TP 3 - SIMULACIÓN DE UN SISTEMA M/M/1

Leilen Avila
Legajo: 41610
Mail: leilenavila@gmail.com
UTN - FRRO
Zeballos 1341, S2000

Natalia Fernandez
Legajo: 44758
Mail: nata.fernandez77@gmail.com
UTN - FRRO
Zeballos 1341, S2000

9 de julio de 2020

ABSTRACT

Existen muchos modelos en la teoría de colas y en este trabajo se explicara teórica y prácticamente un modelo de colas simples que constará con una única fuente de arribo al sistema y un único servidor. Nuestro estudio se basara fundamentalmente en modelos donde el tiempo entre llegadas de los clientes y el tiempo de servicio seguirán una distribución exponencial, ya que es la distribución que más se da en el sistema de colas debido a su propiedad de pérdida de memoria, donde las llegadas solo depende del momento en el que llegan y no del tiempo entre llegadas. Para poder entender estos modelos previamente clasificaremos el sistema de cola según su estructura: Fuente de entrada, disciplina de la cola y mecanismo de servicio. Mediante la variación de los diferentes parámetros de nuestro sistema obtendremos diferentes comportamientos que serán descriptos con tres enfoques diferentes: analíticamente, de forma simulada utilizando Python y de forma simulada utilizando Anylogic. Análogamente compararemos los resultados obtenidos mediante los tres mecanismos con dichas manipulaciones sacando nuestras propias conclusiones en cada apartado.

1. Introducción

Previo al comienzo de nuestro análisis, sobre el comportamiento de un sistema de colas, explicaremos algunos conceptos básicos para que posteriormente sea de mayor sencillez la comprensión de los resultados generados. Explicaremos en detalle los fundamentos que sustentan el proceso a analizar, tanto de manera analítica como de manera gráfica.

Para comenzar, es bueno decir que la teoría de colas aparece a principios del siglo veinte para estudiar los problemas de congestión de tráfico que se presentaban en las comunicaciones telefónicas. Entre 1903 y 1905, Erlang es el primero en tratar el tráfico telefónico de forma científica, y establece la unidad de tráfico telefónico, que recibe su nombre. Posteriormente esta teoría se ha aplicado a multitud de problemas de la vida real, como el tráfico de automóviles, la regulación de semáforos en una ciudad, la determinación de cajeros en los supermercados, o el control de los tiempos de espera de los procesos que acceden al procesador de un ordenador que trabaja en tiempo compartido. El objetivo es el estudio matemático de colas y líneas de espera.

La teoría de colas utiliza los modelos de colas para representar los tipos de sistemas de líneas de espera. Por lo tanto, estos modelos de líneas de espera son muy útiles para determinar cómo operar un sistema de colas de la manera más eficaz. Proporcionar demasiada capacidad de servicio para operar el sistema implica costos excesivos; pero si no se cuenta con suficiente capacidad de servicio surgen esperas excesivas con todas sus desafortunadas consecuencias. Los modelos permiten encontrar un balance adecuado entre el costo de servicio y la cantidad de espera. A continuación, definiremos algunos conceptos básicos que nos ayudaría a entender la teoría de colas, desde una visión matemática.

2. Marco teórico: Estructura de un modelo de colas.

La estructura del modelo de colas permite que el mismo funcione de la siguiente forma: Los clientes que requieren un servicio se generan en una fuente de llegada. Luego ingresa al sistema y se une en la cola. Si hubiese un servidor libre, su servicio empieza directamente, caso contrario, se une a la cola. Cuando el servidor pase a estar libre, y el cliente sea seleccionado para ser servidor, empezará su servicio. Cuando haya acabado su servicio, el cliente sale del sistema.

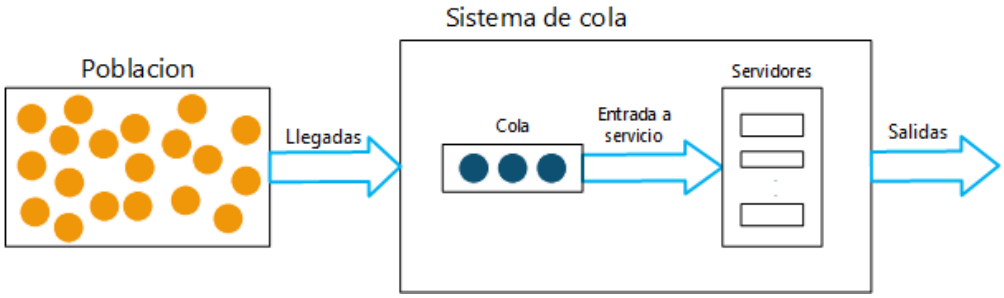

Figura 1: Sistema de colas M/M/1.

Brevemente se detallan cada una de las características que posee el sistema y que se presentaron en la figura anterior. Dichas características principales son:

1. **Fuente de entrada o población.** Su característica es el tamaño. Llamamos tamaño al número total de clientes que pueden requerir servicio en un determinado momento. Podemos suponer que el tamaño es finito o infinito. Se debe especificar el patrón estadístico mediante el cual se generan los clientes a través del tiempo.

2. **Cola.** La cola es el lugar donde los clientes esperan antes de recibir el servicio. Esta posee dos características principales, en primer lugar la capacidad de la cola, es decir, el número máximo de clientes que puede llegar a soportar. Esta capacidad puede ser finita o infinita, el supuesto de cola infinita es el estándar en la mayoría de modelos ya que poner un límite a la cola puede complicar bastante el análisis, solo será necesario el supuesto contrario cuando el límite de cola sea bastante pequeño y se llegue a él con regularidad. El otro factor determinante de la cola es la disciplina que sigue, que la veremos a continuación.

3. **Disciplina de la cola.** La disciplina de la cola se refiere al orden en el que sus miembros se seleccionan para recibir el servicio.
   - FIFO (First-In-First-Out): Se le da servicio al primero que ha llegado, de forma que la cola esta ordenada según el orden de llegada de los usuarios.
   - LIFO (Last-In-First-Out): Se le da servicio al último que ha llegado, de forma que la cola esta ordenada en orden inverso al de llegada de los usuarios.
   - SIRO (Service-In-Random-Order): se sortea aleatoriamente cuál de los usuarios en espera accederá al servicio
   - PS (Processor Sharing): Se otorga a cada cliente un pequeño tiempo de servicio de forma secuencial, hasta completar el tiempo de servicio.
   - PR (Priority): Cada cliente tiene una prioridad de servicio, y los que tienen mayor prioridad son los primeros en ser servidos.

   En sistemas finitos, en los que el número de usuarios en espera es limitado, es necesario establecer además qué sucede con aquellos usuarios que acceden al sistema cuando la cola de espera está completa. Por último, en los sistemas en que los usuarios son humanos, hay que tener en cuenta otros factores propios del comportamiento humano como el hecho de que hay individuos que no respetan el orden establecido en la cola o bien que hay usuarios que, a la vista de la cola, renuncian a acceder al sistema.

4. **El mecanismo de servicio** consiste en una o más estaciones de servicio, cada una de ellas con uno o más servidores o canales de servicio paralelos, llamados servidores. Los modelos de colas deben especificar el número de servidores. Si el tiempo que tardan los usuarios en salir del sistema es mayor que el intervalo entre

llegadas, la cola aumentará indefinidamente y el sistema puede llegar a colapsarse. Por tanto, es necesario diseñar el sistema de forma que el tiempo de servicio sea igual o menor que el intervalo entre llegadas. En esta situación es importante saber cuánto tiempo va a estar un servidor inactivo, tiempo que ha de ser mínimo para optimizar el rendimiento del sistema. No obstante, en la mayora de los sistemas la duración del servicio es también una magnitud aleatoria.

5. **Tiempo de servicio.** Se llama tiempo de servicio (o duración del servicio) al tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una estación. Un modelo de un sistema de colas determinado debe especificar la distribución de probabilidad de los tiempos de servicio de cada servidor (y tal vez de los distintos tipos de clientes), aunque es común suponer la misma distribución para todos los servidores. La distribución del tiempo de servicio que más se usa en la práctica por ser más manejable que cualquier otra es la distribución exponencial. Otras distribuciones de tiempos de servicio importante son la distribución degenerada (tiempos de servicio constante) y la distribución Erlang.

6. La **disciplina del servicio**: es una regla para seleccionar clientes de la linea de espera al inicio del servidor. las disciplinas más utilizadas son:
   - FIFO [First In First Out]: Los primeros en entrar seran los primeros en salir
   - LIFO [Last In First Out]: Los últimos en llegar serán los primeros en salir.
   - Existen otras disciplinas denominadas al azar y de prioridad.

7. **Capacidad del sistema.** Número máximo de clientes que pueden estar en la cola. Si la misma es infinita entonces admite cualquier cantidad de llegadas, caso contrario, si la cola es finita, cualquier nuevo cliente que llegue y supere este límite deberá retirarse.

8. Los **canales de servicio**: Consiste en identificar la cantidad de servidores y el número de etapas o fases de servicio por las que tiene que pasar una entidad.
   Sistema de un sólo canal con una sola fase:

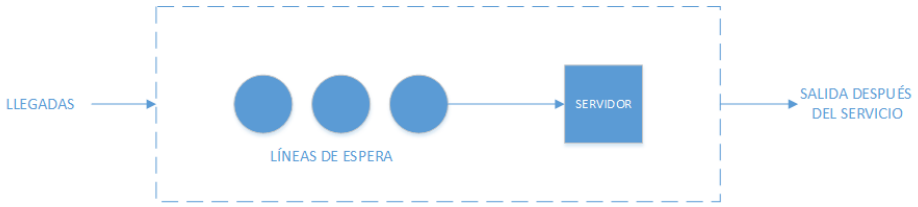

Figura 2: Sistema de un solo canal con una sola fase.

Sistema multicanal con una sola fase:

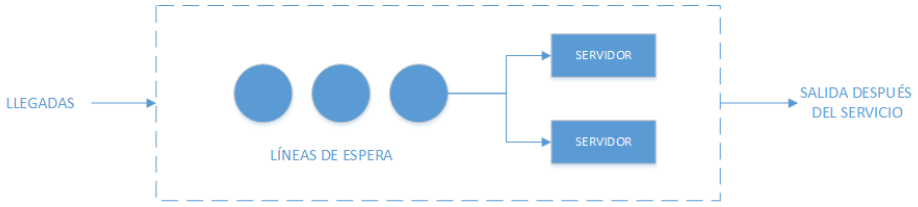

Figura 3: Sistema de un solo canal con una sola fase.

3. Procesos que intervienen en un sistema de colas.

Un sistema de colas está compuesto por un conjunto de procesos que operan de manera conjunta para dar como resultado el modelo de la simulación. En la figura que se encuentra debajo se puede comprender como interactúan entre ellos y en las secciones siguientes detallaremos cada uno en concreto ya que contendrán los principios básicos que necesitaremos para posteriormente comenzar la simulación.

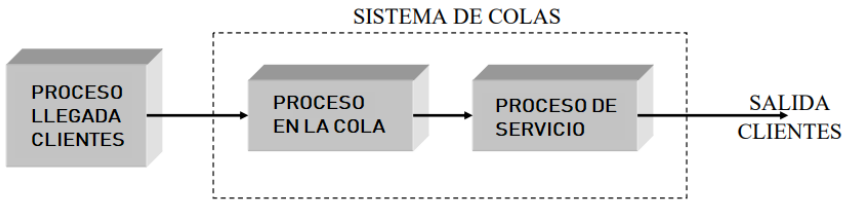

Figura 4: Procesos de un sistema de colas

3.1. El proceso de llegada de los clientes

Se mide a través del número de llegadas por unidad de tiempo y el tiempo que existe entre las sucesivas llegadas. El tiempo que transcurre entre dos llegadas sucesivas en el sistema se conoce como tiempo entre llegadas. Se supone que los usuarios entran al sistema en los tiempos $t_0 < t_1 < t_2 < ... < t_n$.

Cada cliente llegará en un tiempo correspondiente a una variable aleatoria con una determinada distribución. Se asume que todos los clientes dispondrán de una misma distribución de probabilidad, es por esto que el comportamiento individual de cada uno quedará definido por una variable aleatoria idénticamente distribuida(V.A.I.D). Las variables aleatorias $\tau_k = t_k - t_{(k-1)}$ con $k \geq 1$ se llaman intervalos entre arribos y estos se pueden representar gráficamente a continuación:

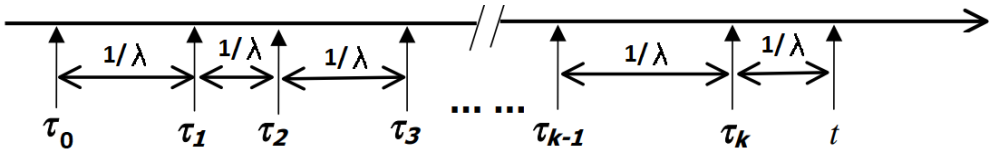

Figura 5: Intervalos de arribo de los clientes

Este tiempo puede ser:

- Determinista: Es constante, es decir, se que cada cierto intervalo de tiempo llega un cliente.
- Probabilístico: La llegada de los clientes es estocástica. Es decir, el tiempo entre los intervalos es incierto y variable.

Sea

- λ : Número promedio de llegadas / Unidad de tiempo, que mide la velocidad de arribo de los clientes.
- $1/\lambda$: Tiempo promedio entre la llegada de cada usuario al sistema.

Sera necesario estimar el número de llegadas por intervalo de tiempo, tomaremos como que este sigue una distribución de Poisson. Para describir los intervalos aleatorios entre los que llegan se utiliza la distribución exponencial.

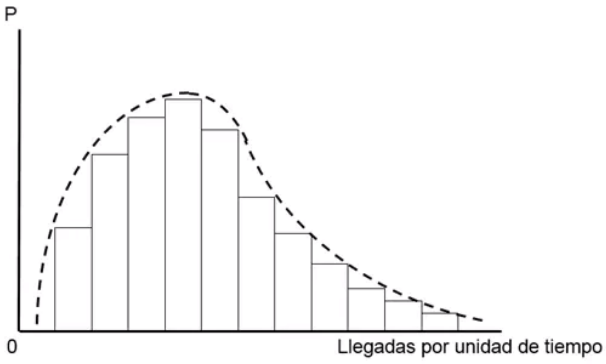

Figura 6: Proceso de poisson. Arribos. Imagen tomada del sitio web: www.cartagena99.com/recursos/teroiacola.pdf

Cada rectángulo indica un intervalo continuo de tiempo en el que arriba un cliente al sistema. Estos intervalos son los que tienen una distribución exponencial y el número de clientes que llega en cada intervalo sigue la distribución de Poisson. Entonces, la probabilidad de que ocurra un número n de llegadas en un intervalo I de tiempo es de:

$$P(X \leq n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!} \tag{1}$$

Como se distribuyen los intervalos según van llegando los clientes al sistemas está definido por:

$$P(\tau \leq t) = 1 - e^{-\lambda t} \tag{2}$$

El conjunto de estas dos distribuciones es lo que se conoce con el nombre de "Proceso de Poisson". Este proceso es una de las disciplinas más utilizadas por su simplicidad, propiedades y características generales. Estas características sólo permiten unos análisis más bien simples, que se ajustan a fuentes de datos en general, pero no válidos para los casos en los que tengamos fuentes de datos más o menos complejos.

3.2. El proceso en la cola

El número de clientes en la cola es el número de clientes que esperan el servicio. Una cola se caracteriza por el número máximo permisible de clientes que puede admitir. Las colas pueden ser finitas o infinitas, según si este número es finito o infinito.

La suposición de una cola infinita es la estándar para la mayor parte de los modelos, incluso en situaciones en las que de hecho existe una cota superior (relativamente grande) sobre el número permitido de clientes, ya que manejar una cota así puede ser un factor complicado para el análisis. Los sistemas de colas en los que la cota superior es tan pequeña que se llega a ella con cierta frecuencia, necesitan suponer una cola finita.

Cuando un cliente ingresa en el sistema pueden ocurrir dos cosas. La primera es que, en el caso de que exista un único servidor y el mismo esté desocupado, el cliente ingresa a ser atendido directamente por este servidor. El segundo caso es cuando el servidor está ocupado, en este caso el cliente se une a la cola en caso de existir, y en caso de que no, se posiciona en primer lugar a la espera de ser atendido. En este momento es cuando empieza a correr el reloj sobre la cantidad de tiempo que demora dicho cliente en ser atendido.

Sumado a lo descripto anteriormente, existe un estado de sistemas de colas que se basa en que en principio el sistema está en un estado transitorio, en donde los usuarios están esperando para ingresar. Luego el sistema llega a una condición de estado estable, siendo este su nivel normal de operación y a la vez existen condiciones anormales como las horas pico. Nos podemos encontrar con diferentes posibilidades en función de los valores que tomen λ y μ.

- $\lambda < \mu$, El sistema es estable y la cola no se llenará.
- $\lambda > \mu$, El sistema se saturará y se llenará la cola de espera.
- $\lambda = \mu$, Es el límite de estabilidad; se servirán tantos clientes como lleguen.

A fin de que la espera no sea muy larga, se intenta que las colas no queden muy ocupadas; eso quiere decir que $\lambda \leq \mu$. Aunque para optimizar el coste, habitualmente se dimensiona el sistema de manera que se considera la posibilidad de una cierta congestión siempre que se garantice un mínimo nivel de calidad del servicio.

3.3. El proceso de servicio

Describe como son atendidos los clientes, y se caracteriza por el tiempo empleado en dar servicio a un cliente y por el numero de servidores de que se dispone.

El servicio puede ser brindado por un solo servidor o por servidores múltiples. El tiempo de servicio varía de cliente en cliente. Es una variable aleatoria con una determinada distribución, por esta razón cada cliente tendrá un valor distinto en su tasa de servicio (Es decir, en el tiempo en que tardarán en atenderlo).

Es por esta razón que es necesario seleccionar una distribución de probabilidad para los tiempos de servicio. El tiempo promedio de un servicio es E(s), el mismo depende de (velocidad promedio de atención de un servicio) y de $1/\mu$ (Tiempo entre servicios, o sea entre que se atiende un servicio y el siguiente). Donde

$$W_s = 1 - e^{-\mu t} \tag{3}$$

El ritmo de atención al cliente cuando el servidor está ocupado sigue una distribución de Poisson (Numero de clientes que el servidor atiende en un intervalo de tiempo). Mientras que como se distribuyen las duraciones de la atención a los clientes sigue una distribución exponencial. Los tiempos de servicio pueden ser constantes, donde para todos los clientes el servidor tardará lo mismo.

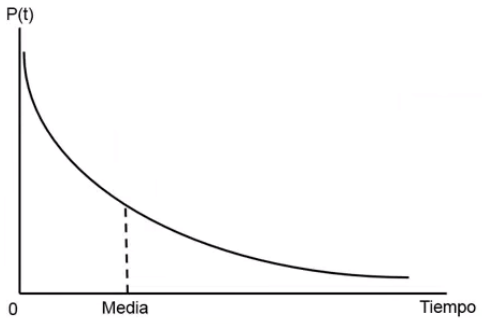

Figura 7: Proceso de poisson. Servicio. Imagen tomada del sitio web: www.cartagena99.com/recursos/teroiacola.pdf

Mientras menor sea el tiempo entre cada servicio más clientes serán atendidos, es decir, la probabilidad aumenta. Este proceso entonces, describe como son atendidos los clientes y se caracteriza por el tiempo empleado en dar servicio a un cliente y el numero de servidores que se dispone. Por lo visto antes, sabemos que el servicio puede ser brindado por un solo servidor o por servidores múltiples.

3.4. Notación Kendall

Todos los sistemas de líneas de espera pueden ser representados de manera más sencilla mediante el uso de la notación desarrollada por A. G. Kendall, la cual tiene el propósito de representar mediante tres caracteres el tipo de sistema a analizar.

La notación Kendall se representa en su forma general, de la siguiente manera:

A/B/c/K/n/Z

A: Distribución del intervalo entre llegadas.
B: Distribución del tiempo de servicio.
c: Numero de servidores.
K: Capacidad del sistema.
n: Numero de usuarios.
Z: Disciplina de la cola.

En este trabajo utilizaremos una notación simplificada de la mencionada anteriormente, para esto tendremos en cuenta:

A/B/c = Distribución de Llegadas / Distribución de tiempos de Servicio / Número de servidores

donde c = 1 y las letras que se despreciaron se asumen que toman los siguientes valores:

K: Cola de espera infinita,
n: Fuente de usuarios infinita,
Z: Disciplina de la cola FIFO.

También utilizaremos un sistema **A/B/c/K** donde variaremos la capacidad del sistema (K) para diferentes valores y mantendremos contante c = 1. En ambos casos tomaremos como que A tiene una distribución markoviana de Poisson y como que B tiene una distribución markoviana exponencial.

Donde:

La Distribución de Llegadas puede ser:

- **M** = distribución de llegadas de tipo Poisson
- **D** = distribución de llegadas es constante
- **G** = distribución de llegadas general con varianza y media conocidas

La Distribución de tiempos de Servicio puede ser:

- M = distribución de tiempos de servicio de tipo exponencial
- D = distribución de tiempos de servicio es constante
- G = distribución de tiempos de servicio general con varianza y media conocidas

3.5. Medidas de eficiencia de un sistema de colas M/M/1

A continuación se presenta el enfoque de análisis que se debe dar al sistema de línea de espera típico con llegadas de tipo Poisson, tiempos de servicio de tipo Exponencial con un sólo servidor. Se supone que en este sistema, la entidad está dispuesta a esperar el tiempo que sea para ser atendido, es decir no hay rechazo. Donde:

λ = número promedio de llegadas al sistema/ unidad de tiempo (velocidad de llegadas)

μ = número promedio de entidades que se atienden en el sistema / unidad de tiempo (velocidad de atención del servidor).

Las medidas de eficiencia que utilizaremos a lo largo del trabajo para comparar los diferentes resultados de nuestras dos fuentes de simulación serán las que se detallarán a continuación:

- N: Número real de clientes en el sistema.
- Nq: Número real de clientes en la cola.
- Ns: Numero de clientes que están recibiendo servicios.

- Ls: Número promedio de unidades en el sistema.

$$Ls = \frac{\lambda}{\mu - \lambda} \tag{4}$$

- Ws: Tiempo promedio en que una unidad esta dentro del sistema.

$$Ws = \frac{1}{\mu - \lambda} \tag{5}$$

- Lq: Número promedio de unidades en la fila de espera.

$$Lq = \frac{\lambda^2}{\mu(\mu - \lambda)} \tag{6}$$

- Wq: Tiempo promedio en que una unidad pasa por la fila de espera.

$$Wq = \frac{\lambda}{\mu(\mu - \lambda)} \tag{7}$$

- ρ: Factor de uso del sistema o del servidor.

$$\rho = \frac{\lambda}{\mu} \tag{8}$$

- P_0: Probabilidad de que ninguna unidad se encuentre en el sistema.

$$P_0 = 1 - \rho = 1 - \frac{\lambda}{\mu} \tag{9}$$

- P_n Probabilidad de que el sistema tenga exactamente "n" unidades.

$$P_n = (1 - \frac{\lambda}{\mu})(\frac{\lambda}{\mu})^n \tag{10}$$

3.6. Formulas de Little para relacionar las medidas de eficiencia.

Según las formulas de Little se puede expresar al numero medio de clientes en el sistema/en la cola como la tasa de llegadas por el tiempo medio de los clientes en el sistema/ en la cola. Esto queda expresado analíticamente como:

$$L = \lambda W \qquad L_q = \lambda W_q \tag{11}$$

Para entender la primer formula sopongase que $W = 2$ horas, $\lambda = 3$ clientes/hora, entonces el número medio de clientes en el sistema es $3 * 2 = 6$, tal y como se muestra en la siguiente figura (la hora 1 es despreciable dado que se supone un sistema estacionario y por lo tanto horas homogéneas):

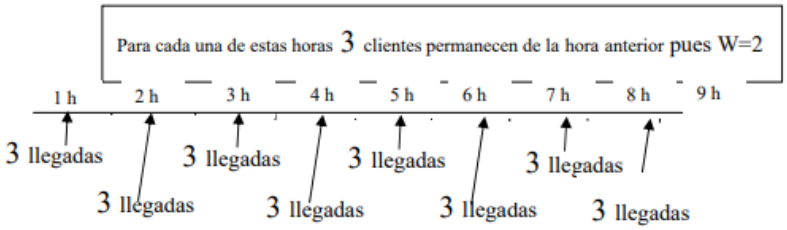

Figura 8: Comportamiento Little. Imagen tomada del sitio web: www.cartagena99.com/recursos/teroiacola.pdf

También podemos expresar al tiempo medio de los clientes en el sistema como el tiempo medio de los clientes en la cola más el tiempo medio de servicio de un servidor. Otra formula de Little es la que indica que el número medio de clientes en el sistema es igual al número medio de clientes en la cola mas el número medio de servidores ocupados. Ellas son, entonces:

$$W = W_q + E(s) \qquad L = L_q + \lambda/\mu \tag{12}$$

4. Previo al proceso de experimentación y análisis.

Antes de comenzar con la simulación del sistema de colas vamos a establecer algunos principios que utilizaremos a lo largo de lo que resta del trabajo. Primeramente aclaramos que, como se dijo en la introducción, estaremos utilizando tres fuentes diferentes para simular el sistema de colas. El primero será analítico y los dos restantes serán a través de Python y Anylogic.

Para obtener soluciones analíticas en torno a un modelo de esta naturaleza; la teoría probabilística de colas, provee de conceptos y expresiones matemáticas que han sido deducidas y permiten calcular los parámetros relevantes y característicos del sistema, tales como: la longitud promedio de clientes en el sistema y en la cola, el tiempo esperado que pasa un cliente en la cola y en el sistema, las probabilidades de estado estable, los porcentajes de ocupación y de ocio, etc.

Planteada esta problemática, se puede acudir a la utilización de un procedimiento numérico y experimental, tal como la simulación Montecarlo. Esta técnica tiene la virtud de poder establecer generadores de procesos que siguen cualquier distribución de probabilidad conocida, e inclusive empírica.

- El **método analítico** contendrá todas las formulas mencionadas en el marco teórico y procederemos a calcular, cada vez que variemos un parámetro del sistema, nuevamente dichas formulas con los valores correspondientes.

  Estos cálculos, a modo de simplificar la tarea repetitiva de los mismos, serán obtenidos a través de una calculadora online proveniente de la página web: https://www.supositorio.com/rcalc/rcalclite_esp.htm. La misma nos provee de las opciones necesarias para ingresar tanto la tasa de arribos como la de llegada deseadas y, mediante este ingreso, nos retorna los valores de todas las medidas de rendimiento descriptas en la sección anterior. La calculadora online utiliza las formulas previamente mencionadas.
- A través de **Python** simularemos un sistema de líneas de espera. Para esto utilizaremos el mecanismo de avance en el tiempo, donde haremos que el tiempo avance a tiempos discretos definidos por la ocurrencia de dos diferentes eventos (uno es el de arribo y el otro es el de partida).

  Cada vez que ocurra uno de los dos tipos de eventos avanzaremos en el tiempo y calcularemos las diferentes medidas de desempeño. De esta forma obtendremos para cualquier instante t el estado del sistema. Es decir, en un momento determinado podremos saber cuántos clientes se encuentran en la cola, cuantos en el sistema, la demora de cada uno, etc. Ejecutaremos muchas corridas (cada una contendrá un numero n de eventos), y graficaremos el comportamiento de cada una.

  La idea es aplicar el método de Montecarlo, el cual nos permitirá aproximar los valores obtenidos previamente con la calculadora online a través de calcular el promedio de las diferentes corridas calculadas. Mediante este mecanismo obtendremos los valores de las diferentes medidas de rendimiento y podremos extendernos a realizar otras simulaciones variando dichas medidas.
- En el caso de **Anylogic**, en comparación con otros tipos de modelos, los modelos de simulación son extremadamente versátiles.

  Podemos decir un modelo de simulación es un conjunto de reglas que define el funcionamiento del simulando. Un simulador como AnyLogic nos permite establecer el estado de inicio del modelo y usar las reglas que seleccionamos para simular la evolución del sistema a través del tiempo. Utilizaremos, entre los diferentes métodos que existen, el modelado de eventos discretos. El modelado centrado en procesos utiliza una abstracción de nivel inferior que se denomina modelado de eventos discretos para rastrear y cuantificar dinámicamente el estado del proceso.

  El modelado de eventos discretos es un paradigma integral poderoso que es capaz de modelar casi cualquier sistema que cambie con el tiempo a través de eventos. Es por esta razón que consideramos que es el que mejor se adapta a nuestras necesidades. Iremos indicando las salidas que obtuvimos del modelo y realizaremos la simulación a través de la parametrización de los objetos que representarán nuestro servidor y nuestros clientes. Además, variaremos nuevamente los parámetros para obtener las diferentes medidas de rendimiento.

5. Experimentación y análisis.

5.1. Simulación de un sistema de colas M/M/1.

Previo a comenzar cada simulación calcularemos de manera analítica (utilizando las formulas descriptas en el marco teórico y a través de la pagina web mencionada) los valores de las medidas de rendimiento que deseamos estimar. Utilizaremos estos resultados para comprobar si se ajustan a los valores esperados, para esto comenzaremos calculando las diferentes medidas de rendimiento para un sistema que posee una tasa de arribos y una tasa de servicio con distribuciones de Poisson, o lo que es lo mismo, que sus tiempos de arribo y de servicio poseen una distribución exponencial.

5.1.1. Promedio de clientes en la cola.

Para obtener el valor adecuado de esta medida de rendimiento efectuaremos una simulación que constará de diez corridas, quienes contendrán individualmente un número muy grande de iteraciones. Para resolverlo utilizaremos el método de Montecarlo. Así pues, el objetivo principal de la simulación de Montecarlo es intentar imitar el comportamiento de variables reales para, en la medida de lo posible, analizar o predecir cómo van a evolucionar. Luego de ejecutadas las diez corridas obtendremos el promedio de todas para poder estimar el estadístico deseado. Gracias a las formulas de Little mencionadas con anterioridad obtendremos el valor esperado de la variable aleatoria:

L_q: Longitud de la cola de un sistema de espera M/M/1.

Entonces:

$$L_q = \sum_{n=1}^{\infty}(n-1) * P_n = \sum_{n=1}^{\infty} P_n * n - \sum_{n=1}^{\infty} P_n = L - (1 - P_0) = L - \rho = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{13}$$

Diremos entonces, que dicho valor estimado será el que tendrá que asemejarse al valor esperado y calculado por la página web. Previamente calculamos el valor con dicha página para tres escenarios posibles:

Primer caso: $\mu > \lambda$**.** Tomaremos que μ representa la tasa de servicio de un cliente por unidad de tiempo y que λ tomará un valor igual a la mitad de la tasa de servicio. Es decir:

$$\mu = 1\frac{clientes}{unid.tiempo} \qquad \lambda = 0{,}5\frac{cliente}{unid.tiempo} \tag{14}$$

Analíticamente podemos calcular que el promedio de clientes en cola Lq será igual a:

$$Lq = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{0{,}5^2}{1(1-0{,}5)} = 0{,}5 \tag{15}$$

Segundo caso: $\mu = \lambda$**.** Tomaremos que μ representa la tasa de servicio de un cliente por unidad de tiempo y que λ tomará un valor igual esta última. Es decir que vamos a calcular que el promedio de clientes en cola L_q será igual a:

$$Lq = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{1^2}{1(1-1)} = \frac{1}{0} = \infty \tag{16}$$

Tercer caso: $\mu < \lambda$**.** Tomaremos que μ representa la tasa de servicio de un cliente por unidad de tiempo y que λ tomará un valor igual al doble de esta última. Es decir que podemos calcular el valor de L_q y obtendremos como resultado que el valor de la cola es negativo. Esto claramente no tiene sentido ni lógico ni físico. Es debido a que lógicamente si la tasa de arribos es superior a la de servicios y se cuenta con un solo servidor, entonces este estará imposibilitado de asistir a todas las peticiones de los clientes y por lo tanto ocurre un absurdo.

$$Lq = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{2^2}{1(1-2)} = -4 \tag{17}$$

Una vez calculados los tres valores para la misma medida de rendimiento podemos notar como obviamente las ultimas dos, al sobrepasar o ser igual a la capacidad de trabajo del sistema de servicio, ocurrirá que el promedio de afluencia de clientes al sistema no podrá ser cubierto por el mismo. Sin embargo, cuando el sistema tiene suficiente capacidad de

trabajo frente a la afluencia de clientes se puede ver como el número de clientes promedio crecerá en ocasiones pero siempre retornará a 0 y es este el valor que provoca que el promedio ronde en torno al 0.5.

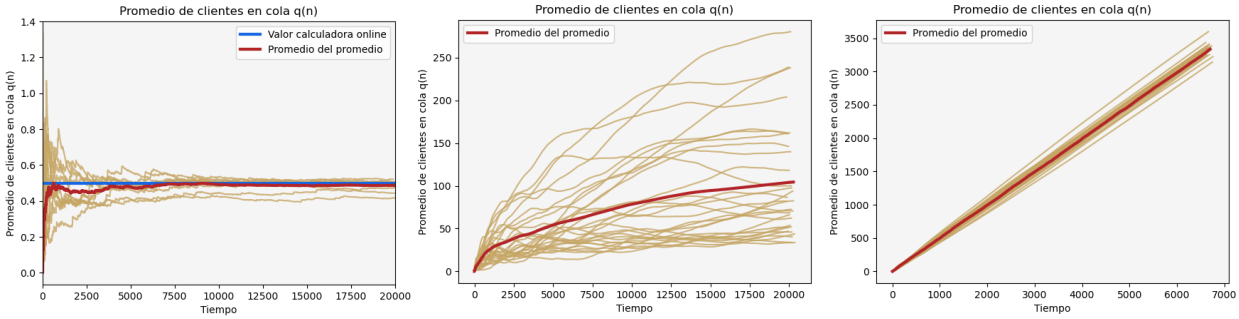

Figura 9: 10 corridas para diferentes parametros.

En las figuras puede verse que conforme la tasa de arribo sobrepasa a la de servicio, este ultimo no da a basto con las peticiones y no logra atender a todos los clientes. Lograndosé así, en la tercer figura, alcanzar un valor que crece de manera lineal hacia el infinito, haciendo referencia a la imposibilidad de cumplir con los requerimientos. En las dos ultimas figuras no se graficó el valor de la calculadora porque como se indicó, el resultado de los mismos tendía al infinito.

Mediante la ejecución de la simulación en Python se obtuvo, para la primer figura, que $L_q = 0{,}49998$. Por lo tanto se puede decir que la simulación esta funcionando correctamente ya que se adecúa al valor esperado y previamente calculado. Procedimos a realizar el mismo procedimiento de variar los parámetros, pero esta vez quisimos ver, ya que contamos con la posibilidad, como variaba el flujo de clientes en la cola de manera real"por cada corrida. El resultado fue el siguiente:

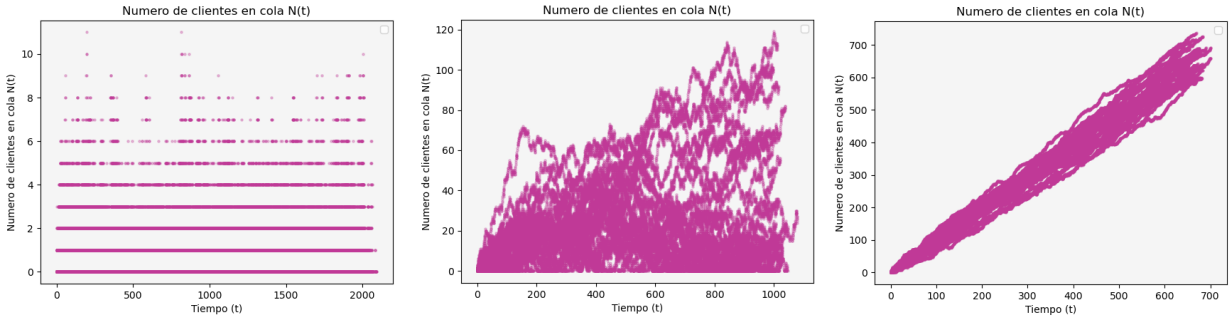

Figura 10: Corridas para diferentes parametros.

Puede verse como la mayor densidad de clientes se encuentra entre cero y tres, y a medida que la tasa de servicio es superada por la de arribo se ve la tendencia descripta anteriormente a irse, la cantidad de clientes que llegan a la cola, al infinito. Puede parecer engañoso que nos de un resultado de 0.5 cuando el numero de clientes oscila entre cero y tres, pero lo cierto es que en realidad estamos simulando un numero muy grande de iteraciones y, a la hora de contabilizar, es claro que la mayor densidad se encuentra en torno al cero en la primer figura.

Concretamente se puede entender a continuación como el sistema de servicio tiene suficiente capacidad de trabajo para afrontar la afluencia de clientes. Entonces N(t) puede crecer en ocasiones pero el sistema siempre retornará al valor 0, es decir al estado vacío.

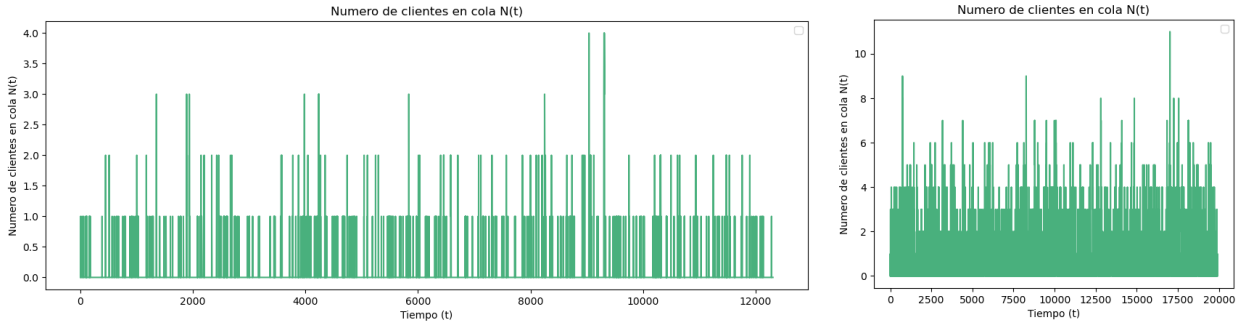

Figura 11: Corridas para diferentes parametros.

Con correspondencia a las gráficas anteriores podemos afirmar que los valores analíticos coinciden con los obtenidos mediante nuestro programa en Python. Ahora evaluaremos si los mismos se corresponden utilizando la simulación en Anylogic. El resultado que obtuvimos fue utilizando el modelo de eventos discretos es:

Dichos estadisticos y los posteriores resultados que obtengamos con Anylogic serán volvados en una tabla con el fin de comparar entre los diferentes métodos. Entonces:

μ	λ	L_q Calculadora	L_q Anylogic	L_q Python
1	0.5	0.5	0.508	0,49998
1	1	∞	86.51	88.991
1	2	∞	12.570	13.560

Finalmente podemos notar como se asemejan correctamente todos los metodos con las medidas esperadas para cada valor de los parametros μ y λ.

5.1.2. Promedio de clientes en el sistema.

La medida de eficiencia de promedio de clientes en el sistema, para nuestro sistema de estudio de una cola M/M/1 será practicamente similar a la medida de desempeño descripta anteriormente debido a que no contamos con multiples servidores. Por lo tanto la variable aleatoria:

L_s: Número promedio de unidades en el sistema.

Analíticamente la formula proviene de haber expresado previamente esta variable en funcion del factor de utilización del servidor. Por esto:

$$L_s = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = \rho*(1-\rho)\sum_{n=0}^{\infty} n*\rho^{n-1} = \rho(1-\rho)\frac{1}{(1-\rho)^2} = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda} \tag{18}$$

En las secciones posteriores explicaremos como expresar analiticamente la probabilidad de tener n clientes en el sistema/cola y de esta forma puede obtenerse el valor de P_n. Generamos el valor de esta variable a través de la calculadora online y el resultado fue de:

$$L_s = \frac{\lambda}{\mu-\lambda} = \frac{0{,}5}{1-0{,}5} = 1 \tag{19}$$

Este resultado se obtuvo para $\lambda < \mu$. Si se expresa en función de $\lambda \geq \mu$ sabemos por lo explicado en el apartado anterior que nuestros valores tenderán al infinito, por esto no haremos dicho cálculo. Sin embargo, estos valores cuando se realizan de manera simulada si pueden obtenerse para cada instante t. Por lo tanto:

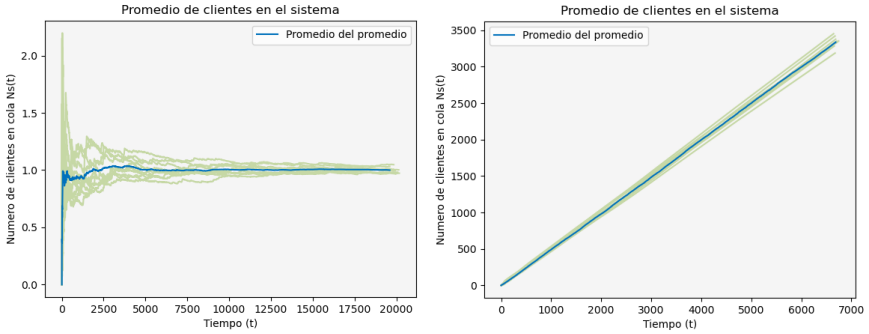

Figura 12: Clientes en el sistema

Obtuvimos que el valor simulado se asemeja exactamente al valor esperado. Resultando un total de $L_s = 1,0002$. Nuevamente nuestro método simulado, además de ser mucho más sencillo, nos proporciona una excelente precisión.

Nos pareció interesante expresar los datos generados en forma de diagramas de barras donde el eje x representara la cantidad de clientes que hubo en total en el sistema (es decir, su frecuencia absoluta). La primer figura hace referencia a cuando la tasa de arribo es menor a la de servicio y el caso contrario esta representado por la figura de la derecha.

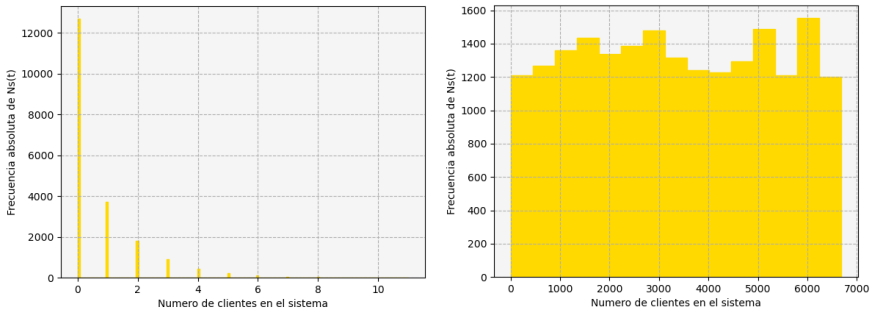

Figura 13: Clientes en el sistema

A continuación simularemos para los mismos parámetros el sistema de colas en Anylogic. El resultado que obtuvimos fue el siguiente:

Estos estadísticos se asemejan con los obtenidos a través del método de simulación con Python y con el método analítico. Ahora volcaremos los resultados obtenidos en una tabla:

μ	λ	L_q Calculadora	L_q Anylogic	L_q Python
1	0.5	1	1.012	0,49998
1	1	∞	87.501	88.9981
1	2	∞	12.571	-

5.1.3. Tiempo promedio en el sistema y en la cola.

Para encontrar los valores del tiempo de espera en la cola y en el sistema, W_q y W, respectivamente, hacemos uso de las Formulas de Little para obtener:

$$Ws = \frac{1}{\mu - \lambda} \qquad Wq = \frac{\lambda}{\mu(\mu - \lambda)} \tag{20}$$

Con todo esto, tenemos bien caracterizado el comportamiento estacionario de la cola. Una caracterıstica interesante del sistema M/M/1 es que, además, es relativamente sencillo proporcionar la distribución del tiempo que un cliente pasa en la cola y en el sistema.

Sea T la variable que denota el tiempo que pasa un cliente en el sistema y W(t) su función de distribución. Para obtener W(t) condicionamos sobre el numero de clientes en el sistema a la llegada del cliente.

$$W(t) = P_r(T \leq t) = \sum_{n=0}^{\infty} P_r(T \leq t \mid N = n) * P_r(N = n) \tag{21}$$

Donde N es la variable que cuenta el numero de clientes en el sistema. Ahora bien,

- Si $n = 0$, el cliente que llega estará en el sistema su tiempo de servicio.
- Si $n \geq 1$, habrá un cliente en servicio y n 1 clientes esperando servicio después de él. Debido a la pérdida de memoria de la exponencial (para el tiempo de servicio del primer cliente), el cliente deberá esperar un tiempo que es la suma de n + 1 exponenciales independientes con parámetro , que se sabe sigue una distribución Gamma con parámetro $\alpha = n + 1$ y $\beta = \mu$ cuya función de densidad es:

  $$f(x) = \frac{\mu^{n+1} e^{-\mu x} x^n}{n!}, x \geq 0 \tag{22}$$

  Reemplazando la ecuación de la función de densidad en W(t) se tiene que:

  $$W(t) = 1 - e^{-(\mu - \lambda)t} \tag{23}$$

  Por lo tanto, el tiempo que un cliente pasa en el sistema es una variable aleatoria exponencial de parámetro $\mu - \lambda$, cuya media es:

  $$E(t) = \frac{1}{(\mu - \lambda)} \tag{24}$$

  Igual al valor que habíamos obtenido para W_s. De manera analoga se procede a obtener W_q.

Con el uso de dichas formulas y para los mismos valores de μ y λ obtuvimos los valores de $W_s = 2$ y $W_q = 1$ cuando la tasa de arribo es menor a la de servicio. Para el caso en el que las tasas son iguales o la de servicio es inferior a la de arribo, nuevamente se produce el efecto de que tenderán al infinito de manera analítica. Sin embargo, a través de la simulación con dichos valores obtuvimos que:

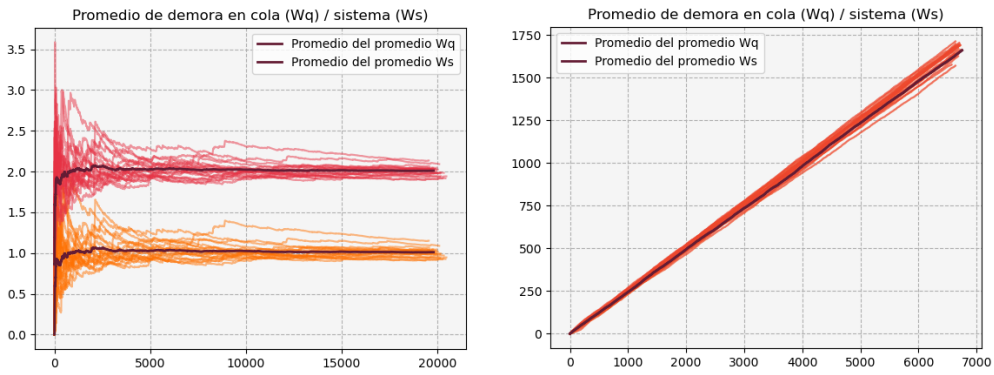

Figura 14: Espera promedio en el sistema y en la cola .

El valor calculado y el valor simulado se ajustan de manera casi exacta. Siendo $W_s = 2,0010$ y $W_q = 1,0181$. Podemos nuevamente asegurar que para nuestros usos prácticos, el método de simulación se ajusta a la perfección.

Continuando con el análisis de los resultados, analizaremos si coinciden los tiempos promedios de espera en la cola y en el sistema utilizando la simulación de Anylogic.

Tiempos en la cola con Anylogic.

Empezando el análisis con los parámetros λ=0.5 y μ=1 , obtendremos una gráfica de histograma, en la cual podremos ver en negro el valor modal de los tiempos de espera.

El promedio conseguido con esta simulación es de 1.01 y el esperado según la calculadora, es de 1, por lo cual podemos decir que los resultados conseguidos en Anylogic son correctos. De la imagen obtenida podemos analizar que el 77,2 % de los clientes que ingresan en una cola de espera, esperan hasta 1.6 unidades de tiempo en dicha cola, y el 12,7 % espera entre 1,6 y 3,2. A medida que se avanza hacia la izquierda del gráfico, podemos ver que disminuye la cantidad de gente en la cola cuanto más grande es el tiempo de espera.

Posteriormente realizaremos la misma simulación, pero dándole a lambda un valor igual o mayor a mu, en este caso los igualaremos a 1. Uniendo los tres gráficos (tasa de arribo igual a las de servicio, tasa de arribo inferior a la de servicio y tasa de arribo superior a la de servicio) obtuvimos el siguiente resultado:

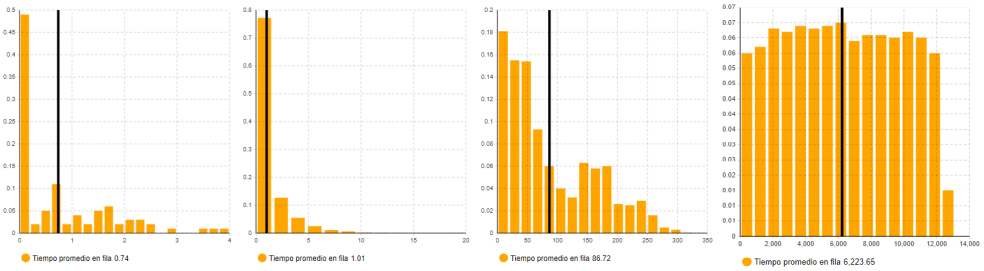

Figura 15: Espera promedio en el la cola .

Como podemos observar, el valor promedio es muy alto, y se puede decir que tiende a infinito, coincidiendo con lo esperado, ya que para cualquier valor de lambda mayor o igual que mu, el tiempo medio tenderá a infinito.

Tiempos en el sistema Ws con Anylogic.

A continuación simularemos para los mismos parámetros el sistema de colas. Realizando la simulación, siempre comenzando con los parámetros λ=0.5 y μ=1, podemos ver que se realizó en 85 segundos y nos muestra que el tiempo promedio de los clientes en el sistema (Ws) es 2.01.

Generamos un histograma para mostrar los datos ordenadamente, y podemos ver la media del tiempo que tarda un cliente en el sistema, marcada en color negro, que es la que mencionamos previamente. El promedio esperado del tiempo de clientes en el sistema según nuestra calculadora es 2, con lo cual la simulación muestra que coincide con el valor esperado. También podemos observar que el tiempo promedio del sistema se concentra en el segmento de tiempo hasta las 3 unidades de tiempo, donde se concentra el 77,4 % de los tiempos medidos.

En la figura de la derecha procedimos a realizar la simulación con valores de lambda mayor e igual a mu. Primero los valores serán λ=1 y μ=1. Nos informa que el tiempo promedio en el sistema es de 87.71. Lo que sucede es que cuando lambda es igual o mayor a mu, las colas tienden a infinito ya que siempre estarán ocupadas.

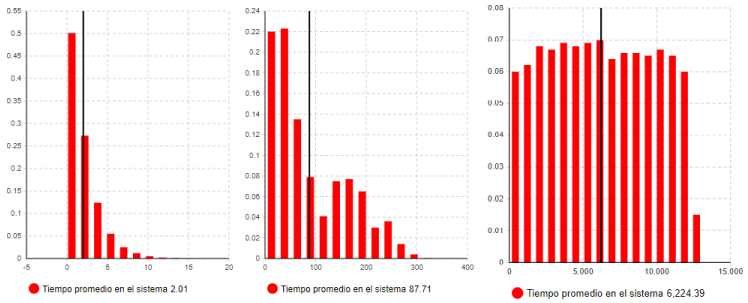

Figura 16: Espera promedio en el sistema.

Los resultados que obtuvimos para los tiempos promedio en el sistema y en la cola fueron los que se encuentran la siguiente tabla. Estos estadísticos se asemejan con los obtenidos a través del método de simulación con Python y con el método analítico.

μ	λ	Calc W_s	Calc W_q	W_s Anylogic	W_q Anylogic	W_s Python	W_q Python
1	0.5	2	1	2,01	1,001	2,0010	1,0181
1	1	∞	∞	87.71	85.72	90.324	89.540

5.1.4. Utilización del servidor.

El factor de utilización ρ es la probabilidad de que el servidor esté ocupado. Dado que:

$$\rho = \frac{\lambda}{\mu} \tag{25}$$

Sabemos que este factor representa un porcentaje y por lo tanto $0 \leq \rho \geq 1$. Cuanto el servidor esté siendo utilizado un factor de 1, o, lo que es lo mismo, un factor del $100\,\%$, diremos que el sistema está saturado. Nuestro propósito siempre será encontrar un equilibro entre estos valores, no queremos que nuestro servidor esté ocioso ni tampoco que no de a basto con la cantidad de requerimientos que recibe.

Basándonos en esta medida de rendimiento podríamos determinar precisamente si nuestro sistema necesita de un servidor más, para poder apaciguar la demanda, o en caso de que tengamos muchos, si es necesario eliminar uno. Aunque no siempre puede resultar bueno agregar un nuevo servidor, esto se conoce como la paradoja de Braess.

La paradoja de Braess es una explicación propuesta para la situación en la que una alteración de una red de carreteras para mejorar el flujo de tráfico en realidad tiene el efecto contrario e impide el tráfico a través de él. La paradoja se postuló en 1968 por el matemático alemán Dietrich Braess , que se percató de que la adición de un camino a una congestionada red de tráfico de carreteras podría aumentar el tiempo total de viaje, y se ha utilizado para explicar los casos de mejora de flujo de tráfico cuando las carreteras principales existentes están cerrados.

La paradoja puede tener analogías en las redes eléctricas y los sistemas biológicos. Se ha sugerido que, en teoría, la mejora de una red mal funcionamiento podría lograrse mediante la eliminación de ciertas partes del mismo.

Por lo tanto, podemos apreciar la importancia de conocer cuanto se está utilizando nuestro servidor. La paradoja mencionada es útil también para cuando se estiman los tiempos de espera en el sistema. Dicho esto, pasemos a calcular de manera análitica los valores de esta medida para el caso en el que la tasa de arribos es menos a la de servicio. Entonces:

$$\rho = \frac{\lambda}{\mu} = \frac{0{,}5}{1} = 0{,}5 \tag{26}$$

En este caso diremos que nuestro sistema está siendo utilizado un $50\,\%$ de su totalidad. Esto es lógico puesto que la tasa de servicio es el doble que la de arribo. Cuando tenemos una tasa de arribos mayor o igual a la de servicios se tiene:

$$\rho = \frac{\lambda}{\mu} = \frac{1}{1} = 1 \tag{27}$$

Comenzamos por nuestra simulación en Python y el resultado que obtuvimos fue el siguiente:

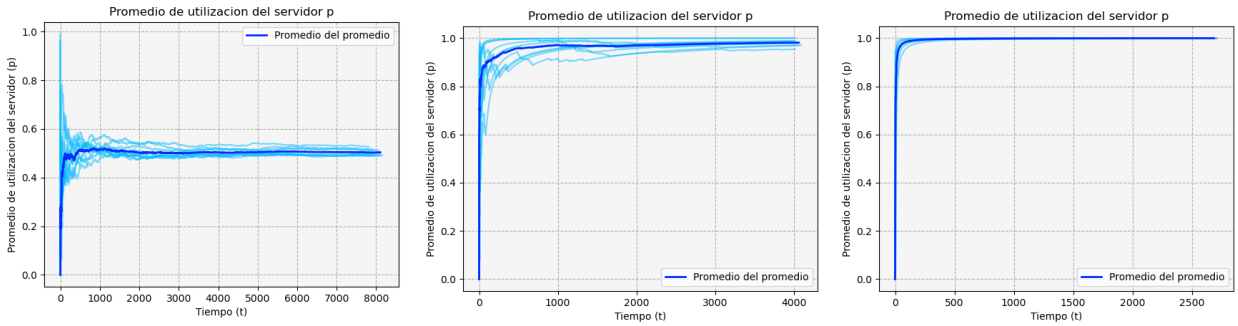

Figura 17: Utilizacion del servidor .

Se puede ver como el conjunto de corridas se estabiliza en el valor de $\rho = 0{,}5$. Estos resultados son similares a los esperados. También graficamos los casos en los que las tasas de servicio son iguales (figura del medio) y cuando la tasa de arribos sobrepasa a la de servicio (última figura). En los últimos dos casos, nuestro sistema se está utilizando al $100\,\%$ de su totalidad, y por esto, cuanto mayor sea la diferencia entre ambos, mayor tenderá a 1 el factor de utilización del servidor.

Decidimos gráficar como influye la cantidad de clientes en el sistema con respecto al factor de utilización del servidor. Esto nos pareció interesante ya que, como vimos, ambas formulas se encuentran relacionadas. La primer figura representa cuando el $\lambda < \mu$ y hacia el lado derecho vamos haciendo tender esta desigualdad para el caso en el que $\lambda > \mu$. Puede verse como estos parametros estan directamente relacionados:

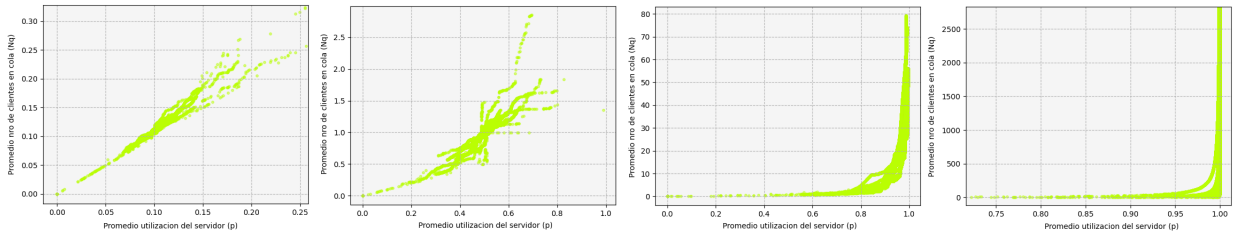

Figura 18: Utilizacion del servidor.

En la segunda figura puede verse como se produce una especie de ”moño” en el cual su centro representa el momento de mayor congestión. Ese será por lo tanto la media de esta medida de rendimiento de todas las corridas.

En la primer figura la densidad de puntos se encuentra distribuida casi de manera similar al rededor de los diferentes factores que puede presentar el sistema en un rango desde $[0; 0{,}25]$. Esto se debe a que al ser baja la tasa de arribos el sistema se encuentra muy lejos de tomar un valor cercano al 1. Caso contrario, en las ultimas figuras cuando el número de clientes por unidad de tiempo es muy alta y sobrepasa al servidor, el factor de utilización tiende a 1.

A la vez nos pareció que podíamos representar esto a través de un diagrama de tortas en el cual se represente cuanto esta siendo utilizado el servidor. Entonces podemos notar todo lo explicado anteriormente en la siguiente imagen:

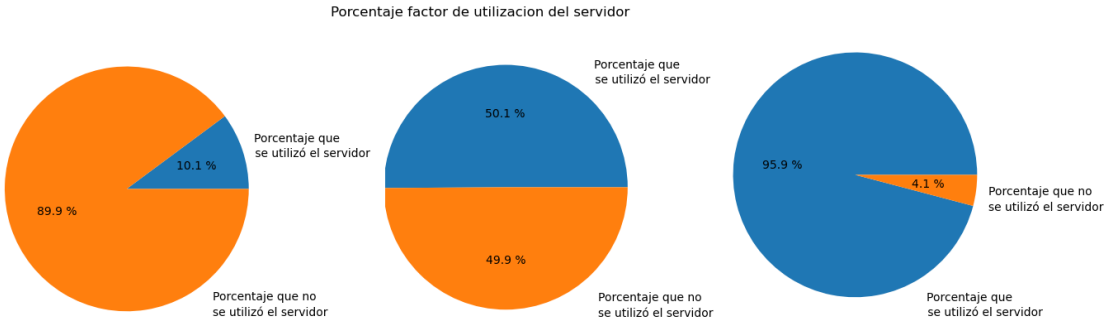

Figura 19: Utilización del servidor.

A continuación simularemos para los mismos parámetros el sistema de colas en Anylogic. El resultado que obtuvimos para las utilizaciones del servidor fueron los siguientes:

Estos estadísticos se asemejan con los obtenidos a través del método de simulación con Python y con el método analítico. Ahora volcaremos los resultados obtenidos en una tabla para comparar los tres metodos entre sí:

μ	λ	ρ Calculadora	ρ Anylogic	ρ Python
1	0.5	0.5	0.50	0,5007
1	1	1	0.99	0.9981
1	2	1	0.99	0.9999

5.2. Probabilidad de n clientes en cola.

Para comprender mas en detalle cuan importante es esta medida debemos introducir un tema muy importante en la teoría de colas, que es el método de los nacimientos y muertes.

Método de nacimientos y muertes.

Dado que sabemos como calcular L y L_q y si le sumamos él tener una noción básica sobre lo que son las cadenas de Markov, estaremos en condiciones de comprender los siguientes enunciados.

Sabemos que existen relaciones básicas entre las formulas de Little. Este nos dice una relación entre el numero de elementos que hay en la cola y el tiempo de estancia en el sistema. Esta relación se da a través de λ. Pero entonces si tenemos que:

$$L = \lambda W \tag{28}$$

$$L_q = \lambda W_q \tag{29}$$

$$W = W_q + 1/\mu \tag{30}$$

Tenemos 3 ecuaciones con 4 incógnitas. Necesitamos una ecuación mas para resolver el sistema. Para esto existe la medida que se llama el número de clientes promedio.

$$L = E(n) = \sum_{0}^{\infty} nP(n) \tag{31}$$

Si fuésemos capaces de calcular esta sumatoria, para los clientes 0, 1, 2 en el sistema hasta el infinito, entonces podremos calcular cuantos hay en promedio. Es decir, si es un $50\,\%$ que no hay nadie, entonces sabremos que la cola no será muy larga. Habiendo entendido esto, procederemos a calcular la probabilidad de n. A continuación podemos entender la siguiente imagen como que muestra los distintos estados de nuestro sistema:

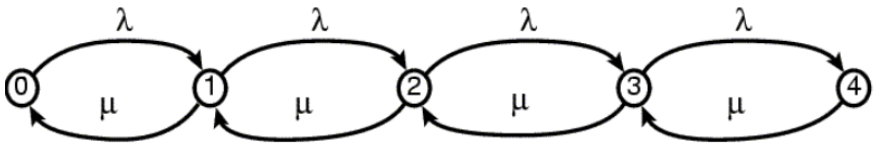

Figura 20: Nacimientos y muertes de un sistema M/M/1 .

Cada nodo representa la cantidad de clientes que se encuentran en el sistema. Si tuviésemos estos estados podemos conseguir la probabilidad de estar en dicho estado. Para esto calculamos las transiciones entre estados.

Para esto, la probabilidad de que un sistema que esta en 0 pase a 1 es λ. Es decir, mientras mayo sea esta, mayor sera la probabilidad de cambiar del estado vacío al 1. La probabilidad de regresar al estado 0 es μ, es decir la probabilidad de que el cliente que esta usando el servidor termine.

Ahora, en el transito desde 1 hacia 2 ocurre exactamente lo mismo, porque solo entra un cliente, en este caso se es independiente del número de servidores. Lo mismo ocurrirá con los demás estados, puesto que contenemos solamente un servidor.

Se dice que este sistema esta en equilibro cuando la probabilidad de entrar en un estado es la misma que salir de dicho estado. Es decir en el primer caso será $\lambda P(0) = \mu P(1)$, luego también $\lambda P(1)\mu P(1) = \lambda P(0)\mu P(2)$. Lo extenderemos hasta n al infinito y nos faltaría una ecuación para completar el sistema:

$$\sum_{n}^{\infty} P(n) = 1 \tag{32}$$

Mediante estas referencias podremos calcular la probabilidad de que haya n clientes en cola. Es decir, despejaremos y resolveremos el sistema hasta obtener la probabilidad deseada.

Procedimos a graficar, para cada número de cliente, la probabilidad de que haya exactamente esa cantidad de unidades en la cola. El resultado que obtuvimos fue el siguiente:

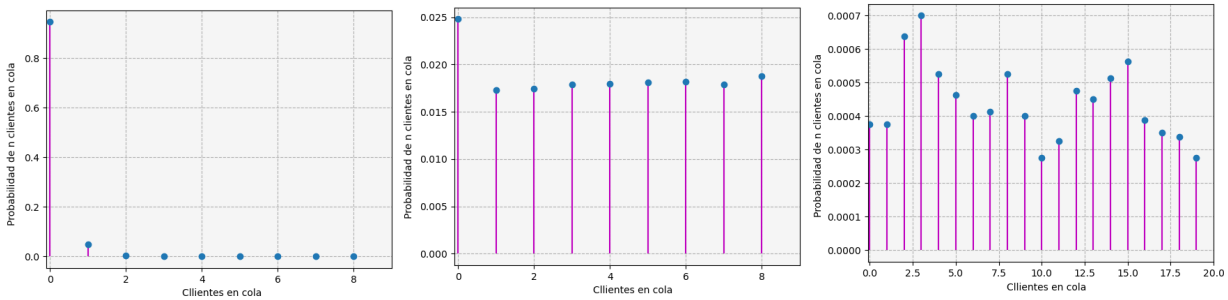

Figura 21: Probabilidad n clientes en cola.

Puede verse que en la primer figura el factor de utilización del servidor es del menos del $10\,\%$. Esto se puede calcular facilmente ya que cuando hay 0 clientes en el sistema simboliza que el servidor no está siendo utilizado. Por lo tanto se puede hacer:

$$P(0) = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho \qquad \rho = 1 - P(0) \tag{33}$$

Obteniendo, de esta forma, cuanto se está utilizando el sistema en cada caso. Los valores obtenidos de manera simulada se ajustan con los obtenidos a través de la calculadora, quien nos indicaba que cuando los valores de las tasas son iguales, las probabilidades de obtener cualquier cantidad de clientes en el sistema tienden a ser iguales (imagen 2 de la figura).

5.3. Simulación de un sistema de colas M/M/1/k.

Representaremos la cadena de Markov para un sistema de colas con limite de capacidad y conoceremos la formula mediante la cual se calcula el tamaño medio de cola en el sistema citado. Se da el problema de colas con limite de capacidad en aquellos sistemas donde la capacidad de la cola esta limitada. En estos sistemas la tasa de llegada de elementos al sistema se hace nula en el momento que el sistema se llena de clientes. Lo explicaremos con un ejemplo. Si tuviésemos un sistema que puede contener solo 6 clientes en el sistema. Entonces diremos que la cantidad de estados del sistema es de 6. Podemos tener 0,1,2,3,4,5 clientes y cuando llegue el número 6 no podrá ingresar al mismo.

Representandolo mediante un proceso de nacimientos y muertes tendremos siempre el mismo valor de μ puesto que tenemos 1 solo servidor.

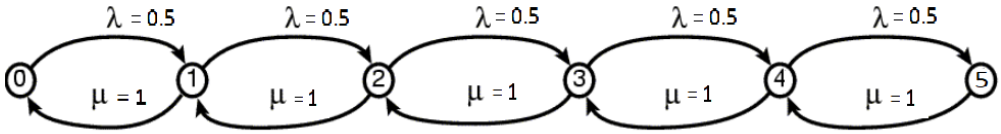

Figura 22: Ejemplo de cadena de Markov.

Gracias a las ecuaciones de equilibrio que nos indican que las entradas a un estado son iguales a las salidas del mismo. Resolviendo este sistema tendremos que:

$$\begin{aligned}
\lambda P_o &= \mu P_1 \\
\lambda P_1 + \mu P_1 &= \lambda P_o + \mu P_2 \\
\lambda P_2 + \mu P_2 &= \lambda P_1 + \mu P_3 \\
\lambda P_3 + \mu P_3 &= \lambda P_2 + \mu P_4 \\
\lambda P_4 + \mu P_4 &= \lambda P_3 + \mu P_5 \\
\lambda P_4 &= \mu P_5
\end{aligned}$$

$$P_0 + P_1 + P_2 + P_3 + P_4 + P_5 = 1$$

Dicho sistema de ecuaciones puede ser resuelto de manera sencilla debido a que tenemos un solo servidor. Dado que nuestro factor de utilización del servidor será:

$$\rho = \frac{\lambda}{\mu} = 0{,}5 \tag{34}$$

Podremos despejar y obtener que la probabilidad de que el sistema se sature debido a que llega a la capacidad máxima K, será de:

$$P_k = \frac{(1-\rho)\rho^k}{1-\rho^{K+1}} \qquad \rho \neq 1 \tag{35}$$

$$P_k = \frac{1}{K+1} \qquad \rho = 1 \tag{36}$$

Entonces en nuestro ejemplo, el la probabilidad de que el sistema se sature cuando la capacidad del mismo es de K = 2:

$$P_k = \frac{(1-\rho)\rho^k}{1-\rho^{K+1}} = \frac{(1-0{,}5)0{,}5^2}{1-\rho^{2+1}} = \mathbf{0.015625} \tag{37}$$

Esto significa que tenemos una probabilidad relativamente baja de que se denegen una gran cantidad de clientes (Solo el $0{,}1\,\%$ será denegado). Procedimos a gráficar, para los mismos valores mencionados en el ejemplo, el valor de la saturación del sistema con el K = 2. Se puede visualizar como la curva de la segunda imagen se ajusta de manera casi perfecta al valor calculado:

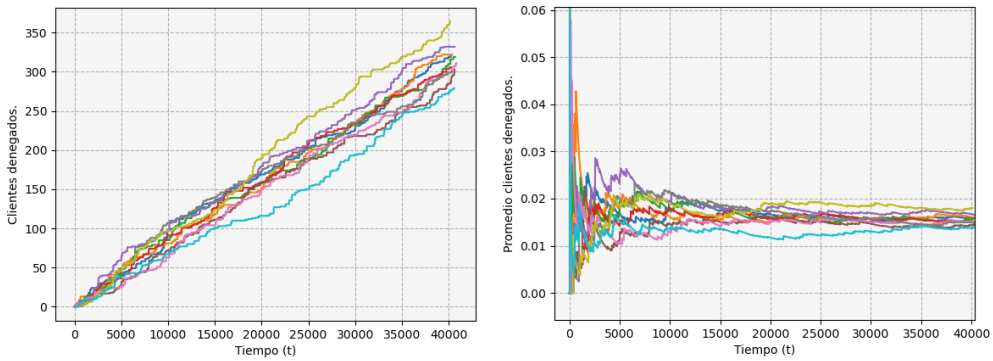

Figura 23: Denegación del servicio.

También graficamos como se incrementaba el numero de clientes por unidad de tiempo. Vemos que para una cola limitada, el número de clientes siempre será creciente, el objetivo cuando se está diseñando es intentar que la pendiente de la misma sea lo más leve posible.

Comprobaremos como se comporta el sistema con diferentes valores para K. Dichos valores serán: 0, 2, 5, 10, 50. Graficados en conjunto se obtuvo el siguiente resultado:

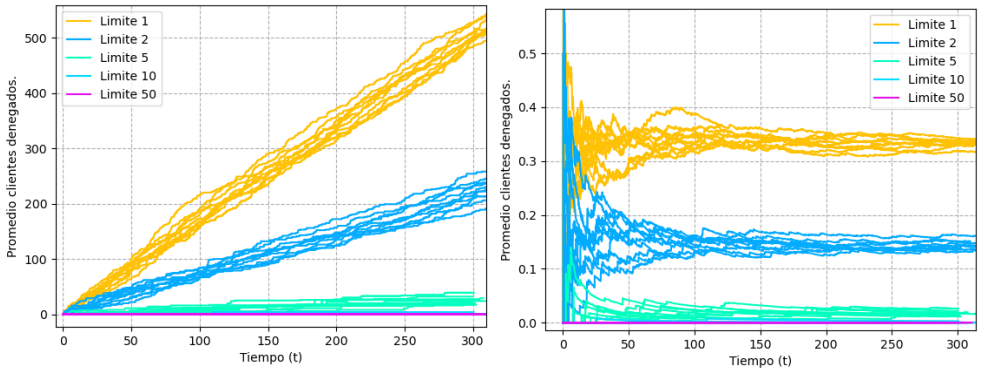

Figura 24: Denegación del servicio.

Puede verse que a medida que aumenta el limite admisible de clientes en el sistema, aumenta consigo la probabilidad de denegar mayor cantidad de clientes y por lo tanto el número de clientes denegados resulta mayor(Corridas en color amarillo, lado izquierdo). La primer figura nos esta representando el valor neto de clientes a quienes se les denegó el servicio, por esta razón siempre crecerán de manera indefinida. Pero como nos parece interesante notar la proporcionalidad que existe entre el limite superior admisible y la tasa de arribos al sistema, y por supuesto, la tasa de servicio.

Es por esta razón que graficamos para diferentes valores de dichas tasas lo que ocurría cambiando los limites del sistema. Puede entenderse de manera intuitiva que si variamos la tasa de arribos lamnbda estaremos variando de manera indirecta

el limite maximo de clientes que permite el sistema. Por esta razón la manera de identificar cada corrida es por su valor de lambda. El resultado fue el siguiente:

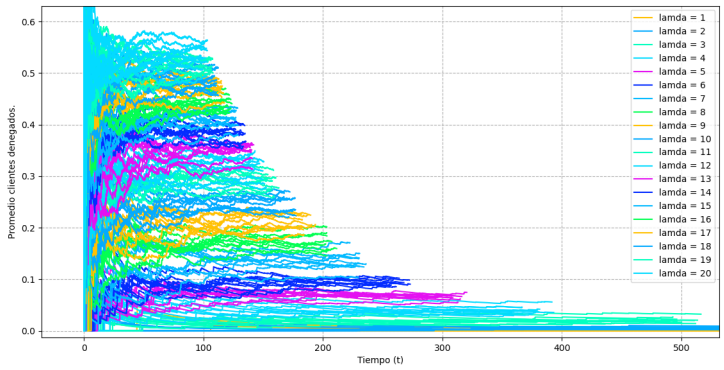

Figura 25: Variando parametros de denegacion de servicio.

Para hacer mas eficiente el análisis procederemos a llenar una tabla con los diferentes valores para el parámetro del limite del sistema. Así, podremos ver la coincidencia que existe entre lo simulado y lo calculado. Para el caso de la simulación, cada resultado corresponde al promedio de todas las corridas generadas.

LIMITE SISTEMA	PROMEDIO DE LA CALCULADORA	PROMEDIO DE LA SIMULACIÓN
0	0.333	0.3333
2	0.0667	0.06828
5	0.0079	0.006776
10	0.0002	0.0001999
50	0	0

El número real de clientes que fueron denegados en el sistema para diferentes valores se encuentra representado debajo, estos están en correspondencia con los valores representados en las tablas. Puede verse como cuando una corrida termina en un tiempo corto es debido a que posee un lambda menor y por lo tanto tardarán menos tiempo en arribar los clientes, comparado con cuando se tiene un lambda mucho mayor.

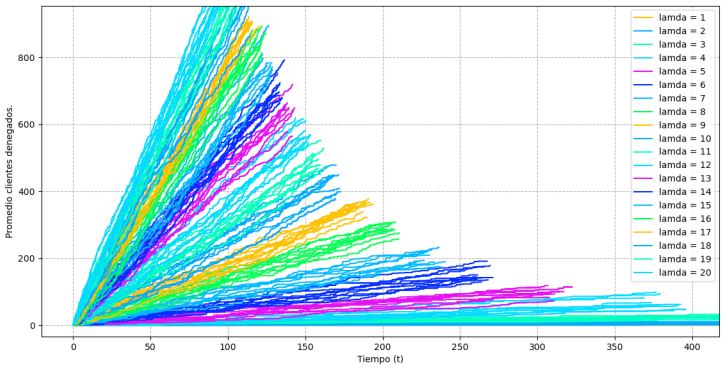

Figura 26: Denegación de servicio clientes reales.

5.4. Tasas de arribo y de llegadas.

Finalmente y en base a todas las variaciones hechas para cada tasa y para cada medida de desempeño, puede resultar útil graficar la variabilidad de todos en un mismo lugar. En esta ocasión preferimos correr el programa las correspondientes 10 veces, pero de cada una de ellas calcularemos su respectivo promedio y será este el que graficaremos a continuación. Dedicimos graficar como varía el número promedio de clientes en cola para tasas que van desde 1 a 20 de 1 en 1. El resultado que obtuvimos fue el siguiente:

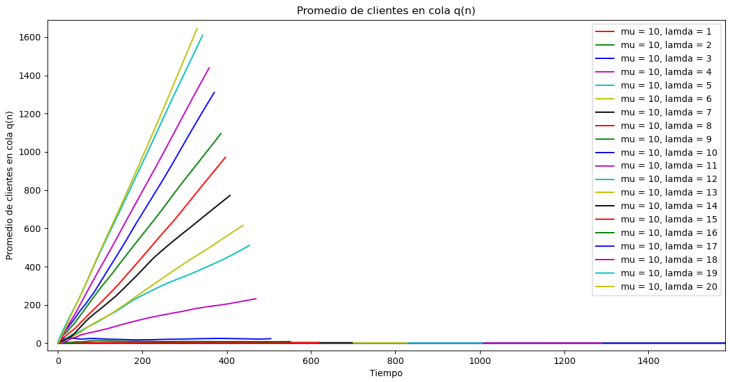

Figura 27: Denegación de servicio clientes reales.

Es interasante recalcar que cuanto mayor es la tasa de arribos menor será el tiempo transucurrido, puesto que todos los clientes llegan en un tiempo menor y la simulación finaliza antes.

Puede verse que conforme aunmenta esta tasa también sucede que el número de clientes que demandan ser atendidos tiende a infinito, y por lo tanto, la utilización del servidor tenderá a 1.

Podemos hacer una comparación entre los resultados obtenidos por ambos métodos en el formato de una tabla. Indicando, para cada medida de desempeño, por que método fue obtenida y su correspondiente valor. Tomaremos solamente los porcentajes del 25 %, 50 %, 75 %, 100 %, 125 % con respecto a la tasa de servicio. Entonces:

Método	**%**	L_s	L_q	W_s	W_q	ρ
Calculadora	25 %	0.3333	0.0833	0.1333	0.0333	0.25
Python	25 %	0.331	0.0867	0.1354	0.0355	0.2443
Anylogic	25 %	0.3333	0.08453	0.1332	0.0333	0.2501
Calculadora	50 %	1	0.5	0.2	0.1	0.5
Python	50 %	1.1101	0.5707	0.2153	0.1106	0.5394
Anylogic	50 %	1.010	0.5	0.2110	0.1091	0.5500
Calculadora	75 %	3	2.25	0.4	0.3	0.75
Python	75 %	3.2395	2.4682	0.4234	0.3225	0.7713
Anylogic	75 %	3.21	2.243	0.412	0.3	0.7610
Calculadora	100 %	∞	∞	∞	∞	1
Python	100 %	30.0589	29.0647	3.0499	2.9465	0.9942
Calculadora	125 %	∞	∞	∞	∞	1
Python	125 %	191.8365	190.8374	15.3967	15.2941	0.9991

La representación gráfica de los valores utilizados para llenar la tabla se muestra a continuación. La misma contiene en primer lugar la representación de cada una de las corridas para cada valor de lambda y luego, en la gráfica de la derecha se encuentra graficado el promedio de cada conjunto de corridas correspondiente a un parámetro lambda:

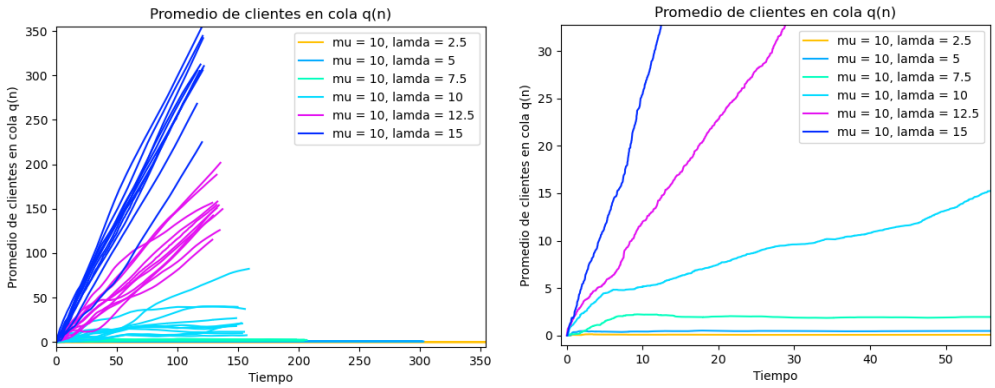

Figura 28: Correspondencia gráfica con las tablas.

Finalmente, con la cantidad de corridas analizadas se puede intuir fácilmente por qué resulta muy importante y necesario contar con un método que nos permita realizar este tipo de simulaciones. La facilidad y sencillez con que se pueden obtener estadísticos de manera simulada nos ahorra del muchisimo tiempo que necesitaríamos para elaborar el método analítico.

6. Conclusión.

Se mencionó a la teoría de colas como un conjunto de modelos matemáticos que describen sistemas de líneas de espera particulares. El objetivo principal fue encontrar el estado estable del sistema y determinar una capacidad de servicio apropiada que garantice un equilibrio entre el factor cuantitativo (referente a costos del sistema, relacionado con no tener un servidor ocioso) y el factor cualitativo (referente a la satisfacción del cliente por el servicio).

Una técnica para ejecutar estudios piloto, con resultados rápidos y a un costo relativamente bajo, está basado en la modelación de escenarios a través de la simulación. El proceso de elaboración del modelo involucra un grado de abstracción y no necesariamente es una réplica de la realidad. Esta realidad intentamos imitarla a través de dos métodos: Un programa en Python y otro en Anylogic.

En pro de presentar las utilidades de aplicar los conceptos de la Teoría de Colas a través de la Simulación, se tomó como caso de estudio la cola con un solo servidor con medidas de desempeño de dicho sistema por medio del modelo matemático respectivo, para luego ser comparado con el modelo de Simulación. Se evaluaron como influían las tasas de arribo y de servicio en el mismo y se comprobó su comportamiento en casos extremos. Se pudo definir los parámetros que estabilizan el sistema así como los que no.

Con lo anterior se puede corroborar una vez más la importancia de la implementación de modelos de simulación, ya que estos permiten profundizar mucho más en el comportamiento del sistema analizado. Igualmente se resalta la importancia de apoyar la simulación con modelos teóricos, esto debido a que es una excelente forma de validar la representación del modelo simulado con respecto al modelo real, tal como se evidencio a lo largo de las diferentes tablas. De esta manera, el investigador podrá realizar cambios y ajustes al modelo con la tranquilidad de que los resultados obtenidos serán muy acordes con la realidad.

Referencias

[1] Detalle de sistemas de cola https://idus.us.es/bitstream/handle/11441/77595/Esteban

[2] Modelos basicos de simulacion https://www.fiwiki.org/images/0/07/3.2Modelosdcolasbasicos.pdf

[3] Detalle de las funciones que tiene precargadas. Su funcionamiento y parámetros http://www.cartagena99.com/recursos/alumnos/apuntes/Modulo02Analisis

[4] Medidas de rendimiento http://www-eio.upc.es/teaching/TCiS/sioe3d.pdf

[5] https://en.wikipedia.org/wiki/M/M/1queue

[6] http://personales.upv.es/jpgarcia/linkeddocuments/teoriadecolasdoc.pdf

[7] https://www.fcfm.buap.mx/assets/docs/docencia/tesis/ma/CarlosCamiloGaray.pdf

[8] https://www.um.es/or/ampliacion/node8.html

[9] https://onedrive.live.com/?cid=3e1e09d7731a234fid=3E1E09D7731A234F

[10] https://www.fing.edu.uy/inco/cursos/mmc/unidad04/sesion10/transp.pdf

[11] http://catarina.udlap.mx/udla/tales/documentos/lem/gardunoaf/capitulo2.pdf

[12] https://www.um.es/or/ampliacion/node1.html