Modelo matemático descriptivo (Tomado de *H. Taha*[1] y otros autores)

La distribución exponencial.

Los tiempos (aleatorios) entre llegadas y de servicio se describen de forma cuantitativa con el propósito de modelar colas mediante la distribución exponencial.

$$f(t) = \lambda e^{-\lambda t}, t > 0 \text{ donde } E(t) = \frac{1}{\lambda}$$

Si t se distribuye en forma exponencial y S es el intervalo de tiempo desde la ocurrencia del último evento, entonces $P\{t > T + S | t > S\} = P\{t > T\}$ "*propiedad de olvido*".

$$P\{t > Y\} = 1 - P\{t < Y\} = 1 - \int_0^Y f(t)dt = 1 - \int_0^Y \lambda e^{-\lambda t} dt = 1 - \left(-e^{-\lambda t}\Big|_0^Y\right) = 1 - \left(-e^{-\lambda Y} + 1\right) = e^{-\lambda Y}$$

$$P\{t > T + S | t > S\} = \frac{P\{t > T + S, t > S\}}{P\{t > S\}} = \frac{P\{t > T + S\}}{P\{t > S\}} = \frac{e^{-\lambda(T+S)}}{e^{-\lambda S}} = e^{-\lambda T} = P\{t > T\}$$

La distribución exponencial se basa en tres condiciones:

1. Dado $N(t)$, el número de eventos durante el intervalo (0,*t*), el proceso de probabilidad que describe $N(t)$ tiene incrementos independientes[2] estacionarios[3], en el sentido de que la probabilidad de un evento que ocurre en el intervalo $(T, T + S)$ depende sólo de la longitud de *S*.

2. La probabilidad de que un evento ocurra en un intervalo de tiempo suficientemente pequeño $h > 0$ es positiva y menor que 1.

3. En un intervalo de tiempo suficientemente pequeño, $h > 0$, como mucho puede ocurrir un evento, es decir, $P\{N(h) > 1\} = 0$ (no hay simultaneidad de eventos).

Las tres condiciones dadas describen un proceso donde el conteo de eventos durante un intervalo de tiempo dado, sigue la distribución de *Poisson* y que equivalentemente, el intervalo de tiempo entre eventos sucesivos es *exponencial*. En tal caso, se dice que las condiciones representan un *Proceso de Poisson*.

---

[1] *Taha H. A.*, "Investigación de Operaciones, una introducción", Prentice-Hall, 1998.

[2] $\forall r > s > t \geq 0$ las variables aleatorias $N(r) - N(s)$ y $N(s) - N(t)$ son independientes.

[3] $\forall s > t \geq 0$ y $h > 0$ las variables aleatorias $N(s) - N(t)$ y $N(s + h) - N(t + h)$ están idénticamente distribuidas (homogéneo en el tiempo, no importa cuando ocurren los eventos sino la longitud).

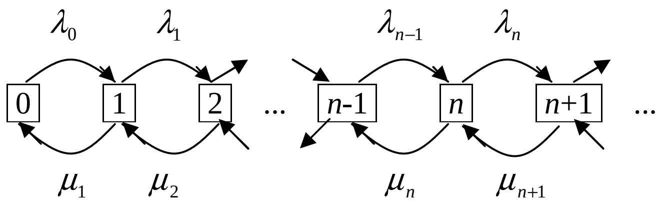
Los modelos de nacimiento y muerte permiten representar situaciones de colas:

- Se permiten llegadas solamente (nacimiento puro). Ejemplo, creación de actas de nacimiento para bebés.
- Se permiten salidas solamente (muerte pura). Ejemplo, retiro aleatorio de artículos del inventario de las existencias.

Modelo de colas de Poisson generalizado.

Modelo base para la derivación de modelos posteriores. Modelo general de colas que combina llegadas y salidas con base en las suposiciones de Poisson (tiempos entre llegadas y de servicio se distribuyen exponencialmente). El modelo supone que las tasas de llegadas y de salida son dependientes del estado del sistema. Definamos:

n = Número de clientes en el sistema.
λ_n = Tasa de llegada de los clientes dados n clientes en el sistema.
μ_n = Tasa de salida de los clientes dados n clientes en el sistema.
p_n = Probabilidad de estado estable de n clientes en el sistema.



Tasa esperada de flujo de entrada al estado n =

$$0(p_0 + \cdots + p_{n-2}) + \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} + 0(p_{n+2} + \cdots) = \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}$$

Tasa esperada de flujo de salida del estado n $= (\lambda_n + \mu_n) p_n$

Igualando las dos tasas se obtiene la ecuación de balance (equilibrio):

$$\lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} = (\lambda_n + \mu_n) p_n, \quad n = 1,2,\ldots$$

$$\lambda_0 p_0 = \mu_1 p_1, \quad n = 0 \text{ (ver figura)}$$

Las ecuaciones de balance (equilibrio) se resuelven en forma recursiva:

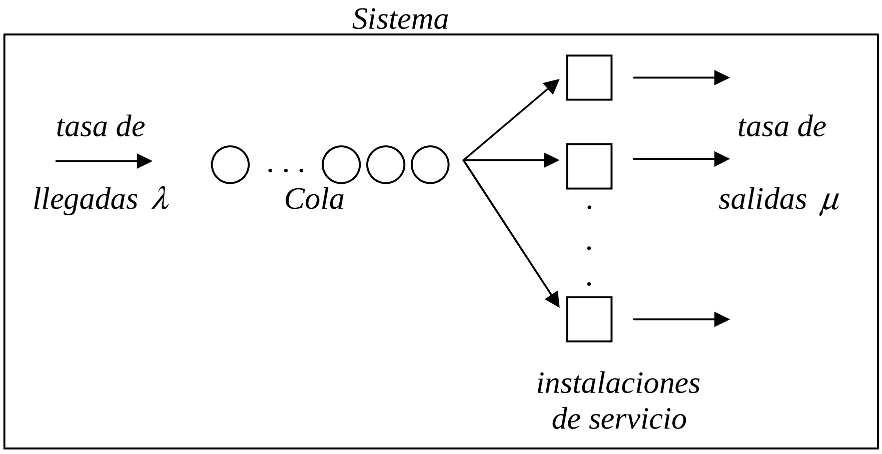
$$n = 0 \Rightarrow p_1 = \left(\frac{\lambda_0}{\mu_1}\right) p_0$$

$$n = 1 \Rightarrow p_2 = \left(\frac{\lambda_1 \lambda_0}{\mu_2 \mu_1}\right) p_0$$

En general, $p_n = \left(\dfrac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}\right) p_0, \quad n = 1,2,\ldots$

El valor de p_0 sale de la ecuación $\sum_{n=0}^{\infty} p_n = 1$

Colas especializadas de Poisson.
(Se modelan líneas de espera con llegadas y salidas combinadas)



Una notación estandarizada universalmente para resumir las características principales de las líneas de espera en paralelo es la introducida por *Kendall-Lee-Taha*.

$$(a/b/c):(d/e/f)^4$$

donde:

- ***a*** describe la distribución de las llegadas.
- ***b*** describe la distribución de las salidas (tiempo de servicio).
- ***c*** representa el número de servidores paralelos.
- ***d*** describe la disciplina de la cola.
- ***e*** representa el número máximo de clientes permitidos en el sistema.
- ***f*** representa el tamaño de la fuente de la que se generan los clientes.

Ejemplo:

$$(M/M/10):(DG/N/\infty)$$

Se tiene un sistema con llegadas y salidas Poisson (*M*), o equivalentemente, los tiempos entre llegadas y de servicio se distribuyen exponencial. Tenemos 10 servidores en paralelo. La disciplina de la cola (*DG*) es general, es decir, FCFS, LCFS, SIRO o cualquier otro procedimiento. El número máximo de clientes en el sistema (cola más servicio) es *N*. El tamaño de la fuente de la que provienen los clientes es infinita.

****

El objetivo final de analizar situaciones de espera consiste en generar *medidas de desempeño* para evaluar los sistemas reales. Todo sistema de espera opera como función del tiempo, por ello se distinguen dos estados:

[4] *a,b,c – D. G. Kendall* (1953); *d,e – A. M. Lee* (1966); *f – H. A. Taha* (1968).

- Operación inicial del sistema (*estado transitorio* o de calentamiento, análisis complejo desde el punto de vista matemático).
- Operación del sistema durante un periodo suficientemente grande (*estado estable*).

Las medidas de desempeño (rendimiento) más frecuentemente utilizadas en una situación de espera o colas son:

- L_s es el número esperado de clientes en el sistema.
- L_q es el número esperado de clientes en la cola.
- W_s es el tiempo aproximado de espera en el sistema.
- W_q es el tiempo aproximado de espera en la cola.
- $\bar{c}$ es el número esperado de servidores ocupados.

Llamemos p_n a la probabilidad (de estado estable) de que haya *n* clientes en el sistema. En consecuencia,

$$L_s = \sum_{n=0}^{\infty} n p_n$$

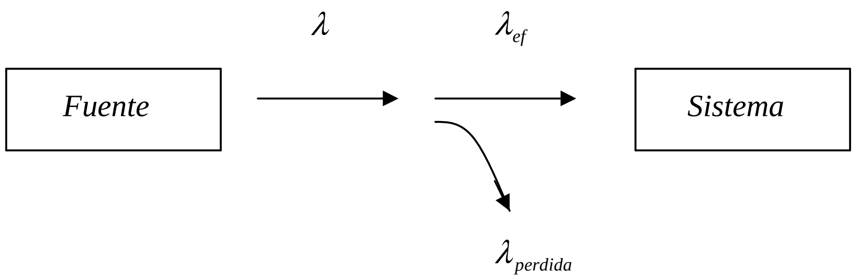
$$L_q = \sum_{n=c}^{\infty} (n-c) p_n$$

Las relaciones entre $L_\bullet$ y $W_\bullet$ son conocidas como fórmulas de *Little*.

$$L_s = \lambda_{ef} W_s$$

$$L_q = \lambda_{ef} W_q$$

λ_{ef} es la tasa de llegadas efectiva al sistema = la tasa (nominal) de llegadas λ cuando todos los clientes que llegan ingresan al sistema. Si algún cliente no puede ingresar al sistema es por que éste está lleno, como por ejemplo, en un estacionamiento, luego $\lambda_{ef} < \lambda$.



$$\lambda = \lambda_{ef} + \lambda_{perdida}$$

$$\lambda_{ef} = \sum_{n=0}^{\infty} \lambda_n p_n$$

λ_n es el número esperado de clientes que llegan al sistema por unidad de tiempo cuando hay *n* clientes en el mismo.

Existe una relación directa entre W_s y W_q. Por definición tenemos lo siguiente:

Tiempo de espera aproximado en el sistema	=	Tiempo de espera aproximado en la cola	+	Tiempo de servicio esperado

$$W_s = W_q + \frac{1}{\mu}$$

$$\mu \text{ = tasa de servicio = } \frac{\#clientes}{t} \Rightarrow t = \frac{\#clientes}{\mu} = \frac{1}{\mu}$$

Al multiplicar ambos lados de la ecuación por λ_{ef} se obtiene,

$$L_s = L_q + \frac{\lambda_{ef}}{\mu}$$

Número promedio de servidores ocupados	=	Número promedio de clientes en el sistema	–	Número promedio de clientes en cola

$$\overline{c} = L_s - L_q = \frac{\lambda_{ef}}{\mu}$$

Porcentaje de utilización de los servidores = $\frac{\overline{c}}{c} * 100$

Modelos de un solo servidor.

Los clientes llegan a una tasa constante de λ clientes por unidad de tiempo. La tasa de servicio es también constante, μ clientes por unidad de tiempo. Se utiliza en la notación el símbolo *DG* (disciplina general de la cola) pues las derivaciones de p_n y todas las medidas de desempeño son totalmente independientes de una disciplina de colas específica.

$\boxed{(M/M/1):(DG/\infty/\infty)}$

$$\left.\begin{aligned}\lambda_n &= \lambda\\ \mu_n &= \mu\end{aligned}\right\}\forall n = 0,1,2,\dots$$

$\lambda_{ef} = \lambda$ y $\lambda_{perdida} = 0$ (todos los clientes pueden unirse al sistema).

Definamos $\rho = \frac{\lambda}{\mu}$, entonces la expresión para p_n en el modelo generalizado[5] se reduce a $p_n = \rho^n p_0 ; n = 0,1,2,\dots$

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 \left(1 + \rho + \rho^2 + \cdots\right) = 1$$

$$\text{Si } \rho < 1 \text{ (serie geométrica) } \Rightarrow p_0 \left(\frac{1}{1-\rho}\right) = 1 \Rightarrow p_0 = 1 - \rho$$

[5] $p_n = \left(\frac{\lambda_{n-1}\lambda_{n-2}\dots\lambda_0}{\mu_n\mu_{n-1}\dots\mu_1}\right)p_0 ; n = 1,2,\dots$ (véase Cap. 17, *Taha H. A.*, "Investigación de Operaciones, una introducción", Prentice-Hall, 1998.)

$$\therefore\ p_n = (1-\rho)\rho^n; n = 1,2,\ldots$$

$\rho = \dfrac{\lambda}{\mu} < 1$ significa que la tasa de llegadas debe ser estrictamente menor que la tasa de servicio para que el sistema alcance condiciones de estado estable. Si $\lambda \geq \mu$, la serie geométrica no convergerá y las probabilidades de estado estable p_n no existirán.

$$L_s = \sum_{n=0}^{\infty} np_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\rho\frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right) = (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right) = \frac{\rho}{1-\rho}$$

Las medidas desempeño restantes se calculan de manera sencilla:

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1-\rho)},\quad W_q = W_s - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)},\quad L_q = \lambda W_q = \frac{\rho^2}{1-\rho},\quad \bar{c} = L_s - L_q = \rho$$

$\boxed{(M/M/1):(DG/N/\infty)}$

Ejemplo de este sistema: Situaciones de manufactura en las que la máquina puede tener un área de espera limitada.

$$\lambda_n = \begin{cases} \lambda, & n = 0,1,2,\ldots,N-1 \\ 0, & n = N, N+1,\ldots \end{cases}$$ (máxima longitud de la cola = N-1)

$$\mu_n = \mu; n = 0,1,2,\ldots$$

Usando $\rho = \dfrac{\lambda}{\mu}$, obtenemos $p_n = \begin{cases} \rho^n p_0, & n \leq N \\ 0, & n > N \end{cases}$

$$p_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{N+1}}, & \rho \neq 1 \\ \dfrac{1}{N+1}, & \rho = 1 \end{cases} \qquad n = 0,1,2,\ldots,N$$

$$L_s = \begin{cases} \dfrac{\rho\left(1-(N+1)\rho^N + N\rho^{N+1}\right)}{(1-\rho)\left(1-\rho^{N+1}\right)}, & \rho \neq 1 \\ \dfrac{N}{2}, & \rho = 1 \end{cases}$$

$$\lambda_{perdida} = \lambda p_N \Rightarrow \lambda_{ef} = \lambda - \lambda_{perdida} = \lambda(1-p_N)$$

$$\mu\left(L_s - L_q\right) = \lambda_{ef} = \lambda(1-p_N)$$

$$L_q = L_s - \frac{\lambda_{ecf}}{\mu} = L_s - \frac{\lambda(1-p_N)}{\mu}$$

$$W_q = \frac{L_q}{\lambda_{ecf}} = \frac{L_q}{\lambda(1-p_N)}, \qquad W_q + \frac{1}{\mu} = W_s = \frac{L_s}{\lambda(1-p_N)}$$

Modelos de servidores múltiples.

Se trata de versiones de múltiples servidores de los modelos de la sección anterior.

$\boxed{(M/M/c):(DG/\infty/\infty)}$

c servidores paralelos, tasa de llegadas λ y de servicio μ.

No hay límite en el número de clientes en el sistema.

$$\lambda_n = \lambda; n \geq 0$$

$$\mu_n = \begin{cases} n\mu, & n \leq c \\ c\mu, & n > c \end{cases}$$

$$p_n = \begin{cases} \left(\dfrac{\rho^n}{n!}\right) p_0, & 0 \leq n \leq c \\ \left(\dfrac{\rho^n}{c!\,c^{n-c}}\right) p_0, & n > c \end{cases}$$

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

$$L_q = \left(\frac{c\rho}{(c-\rho)^2} \right) p_c$$

$$\lambda_{ecf} = \lambda \Rightarrow L_s = L_q + \rho, \qquad W_q = \frac{L_q}{\lambda}, \qquad \frac{L_s}{\lambda} = W_s = W_q + \frac{1}{\mu}$$

$\boxed{(M/M/c):(DG/N/\infty), c \leq N}$

Capacidad del sistema = N. Tamaño de la cola = *N-c*

$$\lambda_n = \begin{cases} \lambda, & 0 \leq n < N \\ 0, & n \geq N \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n < c \\ c\mu, & c \leq n \leq N \end{cases}$$

$$p_n = \begin{cases} \left(\dfrac{\rho^n}{n!}\right) p_0, & 0 \le n \le c \\ \left(\dfrac{\rho^n}{c!c^{n-c}}\right) p_0, & c \le n \le N \end{cases}$$

$$p_0 = \begin{cases} \left(\displaystyle\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c\left(1-\left(\frac{\rho}{c}\right)^{N-c+1}\right)}{c!\left(1-\rho/c\right)} \right)^{-1}, & \dfrac{\rho}{c} \ne 1 \\ \left(\displaystyle\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!}(N-c+1) \right)^{-1}, & \dfrac{\rho}{c} = 1 \end{cases}$$

$$L_q = \begin{cases} \dfrac{\rho^{c+1}}{(c-1)!(c-\rho)^2} \left\{ 1-\left(\dfrac{\rho}{c}\right)^{N-c+1} - (N-c+1)\left(1-\dfrac{\rho}{c}\right)\left(\dfrac{\rho}{c}\right)^{N-c} \right\} p_0, & \dfrac{\rho}{c} \ne 1 \\ \dfrac{\rho^c (N-c)(N-c+1)}{2c!} p_0, & \dfrac{\rho}{c} = 1 \end{cases}$$

$$L_s = L_q + (c - c') = L_q + \frac{\lambda_{ecf}}{\mu}$$

donde $c' =$ número estimado de servidores inactivos $= \sum_{n=0}^{c} (c-n) p_n$

Modelo de autoservicio.

$\boxed{(M/M/\infty):(DG/\infty/\infty)}$

El número de servidores es ilimitado $(c = \infty)$ pues el cliente es un servidor. Las gasolineras de autoservicio y los cajeros automáticos de los bancos no entran dentro de esta descripción, pues los servidores son en realidad las mismas bombas de gasolina y los cajeros automáticos. Ejemplo de este modelo, un restaurante de autoservicio (no intervienen autómatas).

$$\left.\begin{aligned} \lambda_n &= \lambda \\ \mu_n &= n\mu \end{aligned}\right\} \forall n = 0,1,2,\ldots$$

$$p_n = \frac{e^{-\rho}\rho^n}{n!};\, n = 0,1,2,\ldots \text{ Distribución de Poisson}$$

$$L_s = \rho, \qquad W_s = \frac{1}{\mu}, \qquad L_q = 0 = W_q \text{ (cada cliente se atiende a sí mismo)}$$

Modelo de servicio de máquinas.

$\boxed{(M/M/R):(DG/K/K), R<K}$

Se tiene un taller con K máquinas y R mecánicos (servidores). La tasa de averías por máquina es λ averías por unidad de tiempo (se supone que las averías y servicios siguen una distribución de Poisson).

Tener n máquinas en el sistema significa que las n máquinas están averiadas. Luego, la tasa de averías para el taller es proporcional al número de máquinas que funcionan.

$$\lambda_n = \begin{cases} (K-n)\lambda, & 0 \le n < K \\ 0, & n \ge K \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \le n < R \\ R\mu, & R \le n < K \\ 0, & n \ge K \end{cases}$$

$$p_n = \begin{cases} \binom{K}{n} \rho^n p_0, & 0 \le n \le R \\ \binom{K}{n} \dfrac{n!\rho^n}{R!R^{n-R}} p_0, & R \le n \le K \end{cases}$$

$$p_0 = \left\{ \sum_{n=0}^{R} \binom{K}{n} \rho^n + \sum_{n=R+1}^{K} \binom{K}{n} \frac{n!\rho^n}{R!R^{n-R}} \right\}^{-1}$$

$$R' = \text{ número estimado de técnicos no ocupados } = \sum_{n=0}^{R} (R-n) p_n$$

$$\lambda_{ecf} = E\{\lambda(K-n)\} = \lambda(K-L_s)$$