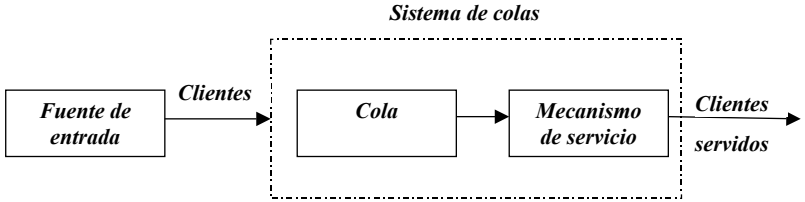
## *TEMA 4: MODELOS DE COLAS*

*INTRODUCCIÓN*: La formación de líneas de espera es un fenómeno común que ocurre siempre que la demanda actual de un cierto servicio excede a la capacidad actual de proporcionarlo. Con frecuencia, deben tomarse decisiones respecto a la cantidad de servicio que debe proporcionarse. Sin embargo, muchas veces es imposible predecir con exactitud cuándo llegarán las unidades que buscan servicio y/o cuánto tiempo será necesario para dar ese servicio. Proporcionar demasiado servicio implica *costos de servicio* excesivos. Por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas en ciertos momentos, incurriéndose en un *costo de espera*. Entonces, la meta final es lograr un *balance económico* entre el costo de servicio y el costo asociado con la espera por ese servicio. Antes de dar respuesta a estas cuestiones, es de interés, describir mediante modelos matemáticos, el desarrollo del sistema de colas, los cuales nos proporcionarán información sobre el mismo, tales como: tiempo medio de espera, número medio de clientes en el sistema, probabilidad de que el sistema esté ocioso,....

*EJEMPLO*: Supóngase que estamos interesado en comprar una impresora para ser utilizada en línea por un grupo de personas. Así pues, cada una de las personas que integran dicho grupo, enviaran sus documentos a imprimir. Si la impresora, está ociosa, el trabajo será inmediatamente procesado e impreso. En caso contrario, el trabajo será almacenado en una línea de espera hasta que llegue su turno de impresión. Para esta situación sería de interés el comportamiento del sistema, con el fin de determinar las características de la impresora a comprar, de tal manera que exista un equilibrio entre el costo de servicio y el costo de espera.

*ESTRUCTURA BÁSICA DE LOS MODELOS DE COLAS*: Una línea de espera, está constituida por un cliente que requiere de un *servicio* (proporcionado por un servidor) en un determinado periodo. Los clientes entran aleatoriamente al sistema y forman una o varias *colas* para ser atendidos. Si el servidor está desocupado, de acuerdo a ciertas reglas preestablecidas, conocidas con el nombre de *disciplina del servicio*, se proporcionan el servicio a los elementos de la cola. El cliente será atendido en un periodo determinado de tiempo, llamado *tiempo de ser*vicio. Al finalizar éste, el cliente abandona el sistema y el servidor, el cual será ocupado por otro cliente, si es que existe.




*ELEMENTOS BÁSICOS DE LOS MODELOS DE COLAS*: Atendiendo a la estructura básica de los modelos de colas, descrita anteriormente, se puede distinguir los principales elementos del sistema:



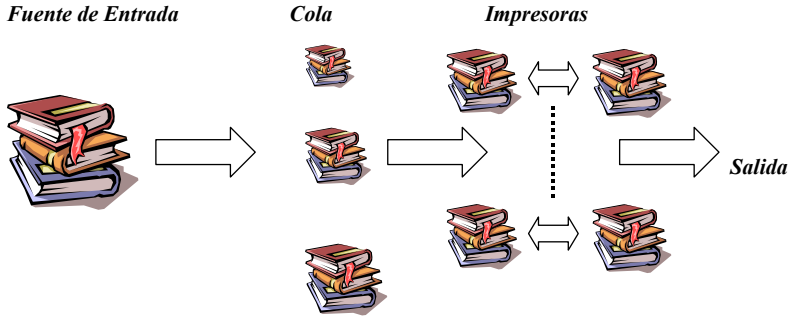


- ***Fuente de entrada***: Generación de nuevos clientes al sistema. La población a partir de la cual surgen las unidades, se conoce como *población de entrada*. Atendiendo al tamaño de esta población, individuos potencialmente distintos, se distingue entre *población finita o infinita*. Asimismo, una característica de interés, es determinar el patrón estadístico mediante el cual se generan los clientes a través del tiempo. La suposición normal, es que se generan de acuerdo a un *proceso de Poisson*, es decir, el número de clientes que llegan hasta un tiempo específico tienen una distribución Poisson.

- ***Cola:*** Una cola se caracteriza por el número máximo permisible de clientes que puede admitir. Según este número, pueden ser infinitas o finitas. Una propiedad relevante en el sistema de colas será la *disciplina de cola*, es decir, el orden en el que se seleccionan sus miembros para recibir el servicio. Por ejemplo, puede ser: primero en entrar, primero en salir *(FIFO)*, aleatoria, de acuerdo a algún criterio de prioridad o algún otro orden.

- ***Mecanismo de servicio***: Consiste en un grupo de *servidores*, distribuidos en una o varias instalaciones, en las cuales entra el individuo y el servidor le presta el servicio completo. Los modelos más usuales, suponen una instalación, ya sea con un servidor o con un número finito de servidores. El tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación, normalmente será aleatorio, se llama *tiempo de servicio*. Es comúnmente suponer la misma distribución a todos los servidores. Asimismo, la distribución del tiempo de servicio que más se usa en la práctica es la *distribución exponencial*.

Atendiendo a los distintos elementos de un modelo de cola, se puede realizar una clasificación del mismo, a partir de sus características fundamentales que los describe. Uno de los modelos más utilizado, es aquel en el que se considera una única línea de espera (que puede estar vacía en ciertos tiempos) frente a una instalación de servicio, dentro de la cual se encuentra uno o más servidores. Cada cliente generado por una fuente de entrada recibe el servicio de uno de los servidores, quizás después de esperar un poco en la cola (línea de espera). En el siguiente gráfico, se ilustra el esquema de dicho sistema:



Con el fin de identificar al modelo de colas estudiado, utilizaremos la siguiente notación:

***Distribución de tiempos entre llegadas/Distribución de tiempos entre servicio/Número de servidores***

donde *M* denota a la distribución exponencial y *G* a una distribución general.

*PARÁMETROS DE INTERÉS EN UN MODELO DE COLA*: Descrito el funcionamiento de un sistema de colas, a continuación enumeraremos los principales parámetros del mismo, que nos informará sobre su comportamiento a lo largo del tiempo.

$N(t)$: Número de clientes en el sistema de colas en el tiempo *t*, $t \geq 0$

$P_n(t)$: Probabilidad de que exactamente *n* clientes estén en el sistema en tiempo *t*.

$s$: número de servidores en el sistema de colas

$\lambda_n$: tasa media de llegadas (número esperado de llegadas por unidad de tiempo) de nuevos clientes en el sistema cuando hay *n* clientes en el sistema.

$\mu_n$: tasa media de servicio para todo el sistema cuando hay *n* clientes en el sistema.

Cuando $\lambda_n$ es constante para todo *n*, esta constante se denota por $\lambda$. Asimismo, si la tasa media de servicio por servidor ocupado es constante para todo *n*, esta constante se denota por $\mu$, donde $\mu_n = s\mu$ si $n \geq s$. En esta circunstancias, $1/\lambda$ y $1/\mu$ son los *tiempos entre llegadas esperados* y los *tiempos de servicio esperados*, respectivamente. Por tanto, $\rho = \dfrac{\lambda}{s\mu}$ es el *factor de utilización* para la instalación de servicio, es decir, la fracción esperada de tiempo que los servidores individuales están ocupados. ¿Qué sucederá si $\rho > 1$?

*OBSERVACIÓN*: Cuando un sistema de colas apenas inicia su operación, el estado del sistema (el número de clientes en el sistema) se encuentra bastante afectado por el estado inicial y el tiempo que ha pasado desde el inicio. Se dice entonces que el sistema se encuentra en *condición transitoria*. Después de que ha pasado un tiempo suficiente, el estado del sistema se vuelve, en esencia, independiente del estado inicial y del tiempo transcurrido. Así, se puede decir que el sistema ha alcanzado su *condición de estado estable*, en la que la distribución de probabilidad del estado del sistema se conserva (la distribución estacionaria o de estado estable) a través del tiempo.

A partir de ahora, suponemos que el sistema ha alcanzado su condición de estado estable, es decir no depende del tiempo. En esta situación, sea:

$P_n$: Probabilidad de que haya exactamente *n* clientes en el sistema.

$L$: Número esperado de clientes en el sistema.

$L_q$: Longitud esperada de la cola (excluye los clientes que están en el servicio).

$\varpi$: Tiempo de espera en el sistema (incluye el tiempo de servicio) para cada cliente.

$W: E[\varpi]$

$\varpi_q$: Tiempo de espera en la cola (excluye tiempo de servicio) para cada cliente.

$W_q: E[\varpi_q]$





Se verifican las siguientes relaciones entre los parámetros anteriormente definidos:

$$L = \lambda W \quad \textit{Fórmula de Little}; \qquad L_q = \lambda W_q; \qquad W = W_q + \frac{1}{\mu}$$

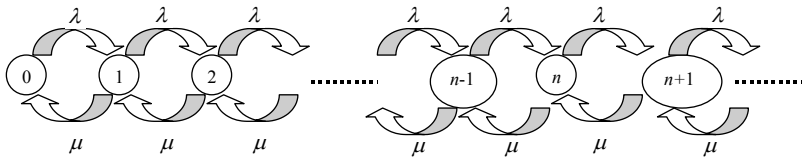
(interpretar las relaciones anteriores).

A continuación, se realizará un estudio analítico de los principales modelos de inventarios, para ciertas restricciones de $\rho$ y los tiempos de llegada y de servicios son independientes. En primer lugar abordaremos el problema bajo el supuesto de la distribución exponencial de dichos tiempos, apoyándonos en los procesos de nacimiento-muerte. Finalmente se generalizan dichos resultados cuando o bien los tiempos de llegadas o de servicio no se comportan exponencialmente.

*MODELOS DE COLAS BASADOS EN EL PROCESO DE NACIMIENTO-MUERTE*: En esta sección, supondremos que los tiempos de llegadas y de servicio siguen modelos exponenciales, de parámetros $\lambda$ y $\mu$, respectivamente. Por tanto, el número de llegada al sistema, se describe a partir de un proceso de Poisson de intensidad $\lambda$. Así pues, bajo la suposición de independencia entre los sucesos llegada y servicio, se puede describir el sistema de cola, como un proceso de nacimiento-muerte, en el que una llegada al sistema es considerada como un nacimiento, mientras que una salida del mismo, es considerada como una muerte.

A continuación, se estudia analíticamente algunos modelos de colas de interés:

***Modelo M/M/1***: Este es el modelo más básico que se puede considerar con un único servidor en el sistema. En este caso la tasa de llegada (nacimiento) es $\lambda_n = \lambda$ y la tasa de servicio (muerte) $\mu_n = \mu$, obteniéndose el siguiente esquema:




Teniendo en cuenta la teoría sobre procesos de nacimiento-muerte se tiene que la tasa media de entrada en el estado *k* es igual a la tasa media de salida de *k.* Así pues:

$$\lambda P_0 = \mu P_1 \Rightarrow P_1 = \frac{\lambda}{\mu} P_0$$

$$\lambda P_0 + \mu P_2 = (\lambda + \mu) P_1 \Rightarrow P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0$$

$$\lambda P_{n-1} + \mu P_{n+1} = (\lambda + \mu) P_n, n \geq 2 \Rightarrow P_{n+1} = \left(\frac{\lambda}{\mu}\right)^{n+1} P_0$$

$$\sum P_n = 1 \Rightarrow P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1}$$

Por tanto, si $\rho \geq 1$ el sistema no alcanzará el estado estable pues la cola "explota" y crece sin límite, incluso cuando $\lambda = \mu$, caso crítico, pues el número esperado de clientes en el sistema crecerá sin límite lentamente con el tiempo, aunque siempre es posible un regreso temporal a no tener clientes, las probabilidades de tener número grandes de clientes crecen significativamente con el tiempo. Si $\rho < 1$, se obtienen los siguientes valores de los parámetros:

$$P_0 = 1-\rho\,;\;\; P_n = (1-\rho)\rho^n\,;\;\; L = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\lambda}{\mu-\lambda}\,;\; L_q = \sum_{n=1}^{\infty}(n-1)P_n = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Además, a partir de las relaciones vista anteriormente se tiene:

$$W = E[\varpi] = \frac{1}{\mu-\lambda}\,;\quad W_q = E[\varpi_q] = \frac{\lambda}{\mu(\mu-\lambda)}$$

y distribuciones de tiempos de espera $\varpi \mapsto \exp(\mu-\lambda)$ y $\varpi_q \mid \varpi_q > 0 \mapsto \exp(\mu-\lambda)$.

*EJEMPLO*:

| *System: M/M/1* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *1,0000* |
| *Service rate per server (mu) per hour =* | *5,0000* |
| *Overall system effective arrival rate per hour =* | *1,0000* |
| *Overall system effective service rate per hour =* | *1,0000* |
| *Overall system utilization =* | *20,0000 %* |
| *Average number of customers in the system (L) =* | *0,2500* |
| *Average number of customers in the queue (Lq) =* | *0,0500* |
| *Average time customer spends in the system (W) =* | *0,2500 hours* |
| *Average time customer spends in the queue (Wq) =* | *0,0500 hours* |
| *The probability that all servers are idle (Po) =* | *80,0000 %* |
| *The probability an arriving customer waits (Pw or Pb) =* | *20,0000 %* |

| *n* | *n Customers in the System* | *Cumulative Probability* |
|---|---|---|
| *0* | *0,8000* | *0,8000* |
| *1* | *0,1600* | *0,9600* |
| *2* | *0,0320* | *0,9920* |
| *3* | *0,0064* | *0,9984* |
| *4* | *0,0013* | *0,9997* |
| *5* | *0,0003* | *0,9999* |
| *6* | *0,0001* | *1,0000* |

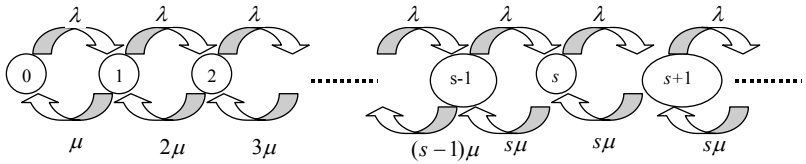
***Modelo M/M/s***: Este modelo generaliza al anterior, en el cual se consideran *s>1* servidores en el sistema. En este caso, la tasa de llegada (nacimiento) es $\lambda_n = \lambda$ y la tasa de servicio (muerte) $\mu_n$, depende de *n* como sigue:

$$\mu_n = \begin{cases} n\mu, si\ n \leq s \\ s\mu, si\ n > s \end{cases}$$





obteniéndose el siguiente esquema:




Teniendo en cuenta la teoría sobre procesos de nacimiento-muerte se tiene que la tasa media de entrada en el estado *k* es igual a la tasa media de salida de *k.* Así pues:

$$\lambda P_0 = \mu P_1 \Rightarrow P_1 = \frac{\lambda}{\mu} P_0$$

$$\lambda P_{n-1} + (n+1)\mu P_{n+1} = (\lambda + n\mu)P_n, n+1 \leq s$$

$$\lambda P_{n-1} + s\mu P_{n+1} = (\lambda + s\mu)P_n, n \geq s$$

Por tanto, si $\rho \geq 1$ $(\lambda \geq s\mu)$el sistema no alcanzará el estado estable. Si $\rho < 1$, se obtienen los siguientes valores de los parámetros:

$$\begin{cases} P_0 = 1\Big/\left[\sum_{n=0}^{s-1}\frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\frac{1}{1-(\lambda/\mu s)}\right] \\ P_n = \frac{(\lambda/\mu)^n}{n!}P_0, si\ n \leq s \qquad ;\; L_q = \sum_{n=s}^{\infty}(n-s)P_n = \frac{P_o(\lambda/\mu)^s\rho}{s!(1-\rho^2} \\ P_n = \frac{(\lambda/\mu)^n}{s!s^{n-s}}P_0, si\ n \geq s \end{cases}$$

Además, a partir de las relaciones vista anteriormente se tiene:

$$W_q = \frac{L_q}{\lambda}\,;\qquad W = W_q + \frac{1}{\mu}\,;\qquad L = \lambda\left(W_q + \frac{1}{\mu}\right) = L_q + \frac{\lambda}{\mu}$$

*EJEMPLO*:

| *System: M/M/2* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *1,0000* |
| *Service rate per server (mu) per hour =* | *2,0000* |
| *Overall system effective arrival rate per hour =* | *1,0000* |
| *Overall system effective service rate per hour =* | *1,0000* |
| *Overall system utilization =* | *25,0000 %* |
| *Average number of customers in the system (L) =* | *0,5333* |
| *Average number of customers in the queue (Lq) =* | *0,0333* |
| *Average time customer spends in the system (W) =* | *0,5333 hours* |
| *Average time customer spends in the queue (Wq) =* | *0,0333 hours* |
| *The probability that all servers are idle (Po) =* | *60,0000 %* |
| *The probability an arriving customer waits (Pw or Pb) =* | *10,0000 %* |

| *n* | *n Customers in the System* | *Cumulative Probability* |
|---|---|---|
| *0* | *0,6000* | *0,6000* |
| *1* | *0,3000* | *0,9000* |
| *2* | *0,0750* | *0,9750* |
| *3* | *0,0188* | *0,9938* |
| *4* | *0,0047* | *0,9984* |
| *5* | *0,0012* | *0,9996* |
| *6* | *0,0003* | *0,9999* |
| *7* | *0,0001* | *1,0000* |

***Variación de cola finita al modelo M/M/s***: Este modelo considera la finitud de la línea de espera, es decir, no se permiten que el *número de clientes en el sistema* exceda un número especificado, denotado por *K,* considerándose que la cola es de longitud *K-s*. A cualquier cliente que llega cuando la cola está "llena" se le niega la entrada al sistema y este cliente lo deja para siempre. Desde el punto de vista del proceso de nacimiento-muerte, la tasa de entrada al sistema se hace cero en estos momentos. Por tanto, la única modificación necesaria en el modelo *M/M/s* para introducir una cola finita es considerar los parámetros:

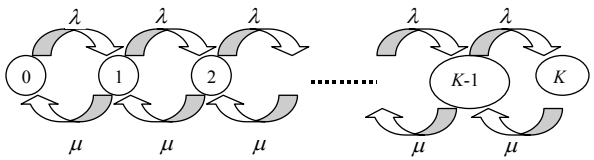
$$\lambda_n = \begin{cases} \lambda, si\ n = 0,...,K-1 \\ 0, si\ n \geq K \end{cases}$$

*OBSERVACIÓN*: Como $\lambda_n = 0$ para $n \geq K$, el sistema de colas alcanzará la condición estable, pues no habrá más de *K* individuos en el sistema. Por tanto, para cualquier tasa de utilización $\rho$ será válido.

Este modelo es etiquetado como *M/M/s/K*, donde *K* indica que la cola es finita de longitud *K-s*. Obsérvese que el modelo *M/M/s,* se ajusta al modelo $M/M/s/\infty$.

A continuación se detallará el estudio del modelo para *s=1*, generalizándose después para *s>1*.

***s=1***: Para esta situación se obtiene el siguiente esquema:




Obteniéndose, si $\rho \neq 1$:

$$P_0 = \frac{1}{\sum_{n=0}^{K} \rho^n};\ P_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n,\ n = 0,...,K;\ L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}};\ L_q = L - (1-P_0)$$





Si $\rho = 1$, entonces $P_n = \frac{1}{K+1}$, $n = 0,...,K$; $L = K/2$; $L_q = L - (1-P_0)$

Además, en cualquier situación, se verifica:

$$W = \frac{L}{\lambda(1-P_K)};\ W_q = \frac{L_q}{\lambda(1-P_K)}$$

*EJEMPLO*:

| *System: M/M/1/2* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *1,0000* |
| *Service rate per server (mu) per hour =* | *5,0000* |
| *Overall system effective arrival rate per hour =* | *0,9677* |
| *Overall system effective service rate per hour =* | *0,9677* |
| *Overall system utilization =* | *19,3548 %* |
| *Average number of customers in the system (L) =* | *0,2258* |
| *Average number of customers in the queue (Lq) =* | *0,0323* |
| *Average time customer spends in the system (W) =* | *0,2333 hours* |
| *Average time customer spends in the queue (Wq) =* | *0,0333 hours* |
| *The probability that all servers are idle (Po) =* | *80,6452 %* |
| *Average number of customers being balked per hour =* | ***0,0323*** |

| *n* | *n Customers in the System* | *Cumulative Probability* |
|---|---|---|
| *0* | *0,8065* | *0,8065* |
| *1* | *0,1613* | *0,9677* |
| *2* | *0,0323* | *1,0000* |

Obsérvese que si aumentamos el tamaño de la cola, obtenemos el modelo *M/M/1*.

| *System: M/M/1/10* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *1,0000* |
| *Service rate per server (mu) per hour =* | *5,0000* |
| *Overall system effective arrival rate per hour =* | *1,0000* |
| *Overall system effective service rate per hour =* | *1,0000* |
| *Overall system utilization =* | *20,0000 %* |
| *Average number of customers in the system (L) =* | *0,2500* |
| *Average number of customers in the queue (Lq) =* | *0,0500* |
| *Average time customer spends in the system (W) =* | *0,2500 hours* |
| *Average time customer spends in the queue (Wq) =* | *0,0500 hours* |
| *The probability that all servers are idle (Po) =* | *80,0000 %* |
| *Average number of customers being balked per hour =* | *0,0000* |

***s>1***: Se obtienen las siguientes expresiones para los principales parámetros:

$$\begin{cases} P_0 = 1 \Big/ \left[ \sum_{n=0}^{s} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s+1}^{K} \left( \frac{\lambda}{s\mu} \right)^{n-s} \right] \\ P_n = \frac{(\lambda/\mu)^n}{n!} P_0, si\ n = 1,...,s \\ P_n = \frac{(\lambda/\mu)^n}{s!\, s^{n-s}} P_0, si\ n = s, s+1,...,K \\ P_n = 0, si\ n > K \end{cases}$$

$$L_q = \frac{P_0 (\lambda/\mu)^s \rho}{s!(1-\rho)^2} \left[ 1 - \rho^{K-s} - (K-s)\rho^{K-s}(1-\rho) \right]$$

$$L = \sum_{n=0}^{s-1} n P_n + L_q + s\left( 1 - \sum_{n=0}^{s-1} P_n \right)$$

*EJEMPLO*:

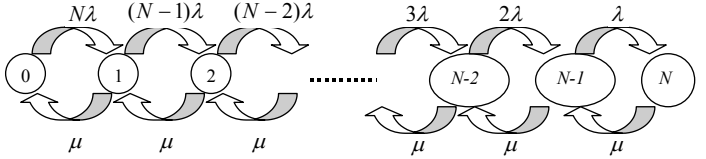
| *System: M/M/2/3* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *1,0000* |
| *Service rate per server (mu) per hour =* | *2,0000* |
| *Overall system effective arrival rate per hour =* | *0,9811* |
| *Overall system effective service rate per hour =* | *0,9811* |
| *Overall system utilization =* | *24,5283 %* |
| *Average number of customers in the system (L) =* | *0,5094* |
| *Average number of customers in the queue (Lq) =* | *0,0189* |
| *Average time customer spends in the system (W) =* | *0,5192 hours* |
| *Average time customer spends in the queue (Wq) =* | *0,0192 hours* |
| *The probability that all servers are idle (Po) =* | *60,3774 %* |
| *Average number of customers being balked per hour =* | *0,0189* |

| *n* | *n Customers in the System* | *Cumulative Probability* |
|---|---|---|
| *0* | *0,6038* | *0,6038* |
| *1* | *0,3019* | *0,9057* |
| *2* | *0,0755* | *0,9811* |
| *3* | *0,0189* | *1,0000* |

***Variación de población finita al modelo M/M/s***: Este modelo incorpora la novedad con respecto al estudiado *M/M/s* en que la *fuente de entrada está limitada*, es decir, el tamaño de la población potencial es finito. Para este caso, sea *N* el tamaño de esa población, denotándose *M/M/s/N/N*. Darse cuenta que cuando el número de clientes en el sistema de colas es *n*, existe sólo *N-n* clientes potenciales restantes en la fuente de entrada. La aplicación más importante de este modelo es el *problema de reparación de máquinas*. En este caso, todos los miembros de la población potencial se encuentran alternativamente *dentro* y *fuera* del sistema de colas, donde el tiempo fuera de cada miembro, esto es, el tiempo que pasa desde que deja el sistema hasta que regresa, tiene una distribución exponencial con parámetro $\lambda$. La situación, anteriormente planteada para un único servidor, se representa en el siguiente esquema:








Las expresiones para los principales parámetros son dadas:

$$\begin{cases} P_0 = 1 \Big/ \left[ \sum_{n=0}^{N} \frac{N!}{(N-n)!} \left( \frac{\lambda}{\mu} \right)^n \right] \\ P_n = \frac{N!}{(N-n)!} \left( \frac{\lambda}{\mu} \right)^n P_0, si\ n = 1,...,N \end{cases} ;\ L_q = N - \frac{\lambda+\mu}{\lambda}(1-P_0);\ L = N - \frac{\mu}{\lambda}(1-P_0)$$

$$W = \frac{L}{\lambda(N-L)};\ W_q = \frac{L_q}{\lambda(N-L)}$$

*EJEMPLO*: Supóngase un modelo de cola con población finita *N=10,* $\lambda = 2$, $\mu = 4$.

| *System: M/M/1/10/10* | *From Formula* |
|---|---|
| *Customer arrival rate (lambda) per hour =* | *2,0000* |
| *Service rate per server (mu) per hour =* | *4,0000* |
| *Overall system effective arrival rate per hour =* | *3,9998* |
| *Overall system effective service rate per hour =* | *3,9998* |
| *Overall system utilization =* | *99,9962 %* |
| *Average number of customers in the system (L) =* | *8,0001* |
| *Average number of customers in the queue (Lq) =* | *7,0001* |
| *Average time customer spends in the system (W) =* | *2,0001 hours* |
| *Average time customer spends in the queue (Wq) =* | *1,7501 hours* |
| *The probability that all servers are idle (Po) =* | *0,0038 %* |

| *n* | *n Customers in the System* | *Cumulative Probability* |
|---|---|---|
| *0* | *0,0000* | *0,0000* |
| *1* | *0,0002* | *0,0002* |
| *2* | *0,0009* | *0,0011* |
| *3* | *0,0034* | *0,0045* |
| *4* | *0,0120* | *0,0166* |
| *5* | *0,0361* | *0,0526* |
| *6* | *0,0902* | *0,1429* |
| *7* | *0,1804* | *0,3233* |
| *8* | *0,2707* | *0,5940* |
| *9* | *0,2707* | *0,8647* |
| *10* | *0,1353* | *1,0000* |

*OBSERVACIÓN*: En ocasiones, si el tamaño de la población es excesivamente grande, se puede considerar como infinita y aplicar el modelo *M/M/1*.

*OBSERVACIÓN*: Se puede considerar la situación anterior con más de un servidor, *s>1*. Incluso, combinar la situación de finitud en la cola, con la finitud en la población.

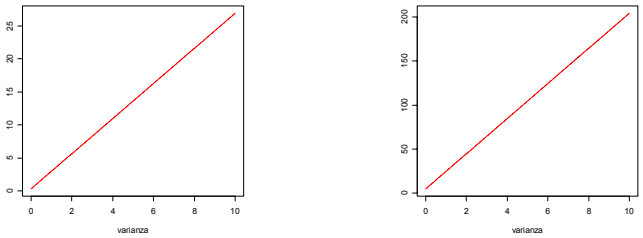
*MODELOS DE COLAS CON DISTRIBUCIONES NO EXPONENCIALES*: Todos los modelos presentados anteriormente, suponen como hipótesis la distribución exponencial tanto en los tiempos de llegada como de servicio. En múltiples ocasiones, estas suposiciones, no se puede asumir. Es importante por ello, disponer de otros modelos de colas que usen otras distribuciones de probabilidad. Desafortunadamente, el análisis de los modelos de colas con distribuciones no exponenciales es mucho más difícil. Por tanto, en ocasiones, dicho análisis se realiza a través de un *proceso de simulación*. A continuación se exponen algunos resultados útiles para algunos modelos concretos.

***Modelo M/G/1***: Sistema de colas con un servidor y proceso de entrada de Poisson con una tasa media de llegada fija $\lambda$. Se suponen que los clientes tienen tiempo de servicio independientes con la misma distribución de probabilidad, no necesariamente exponencial. Conocida la media $1/\mu$ y varianza $\sigma^2$, cualquier sistema de líneas de espera de este tipo podrá alcanzar, en algún momento, una condición de estado estable si $\rho = \lambda/\mu < 1$. Para esta situación, se obtienen las siguientes expresiones de los parámetros:

$$P_0 = 1-\rho;\ L_q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)};\ L = \rho + L_q;\ W_q = \frac{L_q}{\lambda};\ W = W + \frac{1}{\mu}$$

Darse cuenta que se obtienen las mismas expresiones que para el modelo *M/M/1* cuando $\sigma^2 = \frac{1}{\mu^2}$, la varianza de la distribución exponencial. Así mismo, si la tasa de servicio está fija, pero aumenta la varianza, los parámetros $L, L_q, W, W_q$, también se incrementan.

$L_q$ $W_q$




Asimismo, si la distribución es degenerada, es decir el tiempo de servicio es constante, entonces $\sigma^2 = 0$, verificándose que $L_q = \frac{\rho^2}{2(1-\rho)}$, que es la mitad del obtenido para el caso exponencial, $L_q = \frac{\rho^2}{(1-\rho)}$. Lo mismo ocurre para $W_q$.

*OBSERVACIÓN*: Se podría realizar un estudio similar, cuando el tiempo de llegadas no es exponencial, pero es bastante complicado obtener una solución analítica para esta situación. Es por ello que la simulación, jugará un papel fundamental.





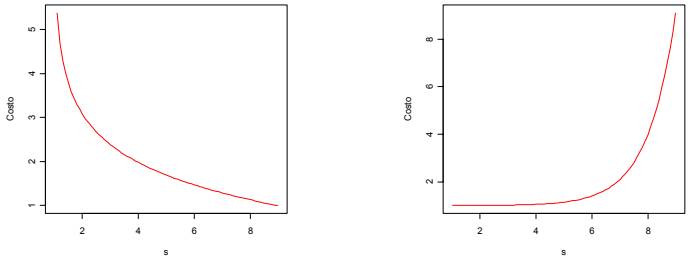
*OBSERVACIÓN*: Todos los modelos de colas estudiados, suponen como disciplina de cola, *FIFO*, es decir primero en entrar primero en salir. Existen otros modelos, más complejo con otro tipo de disciplina, como por ejemplo, una disciplina con prioridades entre varios grupos de clientes.

*PROCESO DE DECISIÓN EN LOS SISTEMAS DE COLAS*: Una vez estudiado el comportamiento de un sistema de colas determinado, será de interés seleccionar el mejor modelo para una situación particular, tal que se *minimice el costo total*, es decir, el costo de servicio y de espera. Con el fin de lograr un equilibrio entre ambos costos, las decisiones más comunes en estos modelos, se refieren a fijar el nivel de los parámetros del sistema, sean:

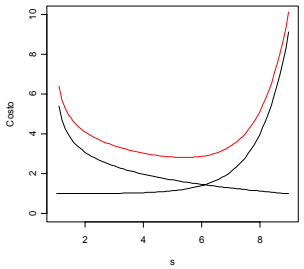
- Número requerido de servidores en una unidad de servicio, *s*.
- Eficacia de servicio, $\mu$.
- Número requerido de unidades de servicios, $\lambda$.

Es conocido, que si aumentamos el número de servidores, se reduce el gasto de espera incrementándose el costo de servicio. Este comportamiento se invierte si disminuimos el número de servidores. Análogo sucederá con la tasa de servicio.

***Costo de espera*** ***Costo de servicio***




***Costo total = Costo de espera + Costo de servicio***




En múltiples ocasiones, la dificultad estriba en la determinación explícita de los costos de espera, no tanto así los costos de servicio correspondiente a la operación y mantenimiento del sistema, que dependerán de la tasa de servicio. En realidad, dadas

las características aleatorias de una línea de espera, se debe hablar de *valores esperados de costos* y no del costo en sí. A partir de ahora denotaremos por $E[CW]$ al *costo esperado por espera del servicio*. Con el fin de obtener una expresión del costo de espera, es comúnmente aceptado que o bien depende del *número de individuos* en el sistema $g(N)$, o bien del *tiempo que transcurren* cada individuo en el sistema, $h(\varpi)$, verificándose:

$$E[CW] = E[g(N)] = \sum_{n=0}^{\infty} g(n)P_n \; ; \; E[h(\varpi)] = \int_0^{\infty} h(w) f_{\varpi}(w) dw$$

Un caso sencillo, es cuando $g(N)$ es una función de tipo lineal, es decir, la tasa de costo esperado es proporcional a *N*: $g(N) = C_W N$, donde $C_W$ es el costo de espera por unidad de tiempo para cada cliente. En este caso, se obtiene que $E[CW] = E[g(N)] = C_W L$.

Así pues, a continuación, nos plantearemos resolver los siguientes problemas de decisión de líneas de espera:

- Determinación del número óptimo de servidores, dada una función de costo de espera y servicio, así como las tasas $\lambda$ y $\mu$.
- Determinación de la tasa de servicio $\mu$, dada una función de costo de espera y servicio, así como las tasas $\lambda$ y el número de servidores *s*.
- Determinación de la tasa de servicio $\mu$ y el número *s* de servidores, dada una función de costo de espera y servicio, así como las tasas $\lambda$.

*DETERMINACIÓN DEL NÚMERO s ÓPTIMO DE SERVIDORES*: En esta situación, se debe determinar el número óptimo de servidores, en función del costo total incurrido. Para ello será suficiente calcular el número óptimo $S^*$, tal que:

$$E_{S^*}[CT] = \min_s \{sC_s + E[CW]\}$$

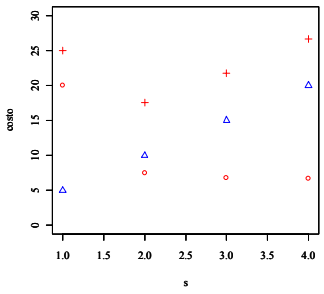
donde $C_s$ denota el costo marginal de una servidor por unidad de tiempo. En general será suficiente considerar unos cuantos valores de *s*.

*EJEMPLO*: Determinar el número de impresoras a comprar, que estarán conectadas en línea, con una tasa de servicio $\mu = 3$ si la tasa de llegada es $\lambda = 2$, el costo de servicio es de 5 euros y el costo por trabajo en el sistema es de 10 euros.

| ***Servidores*** | ***L*** | ***E[CW]*** | ***E[CE]*** | ***E[CT]*** |
|---|---|---|---|---|
| ***1*** | *2* | | | |
| ***2*** | ***0.75*** | | | |
| ***3*** | *0.6760* | | | |
| ***4*** | *0.6677* | | | |







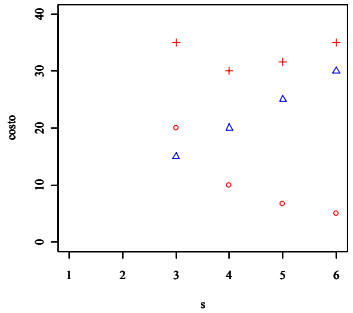

*DETERMINACIÓN DE LA TASA DE SERVICIO* $\mu$: En esta situación, se debe determinar la tasa óptima de servicio, en función del costo total incurrido. Para ello será suficiente calcular la tasa número óptimo $\mu^*$, tal que:

$$E[CT] = \min_{\mu} \{sf(\mu) + E[CW]\}$$

donde $f(\mu)$ denota el costo marginal de una servidor por unidad de tiempo cuando la tasa media de servicio es $\mu$.

*EJEMPLO*: Determinar la tasa de servicio, unidades enteras, de la impresora a comprar, si la tasa de llegada es $\lambda = 2$, el costo de servicio es de 5 $\mu$ euros y el costo por trabajo en el sistema es de 10 euros.

| ***Tasa de servicio*** | ***L*** | ***E[CW]*** | ***E[CE]*** | ***E[CT]*** |
|---|---|---|---|---|
| ***1*** | *Situación no estable* | | | |
| ***2*** | | | | |
| ***3*** | *2* | | | |
| ***4*** | ***1*** | | | |
| ***5*** | *0.6667* | | | |
| ***6*** | *0.5* | | | |



*DETERMINACIÓN DE LA TASA DE SERVICIO $\mu$ Y NÚMERO s DE SERVIDORES*: En esta situación, se debe determinar tanto la tasa óptima de servicio como el número de servidores óptimos, en función del costo total incurrido. Se obtiene que, si:

a) existe un valor $\mu$ factible para *s=1* que minimice el costo total esperado y
b) $f(\mu)$, costo marginal de una servidor por unidad de tiempo, es o una función lineal

la solución óptima, será la selección de *un único servidor*, es decir es mejor *concentrar la capacidad de servicio* en un servidor rápido que dispersarla entre varios lentos. Además la condición *b)*, nos indica que esta concentración de cierta cantidad de servicio se puede realizar sin aumentar el costo del mismo. Es por ello que la solución $(s^*, \mu^*)$, $s^* > 1$, tiene asociado un costo esperado total, mayor que la solución con un único servidor $(1, s^*\mu^*)$, pues se tiene mayor capacidad de servicio, por tanto menor costo de espera y $s^* f(\mu) = f(s^*\mu^*)$, pues la función es lineal, es decir se obtiene el mismo costo de servicio.

| ***N=n*** | ***Tasa media de servicios terminados*** |
|---|---|
| | $(s^*, \mu^*)$ *contra* $(1, s^*\mu^*)$ |
| $n=0$ | $0=0$ |
| $n = 1,2,...,s^* - 1$ | $n\mu^* < s^*\mu^*$ |
| $n \geq s^*$ | $s^*\mu^* = s^*\mu^*$ |

Por tanto, será suficiente aplicar el modelo 2 planteado anteriormente cuando *s=1*.

*OBSERVACIÓN*: Darse cuenta que en la resolución analítica planteada para la decisión en los modelos de cola, suponen el conocimiento del comportamiento del mismo. Así pues, como muchas ocasiones esto no es posible, el *proceso de simulación*, jugará un papel fundamental en esta situación.