El presente material fue extraído de www.auladeeconomia.com, y persigue el objetivo de ser informativo y orientativo para los estudiantes de Investigación Operativa. Se recomienda ampliar los contenidos aquí detallados consultando: Taha (2001), "Investigación de Operaciones". Alfaomega; y Winston, W. (2005) "Investigación de operaciones: aplicaciones y algoritmos". Cengage Learning Editores.

Filas de Espera

Introducción

Las filas de espera o colas son frecuentes en nuestra vida cotidiana:

- En un banco
- En un restaurante de comidas rápidas
- Al matricular en la universidad

Es sabido que generalmente, a nadie le gusta esperar y cuando la paciencia llega a su límite, la gente se va a otro lugar. Sin embargo, un servicio muy rápido tendría un costo muy elevado por lo que es necesario encontrar un balance adecuado.
La teoría de colas es un conjunto de modelos matemáticos que describen sistemas de líneas de espera particulares siendo su objetivo encontrar el estado estable del sistema y determinar una capacidad apropiada.
Por supuesto, existen muchos sistemas de colas distintos, algunos son muy especiales y otros se ajustan a modelos más generales. Nosotros nos estaremos ocupando de estos últimos.

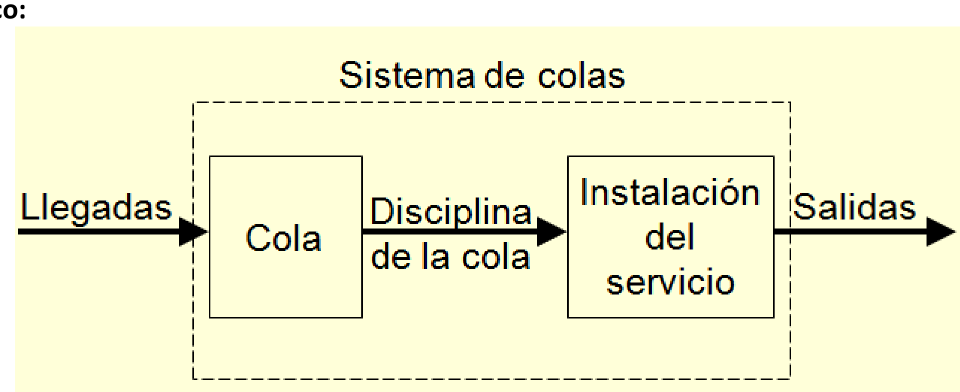
Sistemas de colas: Modelos básicos

En estos casos, cuando un cliente se puede encontrar con 2 situaciones:

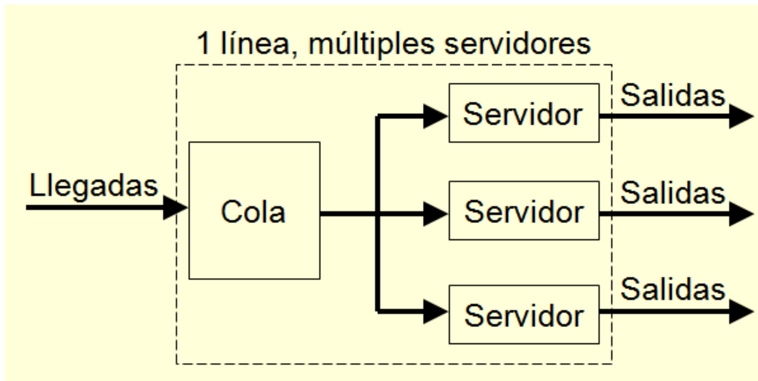
1. Llegar y que no haya nadie en la cola: pasa de una vez a recibir el servicio
2. Llegar y que sí haya gente en la cola: se une a la cola

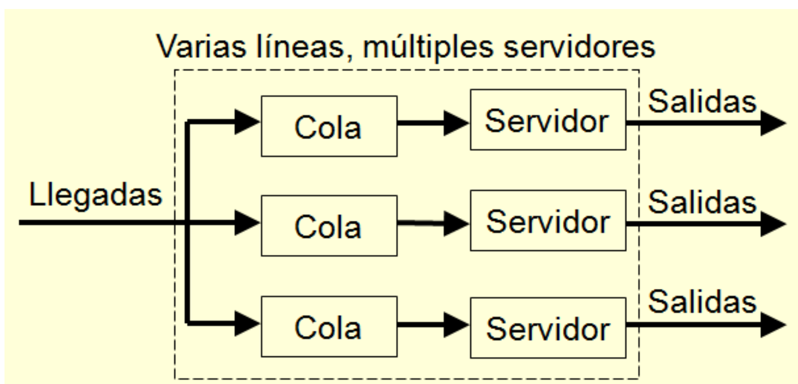
Es importante señalar que la cola no incluye a quien está recibiendo el servicio y que las llegadas van a la instalación del servicio de acuerdo con la disciplina de la cola. Generalmente ésta es *primero en llegar, primero en ser servido*. Pero puede haber otras reglas o colas con prioridades

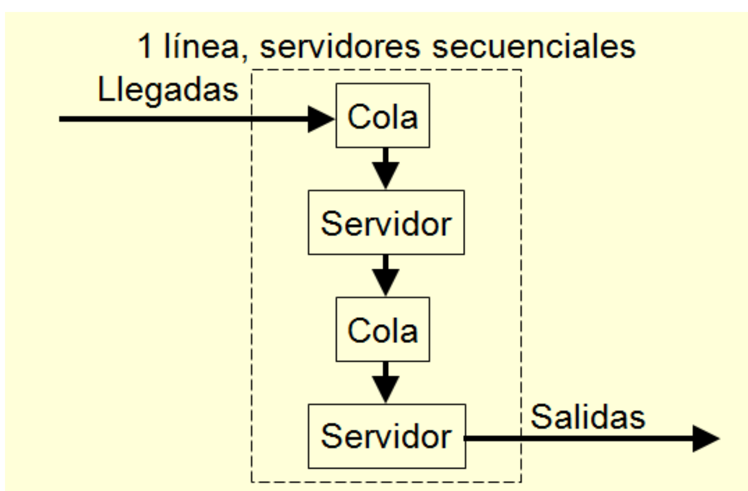
Esquema básico:



Como se puede observar en el diagrama anterior, este modelo en particular posee 1 linea, es decir una fila de espera y 1 servidor pero también se pueden dar las siguientes variantes:







Particularidades

Costos de un sistema de colas

Se pueden identificar 2 costos importantes:

1. Costo de espera: Es el costo que le representa al cliente esperar, es decir representa el costo de oportunidad del tiempo perdido. Un sistema con un bajo costo de espera es una fuente importante de competitividad.
2. Costo de servicio: Es el costo de operación del servicio brindado y es más fácil de estimar ya que el objetivo de un sistema de colas es encontrar el sistema del costo total mínimo.

Las llegadas

Se define al número esperado de llegadas por unidad de tiempo como **tasa media de llegadas (λ)** y el **tiempo esperado entre llegadas es 1/λ** (tiempo que transcurre entre dos llegadas sucesivas)

Estos tiempos pueden ser muy variables por lo que es necesario estimar la distribución de probabilidad de los tiempos entre llegadas. Generalmente se supone una distribución exponencial o de Poisson, pero siempre dependerá del comportamiento de las llegadas.

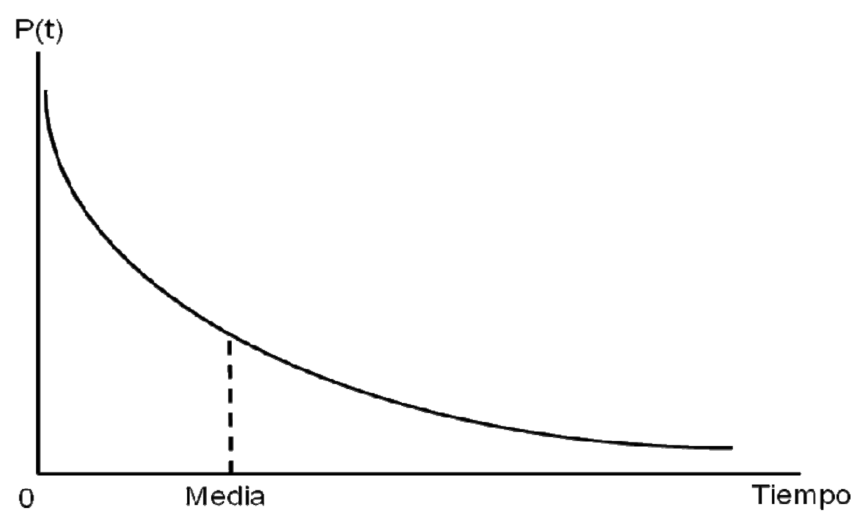
Distribución Exponencial

Este tipo de distribución supone una mayor probabilidad para tiempos entre llegadas pequeños. En general, se considera que las llegadas son aleatorias y que la última llegada no influye en la probabilidad de llegada de la siguiente.

Su forma algebraica es:

$$P(tiempo\,de\,servicio \leq t) = 1 - e^{-\mu t}$$

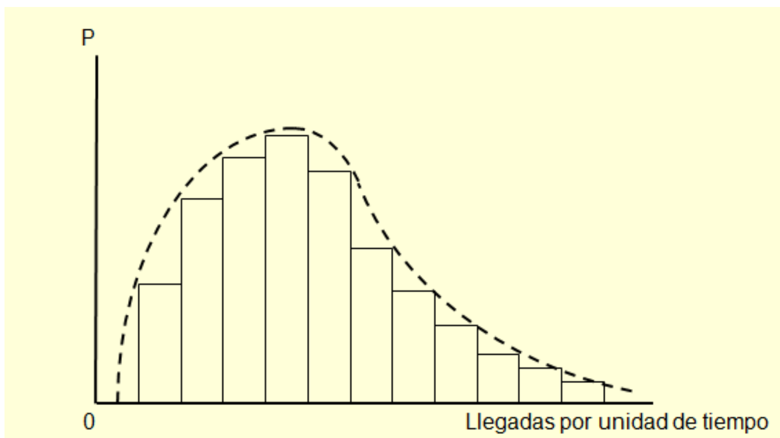
Donde ***t*** representa una cantidad expresada unidades de tiempo (horas, minutos, etc.)



Distribución Poisson
Este tipo de distribución es para tasas medias de llegadas pequeñas asimétrica y se hace más simétrica, aproximándose a la binomial, para tasas de llegadas altas
Su forma algebraica es:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Donde:
P(k) : probabilidad de k llegadas por unidad de tiempo
λ : tasa media de llegadas
e *= 2,7182818…*



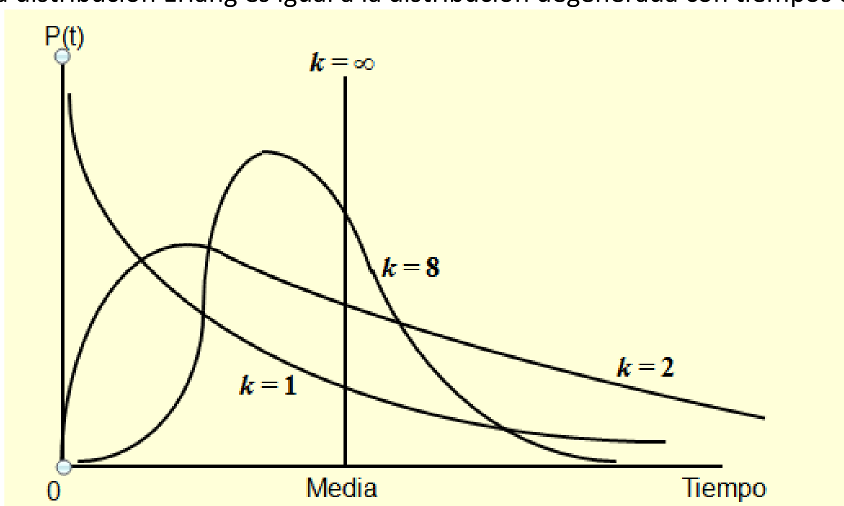
La cola
El número de clientes en la cola es el número de clientes que esperan el servicio, mientras que el número de clientes en el sistema es el número de clientes que esperan en la cola más el número de clientes que actualmente reciben el servicio.
Su capacidad es el número máximo de clientes que pueden estar en la cola, si bien generalmente se supone que la cola es infinita puede darse el caso que la misma sea finita.
Asimismo, su disciplina es el orden en que se seleccionan los miembros de la cola para comenzar el servicio, la más común es PEPS (primero en llegar, primero en servicio), pero puede darse que sea por selección aleatoria, prioridades, UEPS (último en llegar, primero en servicio), entre otras.

El servicio
El servicio puede ser brindado por un servidor o por servidores múltiples y si bien el tiempo de servicio varía de cliente a cliente este dependerá de la **tasa media de servicio (μ)**, por lo que el **tiempo esperado de servicio equivale a 1/μ**

Al igual que en el caso de las llegadas, estos tiempos pueden ser muy variables por lo que es necesario estimar la distribución de probabilidad de los tiempos de servicio. Hay dos distribuciones que representarían puntos extremos, la distribución exponencial (σ=media) y tiempos de servicio constantes (σ=0). Una distribución intermedia es la distribución Erlang, que posee un parámetro de forma ***k*** que determina su desviación estándar:

$$\sigma = \frac{1}{\sqrt{k}}\, media$$

Si k = 1, entonces la distribución Erlang es igual a la exponencial
Si k = ∞, entonces la distribución Erlang es igual a la distribución degenerada con tiempos constantes



Distribución	Desviación estándar
Constante	0
Erlang, k = 1	*media*
Erlang, k = 2	
Erlang, k = 4	1/2 *media*
Erlang, k = 8	
Erlang, k = 16	1/4 *media*
Erlang, cualquier k	

Identificación del modelo
Para identificar cada uno de los modelos, normalmente se utiliza la Notación de Kendall: ***A/B/c***
Siendo:

- ***A***: Distribución de tiempos entre llegadas
- ***B***: Distribución de tiempos de servicio
  - ***M***: distribución exponencial
  - ***D***: distribución degenerada
  - E_k: distribución Erlang
- ***c***: Número de servidores

Desempeño

Para evaluar el desempeño se busca conocer dos factores principales:

1. El número de clientes que esperan en la cola
2. El tiempo que los clientes esperan en la cola y en el sistema

Para poder medirlo se utilizará:

1. **Número esperado de clientes en la cola L_q**
   - $L_q = \lambda W_q$
2. **Número esperado de clientes en el sistema L_s**
   - $L_s = \lambda W_s = L_q + \frac{\lambda}{\mu}$
3. **Tiempo esperado de espera en la cola W_q**
4. **Tiempo esperado de espera en el sistema W_s**
   - $W_s = W_q + \frac{1}{\mu}$

Ejemplo

Suponga una estación de gasolina a la cual llegan en promedio 45 clientes por hora. Se tiene capacidad para atender en promedio a 60 clientes por hora y se sabe que los clientes esperan en promedio 3 minutos en la cola. Calcule las medidas de desempeño del sistema.

La tasa media de llegadas λ es 45 clientes por hora o 45/60 = 0.75 clientes por minuto
La tasa media de servicio μ es 60 clientes por hora o 60/60 = 1 cliente por minuto

$$W_q = 3\min$$
$$W_s = W_q + \frac{1}{\mu} = 3 + \frac{1}{1} = 4\min$$
$$L_s = \lambda W_s = 0.75 \times 4 = 3\,clientes$$
$$L_q = \lambda W_q = 0.75 \times 3 = 2.25\,clientes$$

También se pueden usar las probabilidades como medida de desempeño, sus beneficios son que las mismas permiten evaluar escenarios y establecer metas.
A saber:

- P_n : probabilidad de tener n clientes en el sistema
- $P(W_s \le t)$: probabilidad de que un cliente no espere en el sistema más de t horas

Por último, dada la tasa media de llegadas λ y la tasa media de servicio μ, se define el **factor de utilización del sistema ρ**. Generalmente se requiere que $\rho < 1$

Su fórmula, con un servidor y con ***s*** servidores, respectivamente, es:

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

Modelos

Modelo M/M/1 (Un servidor con llegadas de Poisson y tiempos de servicio exponenciales)

Formulas

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

$$W_s = \frac{1}{\mu - \lambda} \qquad W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$P_n = (1-\rho)\rho^n \qquad P(L_s > n) = \rho^{n+1}$$

$$P(W_s > t) = e^{-\mu(1-\rho)t} \qquad P(W_q > t) = \rho e^{-\mu(1-\rho)t}$$

$$t \geq 0, \rho < 1$$

Ejemplo

Un lavacar puede atender un auto cada 5 minutos y la tasa media de llegadas es de 9 autos por hora. Obtenga las medidas de desempeño y la probabilidad de tener 0 clientes en el sistema, la probabilidad de tener una cola de más de 3 clientes y la probabilidad de esperar más de 30 min en la cola y en el sistema.

$$\lambda = 9, \mu = 12, \rho = \frac{9}{12} = 0.75$$

$$L_s = \frac{\lambda}{\lambda - \mu} = 3\,clientes \qquad L_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = 2.25\,clientes$$

$$W_s = \frac{1}{\mu - \lambda} = 0.33\,hrs = 20\min$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = 0.25\,hrs = 15\min$$

$$P_0 = (1-\rho)\rho^0 = 0.25 \qquad P(L_s > 3) = \rho^{3+1} = 0.32$$

$$P(W_s > 30/60) = e^{-\mu(1-\rho)t} = 0.22$$

$$P(W_q > 30/60) = \rho e^{-\mu(1-\rho)t} = 0.17$$

Modelo M/G/1 (Un servidor con tiempos entre llegadas exponenciales y una distribución general de tiempos de servicio)

Formulas

$$L_s = L_q + \rho \qquad L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}$$

$$W_s = W_q + \frac{1}{\mu} \qquad W_q = \frac{L_q}{\lambda}$$

$$P_0 = 1-\rho \qquad P_w = \rho$$

$$\rho < 1$$

Ejemplo

Un lavacar puede atender un auto cada 5 min. y la tasa media de llegadas es de 9 autos/hora siendo σ = 2 min. Obtenga las medidas de desempeño, la probabilidad de tener 0 clientes en el sistema y la probabilidad de que un cliente tenga que esperar por el servicio

$$L_s = L_q + \rho = 1.31 + .75 = 2.06\,clientes$$

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} = 1.31\,clientes$$

$$W_s = W_q + \frac{1}{\mu} = 0.228\,hrs = 13.7\min$$

$$W_q = \frac{L_q}{\lambda} = 0.145\,hrs = 8.7\min$$

$$P_0 = 1-\rho = 0.25 \qquad P_w = \rho = 0.75$$

Modelo M/D/1 (Un servidor con tiempos entre llegadas exponenciales y una distribución degenerada de tiempos de servicio)

Formulas

$$L_s = \lambda W_s \qquad L_q = \frac{\rho^2}{2(1-\rho)}$$

$$W_s = W_q + \frac{1}{\mu} \qquad W_q = \frac{L_q}{\lambda}$$

$$\rho < 1$$

Ejemplo

Un lavacar puede atender un auto cada 5 min siendo la tasa media de llegadas de 9 autos/hora de acuerdo a una distribución exponencial. Obtenga las medidas de desempeño.

$$L_s = \lambda W_s = 1.875\,clientes$$

$$L_q = \frac{\rho^2}{2(1-\rho)} = 1.125\,clientes$$

$$W_s = W_q + \frac{1}{\mu} = 0.21\,hrs = 12.5\min$$

$$W_q = \frac{L_q}{\lambda} = 0.125\,hrs = 7.5\min$$

Modelo M/E$_k$/1 (Un servidor con tiempos entre llegadas exponenciales y una distribución Erlang de tiempos de servicio)

Formulas

$$L_s = \lambda W_s \qquad L_q = \frac{\rho^2(k+1)}{2k(1-\rho)}$$

$$W_s = W_q + \frac{1}{\mu} \qquad W_q = \frac{L_q}{\lambda}$$

$$\rho < 1$$

Ejemplo

Un lavacar puede atender un auto cada 5 min siendo la tasa media de llegadas de 9 autos/hora. Suponga σ = 3.5 min (aprox.) Obtenga las medidas de desempeño.

$$L_s = \lambda W_s = 2.437\,clientes$$

$$L_q = \frac{\rho^2(k+1)}{2k(1-\rho)} = 1.6875\,clientes$$

$$W_s = W_q + \frac{1}{\mu} = 0.2708\,hrs = 16.25\min$$

$$W_q = \frac{L_q}{\lambda} = 0.1875\,hrs = 11.25\min$$

Modelo M/M/s (s servidores con llegadas de Poisson y tiempos de servicio exponenciales)

Formulas

$$P_0 = \frac{1}{\frac{\rho^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)+\sum_{n=0}^{s-1}\frac{\rho^n}{n!}}$$

$$L_q = \frac{\rho^s \lambda\mu}{(s-1)!(s\mu-\lambda)^2}P_0 \qquad L_s = L_q + \frac{\lambda}{\mu} \qquad W_q = \frac{L_q}{\lambda}$$

$$W_s = W_q + \frac{1}{\mu} \qquad P_n = \frac{\rho^n}{n!}P_0\, , si\ \ n \le k$$

$$P_n = \frac{\rho^n}{s!s^{n-s}}P_0\, , si\ \ n > k \qquad P_w = \frac{1}{s!}\rho^s\left(\frac{s\mu}{s\mu-\lambda}\right)P_0$$

$$Si\ \ s = 2$$

$$L_q = \frac{\rho^3}{4-\rho^2}$$

$$Si\ \ s = 3$$

$$L_q = \frac{\rho^4}{(3-\rho)(6-4\rho+\rho^2)}$$