot\rho^2$: The probability that there are exactly 2 customers in the system.

What is ρ in the M/M/1 Queuing Model?

In the M/M/1 queuing model, ρ (the utilization factor) is the ratio of the arrival rate to the service rate. It represents the fraction of time that the server is busy.
The utilization factor ρ is defined as:

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

where:

- λ is the arrival rate, i.e., the rate at which customers arrive at the system (typically in customers per unit time).
- μ is the service rate, i.e., the rate at which the server can serve customers (typically in customers per unit time).

Interpretation of ρ:

- $\rho = 1$: The system is fully utilized. The server is always busy, with no idle time.
- $\rho < 1$: The system is under-utilized. The server has idle time, and the queue tends to empty over time.
- $\rho > 1$: The system is overloaded. The arrival rate exceeds the service rate, leading to an ever-growing queue, causing instability.

Stability Condition: For the system to remain stable (i.e., to avoid an infinitely growing queue), the arrival rate must be less than the service rate:

$$\rho = \frac{\lambda}{\mu} < 1 \quad \Rightarrow \quad \lambda < \mu$$

When $\rho \geq 1$, the system becomes unstable, and the queue will grow without bound.

Properties of M/M/1 Queuing Model

Key properties of the M/M/1 queue include:

- **Utilization**: The server utilization is given by:

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

where λ is the arrival rate and μ is the service rate. The server is fully utilized when $\rho = 1$.

- **Average Number of Customers in the System** (L): The expected number of customers in the system (including those in service) is:

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

- **Average Time in the System** (W): The expected time a customer spends in the system (waiting + service time) is:

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

- **Average Number of Customers in the Queue** (L_q): The expected number of customers waiting in the queue (excluding the one being served) is:

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

- **Average Waiting Time in the Queue** (W_q): The expected time a customer spends waiting in the queue is:

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

- **Probability of n Customers in the System** (P_n): The probability of having exactly n customers in the system is:

$$P_n = (1 - \rho)\rho^n, \quad n = 0, 1, 2, \ldots$$

- **Stability Condition**: The system is stable (i.e., the queue does not grow indefinitely) if:

$$\lambda < \mu$$

These properties are essential for performance analysis and system design in various real-world applications like telecommunications, service systems, and computer networks.

Average Number of Entities in the System

- Average number of entities in the system (L):

$$L = E(N) = \sum_{n=0}^{\infty} n \cdot P_n = \sum_{n=0}^{\infty} n \cdot (1-\rho)\rho^n$$

Simplifying the series:

$$L = (1-\rho) \sum_{n=0}^{\infty} n \cdot \rho^n = (1-\rho) \cdot \frac{\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

Substituting $\rho = \frac{\lambda}{\mu}$, we get:

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

which is the average number of customers in the M/M/1 system.

Variance $V(N)$ of the Number of Entities in the System

The variance $V(N)$ of the number of entities in the system is given by:

$$V(N) = E[N^2] - (E[N])^2$$

where $E[N^2]$ is calculated as:

$$E[N^2] = \sum_{n=0}^{\infty} P_n \cdot n^2$$

Substituting the expression for P_n and performing the necessary simplifications, we find:

$$V(N) = \frac{\rho}{(1-\rho)^2} = \frac{\lambda}{(\mu-\lambda)^2}$$

This is the variance of the number of entities in the system.

Time Spent in the System

In an M/M/1 queue, the total time T an entity spends in the system is the sum of:

- The waiting time in the queue, W_q,
- The service time, S.

Therefore, the total time in the system is:

$$T = W_q + S$$

We need to find the density function of T, i.e., $f_T(t)$, where T is the total time spent in the system.

Step 1: Service Time Distribution

The service time S is exponentially distributed with rate μ, so its probability density function (PDF) is:

$$f_S(t) = \mu e^{-\mu t}, \quad t \geq 0$$

This describes the time a customer spends being served.

Step 2: Waiting Time Distribution

The waiting time in the queue, W_q, in the M/M/1 queue is exponentially distributed with rate $\mu - \lambda$, where λ is the arrival rate. The PDF of the waiting time in the queue is:

$$f_{W_q}(t) = (\mu - \lambda)e^{-(\mu-\lambda)t}, \quad t \geq 0$$

This describes how long a customer waits in the queue before being served.

Step 3: Convolution of the Densities

Since the total time spent in the system is the sum of two independent random variables, $T = W_q + S$, the PDF of T, denoted as $f_T(t)$, is the convolution of $f_{W_q}(t)$ and $f_S(t)$:

$$f_T(t) = \int_0^t f_{W_q}(\tau) f_S(t-\tau)\, d\tau$$

Substituting the expressions for $f_{W_q}(t)$ and $f_S(t)$, we get:

$$f_T(t) = \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)\tau} \mu e^{-\mu(t-\tau)}\, d\tau$$

Step 4: Simplifying the Integral

Factor out the terms that do not depend on τ:

$$f_T(t) = \mu(\mu - \lambda)e^{-\mu t} \int_0^t e^{(\lambda)\tau}\, d\tau$$

The integral becomes:

$$\int_0^t e^{(\lambda)\tau}\, d\tau = \frac{e^{\lambda t} - 1}{\lambda}$$

Substituting this result back into the expression for $f_T(t)$, we get:

$$f_T(t) = \mu(\mu - \lambda)e^{-\mu t} \cdot \frac{e^{\lambda t} - 1}{\lambda}$$

Step 5: Final Expression for $f_T(t)$

Simplifying further:

$$f_T(t) = \frac{\mu(\mu - \lambda)}{\lambda} \left(e^{-\lambda t} - e^{-\mu t} \right), \quad t \geq 0$$

Hence the probability density function $f_T(t)$ of the total time T an entity spends in the system is:

$$f_T(t) = \frac{\mu(\mu - \lambda)}{\lambda} \left(e^{-\lambda t} - e^{-\mu t} \right), \quad t \geq 0$$

where:

- λ is the arrival rate,
- μ is the service rate.

This function represents the time spent in the system as the sum of the waiting time in the queue and the service time.

Mean Time in the System

The expected total time $E(T)$ spent in the system is the integral of $t \cdot f_T(t)$:

$$E(T) = \int_0^\infty t \cdot f_T(t)\, dt$$

Using standard results for the M/M/1 queue, the expected total time spent in the system is:

$$E(T) = \frac{1}{\mu - \lambda}$$

Hence, an average time an entity spends in the system (W):

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

Average Number of Entities in the Queue L_q

In the M/M/1 queuing model, the average number of entities in the queue L_q is given by:

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

where:

- λ is the arrival rate,
- μ is the service rate.

Derivation of L_q

The total number of entities in the system is $L = \frac{\lambda}{\mu - \lambda}$. The total number of entities in the system is the sum of:

$$L = L_q + \rho$$

where $\rho = \frac{\lambda}{\mu}$ is the server utilization.
Solving for L_q:

$$L_q = L - \rho = \frac{\lambda}{\mu - \lambda} - \frac{\lambda}{\mu}$$

Simplifying:

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

- Average waiting time in the queue (W_q):

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

Problem - 1: M/M/1 Queue Analysis

Given an M/M/1 queuing system with:

- Arrival rate: $\lambda = 5$ customers per hour,
- Service rate: $\mu = 8$ customers per hour.

Calculate the following:

1. The expected waiting time in the queue W_q,
2. The expected number of customers in the queue L_q,
3. The expected total time a customer spends in the system W,
4. The expected number of customers in the system L.

Solution: Part 1 - Expected Waiting Time in the Queue W_q

The formula for the expected waiting time in the queue is:

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

Substituting the given values of $\mu = 8$ and $\lambda = 5$:

$$W_q = \frac{1}{8-5} = \frac{1}{3} \text{ hours.}$$

Therefore, the expected waiting time in the queue is:

$$W_q = 0.3333 \text{ hours} = 20 \text{ minutes.}$$

Solution: Part 2 - Expected Number of Customers in the Queue L_q

The formula for the expected number of customers in the queue is:

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

Substituting the given values of $\mu = 8$ and $\lambda = 5$:

$$L_q = \frac{5}{8(8-5)} = \frac{5}{8 \times 3} = \frac{5}{24}.$$

Therefore, the expected number of customers in the queue is:

$$L_q = 0.2083 \text{ customers.}$$

Solution: Part 3 - Expected Total Time in the System W

The total time a customer spends in the system is the sum of the waiting time in the queue and the time spent being served:

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

We already know that $W_q = 0.3333$ hours and $\mu = 8$ customers per hour, so:

$$W = 0.3333 + \frac{1}{8} = 0.3333 + 0.125 = 0.4583 \text{ hours}.$$

Therefore, the expected total time a customer spends in the system is:

$$W = 0.4583 \text{ hours} = 27.5 \text{ minutes}.$$

Solution: Part 4 - Expected Number of Customers in the System L

The formula for the expected number of customers in the system is:

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

Substituting the given values of $\mu = 8$ and $\lambda = 5$:

$$L = \frac{5}{8-5} = \frac{5}{3} = 1.6667 \text{ customers.}$$

Therefore, the expected number of customers in the system is:

$$L = 1.6667 \text{ customers.}$$

Problem - 2: M/M/1 Queue Analysis

Consider a customer service system with the following parameters:

- Arrival rate: $\lambda = 10$ customers per hour,
- Service rate: $\mu = 15$ customers per hour,
- Single server.

We are asked to calculate the following:

1. The probability that the system is empty, i.e., no customers are in the system (P_0),
2. The probability that there are exactly two customers in the system (P_2),
3. The probability that there are more than 3 customers in the system ($P_{n>3}$),
4. The average number of customers in the system (L),
5. The average time a customer spends in the system (W).

Solution: Part 1 - Probability that the system is empty P_0

The formula for the probability that the system is empty (i.e., no customers are in the system) is:

$$P_0 = 1 - \rho$$

where $\rho = \frac{\lambda}{\mu}$ is the utilization factor. Given:

$$\rho = \frac{10}{15} = \frac{2}{3}$$

Substituting into the formula:

$$P_0 = 1 - \frac{2}{3} = \frac{1}{3}$$

Therefore, the probability that the system is empty is:

$$P_0 = \frac{1}{3}.$$

Solution: Part 2 - Probability of exactly two customers P_2

The probability of having exactly n customers in the system is given by:

$$P_n = (1 - \rho)\rho^n \quad \text{for} \quad n = 0, 1, 2, \ldots$$

For $n = 2$, we substitute $\rho = \frac{2}{3}$ into the formula:

$$P_2 = \left(1 - \frac{2}{3}\right)\left(\frac{2}{3}\right)^2$$

Simplifying:

$$P_2 = \frac{1}{3} \times \frac{4}{9} = \frac{4}{27}$$

Therefore, the probability of having exactly two customers in the system is:

$$P_2 = \frac{4}{27}.$$

Solution: Part 3 - Probability of more than 3 customers $P_{n>3}$

The probability that there are more than 3 customers in the system is the complement of the probability that there are 0, 1, 2, or 3 customers in the system:

$$P_{n>3} = 1 - (P_0 + P_1 + P_2 + P_3)$$

Using the formula for P_n:

$$P_0 = \frac{1}{3}, \quad P_1 = \frac{1}{3} \times \frac{2}{3} = \frac{2}{9}, \quad P_2 = \frac{4}{27}, \quad P_3 = \frac{8}{81}$$

Therefore:

$$P_{n>3} = 1 - \left(\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \frac{8}{81}\right)$$

We first find a common denominator (81):

$$P_{n>3} = 1 - \left(\frac{27}{81} + \frac{18}{81} + \frac{12}{81} + \frac{8}{81}\right)$$

Simplifying:

$$P_{n>3} = 1 - \frac{65}{81} = \frac{16}{81}$$

Therefore, the probability that there are more than 3 customers in the system is:

$$P_{n>3} = \frac{16}{81}.$$

Solution: Part 4 - Average number of customers in the system L

The formula for the average number of customers in the system is:

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

Substituting the given values $\lambda = 10$ and $\mu = 15$:

$$L = \frac{10}{15 - 10} = \frac{10}{5} = 2.$$

Therefore, the average number of customers in the system is:

$$L = 2 \text{ customers.}$$

Solution: Part 5 - Average time a customer spends in the system W

The formula for the average time a customer spends in the system is:

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

Substituting the given values $\lambda = 10$ and $\mu = 15$:

$$W = \frac{1}{15 - 10} = \frac{1}{5} = 0.2 \text{ hours.}$$

Converting to minutes:

$$W = 0.2 \times 60 = 12 \text{ minutes.}$$

Therefore, the average time a customer spends in the system is:

$$W = 12 \text{ minutes.}$$

Applications of Queuing Theory

Queuing theory is widely used in various fields to optimize system performance:

- Telecommunications: Managing data packet flow in network routers.
- Customer Service: Optimizing staff levels and reducing wait times in call centers or banks.
- Healthcare: Managing patient flow through hospitals or clinics.
- Manufacturing: Streamlining production lines and workstation management.