UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

TRABAJO DE FIN DE GRADO

MODELOS DE COLAS CON DISCIPLINA DE PRIORIDADES

Tutor: Miguel Rodríguez Rosa

Autora: Lucía Moreira García

Grado en Estadística

Curso académico 2023-24

UNIVERSIDAD DE SALAMANCA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

VNiVERSiDAD D SALAMANCA

CAMPUS DE EXCELENCIA INTERNACIONAL

TRABAJO DE FIN DE GRADO

MODELOS DE COLAS CON DISCIPLINA DE PRIORIDADES

Firmado por:

MOREIRA GARCIA LUCIA - 11867650H

Firmado digitalmente por MOREIRA GARCIA LUCIA - 11867650H
Fecha: 2024.07.04 17:11:31 +02'00'

RODRIGUEZ ROSA MIGUEL - 70821913Z

Firmado digitalmente por RODRIGUEZ ROSA MIGUEL - 70821913Z
Fecha: 2024.07.04 17:15:29 +02'00'

Autora: Lucía Moreira García

Tutor: Miguel Rodríguez Rosa

Grado en Estadística

Curso académico 2023-24

Certificado del tutor TFG Grado en Estadística

D. Miguel Rodríguez Rosa, profesor del Departamento de Estadística de la Universidad de Salamanca,

HACE CONSTAR:

Que el trabajo titulado "*Modelos de colas con disciplina de prioridades*", que se presenta, ha sido realizado por D.ª Lucía Moreira García, con DNI 11867650H y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, a 4 de julio de 2024.

RODRIGUEZ ROSA MIGUEL - 70821913Z

Firmado digitalmente por RODRIGUEZ ROSA MIGUEL - 70821913Z
Fecha: 2024.07.04 17:15:46 +02'00'

Fdo.: Miguel Rodríguez Rosa

Índice general

Summary

A queueing model with priority discipline is a methodology used to study and understand the behavior of waiting systems in which multiple elements compete to be served. In this model, priority levels are assigned to each element in the queue so that those with higher priority are served before those with lower priority. This approach is particularly valuable in contexts where it is essential to offer preferential treatment to certain elements, thereby allowing an evaluation of how prioritization decisions affect the efficiency and overall performance of the queueing system.

This work focuses on conducting an exhaustive study on the concept of priority in queueing systems, exploring the different types of priorities that exist, including both non-preemptive and preemptive systems. The goal is to formulate an appropriate theory that encompasses the necessary conditions to avoid system saturation, as well as to develop the corresponding traffic equations and performance measures.

Additionally, the theory will be illustrated with practical examples that reflect real-world situations, demonstrating how priority queueing systems are effectively implemented in various contexts.

Introducción

Este trabajo ha surgido de la asignatura de Procesos Estocásticos en Tiempo Discreto (Rodríguez-Rosa, 2021), donde se estudia Teoría de Colas con disciplina FIFO, la cual tiene una gran aplicación práctica en la vida real, ya que muchas operaciones cotidianas se sustentan en los principios fundamentales de la teoría de colas, desde la formación de filas en establecimientos comerciales, hasta la gestión de servicios al cliente y servidores web, se evidencia su relevancia en diversos contextos prácticos.

Según Cao-Abad (2002) "La teoría de colas es una disciplina, dentro de la Investigación Operativa, que tiene por objeto el estudio y análisis de situaciones en las que existen entes que demandan cierto servicio".

El investigador pionero fue el matemático danés A. K. Erlang, quien, en 1909, publicó "La teoría de probabilidades y las conversaciones telefónicas". En trabajos posteriores (Gross et al., 2008), se observó que un sistema telefónico se caracterizaba generalmente por (1) entrada de Poisson, tiempos de espera (servicio) exponenciales y múltiples canales (servidores), o (2) entrada de Poisson, tiempos de espera constantes y un solo canal.

Las colas surgen cuando la demanda de un servicio excede la capacidad disponible para proporcionarlo. Este fenómeno puede atribuirse a diversos factores, como la escasez de servidores disponibles, limitaciones de espacio o restricciones económicas.

La teoría de colas, al profundizar en estos escenarios, no solo ofrece un análisis detallado de las líneas de espera, sino que también proporciona herramientas valiosas para optimizar la eficiencia y elevar la calidad de los servicios ofrecidos. Con el objetivo fundamental de responder preguntas como "¿Cuánto tiempo debe esperar un cliente?" y "¿Cuántas personas se formarán en la fila?".

Características principales de las colas

1. Llegadas clientes/Población:
    - Clientes: Conjunto de entidades, con un tamaño que puede ser finito o infinito. Es importante destacar que los clientes no necesariamente son personas, ya que este término se refiere a cualquier entidad o elemento que ingrese al sistema en estudio.
    - Llegadas: Se asume que las llegadas al sistema siguen una distribución de

Poisson, lo que implica una aleatoriedad en las llegadas, aunque estas podrían ser estacionarias (constantes), o no estacionarias.

- Las tasas medias de llegadas: Estas son constantes e independientes del número actual de clientes en el sistema. El tiempo transcurrido entre dos llegadas consecutivas, conocido como Tiempo entre Llegadas y siguiendo una distribución exponencial, puede ser constante debido a un patrón establecido o ser programado de manera específica.
- Tipos de clientes: se distinguen dos tipos de clientes: los pacientes, que esperan sin importar el tiempo de espera; y los impacientes, que deciden retirarse antes de entrar al sistema.

2. Servidores:
   - Número de servidores: El sistema cuenta con un número específico de servidores destinados a atender a los clientes. Estos servidores pueden ser finitos o infinitos, dependiendo de la capacidad del sistema.
   - Tiempo de servicio: El tiempo que transcurre desde que el cliente inicia el servicio hasta que lo termina se denomina Tiempo de Servicio, por lo general sigue una distribución exponencial, pero también puede seguir distribuciones degeneradas (tiempos constantes) o de Erlang.
   - Configuración de los servidores: Los servidores pueden variar en su configuración, ya sea con diferentes números de canales para atender múltiples clientes simultáneamente o alineados en una única cola.
   - Número de etapas: Algunos sistemas cuentan con servidores que no solo brindan un servicio único, sino que pasan por varias etapas o fases. Se define como etapa a los procesos distintos por los cuales pasa un servicio hasta que se completa, por ejemplo, en procesos de evaluación como exámenes de oposición que pueden involucrar diferentes pruebas.

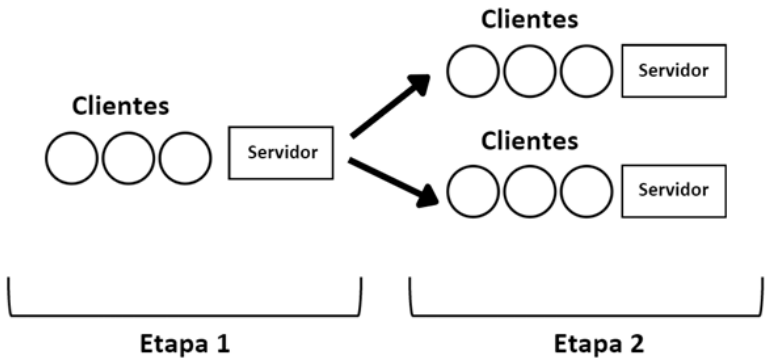

Figura 1: Servidores con dos etapas.

3. Línea de espera:
   - La línea de espera representa el espacio designado donde los clientes esperan para ser atendidos, esta puede tener dimensiones finitas o infinitas.

- La manera en que se selecciona y atiende a los clientes, determina la disciplina de la cola. Siendo las más comunes:
  - FIFO (first in first out): Se atiende al cliente por estricto orden de llegada, el primero que llegue a la cola será el primero en ser atendido. Esta disciplina es la más habitual.
  - LIFO (last in first out): El último cliente en llegar a la cola es el primero en ser atendido. Este método se utiliza en almacenes que manejan productos que no pierden valor con el tiempo y no son susceptibles de caducar o deteriorarse.
  - RSS (random selection of service): La atención se realiza de manera aleatoria, sin seguir un orden preestablecido. Se utiliza a la hora de hacer entrevistas telefónicas.
  - RR (round robin): Los clientes son atendidos en una secuencia cíclica, rotando de manera equitativa la atención entre ellos. Se utiliza en sistemas operativos.

Debemos tener en cuenta que la disciplina de la cola no modifica las probabilidades del estado del sistema, por lo que el tiempo promedio de espera dentro del sistema no se verá alterado. Mientras que sí lo harán los tiempos de espera de la cola.

Las diferentes disciplinas de colas no solo influyen en la eficiencia operativa del sistema, sino que también tienen implicaciones en la percepción del servicio por parte de los clientes. La elección de la disciplina adecuada depende de la naturaleza del servicio y los objetivos específicos del sistema.

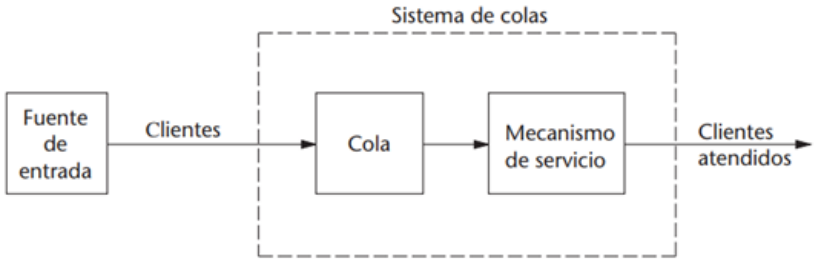

Figura 2: Sistema de colas (Fuente: (Hillier & Lieberman, 2010)).

Proceso de nacimiento-muerte

La mayoría de los modelos básicos de colas parten de la premisa de que las llegadas (entrada de clientes) y las salidas (clientes que abandonan el sistema) siguen un proceso de nacimiento y muerte.

Se trata de un tipo especial de cadenas de Markov en tiempo discreto, que describe la evolución de un sistema con una población variable.

Los “nacimientos” se refieren a la llegada de nuevos clientes al sistema de colas, representando la transición de un estado actual a un estado superior. Cuando el sistema está en el estado n y ocurre un proceso de llegada o nacimiento, la población aumenta a $n + 1$.

Por otro lado, las “muertes” se refieren a la salida de clientes atendidos, representando la transición de un estado actual a un estado inferior. En el estado n, se produce un proceso de salida o muerte, y la población disminuye de n a $n - 1$.

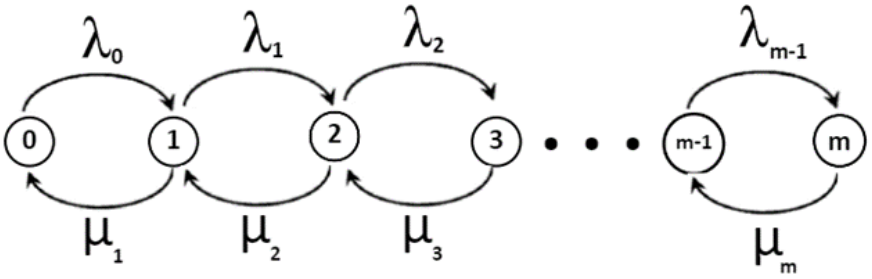

Figura 3: Proceso de nacimiento-muerte.

La representación del estado del sistema en el tiempo “t”, $\forall t \geq 0$, denotada como $N(t)$ refleja el número de clientes presentes en el sistema de colas en el tiempo t.

Al iniciar el proceso ($t = 0$), $N(0) = 0$, indicando que no hay clientes en el sistema.

La distribución de probabilidad actual del tiempo restante para el próximo nacimiento (llegada) es exponencial, caracterizada por el parámetro λ_n, siendo este la tasa media de llegadas cuando ya hay n clientes en el sistema.

La distribución de probabilidad actual del tiempo que falta para la próxima muerte (terminación de servicio) es exponencial con parámetro μ_n, la cual representa la tasa media de salidas cuando todavía hay n clientes en el sistema.

El tiempo que falta hasta el próximo nacimiento y el tiempo que falta hasta la siguiente muerte, son mutuamente independientes.

Existe uniformidad de llegadas y salidas a lo largo del intervalo $n = U(t_1, t_2)$.

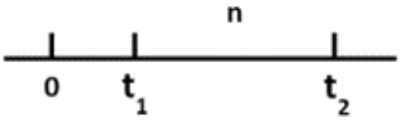

Figura 4: Intervalo de llegadas.

Una vez vistos los fundamentos de la teoría de colas y destacando su aplicación práctica en diversos contextos, desde la gestión de filas en establecimientos comerciales hasta

la administración de servicios en servidores web, podemos observar cómo esta teoría, desarrollada inicialmente para abordar problemas de congestión en sistemas telefónicos, demuestra ser una buena herramienta en la optimización de recursos y la mejora de la eficiencia operativa en múltiples campos.

A medida que avanzamos en este trabajo, nos enfocaremos en un aspecto notable de la teoría de colas: la disciplina de prioridades. Esta área de estudio cobra gran importancia debido a su capacidad para gestionar sistemas donde no todas las solicitudes o clientes tienen la misma urgencia. La disciplina de prioridades permite asignar diferentes niveles de importancia a las tareas, optimizando así el orden de atención y mejorando significativamente la eficiencia y satisfacción del cliente en sistemas con recursos limitados.

Este enfoque en la disciplina de prioridades nos permitirá profundizar en los aspectos teóricos y técnicos de la teoría de colas y nos proporcionará una visión práctica de cómo estas teorías pueden ser utilizadas en el día a día para diseñar sistemas más eficientes y efectivos. Con este trabajo, se busca resaltar la importancia de la teoría de colas con prioridades y su impacto en la optimización de procesos en el mundo real.

Ahora una vez definidos los términos principales de este trabajo veremos su estructura:

Tras un primer apartado con los objetivos del trabajo, nuestro primer capítulo tratará de modelos de colas con prioridad sin interrupción.

Nuestro segundo capítulo hablará de modelos de colas con prioridad con interrupción.

El tercer capítulo constará de un manual de uso de la aplicación web creada para analizar todos estos modelos.

Por último, se ofrecerán las conclusiones, y la bibliografía utilizada para la realización de este trabajo. En esta encontraremos diversos libros, artículos y páginas web.

Objetivos

En este trabajo tendremos los siguientes objetivos:

- Hacer un estudio sobre el concepto de prioridad en sistemas de colas.
- Explorar los tipos de prioridades, incluyendo sistemas sin interrupción y con interrupción.
- Especificar una teoría adecuada, deduciendo condiciones de no saturación, ecuaciones de tráfico y medidas de rendimiento.
- Desarrollar ejemplos que describan la realidad.
- Crear un programa con R para llevar a cabo los estudios descritos.

Notación

Notación de Kendall

$$A|B|C|X|Y|Z$$

- **A**: Patrón de llegadas
- **B**: Patrón de servicios
- **C**: Número de servidores
- **X**: Capacidad del sistema
- **Y**: Disciplina de la cola (Prioridad)
- **Z**: Número de etapas (Por defecto será 1)

Métricas de colas

- λ: Tasa de llegadas
- λ_k: Tasa de llegadas de la prioridad k
- μ: Tasa de salidas
- μ_k: Tasa de salidas de la prioridad k
- ρ: Tasa de ocupación del sistema
- W_q: Tiempo medio que pasa un cliente en la cola
- W: Tiempo medio que un cliente pasa en el sistema
- $W_q^{(k)}$: Tiempo medio que pasa en la cola un cliente de prioridad k
- $W^{(k)}$: Tiempo medio en el sistema de un cliente de prioridad k
- L_q: Número medio de clientes en la cola
- L: Número medio de clientes en el sistema
- $L_q^{(k)}$: Número medio de clientes de prioridad k en la cola
- $L^{(k)}$: Número medio de clientes en el sistema de prioridad k

Una de las aportaciones más importantes en la teoría de colas fue realizada por John D. C. Little en la década de 1960. Little (1961) desarrolló una fórmula que conecta el tamaño promedio de un sistema en estado estacionario con los tiempos de espera promedio de los clientes en ese mismo estado.

Esta fórmula, conocida como la Fórmula de Little, se expresa de la siguiente manera:

$$L = \lambda W \qquad L_q = \lambda W_q$$

En cuanto al tiempo medio de espera, se disponen de las siguientes fórmulas:

$$W_q = W - \frac{1}{\mu} \qquad W = W_q + \frac{1}{\mu}$$

Estas relaciones son muy importantes porque permiten determinar las cuatro medidas de rendimiento fundamentales: L, W, L_q y W_q, desde que se conoce el valor de una de ellas.

Fórmulas de Little en el caso de prioridades

Cuando las tasas de llegada (λ_k) no son iguales entre las diferentes clases de prioridad, se pueden ajustar las fórmulas. De manera que se puede sustituir λ_k por la tasa promedio de llegadas a largo plazo $\bar{\lambda}$ para mantener la coherencia en las ecuaciones.
Así, $L^{(k)}$ y $L_q^{(k)}$ se pueden expresar como:

$$L^{(k)} = \lambda_k W^{(k)} \quad L_q^{(k)} = \lambda_k W_q^{(k)} \qquad \text{para } k = 1, 2, \ldots, n$$

De estas fórmulas podemos obtener también el tiempo medio que un cliente de clase k pasa en la cola o en el sistema:

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu_k} \qquad W^{(k)} = \frac{L^{(k)}}{\lambda_k} \qquad \text{para } k = 1, 2, \ldots, n$$

En conclusión, las contribuciones de John D. C. Little han permitido una comprensión más profunda y práctica de la teoría de colas, facilitando el análisis y la optimización de sistemas complejos en diversas aplicaciones.

Capítulo 1

Modelos con prioridad sin interrupción

El modelo de colas con prioridad sin interrupción, implica que no se permiten interrupciones en el servicio. Esto significa que una vez que un cliente comienza a recibir servicio, este proceso no puede ser interrumpido por la llegada de clientes de igual o mayor prioridad, por lo que los nuevos clientes tendrán que esperar en la cola hasta que el servicio del primero termine. La gran ventaja de este modelo es que simplifica la gestión y el seguimiento de los procesos pudiendo así aumentar la calidad del mismo.

La convención habitual es numerar las clases de prioridad de forma que los números más pequeños correspondan a prioridades más altas.

Los modelos de prioridad sin interrupción son comunes en una gran variedad de contextos, a continuación veremos las diferentes variedades existentes y cómo se pueden aplicar en diversos escenarios.

1.1. Dos Clases

En este escenario los clientes llegan al sistema de forma aleatoria y a cada cliente se le asigna una de clase de prioridad pudiendo ser en este caso 1 o 2.

Denotaremos la tasa de llegadas para la prioridad más alta como λ_1, y λ_2 a la prioridad más baja. Por tanto, la tasa total de llegadas se calcula como $\lambda = \lambda_1 + \lambda_2$.

1.1.1. Tasas de servicio iguales

En situaciones donde las tasas de servicio son iguales para ambas clases de prioridad, el tiempo que lleva servir a cada cliente es constante y no varía, independientemente de la prioridad del cliente.

Este escenario puede ser común en sistemas donde la capacidad de servicio está estandarizada y tienen bastante relevancia en procesos informáticos.

Por tanto, asumimos:

$$\mu = \mu_1 = \mu_2$$

1.1.1.1. M|M|1|Prioridad

Si las distribuciones de los tiempos de servicio de las clases de prioridad son exponenciales y comparten la misma tasa de servicio μ, entonces el número de clientes siendo atendidos seguirá la misma distribución en estado estacionario que una cola M|M|1|FIFO (Gross et al., 2008).

$$p_n = \sum_{m=0}^{n-1}(p_{n-m,m,1} + p_{m,n-m,2}) = (1-\rho)\rho^n \quad (n > 0)$$

donde p_n es la probabilidad de que haya n clientes de cualquier prioridad en el sistema, y $p_{m,n,k}$ es la probabilidad de que en el sistema haya m clientes de prioridad 1, n clientes de prioridad 2, y el cliente que está en el servicio tiene prioridad $k = 1, 2$.

Para cada prioridad la probabilidad de que el sistema este lleno se representa como:

$$\rho_1 = \frac{\lambda_1}{\mu} \qquad \rho_2 = \frac{\lambda_2}{\mu} \qquad \rho = \rho_1 + \rho_2 = \frac{\lambda}{\mu}$$

El número medio de clientes de cada prioridad en la cola y en el sistema es:

$$L_q^{(1)} = \frac{\lambda_1\rho}{\mu - \lambda_1} \qquad L_q^{(2)} = \frac{\lambda_2\rho}{(\mu - \lambda_1)(1-\rho)} \qquad L_q = \frac{\rho^2}{1-\rho} \qquad L^{(k)} = L_q^{(k)} + \rho_k$$

Y el tiempo medio que pasa un cliente cualquiera en la cola, y el que pasa en el sistema un cliente de cada prioridad, es:

$$W_q = \frac{\rho}{(\mu - \lambda_1)(1-\rho)} \qquad W^{(k)} = \frac{L^{(k)}}{\lambda_k}$$

A continuación vamos a aplicar estas formulas con un ejemplo:

Ejemplo: cafetería universitaria

En una concurrida cafetería universitaria, tanto estudiantes como profesores hacen cola para ser atendidos. Se asignan diferentes prioridades a los clientes: Los estudiantes tienen prioridad baja, mientras que los profesores tienen prioridad alta. La tasa de llegada de profesores es $\lambda_1 = 6$ clientes por hora y de estudiantes $\lambda_2 = 8$ clientes por hora. Ambos tipos de clientes son atendidos por el mismo empleado de la cafetería, y sin importar su prioridad, lo que implica que la tasa de servicio es idéntica para ambos: $\mu = 15$ clientes por hora. El problema planteado busca determinar el tiempo promedio que los clientes pasan tanto en la cola como en el sistema, y cuántos clientes, de media, hay en la cola y en el sistema.

$$\lambda_1 = 6 \qquad \lambda_2 = 8 \qquad \mu = 15$$

$$\rho_1 = \frac{6}{15} = 0.4 \qquad \rho_2 = \frac{8}{15} = 0.5333 \qquad \rho = 0.4 + 0.5333 = 0.9333$$

La probabilidad de que en la cafetería haya algún cliente es del 93.33 %.

$$L_q^{(1)} = \frac{6 \cdot 0.9333}{15-6} = 0.6222 \qquad L_q^{(2)} = \frac{8 \cdot 0.9333}{(15-6)(1-0.9333)} = 12.4444$$

$$L_q = \frac{0.9333^2}{1-0.9333} = 13.0667$$

De media en la cola de la cafetería hay 0.6222 profesores y 12.4444 alumnos, por lo que entre profesores y alumnos hay de media 13.0667 clientes en la cola.

$L^{(1)} = 0.6222{+}0.4 = 1.0222 \quad L^{(2)} = 12.4444{+}0.5 = 12.9778 \quad L = 1.0222{+}12.9778 = 14$

El numero medio de profesores en el total de la cafetería es 1.0222, y de alumnos, 12.9777, así que el numero total de clientes en la cafetería es 14.

$$W^{(1)} = \frac{1.0222}{6} = 0.1703 \quad W^{(2)} = \frac{12.9778}{8} = 1.6222$$

$$W_q = \frac{0.9333}{(15-6)(1-0.9333)} = 1.5556$$

El tiempo medio que pasa un cliente clualquiera en la cola es de 1.5555 horas = 93.33 minutos. El tiempo medio que pasa un profesor en la cafetería es de 0.1703 horas = 10.218 minutos y el tiempo medio que pasa un alumno en el sistema es de 1.6222 horas = 97.332 minutos.

En este ejemplo hemos podido observar la gran diferencia que existe entre las prioridades, ya que aunque las tasas de llegadas sean similares, los clientes de prioridad más alta pasan mucho menos tiempo en el sistema que los de prioridad más baja.

1.1.1.2. M|M|c|Prioridad

El modelo M|M|c describe un sistema con múltiples servidores. El parámetro c indica el número de servidores en paralelo que pueden atender a los clientes simultáneamente.

A medida que los clientes llegan al sistema, son atendidos por cualquiera de los c servidores disponibles, siempre respetando el sistema de prioridades.

Si todos los servidores están ocupados, los clientes esperan en la cola hasta que uno de los servidores quede libre. Todos los servidores tiene el mismo tiempo de servicio, el cual sigue una distribución exponencial.

A continuación, estudiaremos las fórmulas y métodos para calcular los tiempos de espera y el número medio de clientes tanto en el sistema como en la cola.

Para calcular ρ lo haremos de la misma manera que en el caso anterior, pero ahora multiplicaremos el numero de servidores por la tasa de llegadas μ.

$$\rho_1 = \frac{\lambda_1}{c\mu} \qquad \rho_2 = \frac{\lambda_2}{c\mu} \qquad \rho = \rho_1 + \rho_2 = \frac{\lambda}{c\mu} \quad \text{siendo} \quad \lambda = \sum_{k=1}^{2} \lambda_k$$

Para calcular el tiempo medio que pasa un cliente de prioridad k en el sistema aplicamos

la siguiente fórmula (Hillier & Lieberman, 2010):

$$W^{(k)} = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu} \quad \text{para} \quad k = 1, 2$$

donde:

$$A = c!\frac{c\mu - \lambda}{\rho^c}\sum_{j=0}^{c-1}\frac{\rho^j}{j!} + c\mu \qquad B_0 = 1 \qquad B_k = 1 - \frac{\sum_{k=1}^{k}\lambda_k}{c\mu}$$

Para calcular el tiempo medio de espera en la cola (sin contar con el tiempo de servicio), utilizamos la fórmula de Little.

La longitud de la cola la calcularemos multiplicando este tiempo por λ_k:

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu} \qquad W_q = \sum_{i=1}^{k}\frac{\lambda_i}{\lambda}W_q^{(i)} \qquad L^{(k)} = \lambda_k W^{(k)} \qquad L_q^{(k)} = W_q^{(k)}\lambda_k$$

Veamos un ejemplo.

Ejemplo: Préstamo de libros

En una biblioteca universitaria se ha implementado un sistema de atención para gestionar las solicitudes de préstamo de libros de manera eficiente. El sistema está diseñado para dar prioridad a las solicitudes de alumnos pertenecientes a la universidad sobre personas externas a la universidad. Las solicitudes de préstamo de alumnos pertenecientes a la universidad tienen una tasa de llegada de $\lambda_1 = 7$ solicitudes por hora, mientras que las solicitudes de los que no son alumnos llegan con una tasa de $\lambda_2 = 12$ por hora. Todas las solicitudes, independientemente de su prioridad, son atendidas con la misma tasa de servicio, $\mu = 5$ solicitudes por hora. La biblioteca cuenta con 4 ordenadores que dan el servicio del préstamo.

Los objetivos de este sistema son calcular el tiempo promedio de espera en la cola y en la biblioteca para cada tipo de solicitud, y calcular el número medio de solicitudes tanto en la cola como en la biblioteca.

$$\lambda_1 = 7 \qquad \lambda_2 = 12 \qquad \lambda = 7 + 12 = 19 \qquad \mu = 5 \qquad c = 4$$

$$\rho_1 = \frac{7}{4 \cdot 5} = 0.35 \qquad \rho_2 = \frac{12}{4 \cdot 5} = 0.6 \qquad \rho = 0.35 + 0.6 = 0.95$$

La probabilidad de que haya alguna solicitud en la biblioteca es del 0.95 %.

Calculamos las medidas de rendimiento correspondientes a los tiempos de espera:

$$W^{(k)} = \frac{1}{AB_{k-1}B_k} + \frac{1}{\mu} \qquad A = c!\frac{c\mu - \lambda}{\rho^c}\sum_{j=0}^{c-1}\frac{\rho^j}{j!} + c\mu \qquad B_k = 1 - \frac{\sum_{k=1}^{k}\lambda_k}{c\mu}$$

$$A = 4!\frac{4\cdot 5 - 19}{0.95^4}\sum_{j=0}^{c-1}\frac{0.95^j}{j!} + 4\cdot 5 = 94.9650$$

$$B_0 = 1 \quad B_1 = 1 - \frac{7}{4\cdot 5} = 0.65 \quad B_2 = 1 - \frac{7+12}{4\cdot 5} = 0.05$$

$$W^{(1)} = \frac{1}{94.9650\cdot 1\cdot 0.65} + \frac{1}{5} = 0.2162 \qquad W^{(2)} = \frac{1}{94.9650\cdot 0.65\cdot 0.05} + \frac{1}{5} = 0.5240$$

Los alumnos pertenecientes a la universidad esperan de media 0.2162 horas = 12.972 minutos en la biblioteca, mientras que los externos esperan 0.5240 horas = 31.44 minutos.

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu}$$

$$W_q^{(1)} = 0.2162 - \frac{1}{5} = 0.0162 \qquad W_q^{(2)} = 0.5240 - \frac{1}{5} = 0.3240$$

$$W_q = \sum_{i=1}^{k}\frac{\lambda_i}{\lambda}W_q^{(i)}$$

Los alumnos de la universidad esperan de media en la cola 0.0162 horas = 0.972 minutos, mientras que los que no pertenecen en la universidad esperan 0.3240 horas = 19.44 minutos.

$$W_q = \frac{7}{19}\cdot 0.0162 + \frac{12}{19}\cdot 0.3240 = 0.2106$$

Entre los dos tipos de alumnos, la media de espera en la cola es de 0.2106 horas = 12.636 minutos.

En cuanto al número medio de clientes:

$$L^{(k)} = \lambda_k W^{(k)} \qquad L_q^{(k)} = W_q^{(k)}\lambda_k$$

$$L^{(1)} = 7\cdot 0.2162 = 1.5134 \qquad L_q^{(1)} = 0.0162\cdot 7 = 0.1134$$

$$L^{(2)} = 12\cdot 0.5240 = 6.2881 \qquad L_q^{(2)} = 0.3240\cdot 12 = 3.8881$$

Como era de esperar, tanto en la cola como en el sistema, de media hay siempre mas alumnos de la clase de prioridad baja, es decir alumnos externos a la universidad.

1.1.2. Tasas de servicio diferentes

Hemos visto que existe la posibilidad de que las tasas de servicio sean iguales para ambas prioridades, pero lo más común es que dichas tasas varíen. En general, las tasas de servicio suelen ser más bajas para las prioridades altas cuando es crucial resolver los problemas rápidamente debido a su impacto, por ejemplo, en emergencias médicas. Sin embargo, en algunos casos, las prioridades altas pueden tener tasas de servicio mayores cuando los servicios requeridos son más largos, como en atención al cliente, donde los clientes VIP reciben atención personalizada y pueden tener consultas más detalladas y prolongadas.

Por lo tanto, las tasas de servicio de prioridad alta y baja serán, respectivamente, μ_1 y μ_2.

1.1.2.1. M|M|1|Prioridad

En este apartado estudiaremos el caso en el que las tasas de servicio son diferentes para cada prioridad y disponemos únicamente de un servidor.

Para calcular ρ_i y ρ aplicamos las mismas formulas que anteriormente:

$$\rho_1 = \frac{\lambda_1}{\mu_1} \qquad \rho_2 = \frac{\lambda_2}{\mu_2} \qquad \rho = \rho_1 + \rho_2$$

A continuación pasamos a definir las fórmulas para obtener el número medio de clientes en la cola para cada prioridad (Gross et al., 2008).

$$L_q^{(1)} = \frac{\lambda_1(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho_1} \qquad L_q^{(2)} = \frac{\lambda_2(\rho_1/\mu_1 + \rho_2/\mu_2)}{(1-\rho_1)(1-\rho)} \qquad L_q = L_q^{(1)} + L_q^{(2)}$$

Siendo $L_q^{(1)}$ y $L_q^{(2)}$ el número medio de clientes en la cola de prioridad 1 y de prioridad 2 respectivamente. Por lo general el segundo valor será más grande que el primero.

Para calcular el tiempo esperado en el sistema de un miembro de prioridad k, recurriremos a la siguiente fórmula (Hillier & Lieberman, 2010):

$$W^{(k)} = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k}$$

$$\text{Siendo:} \quad a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2} \qquad b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i} \qquad b_0 = 1$$

Por último, definimos la fórmula para calcular el tiempo esperado en la cola de cada prioridad.

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu_k}$$

Ejemplo: Compañía de telecomunicaciones

En una oficina de servicios al cliente de una compañía de telecomunicaciones, los clientes pueden realizar consultas generales o solicitar soporte técnico. Se asignan diferentes prioridades a las consultas: las consultas de soporte técnico tienen una prioridad más alta que las consultas generales. La tasa de llegadas de consultas de soporte técnico es de $\lambda_1 = 3$ consultas por hora, y la tasa de llegada de consultas generales a la oficina es de $\lambda_2 = 5$ consultas por hora. La tasa de servicio de soporte técnico es de $\mu_1 = 10$ consultas por hora y la de consultas generales es de $\mu_2 = 12$ por hora. El objetivo es determinar el tiempo promedio que pasan los clientes con consultas generales y consultas de soporte técnico tanto en cola como en el sistema, así como el número promedio de clientes en cola y en el sistema.

$$\lambda_1 = 3 \quad \lambda_2 = 5 \quad \mu_1 = 10 \quad \mu_2 = 12$$

Empezamos calculando ρ y comprobamos que es menor que 1

$$\rho_1 = \frac{\lambda_1}{\mu_1} = \frac{3}{10} = 0.3 \qquad \rho_2 = \frac{\lambda_2}{\mu_2} = \frac{5}{12} = 0.4167$$

$$\rho = \rho_1 + \rho_2 = 0.3 + 0.4167 = 0.7167$$

La probabilidad de que haya al menos una consulta en el sistema es del 71.67 %. Como ρ es menor que 1 el sistema no satura y podemos seguir con los cálculos.

$$L_q^{(1)} = \frac{\lambda_1(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho_1} = \frac{3 \cdot \left(\frac{0.3}{10} + \frac{0.4167}{12}\right)}{1-0.3} = 0.2774$$

$$L_q^{(2)} = \frac{\lambda_2(\rho_1/\mu_1 + \rho_2/\mu_2)}{(1-\rho_1)(1-\rho)} = \frac{5 \cdot \left(\frac{0.3}{10} + \frac{0.4167}{12}\right)}{(1-0.3)(1-0.7167)} = 1.6317$$

$$L_q = L_q^{(1)} + L_q^{(2)} = 0.2774 + 1.6317 = 1.9090$$

Por tanto, el número medio de consultas en cola es de 1.9090, siendo 0.2774 el número medio de consultas de soporte técnico y 1.6317 consultas generales.

Ahora procedemos a calcular el tiempo medio que pasan los clientes en el sistema:

$$W^{(k)} = \frac{a_k}{b_{k-1}b_k} + \frac{1}{\mu_k} \qquad a_k = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2} \qquad b_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i}$$

$$a_1 = \frac{3}{10^2} = 0.03 \qquad a_2 = \frac{3}{10^2} + \frac{5}{12^2} = 0.0647$$

$$b_0 = 1 \qquad b_1 = 1 - \frac{3}{10} = 0.7 \qquad b_2 = 1 - \left(\frac{3}{10} + \frac{5}{12}\right) = 0.2833$$

$$W^{(1)} = \frac{0.03}{1 \cdot 0.7} + \frac{1}{10} = 0.1429 \qquad W^{(2)} = \frac{0.0647}{0.7 \cdot 0.2833} + \frac{1}{12} = 0.4097$$

Los clientes que realizan consultas de soporte técnico pasan de media en el sistema 0.1429 horas = 8.574 minutos, mientras que aquellos que realizan consultas generales permanecen de media 0.4097 horas = 24.582 minutos.

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu_k}$$

$$W_q^{(1)} = 0.1429 - \frac{1}{10} = 0.0429 \qquad W_q^{(2)} = 0.4097 - \frac{1}{12} = 0.3263$$

El tiempo medio que pasan los clientes en cola es de 0.0429 horas = 2.574 minutos para las consultas de soporte técnico y de 0.3263 horas = 19.578 minutos para las generales.

1.1.3. FIFO vs Prioridad

En este apartado, consideremos un modelo con dos clases de clientes donde el orden de atención se basa en el criterio "primero en llegar, primero en ser atendido" (FIFO). En este contexto, tenemos dos clases distintas de clientes, cada una con su propia tasa de llegadas, λ_1 y λ_2, así como su respectiva tasa de servicios, μ_1 y μ_2. Los tiempos de servicios siguen una distribución exponencial y los clientes son atendidos en el orden en que llegan, sin dar prioridad a ninguna clase. Este enfoque, aunque no distingue entre prioridades de clientes, es útil para compararlo con un modelo que sí considera prioridades entre las clases. Esto contrasta con el modelo del apartado 1.1.2.1 anterior de M|M|1|Prioridad, donde los clientes se clasificaban en dos prioridades diferentes y eran atendidos en función de esa prioridad.

Las fórmulas para obtener el número medio de clientes en la cola siguiendo el modelo FIFO, son las siguientes (Gross et al., 2008):

$$L_q^{(1)} = \frac{\lambda_1(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho} \quad L_q^{(2)} = \frac{\lambda_2(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho} \quad L_q = \frac{\lambda(\rho_1/\mu_1 + \rho_2/\mu_2)}{1-\rho}$$

Aplicamos estas fórmulas a los datos del ejemplo anterior: "Compañía de telecomunicaciones":

$$\lambda_1 = 3 \quad \lambda_2 = 5 \quad \mu_1 = 10 \quad \mu_2 = 12$$
$$\rho_1 = 0.3 \qquad \rho_2 = 0.4167 \qquad \rho = 0.7167$$

$$L_q^{(1)} = \frac{3 \cdot (0.3/10 + 0.4167/12)}{1 - 0.7167} = 0.6853 \quad L_q^{(2)} = \frac{5 \cdot (0.3/10 + 0.4167/12)}{1 - 0.7167} = 1.1422$$

$$L_q = \frac{8 \cdot (0.3/10 + 0.4167/12)}{1 - 0.7167} = 1.8275$$

En el modelo con disciplina FIFO, el número medio de solicitudes en cola, L_q, es 1.8275, mientras que para el modelo con diciplina de prioridades, era 1.9090.

El número medio de solicitudes de soporte técnico en cola utilizando la disciplina FIFO $L_q^{(1)}$ es 0.6853, mientras que en el modelo con disciplina de prioridades era 0.2774.

Para las consultas generales, $L_q^{(2)}$, el modelo con disciplina FIFO muestra un número medio de solicitudes en cola de 1.1422, comparado con 1.6317 en el modelo de prioridades.

La comparación de ambos modelos revela que el uso de un sistema de prioridades no proporciona una mejora significativa en la eficiencia de las solicitudes de media. Sin embargo, en el modelo con disciplina de prioridades, las consultas de soporte técnico experimentan una reducción considerable en el número medio de solicitudes en cola, lo que se reflejará en una disminución del tiempo de espera.

1.2. Más de dos clases

En esta situación, los clientes llegan al sistema de forma aleatoria y a cada cliente se le asigna una clase de prioridad, pudiendo ser en este caso $k = 1, 2, 3, \ldots, n$.

Esto permite una mayor flexibilidad y precisión en la gestión de las prioridades, ya que más clases pueden representar diferentes niveles de urgencia o importancia.

Como en el caso anterior, denotaremos la tasa de llegadas para la prioridad más alta como λ_1, la de segunda prioridad será λ_2 y así sucesivamente hasta λ_n. Por tanto, la tasa total de llegadas se calcula como:

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_n.$$

1.2.1. Tasas de servicio iguales

Comenzaremos viendo el caso en el cual las tasas de servicio son iguales para todas las clases de prioridad. En esta situación, el tiempo requerido para atender a cada cliente es fijo y no depende de la prioridad asignada.

1.2.1.1. M|M|c|Prioridad

En este apartado, consideraremos la existencia de múltiples servidores.

La fórmula para calcular si en el sistema hay algún cliente corresponde a:

$$\rho_k = \frac{\lambda_k}{c\mu} \quad (1 \leq k \leq n) \qquad \sigma_k = \sum_{i=1}^{k} \rho_i \quad (\sigma_n = \rho = \lambda/c\mu)$$

La probabilidad de que algún servidor esté ocupado se calcula de la siguiente manera:

$$\sum_{n=c}^{\infty} p_n = p_0 \sum_{n=c}^{\infty} \frac{(c\rho)^n}{c^{n-c}c!} = p_0 \frac{(c\rho)^c}{c!(1-\rho)}$$

donde p_0 se puede calcular como (Gross et al., 2008):

$$p_0 = \left(\frac{r^c}{c!(1-\rho)} + \sum_{i=0}^{c-1} \frac{r^i}{i!} \right)^{-1} \text{ donde r} = \frac{\lambda}{\mu}, \quad \rho = r/c$$

El número medio de clientes de prioridad k que hay en la cola y el número medio de clientes totales en la cola, lo podemos obtener realizando los siguientes cálculos:

$$W_q^{(k)} = \frac{\left[c!(1-\rho)(c\mu) \sum_{i=0}^{c-1} \frac{(c\rho)^{i-c}}{i!} + c\mu \right]^{-1}}{(1-\sigma_{k-1})(1-\sigma_k)} \qquad W_q = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} W_q^{(i)}$$

Para determinar el tiempo de espera en el sistema y el número medio de clientes de prio-

ridad k tanto en la cola como en el sistema haremos uso de las fórmulas de Little.

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu}$$

$$L^{(k)} = \lambda_k W^{(k)} \quad L_q^{(k)} = \lambda_k W_q^{(k)}$$

Realizaremos un ejemplo para aplicar las fórmulas:

Ejemplo: estación de servicio

En una estación de servicio con 4 surtidores, se atienden tres tipos de vehículos con diferentes prioridades: vehículos de emergencia, vehículos de servicio y vehículos particulares.

La tasa de servicio para cada prioridad es constante $\mu = 13$ vehículos por hora. Las tasas de llegada para cada tipo de vehículo son: $\lambda_1 = 14$ vehículos por hora para vehículos de emergencia, $\lambda_2 = 11$ vehículos por hora para vehículos de servicio, y $\lambda_3 = 18$ vehículos por hora para vehículos particulares.

El objetivo es calcular el tiempo promedio de espera en la cola y en la estación, y el número promedio de vehículos en la cola y en la estación para cada tipo de vehículo.

$$\lambda_1 = 14 \qquad \lambda_2 = 11 \qquad \lambda_3 = 18 \qquad \lambda = 43 \qquad \mu = 13 \qquad c = 4$$

Calculamos primero la probabilidad de que la estación esté vacía, para así calcular la probabilidad de que al menos un surtidor esté ocupado.

$$p_0 = \left(\frac{r^c}{c!(1-\rho)} + \sum_{n=0}^{c-1} \frac{r^n}{n!} \right)^{-1} \text{ donde r} = \frac{\lambda}{\mu}, \quad \rho = \frac{r}{c}$$

$$r = \frac{43}{13} = 3.3077, \quad \rho = \frac{3.3077}{4} = 0.8269$$

$$p_0 = \left(\frac{3.3077^4}{4!(1-0.8269)} + \sum_{i=0}^{4-1} \frac{3.3077^i}{i!} \right)^{-1} = 0.0259$$

$$\sum_{n=c}^{\infty} p_n = p_0 \frac{(c\rho)^c}{c!(1-\rho)} = 0.0259 \frac{(4 \cdot 0.8269)^4}{4!(1-0.8269)} = 0.7465$$

La probabilidad de que algún surtidor esté ocupados es del 74.65 %.

El siguiente paso es calcular los tiempos medios de espera en la cola.

$$W_q^{(k)} = \frac{\left[c!(1-\rho)(c\mu) \sum_{i=0}^{c-1} \frac{(c\rho)^{i-c}}{i!} + c\mu \right]^{-1}}{(1-\sigma_{k-1})(1-\sigma_k)}$$

Para ello calculamos ρ_k y σ_k

$$\rho_1 = \frac{14}{4 \cdot 13} = 0.2692 \quad \rho_2 = \frac{11}{4 \cdot 13} = 0.2115 \quad \rho_3 = \frac{18}{4 \cdot 13} = 0.3462$$

$$\rho = 0.2692 + 0.2115 + 0.3462 = 0.8269$$

$$\sigma_0 = 0 \quad \sigma_1 = 0.2692 \quad \sigma_2 = 0.4807 \quad \sigma_3 = \rho = 0.8269$$

Para los vehículos de emergencia:

$$W_q^{(1)} = \frac{\left[4!(1-0.8269)(4 \cdot 13)\sum_{n=0}^{3}\frac{(4 \cdot 0.8269)^{i-4}}{i!} + 4 \cdot 13\right]^{-1}}{(1-0)(1-0.2692)} = 0.0170$$

Para los vehículos de servicio:

$$W_q^{(2)} = \frac{\left[4!(1-0.8269)(4 \cdot 13)\sum_{i=0}^{3}\frac{(4 \cdot 0.8269)^{i-4}}{i!} + 4 \cdot 13\right]^{-1}}{(1-0.2692)(1-0.4807)} = 0.0327$$

Para los vehículos particulares:

$$W_q^{(3)} = \frac{\left[4!(1-0.8269)(4 \cdot 13)\sum_{i=0}^{3}\frac{(4 \cdot 0.8269)^{i-4}}{i!} + 4 \cdot 13\right]^{-1}}{(1-0.4807)(1-0.8269)} = 0.1382$$

Resolvemos también el tiempo medio global de espera:

$$W_q = \sum_{i=1}^{k}\frac{\lambda_i}{\lambda}W_q^{(i)} = \frac{14}{43} \cdot 0.0170 + \frac{11}{43} \cdot 0.0327 + \frac{18}{43} \cdot 0.1382 = 0.0717$$

De media los vehículos de emergencia pasan 0.0170 horas = 1.02 minutos en la cola, los vehículos de servicio 0.0327 horas = 1.962 minutos, y los que más tiempo pasan son los vehículos particulares, que de media esperan 0.1382 horas = 8.292 minutos para ser atendidos. El tiempo medio de espera sin tener en cuenta las prioridades es de 0.0717 horas = 4.302 minutos.

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu}$$

$$W^{(1)} = 0.0170 + \frac{1}{13} = 0.0939 \quad W^{(2)} = 0.0327 + \frac{1}{13} = 0.1097 \quad W^{(3)} = 0.1382 + \frac{1}{13} = 0.2151$$

Los tiempos en minutos que pasan los clientes en el sistema de la prioridad más alta a la más baja son: 5.634 min, 6.582 min y 12.906 min, respectivamente.

$$L_q^{(k)} = \lambda_k W_q^{(k)} \qquad L^{(k)} = \lambda_k W^{(k)}$$

$$L_q^{(1)} = 14 \cdot 0.0170 = 0.2379 \quad L^{(1)} = 14 \cdot 0.0939 = 1.3148$$

$$L_q^{(2)} = 11 \cdot 0.0327 = 0.3600 \quad L^{(2)} = 11 \cdot 0.1097 = 1.2062$$
$$L_q^{(3)} = 18 \cdot 0.1382 = 2.4873 \quad L^{(3)} = 18 \cdot 0.2151 = 3.8719$$

Como era de esperar, tanto en la cola como en el sistema predominan los vehículos particulares. Ya que son los que más tiempo pasan en el sistema.

1.2.2. Tasas de servicio diferentes

Por último, en el análisis sin interrupción, estudiaremos cómo se comporta el sistema cuando las tasas de servicio varían, para más de dos clases de prioridad.

1.2.2.1. M|M|1|Prioridad

En este caso asumiremos que el sistema dispone de un solo servidor, las llegadas siguen un proceso de Poisson del mismo modo que los servicios.

Primero calcularemos la probabilidad de que haya al menos un cliente en el sistema, y calculamos la suma de estas probabilidades para todas las clases, denotada como σ_k.

$$\rho_k = \frac{\lambda_k}{\mu_k} \quad (1 \leq k \leq n) \qquad \sigma_k = \sum_{i=1}^{k} \rho_i, \quad (\sigma_0 = 0, \sigma_n = \rho)$$

Para determinar el número global de clientes en la cola, L_q, y el número promedio de clientes en la cola para cada prioridad, $L_q^{(k)}$, realizaremos los siguientes cálculos:

$$L_q^{(k)} = \frac{\lambda_k \sum_{i=1}^{n} \rho_i/\mu_i}{(1-\sigma_{k-1})(1-\sigma)} \qquad L_q = \sum_{i=1}^{n} L_q^{(i)}$$

Por último, calcularemos los tiempos medios de espera, tanto en la cola como en el sistema (Gross et al., 2008).

$$W_q^{(k)} = \frac{\sum_{i=1}^{n} \rho_i/\mu_i}{(1-\sigma_{k-1})(1-\sigma)} \qquad W_q = \sum_{i=1}^{n} \frac{\lambda_i W_q^{(i)}}{\lambda}$$

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu_k} \qquad W = \sum_{i=1}^{n} \frac{\lambda_k W^{(k)}}{\lambda}$$

A continuación, aplicaremos estas fórmulas a un ejemplo práctico.

Ejemplo: Oficina de correos

Consideramos una oficina de correos en la que los clientes se clasifican en diferentes clases de prioridad, basadas en el tipo de servicio requerido: Prioridad 1: Son clientes que necesitan enviar paquetes urgentes. La tasa de llegada de estos clientes es $\lambda_1 = 2$ y la tasa de servicio es $\mu_1 = 7$. Prioridad 2: Clientes que necesitan enviar cartas o paquetes estándar. La tasa de llegada de estos clientes es $\lambda_2 = 4$ y la tasa de servicio es $\mu_2 = 12$. Prioridad 3:

Clientes que realizan consultas o solicitudes de información. La tasa de llegada de estos clientes es $\lambda_3 = 5$ y la tasa de servicio es $\mu_3 = 14$.

En este escenario, la oficina de correos debe gestionar eficientemente el único servidor para minimizar los tiempos de espera.

Estudiaremos el tiempo promedio de espera en la cola, el tiempo promedio de espera en la oficina, el número promedio de clientes en la cola, y en la oficina.

$$\lambda_1 = 2 \qquad \lambda_2 = 4 \qquad \lambda_3 = 5$$

$$\mu_1 = 7 \quad \mu_2 = 12 \quad \mu_3 = 14$$

Empezamos calculando las tasas de ocupación:

$$\rho_k = \frac{\lambda_k}{\mu_k}\,(1 \le k \le n) \qquad \sigma_k = \sum_{i=1}^{k} \rho_i, \quad (\sigma_0 = 0, \sigma_n = \rho)$$

$$\rho_1 = \frac{2}{7} = 0.2857 \qquad \rho_2 = \frac{4}{12} = 0.3333 \qquad \rho_3 = \frac{5}{14} = 0.3571 \qquad \rho = 0.9762$$

$$\sigma_0 = 0 \qquad \sigma_1 = 0.2857 \qquad \sigma_2 = 0.6190 \qquad \sigma_3 = \rho = 0.9762$$

La probabilidad de que el sistema esté lleno es del 97.62 %.

Con estos datos podemos calcular el numero promedio de clientes en la cola:

$$L_q^{(k)} = \frac{\lambda_k \sum_{i=1}^{n} \rho_i/\mu_i}{(1-\sigma_{k-1})(1-\sigma)} \qquad L_q = \sum_{i=1}^{n} L_q^{(i)}$$

$$L_q^{(1)} = \frac{2 \cdot 0.0941}{(1-0)(1-0.2857)} = 0.2635 \qquad L_q^{(2)} = \frac{4 \cdot 0.0941}{(1-0.2857)(1-0.6190)} = 1.3833$$

$$L_q^{(3)} = \frac{5 \cdot 0.0941}{(1-0.6190)(1-0.9761)} = 51.875 \qquad L_q = 53.5218$$

Se observa que el numero de clientes que esperan a realizar una consulta de información es mucho más grande que los de las prioridades más altas.

Por último calculamos los tiempos medios de espera:

$$W_q^{(k)} = \frac{\sum_{i=1}^{n} \rho_i/\mu_i}{(1-\sigma_{k-1})(1-\sigma)} \qquad W_q = \sum_{i=1}^{n} \frac{\lambda_i W_q^{(i)}}{\lambda}$$

$$W_q^{(1)} = \frac{0.0941}{(1-0)(1-0.9761)} = 3.9524 \qquad W_q^{(2)} = \frac{0.0941}{(1-0,2857)(1-0.9761)} = 5.5333$$

$$W_q^{(3)} = \frac{0.0941}{(1-0.6190)(1-0.9762)} = 10.375 \qquad W_q = \frac{81.913}{11} = 7.4466$$

Los tiempos medios de espera en la cola son bastante elevados para todas las prioridades, llegando a alcanzar más de 10 horas de espera en los clientes con prioridad más baja y siendo el tiempo medio de espera global de 7.4466 horas.

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu_k} \qquad W = \sum_{i=1}^{n} \frac{\lambda_i W^{(i)}}{\lambda}$$

$$W^{(1)} = 3.9524 + \frac{1}{7} = 4.0952 \quad W^{(2)} = 5.5333 + \frac{1}{12} = 5.6167 \quad W^{(3)} = 10.375 + \frac{1}{14} = 10.4464$$

$$W = \frac{82.817}{11} = 7.5354$$

En concordancia con los tiempos de espera en la cola, los tiempos medios que pasan los clientes en la oficina también son elevados, desde casi 4 horas para la prioridad alta, hasta más de 10 horas en la más baja. El tiempo medio global es de 7.5354 horas.

1.2.2.2. M|G|1|Prioridad

Para el servicio exponencial, utilizábamos:

$$\mathbb{E}\left[S_0 \mid \text{el sistema está ocupado con un cliente de prioridad } k\right] = \frac{1}{\mu_k}.$$

donde S_0 es el tiempo de servicio que le falta al cliente que está siendo atendido en el momento en que ocurre una llegada.

Ahora generalizamos los resultados anteriores a una distribución de servicio general (Gross et al., 2008). Sea X_k el tiempo de servicio aleatorio de un cliente de prioridad k. Supongamos que X_k sigue una distribución general con primer momento $\mathbb{E}[X_k] = \frac{1}{\mu_k}$ y segundo momento $\mathbb{E}\left[X_k^2\right]$. Entonces,

$$\mathbb{E}[S_0 \mid \text{el sistema está ocupado con un cliente de prioridad } k] = \frac{\mathbb{E}\left[X_k^2\right]\mu_k}{2}.$$

Este resultado también se refiere al tiempo residual promedio de un proceso de renovación cuya distribución entre llegadas es X_k. De este modo obtenemos:

$$\mathbb{E}[S_0] = \rho \sum_{k=1}^{n} \frac{\mathbb{E}\left[X_k^2\right]\mu_k\lambda_k/\mu_k}{2\rho} = \sum_{k=1}^{n} \frac{\mathbb{E}\left[X_k^2\right]\lambda_k}{2} = \frac{\lambda}{2}\sum_{k=1}^{n}\frac{\lambda_k}{\lambda}\mathbb{E}\left[X_k^2\right] = \frac{\lambda\mathbb{E}\left[S^2\right]}{2}.$$

Sabiendo esto, calcularemos las fórmulas correspondientes al modelo de la siguiente manera:

$$\rho_k = \frac{\lambda_k}{\mu_k} \qquad \sigma_k = \sum_{i=1}^{n} \rho_i \quad \text{siendo} \quad \sigma_o = 0$$

$$\text{Siempre que: } \rho = \sum_{i=1}^{n} \frac{\lambda_i}{\mu_i} < 1$$

El tiempo medio de espera en la cola de la clase de prioridad k se calcula como (Winston, 2004):

$$W_q^{(k)} = \frac{\lambda \mathbb{E}\left[S^2\right]/2}{(1-\sigma_{k-1})(1-\sigma_k)} \quad \text{donde} \quad \mathbb{E}\left[S^2\right] = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \mathbb{E}\left[X_i^2\right]$$

El tiempo medio de espera en el sistema para los miembros de prioridad k:

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu_k}$$

El número medio de clientes en la cola y en el sistema se calcula como:

$$L_q^{(k)} = \lambda_k W_q^{(k)} \qquad L^{(k)} = \lambda_k W^{(k)}$$

Ejemplo: Hospital

En un hospital con tres tipos de pacientes de diferentes prioridades, se desea analizar el sistema de colas. El hospital dispone de 3 doctores. Los pacientes de Prioridad Alta llegan con una tasa de $\lambda_1 = 1$ paciente por hora, los de Prioridad Media llegan con una tasa de $\lambda_2 = 1.5$ pacientes por hora y los de Prioridad Baja con una tasa de $\lambda_3 = 1$ paciente por hora. Las tasas de servicio para cada tipo de paciente son las siguientes: $\mu_1 = 10$ pacientes por hora, $\mu_2 = 8$ pacientes por hora, $\mu_3 = 6$ pacientes por hora. Las varianzas del tiempo de servicio son:

$$\text{Var}(X_1) = 0.05, \text{Var}(X_2) = 0.04, \text{Var}(X_3) = 0.02.$$

Supongamos que el tiempo de servicio de un paciente de prioridad k sigue una distribución general con $\mathbb{E}[X_k] = \frac{1}{\mu_k}$ y segundo momento $\mathbb{E}\left[X_k^2\right]$.

Queremos calcular el tiempo medio de espera en la cola y en el hospital, y el número medio de pacientes en la cola y en el hospital para cada tipo de paciente.

Para Prioridad Alta ($k = 1$):

$$\lambda_1 = 1 \quad \mu_1 = 5, \quad \mathbb{E}[X_1] = \frac{1}{5}, \quad \text{Var}(X_1) = 0.05$$

$$\mathbb{E}\left[X_1^2\right] = 0.05 + \left(\frac{1}{5}\right)^2 = 0.05 + 0.04 = 0.09$$

Para Prioridad Media ($k = 2$):

$$\lambda_2 = 1.5 \quad \mu_2 = 4, \quad \mathbb{E}[X_2] = \frac{1}{4}, \quad \text{Var}(X_2) = 0.04$$

$$\mathbb{E}\left[X_2^2\right] = 0.04 + \left(\frac{1}{4}\right)^2 = 0.04 + 0.0625 = 0.1025$$

Para Prioridad Baja ($k = 3$):

$$\lambda_3 = 1 \quad \mu_3 = 6, \quad \mathbb{E}[X_3] = \frac{1}{6}, \quad \text{Var}(X_3) = 0.02$$

$$\mathbb{E}\left[X_3^2\right] = 0.02 + \left(\frac{1}{6}\right)^2 = 0.02 + 0.0278 = 0.0478$$

Primero calcularemos las tasas de ocupación

$$\rho_1 = \frac{\lambda_1}{\mu_1} = 0.2, \quad \rho_2 = \frac{\lambda_2}{\mu_2} = 0.375, \quad \rho_3 = \frac{\lambda_3}{\mu_3} = 0.1667 \quad \rho = 0.7417$$

La probabilidad de que haya al menos un paciente en el hospital es de 74.17 %.

$$\sigma_k = \sum_{i=1}^{n} \rho_i, \quad \text{siendo} \quad \sigma_0 = 0$$

$$\sigma_1 = \rho_1 = 0.2 \quad \sigma_2 = \rho_1 + \rho_2 = 0.575 \quad \sigma_3 = \rho_1 + \rho_2 + \rho_3 = 0.7417$$

El tiempo medio de espera en la cola de la clase de prioridad k se calcula como:

$$W_q^{(k)} = \frac{\lambda E\left[S^2\right]/2}{(1-\sigma_{k-1})(1-\sigma_k)}$$

donde

$$\mathbb{E}\left[S^2\right] = \frac{1 \cdot 0.09 + 1.5 \cdot 0.1025 + 1 \cdot 0.0478}{3.5} = 0.0833$$

$$W_q^{(1)} = \frac{3.5 \cdot 0.0833/2}{(1-0)(1-0.2)} = 0.1822$$

$$W_q^{(2)} = \frac{3.5 \cdot 0.0833/2}{(1-0.2)(1-0.575)} = 0.4287$$

$$W_q^{(3)} = \frac{3.5 \cdot 0.0833/2}{(1-0.575)(1-0.7417)} = 1.3276$$

Podemos observar como el tiempo de espera en la cola aumenta a la vez que lo hace la prioridad empezando en 0.1822 horas = 10.932 minutos para la prioridad alta, y llegando a alcanzar 1.3276 horas = 79.656 minutos en el caso de la prioridad más baja.

El tiempo medio de espera en el sistema para los miembros de prioridad k:

$$W^{(k)} = W_q^{(k)} + \frac{1}{\mu_k}$$

$$W^{(1)} = 0.1822 + \frac{1}{5} = 0.3822 \quad W^{(2)} = 0.4287 + \frac{1}{4} = 0.6787 \quad W^{(3)} = 1.3276 + \frac{1}{6} = 1.4943$$

Al calcular el tiempo medio de los pacientes en el hospital, vemos cómo llegan a pasar de media 22.932, 40.722 y 89.658 minutos respectivamente.

El número medio de pacientes en la cola y en el hospital se calcula como:

$$L_q^{(k)} = \lambda_k W_q^{(k)} \qquad L_k = \lambda_k W^{(k)}$$

$$L_q^{(1)} = 1 \cdot 0.1822 = 0.1822 \quad L^{(1)} = 1 \cdot 0.3822 = 0.3822$$

$$L_q^{(2)} = 1.5 \cdot 0.4287 = 0.6431 \quad L^{(2)} = 1.5 \cdot 0.6787 = 1.0181$$

$$L_q^{(3)} = 1 \cdot 1.3276 = 1.3276 \quad L^{(3)} = 1 \cdot 1.4943 = 1.4943$$

Llama la atención que en los tres casos el número medio de pacientes es próximo a 1, tanto en el número medio de pacientes en la cola como en el hospital en cada una de las prioridades.

Capítulo 2

Modelos con prioridad con interrupción

En los modelos de prioridad con interrupción, los clientes que estén siendo atendidos pueden ser interrumpidos y devueltos a la cola si llega un cliente de mayor prioridad. El servidor se libera inmediatamente para atender al nuevo cliente de alta prioridad.

Los clientes de menor prioridad expulsados del servicio no pueden volver a entrar hasta que el sistema esté libre de todos los clientes de mayor prioridad. Generalmente, debemos especificar cómo el sistema maneja esto. Las dos suposiciones comunes son: (1) las unidades expulsadas deben comenzar de nuevo, perdiendo el trabajo ya completado, o (2) las unidades expulsadas reanudan el servicio desde el punto de interrupción. Pero debido a la propiedad de pérdida de memoria de la distribución exponencial de los tiempos de servicio, no es necesario definir el punto en el que se reanuda el servicio para un cliente interrumpido, la distribución del tiempo de servicio restante es siempre la misma. Esto simplifica el análisis del sistema y permite una mayor flexibilidad en la gestión de las prioridades.

Debido a que las unidades de mayor prioridad pueden interrumpir a las de menor prioridad, el estado del sistema se determina por el número de clientes de cada clase. A diferencia de las colas sin interrupción, no es necesario especificar la clase del cliente en servicio, ya que siempre será la clase de mayor prioridad presente en el sistema.

2.1. Dos clases

2.1.1. Tasas de servicio iguales

Comenzaremos este nuevo capítulo en el escenario en el cual disponemos de dos clases de prioridad, donde se tienen las mismas tasas de servicio, y los tiempos de servicio siguen una distribución exponencial.

Para calcular el tiempo medio de espera en el sistema para las dos prioridades utilizaremos las siguientes fórmulas (Hillier & Lieberman, 2010):

$$W^{(1)} = \frac{1/\mu}{B_0 B_1}, \quad W^{(2)} = \frac{1/\mu}{B_1 B_2} \quad \text{siendo: } B_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu} \quad B_0 = 1$$

Con las fórmulas de Little calcularemos el tiempo medio de espera en la cola para cada prioridad.

$$W_q^{(1)} = W^{(1)} - \frac{1}{\mu} \qquad W_q^{(2)} = W^{(2)} - \frac{1}{\mu} \qquad W_q = \sum_{k=1}^{2} \frac{\lambda_k}{\lambda} W_q^{(k)}$$

También aplicaremos las fórmulas de Little para calcular el número medio de clientes en el sistema y en la cola.

$$L^{(1)} = \lambda_1 W^{(1)}, \quad L^{(2)} = \lambda_2 W^{(2)}, \quad L_q^{(1)} = W_q^{(1)} \lambda_1 \quad L_q^{(2)} = W_q^{(2)} \lambda_2$$

Ejemplo: centro de datos

En un centro de datos, se procesan dos tipos de tareas con diferentes prioridades: Tareas Críticas y Tareas Regulares, utilizando un sistema de colas con un único servidor. Las tareas críticas llegan a una tasa de $\lambda_1 = 2$ tareas por hora, mientras que las tareas regulares llegan a una tasa de $\lambda_2 = 4$ tareas por hora. El tiempo de servicio para todas las tareas sigue una distribución exponencial con una tasa de $\mu = 10$ tareas por hora. Se busca determinar los tiempos medios de espera en la cola y en el centro, así como el número medio de tareas en espera y en proceso para cada tipo de prioridad. Considerando que siempre que haya una tarea regular en curso esta será interrumpida en el caso de que llegue una tarea crítica.

$$\lambda_1 = 2 \quad \lambda_2 = 4 \quad \mu = 10$$

Calculamos las tasas de ocupación para cada prioridad:

$$\rho_1 = \frac{\lambda_1}{\mu} = \frac{2}{10} = 0.2 \qquad \rho_2 = \frac{\lambda_2}{\mu} = \frac{4}{10} = 0.4 \qquad \rho = 0.2 + 0.4 = 0.6$$

La probabilidad de que el sistema no este vacío es del 60 %.

Calculamos el tiempo medio de espera en el centro para las tareas críticas y las regulares:

$$W^{(k)} = \frac{1/\mu}{B_{k-1} B_k}$$

$$B_0 = 1 \quad B_1 = 1 - \frac{2}{10} = 0.8 \quad B_2 = 1 - \left(\frac{2}{10} + \frac{4}{10}\right) = 1 - 0.6 = 0.4$$

$$W^{(1)} = \frac{1/10}{1 \cdot 0.8} = 0.125 \text{ horas} \quad W^{(2)} = \frac{1/10}{0.8 \cdot 0.4} = 0.3125 \text{ horas}$$

El tiempo medio de espera en la cola para cada prioridad y en general:

$$W_q^{(1)} = W^{(1)} - \frac{1}{\mu} = 0.125 - 0.1 = 0.025 \text{ horas}$$

$$W_q^{(2)} = W^{(2)} - \frac{1}{\mu} = 0.3125 - 0.1 = 0.2125 \text{ horas}$$

$$W_q = \sum_{k=1}^{2} \frac{\lambda_k}{\lambda} W_q^{(k)} \quad \text{Donde: } \lambda = \lambda_1 + \lambda_2 = 2 + 4 = 6$$

$$W_q = \frac{2}{6} \cdot 0.025 + \frac{4}{6} \cdot 0.2125 = 0.0083 + 0.1417 = 0.15 \text{ horas}$$

El tiempo medio de espera para las tareas críticas es de 0.025 horas = 1.5 minutos, para las tareas regulares es de 0.2125 horas = 12.75 minutos y el tiempo medio en cola independientemente de la prioridad es de 0.15 horas = 9 minutos.

Por último, calculamos el número medio de tareas en el centro y en cola.

$$L^{(1)} = \lambda_1 W^{(1)} = 2 \cdot 0.125 = 0.25 \quad L^{(2)} = \lambda_2 W^{(2)} = 4 \cdot 0.3125 = 1.25$$

$$L_q^{(1)} = W_q^{(1)} \lambda_1 = 0.025 \cdot 2 = 0.05 \quad L_q^{(2)} = W_q^{(2)} \lambda_2 = 0.2125 \cdot 4 = 0.85$$

Hay más tareas regulares tanto en el centro como en cola.

2.1.2. Tasas de servicio diferentes

En el caso en el que las tasas de servicio sean diferentes para las dos prioridades, calcularemos el número medio de clientes de mayor prioridad en el sistema, $L^{(1)}$, de la misma manera que si estuviéramos en un modelo M|M|1|FIFO (Gross et al., 2008). Mientras que para calcular el número medio de clientes de la prioridad más baja, tendremos en cuenta tanto las tasas de ocupación como las tasas de servicio de ambas prioridades.

$$L^{(1)} = \frac{\rho_1}{1 - \rho_1} \quad L^{(2)} = \frac{\rho_2 - \rho_1\rho_2 + \rho_1\rho_2 \, (\mu_2/\mu_1)}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Ejemplo: Supermercado

Queremos determinar el número de clientes que hay en un supermercado, el cual dispone de un modelo de colas con prioridad para dos clases de clientes: Clientes con compras de un solo artículo y clientes con compras grandes. Los clientes con compras pequeñas llegan al supermercado con una tasa promedio de $\lambda_1 = 5$ clientes por hora y son atendidos con una tasa promedio de $\mu_1 = 20$ clientes por hora por cajero, mientras que los clientes con compras grandes llegan a una tasa promedio de $\lambda_2 = 10$ clientes por hora y son atendidos con una tasa promedio de $\mu_2 = 14$ clientes por hora por cajero. Los clientes que solo compren un articulo tienen prioridad sobre los clientes con compras grandes, pudiendo interrumpir el servicio de estos últimos. Nos interesa saber cuántos clientes de cada prioridad hay en el supermercado, considerando las diferentes tasas de llegada y servicio.

$$\lambda_1 = 5 \quad \lambda_2 = 10 \quad \mu_1 = 20 \quad \mu_2 = 14$$

Para calcular el número medio de clientes en el supermercado, primero calcularemos las tasas de ocupación de cada prioridad y la suma de ambas.

$$\rho_1 = \frac{5}{20} = 0.25 \qquad \rho_2 = \frac{10}{14} = 0.7143 \qquad \rho = 0.25 + 0.7143 = 0.9643$$

$$L^{(1)} = \frac{0.25}{1-0.25} = 0.3333 \quad L^{(2)} = \frac{0.7143 - 0.25 \cdot 0.7143 + 0.25 \cdot 0.7143\,(14/20)}{(1-0.25)(1-0.25-0.7143)} = 24.6667$$

Para los clientes que compran un solo artículo, el número medio de estos en el supermercado es de 0.3333, mientras que para los clientes con compras más grandes, el número medio de clientes es de 24.6667.

Esto indica que estos clientes con compras pequeñas son atendidos de manera más eficiente y no se acumulan en el supermercado.

Sin embargo, esto resulta en un mayor tiempo de espera y un mayor número de clientes en el supermercado para aquellos con compras grandes.

2.2. Más de dos clases

2.2.1. Tasas de servicio iguales

2.2.1.1. M|M|1

En este apartado estudiaremos el modelo con interrupción que tiene tasas de servicio iguales y un único servidor en el sistema.

Para calcular el tiempo medio de espera en el sistema para la prioridad k aplicamos la siguiente fórmula:

$$W^{(k)} = \frac{1/\mu}{B_{k-1}B_k}, \quad \text{para } k = 1, 2, \ldots, n \quad \text{siendo } B_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu} \qquad B_0 = 1$$

El tiempo medio de espera en la cola para la prioridad k y tiempo medio de espera global en la cola considerando todas las prioridades, se define como:

$$W_q^{(k)} = W^{(k)} - \frac{1}{\mu} \quad W_q = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda} W_q^{(i)}$$

Los resultados correspondientes a la cola también se pueden obtener a partir de $W^{(k)}$ y $L^{(k)}$ igual que en el caso de prioridades sin interrupción.

Número medio de clientes en el sistema y en la cola para la prioridad k:

$$L^{(k)} = \lambda_k W^{(k)}, \quad L_q^{(k)} = W_q^{(k)} \lambda_k$$

Aplicamos las fórmulas con el siguiente ejemplo:

Ejemplo: Rehabilitación de animales

En un centro de rehabilitación de fauna silvestre, se llevan a cabo tres tipos de actividades con diferentes prioridades: Rehabilitación de Animales en Peligro de Extinción, Rehabilitación de Animales Heridos y Cuidados Generales. Se desea analizar el sistema de colas con un único cuidador disponible. Las actividades de rehabilitación de animales en peligro de extinción ocurren a una tasa de $\lambda_1 = 0.5$ veces por hora, las actividades de reha-

bilitación de animales heridos a una tasa de $\lambda_2 = 1$ vez por hora y los cuidados generales a una tasa de $\lambda_3 = 1.5$ veces por hora. El tiempo de servicio para todas las actividades sigue una distribución exponencial con una tasa de $\mu = 5$ actividades por hora. Se busca determinar los tiempos medios de espera en la cola y en el centro, así como el número medio de actividades para cada tipo de prioridad, considerando que siempre que haya que realizar una actividad con mayor prioridad que otra en curso, esta será interrumpida.

$$\lambda_1 = 0.5 \quad \lambda_2 = 1 \quad \lambda_2 = 1.5 \quad \mu = 5$$

Calculamos la probabilidad de que el sistema no esté vacío:

$$\rho_1 = \frac{\lambda_1}{\mu} = \frac{0.5}{5} = 0.1 \quad \rho_2 = \frac{\lambda_2}{\mu} = \frac{1}{5} = 0.2 \quad \rho_3 = \frac{\lambda_3}{\mu} = \frac{1.5}{5} = 0.3$$

$$\rho = 0.1 + 0.2 + 0.3 = 0.6$$

Para poder obtener el tiempo que pasan los animales en el centro calculamos los factores de suma B_k

$$W^{(k)} = \frac{1/\mu}{B_{k-1}B_k}, \quad B_k = 1 - \sum_{i=1}^{k} \frac{\lambda_i}{\mu} \qquad B_0 = 1$$

$$B_1 = 1 - \rho_1 = 1 - 0.1 = 0.9$$

$$B_2 = 1 - (\rho_1 + \rho_2) = 1 - (0.1 + 0.2) = 1 - 0.3 = 0.7$$

$$B_3 = 1 - (\rho_1 + \rho_2 + \rho_3) = 1 - (0.1 + 0.2 + 0.3) = 1 - 0.6 = 0.4$$

El tiempo medio en el centro para las diferentes prioridades es:

Para $k = 1$:

$$W^{(1)} = \frac{1/\mu}{B_0 B_1} = \frac{1/5}{1 \cdot 0.9} = 0.2222 \text{ horas}$$

Para $k = 2$:

$$W^{(2)} = \frac{1/\mu}{B_1 B_2} = \frac{1/5}{0.9 \cdot 0.7} = 0.3175 \text{ horas}$$

Para $k = 3$:

$$W^{(3)} = \frac{1/\mu}{B_2 B_3} = \frac{1/5}{0.7 \cdot 0.4} = 0.7143 \text{ horas}$$

Los animales en peligro de extinción pasan de media en el centro 13.332 minutos, los animales heridos 19.05 minutos y los que se clasifican en cuidados generales pasan de media 42.858 minutos en el centro.

El tiempo medio de espera en la cola para actividades de prioridad k se calcula como:

$$W_q^{(k)} = W_k - \frac{1}{\mu}$$

Para $k = 1$:

$$W_q^{(1)} = 0.2222 - \frac{1}{5} = 0.2222 - 0.2 = 0.0222 \text{ horas}$$

Para $k = 2$:

$$W_q^{(2)} = 0.3175 - \frac{1}{5} = 0.3175 - 0.2 = 0.1175 \text{ horas}$$

Para $k = 3$:

$$W_q^{(3)} = 0.7143 - \frac{1}{5} = 0.7143 - 0.2 = 0.5143 \text{ horas}$$

Respecto al tiempo de espera en la cola, también se observa un patrón ascendente. El tiempo de espera para los animales en peligro de extinción es de 1.332 minutos, para los heridos es de 7.05 minutos y, por último, para los cuidados generales es de 30.858 minutos.

Por último, calculamos el número medio de animales en el centro y en la cola para prioridad k:

$$L^{(k)} = \lambda_k W^{(k)} \quad L_q^{(k)} = W_q^{(k)} \lambda_k$$

Para $k = 1$:

$$L^{(1)} = 0.5 \cdot 0.2222 = 0.1111 \quad L_q^{(1)} = 0.0222 \cdot 0.5 = 0.0111$$

Para $k = 2$:

$$L^{(2)} = 1 \cdot 0.3175 = 0.3175 \quad L_q^{(2)} = 0.1175 \cdot 1 = 0.1175$$

Para $k = 3$:

$$L^{(3)} = 1.5 \cdot 0.7143 = 1.0714 \quad L_q^{(3)} = 0.5143 \cdot 1.5 = 0.7714$$

Los resultados muestran que el número medio de animales en el centro y en la cola varía ligeramente según la prioridad. Para los animales en peligro de extinción, el número medio de animales en el centro es 0.1111 y en la cola es 0.0111. Para los animales heridos, estos valores son 0.3175 y 0.1175, respectivamente. Por último, para los cuidados generales ($k = 3$), los valores son 1.0714 en el centro y 0.7714 en la cola. Aunque hay un aumento en el número medio de animales que necesitan cuidados generales, la diferencia es relativamente pequeña.

2.2.1.2. M|M|c|c

En esta sección estudiaremos los sistemas $M|M|c|c$ donde c es el número de servidores y la capacidad del sistema, por lo que ambos valores son iguales. Esto significa que no hay posibilidad de cola.

Denotaremos como A_k a:

$$A_k = \frac{\lambda_k}{\mu}, \quad k = 1, 2, \ldots, n$$

La fórmula de Erlang B (Erlang, 1917) se utiliza para calcular la probabilidad de que un cliente sea bloqueado en un sistema de colas sin cola de espera, $M|M|c|c$. Esta fórmula devuelve la probabilidad de que un nuevo cliente sea bloqueado debido a la falta de servidores disponibles (Zukerman, 2022).

Se define como:

$$E_c\left(A_k\right) = \frac{\frac{A_k^c}{c!}}{\sum_{i=0}^{c} \frac{A_k^i}{i!}}$$

Probabilidad de Bloqueo para Prioridad 1:

$$P_b(1) = E_c\left(A_1\right)$$

Donde $E_c\left(A_1\right)$ representa la probabilidad de bloqueo para los clientes de prioridad 1. Estos clientes tienen la capacidad de interrumpir el servicio de clientes de prioridades más bajas, asegurando que siempre tengan acceso al sistema independientemente de la carga generada por las otras prioridades.

Probabilidad de Bloqueo para Prioridades $k > 1$:

$$P_b(k) = \frac{\left(\sum_{j=1}^{k} A_j\right) E_c\left(\sum_{j=1}^{k} A_j\right) - \left(\sum_{j=1}^{k-1} A_j\right) E_c\left(\sum_{j=1}^{k-1} A_j\right)}{A_k}$$

Esta fórmula calcula la probabilidad de bloqueo para clientes de prioridad k como la relación entre el tráfico perdido de prioridad k y el tráfico ofrecido de prioridad k.

La probabilidad de bloqueo es una medida de rendimiento que indica la probabilidad de que un cliente sea rechazado o bloqueado cuando intenta acceder al sistema porque todos los servidores están ocupados.

Ejemplo: Fábrica

En una fábrica, se gestionan tres tipos de pedidos con diferentes prioridades (alta, media y baja) utilizando un sistema con $c = 8$ máquinas y una capacidad del sistema de $c = 8$. Los pedidos de alta prioridad llegan a una tasa de $\lambda_1 = 15$ pedidos por minuto, los pedidos de prioridad media llegan a una tasa de $\lambda_2 = 22$ pedidos por minuto, y los pedidos de baja prioridad llegan a una tasa de $\lambda_3 = 18$ pedidos por minuto. El tiempo de procesamiento para todos los pedidos sigue una distribución exponencial con una tasa $\mu = 5$ pedidos por minuto. Utilizando las fórmulas de Erlang B, se pide calcular la probabilidad de bloqueo para cada una de las prioridades.

$$\lambda_1 = 15 \quad \lambda_2 = 22 \quad \lambda_3 = 18 \quad \mu = 5 \quad c = 8$$

Cálculo las A_k:

$$A_k = \frac{\lambda_k}{\mu}, \quad k = 1, 2, 3$$

$$A_1 = \frac{15}{5} = 3 \qquad A_2 = \frac{22}{5} = 4.4 \qquad A_3 = \frac{18}{5} = 3.6$$

Probabilidad de Bloqueo para la Prioridad alta:

$$P_b(1) = E_8(A_1) \qquad E_8(A_1) = \frac{\frac{A_1^c}{c!}}{\sum_{i=0}^{c} \frac{A_1^i}{i!}}$$

$$E_8(3) = \frac{\frac{3^8}{8!}}{\sum_{i=0}^{8} \frac{3^i}{i!}} \qquad E_8(3) = 0.0081 \qquad P_b(1) = 0.0081$$

La probabilidad de bloqueo para los pedidos de alta prioridad es prácticamente nula, indicando que estos pedidos tienen muy pocas probabilidades de ser rechazados.

Probabilidad de Bloqueo para la Prioridad media:

$$P_b(2) = \frac{(A_1 + A_2) E_8(A_1 + A_2) - A_1 E_8(A_1)}{A_2} \qquad E_8(A_1 + A_2) = \frac{\frac{(A_1 + A_2)^8}{8!}}{\sum_{i=0}^{8} \frac{(A_1 + A_2)^i}{i!}}$$

$$E_8(7.4) = \frac{\frac{7.4^8}{8!}}{\sum_{i=0}^{8} \frac{7.4^i}{i!}} \qquad E_8(7.4) = 0.2018$$

$$P_b(2) = \frac{7.4 \cdot 0.2018 - 3 \cdot 0.0081}{4.4} = 0.3338$$

La probabilidad de bloqueo para los pedidos de prioridad media es significativamente alta (33.38 %), lo que indica que aproximadamente la tercera parte de estos pedidos podrían ser bloqueados debido a la falta de capacidad del sistema.

Probabilidad de Bloqueo para la Prioridad baja:

$$P_b(3) = \frac{(A_1 + A_2 + A_3) E_8(A_1 + A_2 + A_3) - (A_1 + A_2) E_8(A_1 + A_2)}{A_3}$$

$$E_8(A_1 + A_2 + A_3) = \frac{\frac{(A_1 + A_2 + A_3)^8}{8!}}{\sum_{i=0}^{8} \frac{(A_1 + A_2 + A_3)^i}{i!}} \qquad E_8(11) = \frac{\frac{11^8}{8!}}{\sum_{i=0}^{8} \frac{11^i}{i!}} = 0.3828$$

$$P_b(3) = \frac{11 \cdot 0.3828 - 7.4 \cdot 0.2018}{3.6} = 0.7548$$

La probabilidad de bloqueo para los pedidos de baja prioridad es del 75.48 %, lo que también es considerablemente alta, y mucho mayor que la de los pedidos de prioridad

media. Esto se debo a que los pedidos de baja prioridad siempre van a depender de que no haya pedidos de las otras prioridades.

2.2.2. Tasas de servicio diferentes

2.2.2.1. M|M|1

En esta sección, analizaremos un sistema de colas con diferentes tasas de servicio para cada tipo de cliente, y en el cual dispondremos de un solo servidor, y tanto los tiempos entre llegadas como los tiempos de servicio siguen una distribución exponencial.

Estudiaremos el llamado "Retraso Medio de un Cliente" (Zukerman, 2022).

Para calcular las tasas de ocupación, lo haremos de la siguiente manera:

$$\rho_k = \frac{\lambda_k}{\mu_k}, \quad k = 1, 2, \ldots, n$$

Donde ρ_k es la demanda ofrecida a los clientes de prioridad k. Representa la tasa de llegadas de clientes de prioridad k relativa a la tasa de servicio.

$$R(k) = \sum_{i=1}^{k} \frac{\lambda_i}{\mu_i^2} = \sum_{i=1}^{k} \frac{\rho_i}{\mu_i}$$

siendo $R(k)$ el tiempo residual medio para todos los clientes hasta prioridad k.

La fórmula para calcular el tiempo medio de retraso para las diferentes prioridades es la siguiente:

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1) + R(1)}{1-\rho_1}$$

$$E[D(k)] = \frac{(1/\mu_k)\left(1 - \sum_{i=1}^{k} \rho_i\right) + R(k)}{\left(1 - \sum_{i=1}^{k-1} \rho_i\right)\left(1 - \sum_{i=1}^{k} \rho_i\right)}$$

Sustituyendo $R(1)$ en la primera, obtenemos:

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1) + \rho_1/\mu_1}{1-\rho_1} = \frac{1}{\mu_1(1-\rho_1)}$$

Ejemplo: Muelle de Carga

En una terminal de carga, se gestionan tres tipos de vehículos con diferentes prioridades para ser cargados: camiones de alta, media y baja prioridad. Los camiones de prioridad alta llegan a una tasa de $\lambda_1 = 1$ vehículos por hora, los camiones de prioridad media llegan a una tasa de $\lambda_2 = 2$ vehículos por hora, y los camiones de prioridad baja llegan a una tasa de $\lambda_3 = 3$ vehículos por hora. Los tres tipos de vehículos son atendidos a diferentes tasas de servicio de $\mu_1 = 5, \mu_2 = 8$ y $\mu_3 = 13$ vehículos por hora. En esta terminal, los camiones de mayor prioridad interrumpen a los de menor prioridad si la estación de carga

está ocupada. Se pide calcular el tiempo medio de retraso para cada una de las prioridades.

$$\lambda_1 = 1 \quad \lambda_2 = 2 \quad \lambda_3 = 3 \quad \mu_1 = 5 \quad \mu_2 = 8 \quad \mu_3 = 13$$

$$\rho_1 = \frac{1}{5} = 0.2 \quad \rho_2 = \frac{2}{8} = 0.25 \quad \rho_3 = \frac{3}{13} = 0.2308 \quad \rho = 0.2 + 0.25 + 0.2308 = 0.6808$$

$$R(1) = \frac{\rho_1}{\mu} = \frac{0.2}{5} = 0.04$$

$$R(2) = \frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2} = \frac{0.2}{5} + \frac{0.25}{8} = 0.0713$$

$$R(3) = \frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2} + \frac{\rho_3}{\mu_3} = \frac{0.2}{5} + \frac{0.25}{8} + \frac{0.2308}{13} = 0.0890$$

$$E[D(1)] = \frac{1}{\mu_1 (1 - \rho_1)} = \frac{1}{5 \cdot (1 - 0.2)} = 0.25$$

$$E[D(2)] = \frac{(1/\mu_2)(1 - \rho_1 - \rho_2) + R(2)}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} = \frac{(1/8)(1 - 0.2 - 0.25) + 0.0713}{(1 - 0.2)(1 - 0.2 - 0.25)} = 0.3182$$

$$E[D(3)] = \frac{(1/\mu_3)(1 - \rho_1 - \rho_2 - \rho_3) + R(3)}{(1 - \rho_1 - \rho_2)(1 - \rho_1 - \rho_2 - \rho_3)} =$$

$$= \frac{(1/13)(1 - 0.2 - 0.25 - 0.2308) + 0.0890}{(1 - 0.2 - 0.25)(1 - 0.2 - 0.25 - 0.2308)} = 0.6468$$

Los camiones de alta prioridad experimentan un tiempo de retraso relativamente corto, tan solo 0.25 horas = 15 minutos, los camiones de prioridad media, 0.3182 horas = 19.092 minutos, mientras que los camiones de baja prioridad tienen un tiempo de retraso significativamente mayor, siendo este de 38.808 minutos.

2.2.2.2. M|G|1

En este apartado, estudiaremos el caso en el cual los tiempos de servicio en el sistema no siguen una distribución exponencial. Este caso, que llamaremos general, representa una generalización del modelo M|M|1 que hemos visto en el apartado anterior.

El retraso medio de un cliente de prioridad k se expresa mediante la siguiente ecuación:

$$E[D(k)] = \frac{1}{\mu_k} + \frac{R(k)}{1 - \sum_{i=1}^{k} \rho_i} + E[D(k)] \sum_{i=1}^{k-1} \rho_i$$

donde $R(k)$ representa el tiempo residual medio de todos los clientes de las clases i =

$1, 2, \ldots, k$ (Zukerman, 2022):

$$R(k) = \frac{1}{2}\sum_{i=1}^{k} \lambda_i \mathbb{E}\left[S_i^2\right] \quad \text{siendo: } \mathbb{E}\left[S_i^2\right] = \text{Var}\left(X_i\right) + \left(\frac{1}{\mu_i}\right)^2$$

suponiendo que X_k es el tiempo de servicio de un cliente de prioridad k.

Despejando $E[D(k)]$ de la ecuación obtenemos las fórmulas para calcular el tiempo medio de retraso para las diferentes prioridades.

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1) + R(1)}{1-\rho_1}$$

$$E[D(k)] = \frac{\left(\frac{1}{\mu_k}\right)\left(1 - \sum_{i=1}^{k}\rho_i\right) + R(k)}{\left(1 - \sum_{i=1}^{k-1}\rho_i\right)\left(1 - \sum_{i=1}^{k}\rho_i\right)}.$$

Podemos calcular el número medio clientes de la siguiente manera (Gross et al., 2008):

$$L^{(k)} = \frac{\rho_k}{1-\sigma_{k-1}} + \frac{\lambda_k \sum_{i=1}^{k} \lambda_i \mathbb{E}\left[S_i^2\right]}{2\left(1-\sigma_{k-1}\right)\left(1-\sigma_k\right)}$$

donde $\sigma_k = \sum_{i=1}^{k} \rho_i$.

Vamos a ver la aplicación práctica de estas fórmulas con un ejemplo:

Ejemplo: Teatro

En un teatro se venden entradas para tres tipos de clientes: clientes preferentes altos, clientes preferentes bajos, y clientes regulares. Cada tipo de cliente tiene una prioridad diferente en la cola de espera para comprar entradas. El teatro desea estudiar el tiempo de retraso promedio para cada cliente.

Los parámetros del sistema son los siguientes (con los tiempos en minutos):

Para clientes preferentes altos:

$$\lambda_1 = 1.5 \quad \mu_1 = 12 \quad \mathbb{E}\left[X_1\right] = \frac{1}{12} \quad \text{Var}\left(X_1\right) = 0.04 \quad \rho_1 = 1.5/12 = 0.125$$

$$\mathbb{E}\left[S_1^2\right] = 0.04 + \left(\frac{1}{12}\right)^2 = 0.0469$$

Para clientes preferentes bajos:

$$\lambda_2 = 2 \quad \mu_2 = 8 \quad \mathbb{E}\left[X_2\right] = \frac{1}{8} \quad \text{Var}\left(X_2\right) = 0.1 \quad \rho_2 = 2/8 = 0.25$$

$$\mathbb{E}\left[S_2^2\right] = 0.1 + \left(\frac{1}{8}\right)^2 = 0.1156$$

Y para clientes regulares:

$$\lambda_3 = 2.5 \quad \mu_3 = 5 \quad \mathbb{E}[X_3] = \frac{1}{5} \quad \mathrm{Var}(X_3) = 0.002 \quad \rho_3 = 2.5/5 = 0.5$$

$$\mathbb{E}\left[S_3^2\right] = 0.002 + \left(\frac{1}{5}\right)^2 = 0.042$$

Cálculo de $R(1), R(2), R(3)$:

$$R(k) = \frac{1}{2}\sum_{i=1}^{k} \lambda_i \mathbb{E}\left[S_i^2\right]$$

$$R(1) = \frac{1}{2} \cdot 1.5 \cdot 0.0469 = 0.0352$$

$$R(2) = \frac{1}{2}(1.5 \cdot 0.0469 + 2 \cdot 0.1159) = 0.1511$$

$$R(3) = \frac{1}{2}(1.5 \cdot 0.0469 + 2 \cdot 0.1159 + 2.5 \cdot 0.042) = 0.2036$$

Cálculo del tiempo medio de retraso:

$$E[D(1)] = \frac{(1/\mu_1)(1-\rho_1) + R(1)}{1-\rho_1} = \frac{(1/12)(1-0.125) + 0.0352}{1-0.125} = 0.1236 \text{ minutos}$$

$$E[D(2)] = \frac{(1/\mu_2)(1-\rho_1-\rho_2) + R(2)}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{(1/8)(1-0.125-0.25) + 0.1511}{(1-0.125)(1-0.125-0.25)} = 0.4187 \text{ minutos}$$

$$E[D(3)] = \frac{(1/\mu_3)(1-\rho_1-\rho_2-\rho_3) + R(3)}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} =$$

$$= \frac{(1/5)(1-0.125-0.25-0.5) + 0.2036}{(1-0.125-0.25)(1-0.125-0.25-0.5)} = 2.9227 \text{ minutos}$$

El tiempo medio de retraso para clientes preferentes altos es de 7.416 segundos, para los clientes preferentes bajos es de 25.122 segundos, mientras que para los clientes regulares asciende hasta los 2.9227 minutos.

Por último calculamos el número medio de clientes en la taquilla:

$$L^{(k)} = \frac{\rho_k}{1-\sigma_{k-1}} + \frac{\lambda_k \sum_{i=1}^{k} \lambda_i \mathbb{E}\left[S_i^2\right]}{2(1-\sigma_{k-1})(1-\sigma_k)}$$

$$\sigma_0 = 0 \quad \sigma_1 = 0.125 \quad \sigma_2 = 0.375 \quad \sigma_3 = 0.875$$

$$L^{(1)} = \frac{0.125}{1-0} + \frac{1.5 \cdot (1.5 \cdot 0.0469)}{2(1-0)(1-0.125)} = 0.1854$$

$$L^{(2)} = \frac{0.25}{1-0.125} + \frac{2 \cdot (1.5 \cdot 0.0469 + 2 \cdot 0.1156)}{2(1-0.125)(1-0.375)} = 0.8373$$

$$L^{(3)} = \frac{0.5}{1-0.375} + \frac{2.5 \cdot (1.5 \cdot 0.0469 + 2 \cdot 0.1156 + 2.5 \cdot 0.042)}{2(1-0.375)(1-0.875)} = 7.3067$$

En promedio, hay menos de un cliente preferente (alto o bajo) en la taquilla en cualquier momento, lo que indica un flujo relativamente eficiente para este tipo de clientes. Sin embargo, en cuanto a los clientes regulares el número medio es algo mayor de 7.

Capítulo 3

Manual de la aplicación web

En este capítulo, presentaremos y describiremos el manual de uso de la aplicación web diseñada para obtener las medidas de rendimiento de los modelos de colas con disciplina de prioridad. La aplicación permite analizar diferentes sistemas de colas divididos en dos categorías principales: sistemas con interrupción y sistemas sin interrupción. Además, cada categoría se subdivide en sistemas con dos prioridades y sistemas con más de dos prioridades, y dentro de cada una de estas subcategorías se considera si las tasas de servicio son iguales o diferentes. A lo largo de este manual, detallaremos las funcionalidades de la aplicación y los pasos para realizar el análisis.

Al iniciar la aplicación, se presenta una interfaz gráfica que guía al usuario a través de varias opciones para configurar el modelo de colas que desea analizar. Las opciones y campos de entrada que se muestran dependerán de las elecciones que el usuario haga a lo largo del proceso. El usuario puede seleccionar si su sistema presenta interrupción o no, si el numero de prioridades es 2 o mayor de 2, y especificar las tasas de llegada (λ), las tasas de servicio (μ) y, en algunos casos, las varianzas. Una vez ingresados los parámetros, la aplicación calcula automáticamente las principales medidas de rendimiento del sistema seleccionado.

La aplicación está diseñada para ser flexible y adaptarse a diferentes escenarios de análisis, proporcionando resultados precisos y útiles para la toma de decisiones en la gestión de sistemas de colas con prioridades.

En los siguientes apartados, explicaremos detalladamente cómo funciona la aplicación en algunos de estos escenarios a modo de ejemplo.

3.1. Sin interrupción

En esta primera parte, estudiaremos algunos de los sistemas que no conllevan interrupción.

3.1.1. Dos Clases

3.1.1.1. Tasas de Servicio Iguales

M|M|c|Prioridad

En este escenario, describiremos cómo configurar y utilizar la aplicación para analizar un sistema de colas sin interrupción, con dos clases de prioridad, con tasas de servicio iguales y múltiples servidores.

Paso 1: Seleccionar las opciones del sistema

1. ¿Sin interrupción o con ella?
   - Seleccione la opción “Sin interrupción”.
2. Número de clases:
   - Seleccione “Dos Clases”.
3. ¿Tasas de servicio iguales o diferentes para cada clase?
   - Seleccione “Tasas de servicio iguales”.
4. Modelo del sistema de colas:
   - Seleccione “M|M|c”.

Paso 2: Ingresar los parámetros del sistema

- c: Ingrese el número de servidores. En este ejemplo, ingrese “4”.
- λ_1: Ingrese la tasa de llegadas para la primera clase de prioridad. En este ejemplo: “7”.
- λ_2: Ingrese la tasa de llegada para la segunda clase de prioridad. En este ejemplo: “12”.
- μ: Ingrese la tasa de servicio. En este ejemplo: “5”.

Paso 3: Ejecutar el análisis

Una vez ingresados todos los parámetros, la interfaz gráfica mostrará los resultados calculados, incluyendo las siguientes medidas de rendimiento: $L_q^{(1)}, L_q^{(2)}, L^{(1)}, L^{(2)}, W_q^{(1)}, W_q^{(2)}, W_q, W^{(1)}, W^{(2)}$.

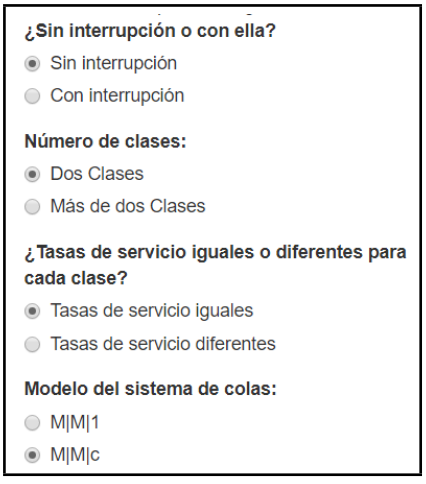

Figura 3.1: Paso 1

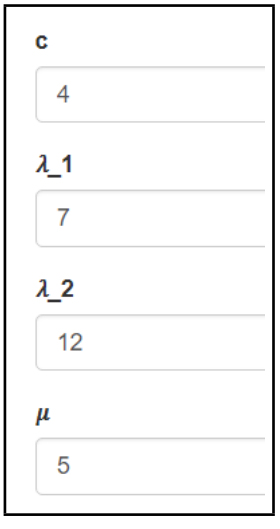

Figura 3.2: Paso 2

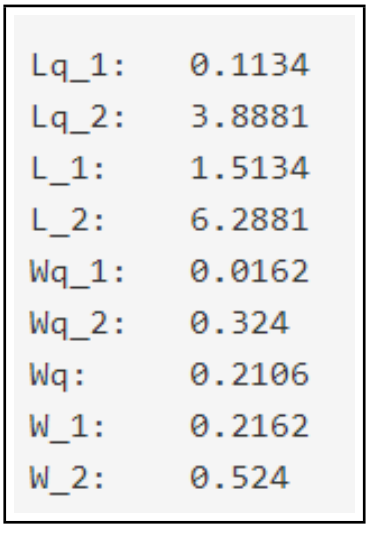

Figura 3.3: Paso 3

3.1.2. Más de Dos Clases

3.1.2.1. Tasas de Servicio Diferentes

M|G|1|Prioridad

En este escenario, describiremos cómo configurar y utilizar la aplicación para analizar un sistema de colas sin interrupción, con más de dos clases de prioridad, con tasas de servicio diferentes y un único servidor, en el cual los tiempos de servicio no siguen una distribución exponencial.

Paso 1: Seleccionar las opciones del sistema

1. ¿Sin interrupción o con ella?
   - Seleccione la opción “Sin interrupción”.
2. Número de clases:
   - Seleccione “Más de dos Clases”.
3. ¿Tasas de servicio iguales o diferentes para cada clase?
   - Seleccione “Tasas de servicio diferentes”.
4. Modelo del sistema de colas:
   - Seleccione “M|G|1”.

Paso 2: Ingresar los parámetros del sistema

- Lista de las λ para cada clase (separadas por comas): Ingrese las tasas de llegadas para cada prioridad, separándolas por comas. En este ejemplo: “1,1.5,1”.
- Lista de las μ para cada clase (separadas por comas): Ingrese las tasas de servicio para cada prioridad, separándolas por comas. En este ejemplo: “5,4,6”.
- Lista de las varianzas para cada clase (separadas por comas): Ingrese las varianzas de las distribuciones de servicio para cada prioridad, separándolas por comas. En este ejemplo: “0.05,0.04,0.02”.

Paso 3: Ejecutar el análisis

Una vez ingresados todos los parámetros, la interfaz gráfica mostrará los resultados calculados, incluyendo las siguientes medidas de rendimiento: $L_q^{(k)}, L^{(k)}, W_q^{(k)}, W^{(k)}$.

Los resultados específicos para cada prioridad k se muestran de izquierda a derecha. Es decir, el primer resultado corresponde a la prioridad 1, el segundo a la prioridad 2, y así sucesivamente hasta la última prioridad introducida.

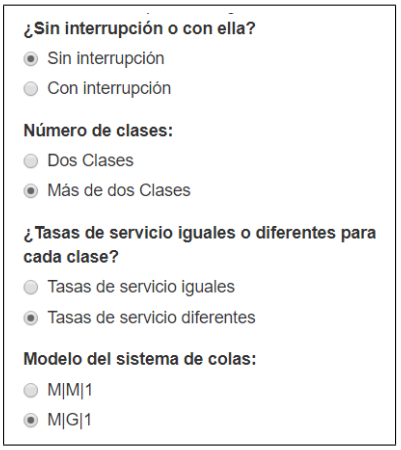

Figura 3.4: Paso 1

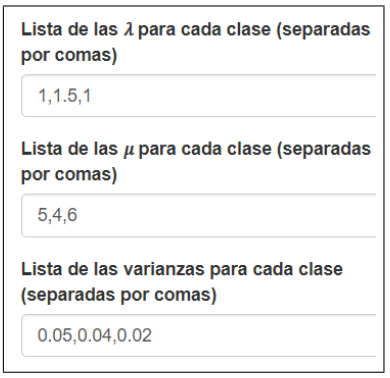

Figura 3.5: Paso 2

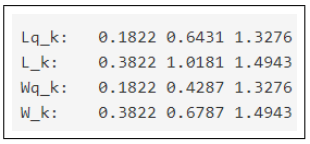

Figura 3.6: Paso 3

3.2. Con interrupción

En esta parte, estudiaremos algunos de los posibles sistemas que conllevan interrupción.

3.2.1. Dos Clases

3.2.1.1. Tasas de Servicio Diferentes

M|M|1|Prioridad

En este escenario, describiremos cómo configurar y utilizar la aplicación para analizar un sistema de colas con interrupción, con dos clases de prioridad, con tasas de servicio diferentes y un único servidor.

Paso 1: Seleccionar las opciones del sistema

1. ¿Sin interrupción o con ella?
   - Seleccione la opción “Con interrupción".
2. Número de clases:
   - Seleccione “Dos Clases”.
3. ¿Tasas de servicio iguales o diferentes para cada clase?
   - Seleccione “Tasas de servicio diferentes”.

Paso 2: Ingresar los parámetros del sistema

- λ_1: Ingrese la tasa de llegada para la primera clase de prioridad. En este ejemplo: “5”.
- λ_2: Ingrese la tasa de llegada para la segunda clase de prioridad. En este ejemplo: “10”.
- μ_1: Ingrese la tasa de servicio para la primera clase de prioridad. En este ejemplo: “20”.

- μ_2: Ingrese la tasa de servicio para la segunda clase de prioridad. En este ejemplo: “14”.

Paso 3: Ejecutar el análisis

Una vez ingresados todos los parámetros, la interfaz gráfica mostrará los resultados calculados, incluyendo las siguientes medidas de rendimiento: $L^{(1)}, L^{(2)}$.

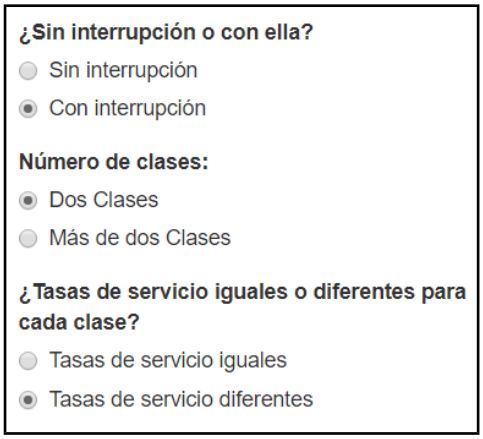

Figura 3.7: Paso 1

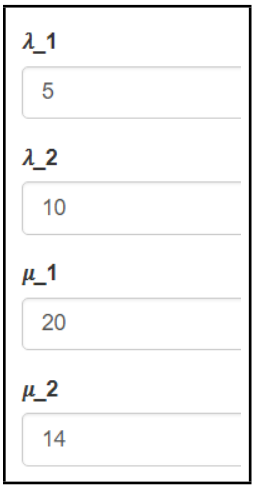

Figura 3.8: Paso 2

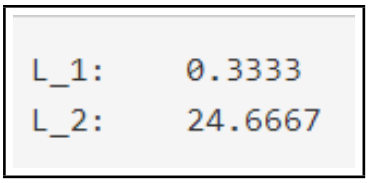

Figura 3.9: Paso 3

3.2.2. Más de Dos Clases

3.2.2.1. Tasas de Servicio Iguales

M|M|c|c|Prioridad

En este escenario, describiremos cómo configurar y utilizar la aplicación para analizar un sistema de colas con interrupción, con más de dos clases de prioridad, con tasas de servicio iguales, múltiples servidores y siendo la capacidad del sistema igual al número de servidores.

Paso 1: Seleccionar las opciones del sistema

1. ¿Sin interrupción o con ella?
   - Seleccione la opción “Con interrupción”.
2. Número de clases:
   - Seleccione “Más de dos Clases”.
3. ¿Tasas de servicio iguales o diferentes para cada clase?
   - Seleccione “Tasas de servicio iguales”.
4. Modelo del sistema de colas:
   - Seleccione “M|M|c|c”.

Paso 2: Ingresar los parámetros del sistema

- c: Ingrese el número de servidores. En este ejemplo, ingrese “8”.

- Lista de las λ para cada clase (separadas por comas): Ingrese las tasas de llegadas para cada prioridad, separando por comas cada valor. En este ejemplo: “15,22,18”.
- μ: Ingrese la tasa de servicio (μ). En este ejemplo: “5”.

Paso 3: Ejecutar el análisis

Una vez ingresados todos los parámetros, la interfaz gráfica mostrará los resultados calculados, incluyendo las siguientes medidas de rendimiento: $P_b^{(k)}$.

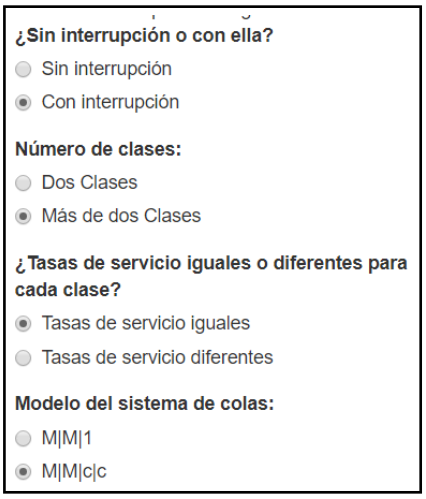

Figura 3.10: Paso 1

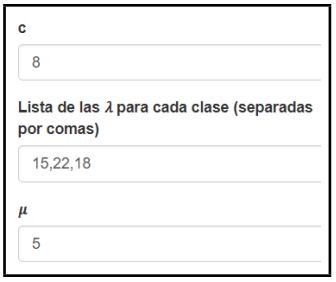

Figura 3.11: Paso 2

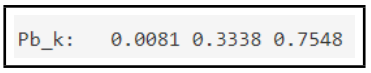

Figura 3.12: Paso 3

Capítulo 4

Conclusiones

En este trabajo, hemos alcanzado satisfactoriamente los propósitos establecidos. A lo largo de nuestra investigación, hemos demostrado que el concepto de prioridad en sistemas de colas es esencial para mejorar el rendimiento y la eficiencia en la gestión de diferentes tipos de clientes y tareas. Nuestro estudio ha abarcado tanto los fundamentos teóricos como su aplicación práctica en diversos contextos.

Hemos explorado en profundidad los sistemas de colas con y sin interrupción, identificando las características distintivas de cada uno y evaluando su impacto en el rendimiento del sistema.

La formulación de una teoría adecuada ha sido muy importante para nuestro análisis. Hemos deducido las condiciones necesarias para evitar la saturación del sistema, formulado las ecuaciones de tráfico pertinentes y calculado diversas medidas de rendimiento, tales como tiempos de espera, probabilidades de bloqueo y número medio de clientes en el sistema.

Para ilustrar la aplicabilidad de nuestra teoría, hemos desarrollado una serie de ejemplos prácticos que reflejan situaciones reales. Estos ejemplos han demostrado cómo los sistemas de colas con prioridades se implementan eficazmente en contextos como fábricas, hospitales y servicios al cliente. Cada ejemplo ha servido para contextualizar la teoría y mostrar su relevancia práctica.

Además, hemos creado una aplicación en R que facilita el análisis y la obtención de medidas de rendimiento para diferentes modelos de colas con prioridades. Esta herramienta es un aporte significativo para investigadores y profesionales del área, permitiendo realizar estudios detallados y obtener resultados precisos de manera eficiente.

En resumen, hemos logrado cumplir con todos los objetivos propuestos, desarrollando un estudio integral y práctico sobre sistemas de colas con prioridades. Nuestra investigación ha proporcionado herramientas prácticas para su aplicación efectiva en diversos contextos reales.

Bibliografía

Cao-Abad, R. (2002). *Introducción a la simulación y a la teoría de colas* (1ª ed.). Netbiblo.

Erlang, A. (1917). Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Engineer's Journal*, *10*, 189-197.

Gross, D., Shortie, J., & Harris, C. (2008). *Fundamentals of Queueing Theory* (4th). John Wiley & Sons.

Hillier, F., & Lieberman, G. (2010). *Introducción a la Investigación de Operaciones* (9th). McGraw-Hill.

Little, J. (1961). A Proof for the Queuing Formula: L = λ W. *Operations Research*, *9*(3), 383-387. https://doi.org/https://doi.org/10.1287/opre.9.3.383

Rodríguez-Rosa, M. (2021). *Apuntes para un curso de Procesos Estocásticos en Tiempo Discreto*. Departamento de Estadística, Universidad de Salamanca.

Winston, W. (2004). *Operations Research: Applications and Algorithms* (4th). Duxbury Press.

Zukerman, M. (2022). *Introduction to Queueing Theory and Stochastic Teletraffic Models*. City University of Hong Kong.