Tema 6

Modelos de colas exponenciales

6.1. La distribución exponencial y los procesos de Poisson

6.1.1. Distribución exponencial

El análisis de los distintos modelos de colas está determinado en gran parte por la distribución de probabilidad de los tiempos entre llegadas y la distribución de los tiempos de servicio. En los sistemas de colas reales, estas distribuciones pueden tomar prácticamente cualquier forma (siempre que sean no negativos). Sin embargo, para formular y analizar un modelo matemático es necesario especificar la forma supuesta de cada una de estas distribuciones. Para que sea útil, la forma expuesta debe ser lo suficientemente realista como para que el modelo proporcione predicciones razonables y al mismo tiempo lo suficientemente manejable para que sea posible tales predicciones. Con estas consideraciones, la distribución exponencial es la distribución de probabilidad más importante en la teoría de colas.

Una v.a. no negativa, X, se dice que sigue una distribución exponencial con parámetro λ, $X \sim \exp(\lambda)$, si

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

donde $\lambda > 0$ es una constante fija.

La función de densidad de X viene dada por:

$$f(x) = \frac{d}{dx} F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ \lambda e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

6.1.1.1. Propiedades de la distribución exponencial

1. $f_X(x)$ es una función estrictamente decreciente. Como consecuencia,

   $$P(0 \leq X \leq \delta x) > P(x \leq X \leq x + \delta x)$$

   Lo que quiere decir que es relativamente probable que X tome valores pequeños. De hecho, se tiene que

   $$P\left(0 \leq X \leq \frac{1}{2\lambda}\right) = 0.393 \text{ mientras que } P\left(\frac{1}{2\lambda} \leq X \leq \frac{3}{2\lambda}\right) = 0.383$$

   de manera que es más probable que X tome un valor cercano a cero que un valor cercano a su media, aún cuando el segundo intervalo tiene el doble de longitud que el primero. ¿Es esta propiedad razonable para un modelo de colas?

   - Si X representa los tiempos de servicio, la respuesta depende de la naturaleza de dicho servicio.
     - Si el servicio requerido es, en esencia, idéntico para cada cliente y el servidor realiza siempre la misma secuencia de operaciones, entonces

los tiempos de servicio reales tienden a aproximarse al tiempo esperado de servicio. Las pequeñas desviaciones de la media que puedan ocurrir se deberían a variaciones en la eficiencia del servidor. En este caso, la distribución exponencial no proporcionaría una buena aproximación a la distribución de los tiempos de servicio.

- Si las tareas que tiene que realizar el servidor difieren de un cliente a otro, entonces es posible que la distribución exponencial pueda constituir un buen ajuste.

- Si X representa los tiempos entre llegadas, la propiedad 1 descarta situaciones en las que los clientes que llegan al sistema tienden a posponer su entrada si ven que otro cliente entra antes que ellos.

2. **Propiedad de pérdida de memoria**. Sea $X \sim \exp(\lambda)$, entonces

$$
\begin{aligned}
P(X > t + s | X > s) &= \frac{P(X > t + s)}{P(X > s)} \\
&= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\
&= e^{-\lambda t} \\
&= P(X > t)
\end{aligned}
$$

Si X representa el tiempo de vida de un organismo, entonces la probabilidad de que un organismo con antigüedad s viva t unidades de tiempo más, coincide con la probabilidad de que un nuevo organismo viva al menos t unidades de tiempo. Es decir, este organismo "no recuerda" que ha vivido ya s unidades de tiempo; la distribución que sigue la cantidad adicional de tiempo que sobrevive el organismo coincide con la distribución original del tiempo de vida.

La distribución exponencial es la única distribución continua que satisface la propiedad de pérdida de memoria.

3. **Mínimo de v.a. exponenciales**. El mínimo de variables aleatorias exponenciales independientes sigue una distribución exponencial. Sean $X_1, \dots, X_n$ v.a. independientes con distribuciones exponenciales de parámetros $\lambda_1, \dots, \lambda_n$, respectivamente. Entonces, la variable aleatoria

$$X = \text{mín}\{X_1, \dots, X_n\}$$

sigue una distribución exponencial de parámetro $\lambda = \lambda_1 + \cdots + \lambda_n$.

6.1.2. Procesos de Poisson

Formalmente, un proceso estocástico se define como una colección de variables aleatorias $\{X(\alpha), \alpha \in T\}$, indexadas por el parámetro α que toma valores en el *conjunto de parámetros* T. El conjunto S en el que toman valores las variables $X(\alpha)$ se denomina **espacio de estados**.

6.1.2.1. El procesos de Poisson como un proceso de conteo

$\{X_n, n \geq 1\}$ secuencia de v.a.'s representando tiempos entre sucesos. Definimos

$$S_0 = 0 \qquad S_n = X_1 + \cdots + X_n$$

S_n es el tiempo en el que ocurre el n-ésimo suceso. Sea

$$N(t) = \text{máx}\{n \geq 0 | S_n \leq t\}, \text{ para todo } t \geq 0$$

$N(t)$ es el número de sucesos ocurridos en el intervalo $(0, t]$. $\{N(t), t \geq 0\}$ se denomina **proceso de conteo**.

Notar que $N(0) = 0$ y $N(t)$ salta con pasos de longitud 1 en los tiempos $t = S_n$, $n = 1, 2, \dots$

Algunas propiedades que satisface un proceso de conteo son las siguientes:

- $N(t) \in \mathbb{Z}_+, \quad \forall t \geq 0$.
- Si $s < t$, entonces $N(s) \leq N(t)$ y $N(t) - N(s)$ es el número de sucesos que han ocurrido en el intervalo $(s, t]$.

Definición 6.1. *Si* $\{X_n, n \geq 1\}$ *es una sucesión de v.a.'s iid* Exp*(*λ*), entonces* $\{N(t), t \geq 0\}$ *se denomina* ***proceso de Poisson*** *de parámetro* λ *y se representa por* $PP(\lambda)$.

Observemos que un proceso de Poisson (como en general, un proceso de conteo) es un proceso estocástico continuo con espacio de estados discreto. Veamos a continuación un par de propiedades sobre el comportamiento transitorio de $\{N(t), t \geq 0\}$.

Proposición 6.2. *Sea* $t \geq 0$ *fijo, entonces*

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

Demostración.
Tenemos que:

$$\begin{aligned} N(t) \geq k &\Leftrightarrow k \text{ o más sucesos ocurren durante } (0, t] \\ &\Leftrightarrow \text{el } k\text{-ésimo suceso tiene lugar en } t \text{ o antes} \\ &\Leftrightarrow S_k \leq t \end{aligned}$$

Luego,

$$P(N(t) \geq k) = P(S_k \leq t),$$

de donde se sigue que

$$\begin{aligned} P(N(t) = k) &= P(N(t) \geq k) - P(N(t) \geq k+1) \\ &= P(S_k \leq t) - P(S_{k+1} \leq t) \end{aligned}$$

Por otro lado, puesto que S_k es suma de v.a.'s exponenciales iid, se tiene que

$$S_k \sim \mathrm{Gamma}(k, \lambda)$$

Finalmente,

$$\begin{aligned} P\big(N(t) = k\big) &= P(S_k \le t) - P(S_{k+1} \le t) \\ &= \Big[1 - \sum_{r=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^r}{r!}\Big] - \Big[1 - \sum_{r=0}^{k} e^{-\lambda t} \frac{(\lambda t)^r}{r!}\Big] \\ &= e^{-\lambda t} \frac{(\lambda t)^k}{k!} \end{aligned}$$

■

La anterior proposición proporciona un motivo para denominar a $\{N(t), t \ge 0\}$ un proceso de Poisson: para cada t fijo, $N(t)$ es una v.a. de Poisson. En consecuencia, para t fijo,

$$E\left(N(t)\right) = \lambda t \qquad \mathrm{Var}\left(N(t)\right) = \lambda t.$$

6.1.3. El procesos de Poisson como un proceso de incrementos estacionarios e independientes

Sea $\{X(t), t \ge 0\}$ un proceso estocástico. Para $s \ge 0$ fijo y $t \ge 0$, $X(t+s) - X(s)$ se denomina *incremento* sobre el intervalo $(s, s+t]$.

Definición 6.3. *Un proceso estocástico $\{X(t), t \ge 0\}$ se dice que tiene **incrementos estacionarios e independientes** si:*

*i) la distribución de $X(t+s) - X(s)$ es independiente de s (**incrementos estacionarios**)*

*ii) los incrementos sobre intervalos no solapados son independientes (**incrementos independientes**)*

Particularizada a procesos de conteo, la propiedad de incrementos independientes significa que el número de sucesos producidos en intervalos de tiempo disjuntos son independientes. Por ejemplo, el número de sucesos que han ocurrido hasta el tiempo 10, $N(10)$, debe ser independiente del número de sucesos que ocurran en el intervalo $(10, 15]$, $N(15) - N(10)$.

En lo que respecta a la propiedad de incrementos estacionarios, ésta significa que la distribución del número de sucesos que ocurren en cualquier intervalo de tiempo sólo depende de la longitud de dicho intervalo. En otras palabras, el número de sucesos ocurridos en el intervalo $(t_1+s, t_2+s]$, $N(t_2+s)-N(t_1+s)$, siguen la misma distribución que el número de sucesos ocurridos en el intervalo $(t_1, t_2]$.

Proposición 6.4. *Un proceso de Poisson tiene incrementos estacionarios e independientes.*

Teorema 6.5. *Un proceso estocástico* $\{N(t), t \geq 0\}$ *es un* $PP(\lambda) \Leftrightarrow$

i) tiene incrementos estacionarios e independientes

ii) $P(N(t) = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$*, para* $t \geq 0$*,* $k = 0, 1, 2, \ldots$

6.1.3.1. Otra caracterización alternativa de un proceso de Poisson

Para obtener esta última caracterización necesitamos previamente la siguiente definición:

Definición 6.6. *Una función* $f : \mathbb{R} \to \mathbb{R}$ *se dice que es* $o(h)$ *si*

$$\lim_{h\to 0} \frac{f(h)}{h} = 0$$

Ejemplo 6.7.

- $f(x) = 2x$ no es $o(h)$, $f(x) = x^2 + 3x^3$ es $o(h)$.

- $f(x) = e^{\lambda x} - 1 - \lambda x$ es $o(h)$.
- Si $f(x)$ y $g(x)$ son $o(h)$, entonces $f(x) + g(x)$ es $o(h)$.
- Si $f(x)$ es $o(h)$ y $c \in \mathbb{R}$, entonces $c \cdot f(x)$ es $o(h)$.

$\triangle$

Teorema 6.8. *Un proceso de conteo $\{N(t), t \geq 0\}$ es un $PP(\lambda) \Leftrightarrow$*

i) tiene incrementos estacionarios e independientes

ii) $N(0) = 0$ y

- $P(N(h) = 0) = 1 - \lambda h + o(h)$
- $P(N(h) = 1) = \lambda h + o(h)$
- $P(N(h) = j) = o(h)$, *para todo* $j \geq 2$

6.2. Algunas consideraciones previas al análisis de los modelos de colas

6.2.1. Notación y terminología

La notación y terminología estándar que utilizaremos a lo largo de los siguientes apartados es la siguiente:

- Estado del sistema = número de clientes en el sistema
- Longitud de la cola = número de clientes en cola = Estado del sistema - número de clientes en servicio

- $n(t)$: número de clientes en el sistema en el instante t
- $p_n(t)$: probabilidad de que exactamente n clientes estén en el sistema en el instante t.
- s: número de servidores en el sistema
- λ_n: tasa media de llegadas (número esperado de llegadas por unidad de tiempo) de nuevos clientes cuando hay n clientes en el sistema.
- μ_n: tasa media de servicio para todo el sistema (número esperado de clientes que completan su servicio por unidad de tiempo) cuando hay n clientes en el sistema.

Cuando λ_n es constante para todo n, se denota por λ. Esto significaría que el número medio de clientes que llega al sistema por unidad de tiempo no depende del estado del sistema. Cuando la tasa media de servicio por servidor ocupado es constante, se denota por μ. En este caso, $\mu_n = s\mu$, cuando $n \geq s$, es decir, cuando los s servidores están ocupados. En estas circunstancias, $\frac{1}{\lambda}$ es el tiempo esperado entre llegadas, $\frac{1}{\mu}$ es el tiempo de servicio esperado y $\rho = \frac{\lambda}{s\mu}$ es el factor de utilización del sistema, es decir, la fracción media de tiempo que los servidores están ocupados.

6.2.2. Estado transitorio vs estado estacionario

Cuando un sistema de colas inicia su operación, los distintos factores del sistema se encuentran bastante influenciados por las condiciones iniciales; se dice que el sistema se encuentra en **estado transitorio**. Una vez que ha pasado suficiente tiempo, usualmente, los factores del sistema se vuelven independientes de las condiciones iniciales y del tiempo transcurrido, y se dice que el sistema se encuentra en **estado estable**. A lo largo de este tema, dedicaremos nuestro análisis al estado estacionario que, afortunadamente, en la práctica es el que guarda mayor interés.

Nuestro objetivo inicial será obtener las probabilidades de estado, $p_n(t)$: probabilidad de que el sistema esté en el estado n (haya n clientes en el sistema) en el instante t. En general, obtener las probabilidades de estado cuando el sistema se encuentra en estado transitorio es bastante complicado, así que nos centraremos en calcular las probabilidades $p_n = \lim_{t\to\infty} p_n(t)$ cuando el sistema se encuentra en estado estable. De algún modo esto significa que esperamos que el sistema se vuelva independiente de las condiciones iniciales y del tiempo que ha transcurrido desde el inicio del mismo. Por lo tanto, la probabilidades estacionaria p_n se puede interpretar como la probabilidad de que haya n clientes en el sistema cuando éste ha alcanzado el estado estacionario. Hay que puntualizar que no todos los sistemas tienen estado estacionario, pues $\lim_{t\to\infty} p_n(t)$ podría no dar lugar a una distribución de probabilidad.

La siguiente notación asume que el sistema se encuentra en estado estacionario:

- p_n : probabilidad de que haya exactamente n clientes en el sistema
- L : número esperado de clientes en el sistema. $L = \sum_{n\geq 0} nP_n$
- L_q : longitud esperada de la cola.
- W : tiempo medio de espera en el sistema para cada cliente
- W_q : tiempo medio de espera en cola

6.3. Modelos de colas exponenciales con un único servidor

El modelo más sencillo de analizar analíticamente es aquel en el que tanto los tiempos de llegada de los clientes como los tiempos de servicio son exponenciales e independientes entre sí, el sistema tiene un único servidor y la disciplina de servicio es FIFO.

6.3.1. Modelo $M/M/1$

En este caso asumiremos que la capacidad del sistema es ilimitada. Empezaremos obteniendo las probabilidades de estado $p_n(t)$. El procedimiento que seguiremos para ello se basa en tres pasos:

1. Obtener las ecuaciones en diferencia para $p_n(t)$
2. Obtener las ecuaciones diferenciales en diferencia para $p_n(t)$
3. Obtener las probabilidades límite p_n para el comportamiento estacionario.

Ecuaciones en diferencia para $p_n(t)$

Para obtener las ecuaciones en diferencia para $p_n(t)$ analizaremos cómo el sistema puede alcanzar el estado n en el instante $t + \Delta t$. Las posibilidades son las siguientes

- Si en el instante t el sistema estaba en el estado n, entonces en $(t, t + \Delta t]$ se produjeron $j \geq 0$ llegadas y j servicios.
- Si en el instante t el sistema estaba en el estado $n + j$, entonces en $(t, t + \Delta t]$ se registraron k llegadas y $j + k$ servicios.
- Si en el instante t el sistema estaba en el estado $n - j$, entonces en $(t, t + \Delta t]$ se produjeron $j + k$ llegadas y k servicios.

Puesto que los tiempos de llegadas y servicio son exponenciales, sabemos que los procesos que rigen el número de llegadas y el número de servicios que se producen hasta un cierto instante son Procesos de Poisson, y por lo tanto la probabilidad de que se den 2 o más llegadas o servicios en un intervalo de longitud Δt es $o(\Delta t)$. Por lo tanto, basta considerar los casos en los que a lo sumo se den una llegada y un servicio. Por lo tanto, para cada $n \geq 1$, las posibilidades son las siguientes:

- En el instante t el sistema estaba en el estado n y
  - en $(t, t + \Delta t]$ no se produjeron ni llegadas ni servicios.
  - en $(t, t + \Delta t]$ se produjeron 1 llegada y 1 servicio.
- En el instante t el sistema estaba en el estado $n+1$ y en $(t, t+\Delta t]$ no se registraron llegadas y sólo finalizó un servicio
- En el instante t el sistema estaba en el estado $n - 1$ y en $(t, t + \Delta t]$ se regitró 1 llegada y no finalizó ningún servicio.

Por lo tanto,

$$
\begin{aligned}
p_n(t + \Delta t) = {} & P(n \text{ clientes en } t, \text{ no llegadas ni servicios en } (t, t + \Delta t])+ \\
& + P(n \text{ clientes en } t, \text{ una llegada } + \text{ un servicio en } (t, t + \Delta t])+ \\
& + P(n + 1 \text{ clientes en } t, \text{ no llegadas } + \text{ un servicio en } (t, t + \Delta t])+ \\
& + P(n - 1 \text{ clientes en } t, \text{ 1 llegada } + \text{ no servicios en } (t, t + \Delta t]) + o(\Delta t).
\end{aligned}
$$

Puesto que los tiempos de llegada y servicio son independientes entre sí y a su vez del estado del sistema en t (propiedad de incrementos independientes), se tiene que para cada $n \geq 1$,

$$\begin{aligned} p_n(t+\Delta t) = {} & p_n(t)P(\text{no llegadas en } (t,t+\Delta t])P(\text{no servicios en } (t,t+\Delta t])+ \\ & + p_n(t)P(\text{1 llegada en } (t,t+\Delta t])P(\text{1 servicio en } (t,t+\Delta t])+ \\ & + p_{n+1}(t)P(\text{no llegadas en } (t,t+\Delta t])P(\text{1 servicio en } (t,t+\Delta t])+ \\ & + p_{n-1}(t)P(\text{1 llegada en } (t,t+\Delta t])P(\text{no servicios en } (t,t+\Delta t]) + o(\Delta t). \end{aligned}$$

Supongamos que los tiempos entre llegadas son exponenciales de parámetro λ y los tiempos de servicio son exponenciales de parámetro μ. Utilizando las propiedades de los correspondientes Procesos de Poisson, podemos escribir:

$$\begin{aligned} p_n(t+\Delta t) = {} & p_n(t)\left(1-\lambda\Delta t + o(\Delta t)\right)\left(1-\mu\Delta t + o(\Delta t)\right) + \\ & + p_n(t)\left(\lambda\Delta t + o(\Delta t)\right)\left(\mu\Delta t + o(\Delta t)\right) + \\ & + p_{n+1}(t)\left(1-\lambda\Delta t + o(\Delta t)\right)\left(\mu\Delta t + o(\Delta t)\right) + \\ & + p_{n-1}(t)\left(\lambda\Delta t + o(\Delta t)\right)\left(1-\mu\Delta t + o(\Delta t)\right) + o(\Delta t). \end{aligned}$$

Uniendo todos los términos $o(\Delta t)$ y teniendo en cuenta que los términos $(\Delta t)^2$ son también $o(\Delta t)$, se tiene:

$$p_n(t+\Delta t) = p_n(t)\left(1-\lambda\Delta t - \mu\Delta t\right) + p_{n+1}(t)\left(\mu\Delta t\right) + p_{n-1}(t)\left(\lambda\Delta t\right) + o(\Delta t). \tag{6.1}$$

Para el caso $n = 0$ el razonamientos es similar teniendo en cuenta simplemente que el caso relativo a $p_{n-1}(t)$ no se puede dar, obteniéndose en este caso la siguiente ecuación:

$$p_0(t+\Delta t) = p_0(t)\left(1-\lambda\Delta t\right) + p_1(t)\left(\mu\Delta t\right) + o(\Delta t). \tag{6.2}$$

Las ecuaciones 6.1 y 6.2 constituyen el sistema de ecuaciones en diferencia para el caso $M/M/1$. Obsérvese que estas ecuaciones en diferencia lo son tanto respecto a t como a n.

Ecuaciones diferenciales en diferencia

El sistema de ecuaciones formado por 6.1 y 6.2 se puede reescribir del siguiente modo:

$$\begin{cases} p_n(t+\Delta t) - p_n(t) = -(\lambda+\mu)p_n(t)\Delta t + \mu p_{n+1}(t)\Delta t + \lambda p_{n-1}(t)\Delta t + o(\Delta t) & n \geq 1 \\ p_0(t+\Delta t) - p_0(t) = -\lambda p_0(t)\Delta t + \mu p_1(t)\Delta t + o(\Delta t). \end{cases}$$

Si dividimos por Δt y tomamos límites cuando $\Delta t \to 0$, tenemos que:

$$\begin{cases} \frac{dp_n(t)}{dt} = -(\lambda+\mu)p_n(t) + \mu p_{n+1}(t) + \lambda p_{n-1}(t), \quad n \geq 1 \\ \frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t) \end{cases} \tag{6.3}$$

Obtención de las probabilidades estacionarias p_n

Supongamos que el sistema alcanza un estado estacionario. Esto significa que cuando $t \to \infty$ la probabilidad $p_n(t)$ se vuelve independiente del tiempo. Por lo tanto, $\frac{dp_n(t)}{dt}$ tendería a cero, y el sistema de ecuaciones 6.3 quedaría como el sistema de ecuaciones en diferencia:

$$\begin{cases} 0 = -(\lambda+\mu)p_n + \mu p_{n+1} + \lambda p_{n-1}, \quad n \geq 1 \\ 0 = -\lambda p_0 + \mu p_1 \end{cases}$$

o escrito de otro modo,

$$\begin{cases} p_{n+1} = \frac{\lambda+\mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1}, \quad n \geq 1 \\ p_1 = \frac{\lambda}{\mu} p_0 \end{cases} \tag{6.4}$$

Este sistema de ecuaciones en diferencia en una variable (n), podemos resolverlo utilizando un procedimiento iterativo. Obsérvese que:

$$\begin{aligned}
p_2 &= \frac{\lambda+\mu}{\mu}p_1 - \frac{\lambda}{\mu}p_0 \\
&= \frac{\lambda+\mu}{\mu}\left(\frac{\lambda}{\mu}p_0\right) - \frac{\lambda}{\mu}p_0 \\
&= \left(\frac{\lambda}{\mu}\right)^2 p_0
\end{aligned}$$

Asimismo,

$$\begin{aligned}
p_3 &= \frac{\lambda+\mu}{\mu}p_2 - \frac{\lambda}{\mu}p_1 \\
&= \frac{\lambda+\mu}{\mu}\left(\frac{\lambda}{\mu}\right)^2 p_0 - \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}p_0\right) \\
&= \left(\frac{\lambda}{\mu}\right)^3 p_0
\end{aligned}$$

Por lo tanto, parece razonable conjeturar que

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \tag{6.5}$$

Demostraremos la relación anterior por inducción matemática. Supongamos 6.5 es cierto para $n = 1, \ldots, k$. Veamos que también se verifica para $k+1$. Del sistema de ecuaciones 6.4 se tiene que

$$p_{k+1} = \frac{\lambda+\mu}{\mu}p_k - \frac{\lambda}{\mu}p_{k-1}$$

Ahora, utilizando la hipótesis de inducción, puesto que 6.5 es cierta para k y $k-1$, se tiene que

$$\begin{aligned}
p_{k+1} &= \frac{\lambda+\mu}{\mu}\left(\frac{\lambda}{\mu}\right)^k p_0 - \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu}\right)^{k-1} p_0 \\
&= \left[\frac{\lambda^{k+1}+\mu\lambda^k}{\mu^{k+1}} - \left(\frac{\lambda}{\mu}\right)^k\right] p_0 \\
&= \left(\frac{\lambda^{k+1}+\mu\lambda^k-\mu\lambda^k}{\mu^{k+1}}\right) p_0 \\
&= \left(\frac{\lambda}{\mu}\right)^{k+1} p_0
\end{aligned}$$

Para completar el análisis falta obtener el valor de p_0. Para ello utilizaremos que $\{p_n\}_{n\geq 0}$ debe ser una función puntual de probabilidad, y por lo tanto,

$$\sum_{n\geq 0} p_n = 1$$

Utilizando 6.5 se tiene que

$$\sum_{n\geq 0} \left(\frac{\lambda}{\mu}\right)^n p_0 = 1,$$

de donde se sigue que

$$\frac{1}{p_0} = \sum_{n\geq 0} \left(\frac{\lambda}{\mu}\right)^n.$$

La expresión $\sum_{n\geq 0} \left(\frac{\lambda}{\mu}\right)^n$ corresponde a la de una serie geométrica que converge si $\left|\frac{\lambda}{\mu}\right| = \frac{\lambda}{\mu} < 1$, o lo que es lo mismo que λ es menor que μ. Esta condición tiene sentido en nuestro contexto, pues si $\lambda > \mu$ significa que el tiempo medio de llegada es mayor que el tiempo medio de servicio, y por lo tanto el estado del sistema crecera indefinidamente. No es tan fácil de interpretar por qué no existe el estado estacionario si $\lambda = \mu$ pero un posible explicación sería que en este caso conforme la cola crece se le hace más difícil al servidor disminuir el tamaño de la cola, puesto que la tasa de servicio no es mayor que la de llegadas.

En definitiva, si $\rho = \frac{\lambda}{\mu} < 1$ se tiene que:

$$\frac{1}{p_0} = \sum_{n\geq 0} \left(\frac{\lambda}{\mu}\right)^n = \sum_{n\geq 0} \rho^n = \frac{1}{1-\rho},$$

de donde se sigue que

$$p_0 = 1 - \rho$$

y por lo tanto, la solución general para el estado estacionario sería

$$p_n = \rho^n(1-\rho).$$

6.3.1.1. Medidas de eficiencia

A través de la distribución de probabilidad del estado del sistema para el caso estacionario podemos obtener algunas medidas para calibrar la eficiencia del sistema, como son el número medio de clientes en el sistema y el número medio de clientes en cola (siempre suponiendo que el sistema se encuentra en estado estacionario).

Sea $\mathcal{L}$ la variable aleatoria "número de clientes en el sistema" y $L = E(\mathcal{L})$.

$$L = \sum_{n\geq 0} np_n = \sum_{n\geq 0} n(1-\rho)\rho^n = \sum_{n\geq 1} n(1-\rho)\rho^n = \rho \sum_{n\geq 1} n(1-\rho)\rho^{n-1} = \rho\frac{1}{(1-\rho)},$$

o equivalentemente,

$$L = \frac{\lambda}{\mu - \lambda}$$

Por otro lado, sea Sea $\mathcal{L}_q$ la variable aleatoria "número de clientes en la cola" y $L_q = E(\mathcal{L})_q$. Obsérvese que

$$\mathcal{L}_q = \begin{cases} 0, & \text{si } \mathcal{L} = 0 \\ \mathcal{L} - 1, & \text{si } \mathcal{L} \geq 1 \end{cases}$$

Por lo tanto, tenemos que:

$$L_q = 0p_o + \sum_{n\geq 1}(n-1)p_n = \sum_{n\geq 1} np_n - \sum_{n\geq 1} p_n = L - (1-p_0) = \frac{\rho}{1-\rho} - \rho$$

o equivalentemente,

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

Otra media de interés es el tamaño esperado de la cola cuando ésta no está vacía. Denotemos esta medida por $\overline{L_q}$

$$\overline{L_q} = E\left(\mathcal{L}_q | \mathcal{L}_q \geq 1\right) = \sum_{n\geq 1}(n-1)\overline{p_n},$$

donde $\overline{p_n}$ es la probabilidad de que en el sistema haya n clientes si hay la cola no está vacía:

$$
\begin{aligned}
\overline{p_n} &= P\left(\mathcal{L} = n | \mathcal{L}_q \geq 1\right) = P\left(\mathcal{L} = n | \mathcal{L} \geq 2\right) \\
&= \frac{P\left(\mathcal{L} = n, \mathcal{L} \geq 2\right)}{P\left(\mathcal{L} \geq 2\right)} \\
&= \frac{p_n}{\sum_{n\geq 2} p_n} \qquad (n \geq 2) \\
&= \frac{p_n}{1 - p_0 - p_1} \\
&= \frac{p_n}{1 - (1-\rho) - (1-\rho)\rho} \\
&= \frac{p_n}{\rho^2}
\end{aligned}
$$

Luego

$$
\overline{p_n} = \begin{cases} 0 & \text{si } n = 0, 1 \\ \frac{p_n}{\rho^2} & \text{si } n \geq 2 \end{cases}
$$

Por lo tanto,

$$
\begin{aligned}
\overline{L_q} &= \sum_{n\geq 1} (n-1)\overline{p_n} \\
&= \sum_{n\geq 2} (n-1)\frac{p_n}{\rho^2} \\
&= \frac{1}{\rho^2}\left(\sum_{n\geq 2} np_n - \sum_{n\geq 2} p_n\right) \\
&= \frac{1}{\rho^2}\left((L - p_1) - (1 - p_0 - p_1)\right) \\
&= \frac{1}{\rho^2}\left(L + p_0 - 1\right) \\
&= \frac{1}{\rho^2}\left(\frac{\rho}{1-\rho} + (1-\rho) - 1\right) \\
&= \frac{\rho - (1-\rho)\rho}{(1-\rho)\rho^2} \\
&= \frac{1}{1-\rho}
\end{aligned}
$$

o, equivalentemente,

$$\overline{L_q} = \frac{\mu}{\mu - \lambda}.$$

6.3.1.2. Distribución de los tiempos de espera

En este apartado obtendremos información acerca del tiempo promedio que debe esperar un cliente en el sistema y en cola para ser servido. Hasta ahora la disciplina de servicio no ha tenido efecto en los resultados obtenidos para el modelo. Sin embargo, esta característica es fundamental en el estudio de los tiempos de espera. En el modelo $M/M/1$ se asume que la disciplina de servicio es FIFO.

Denotemos por $\mathcal{W}_q$ a la variable aleatoria que representa el tiempo que pasa un cliente en cola, y sea W_q su esperanza. La variable $\mathcal{W}_q$ es mixta, pues tiene probabilidad no nula en el valor 0 y sigue un comportamiento continuo en el resto.

Calculamos el tiempo medio que espera un individuo $W_q = E(\mathcal{W}_q)$ condicionando por $\mathcal{L}$, el número de clientes en el sistema cuando llega al mismo.

$$\begin{aligned} W_q &= E(\mathcal{W}_q) = E(E(\mathcal{W}_q|\mathcal{L})) && (6.6) \\ &= \sum_{n=0}^{\infty} E(\mathcal{W}_q|\mathcal{L} = n)\, P(\mathcal{L} = n) && (6.7) \end{aligned}$$

Obtengamos el valor de $E(\mathcal{W}_q|\mathcal{L} = n)$. Si no hay clientes en el sistema $(n = 0)$, claramente el tiempo de espera en cola es 0. Si hay $n \geq 1$ clientes en el sistema, entonces el nuevo cliente tiene que esperar a que se completen los n servicios que tiene delante; el de los $n - 1$ clientes que hay en cola delante de él más el del cliente que se está siriviendo. Los $n - 1$ clientes que están en cola tardarán cada uno un tiempo exponencial de parámetro μ, para el cliente que esta siendo servido, como la propiedad exponencial tiene la propiedad de pérdida de memoria, el tiempo que le queda una vez que se ha producido la llegada del nuevo cliente sigue siendo exponencial de parámetro μ. Por lo tanto, cuando el nuevo cliente llega tiene delante n clientes con tiempos exponenciales de parámetro μ, por lo tanto

$$E\left(\mathcal{W}_q | \mathcal{L} = n\right) = \frac{n}{\mu}.$$

Volviendo de nuevo a la expresión 6.6,

$$\begin{aligned} W_q &= \sum_{n=0}^{\infty} E\left(\mathcal{W}_q | \mathcal{L} = n\right) P\left(\mathcal{L} = n\right) \\ &= \sum_{n=1}^{\infty} \frac{n}{\mu} p_n \\ &= \sum_{n=1}^{\infty} \frac{n}{\mu} \rho^n (1-\rho) \\ &= \frac{\rho}{\mu} \sum_{n=1}^{\infty} \rho^{n-1} (1-\rho) \\ &= \frac{1}{\mu} \frac{\rho}{1-\rho} \\ &= \frac{\lambda}{\mu(\mu-\lambda)} \end{aligned}$$

Aunque usualmente la característica de interés es la esperanza del tiempo de espera en cola, para el cálculo de otras medidas se puede comprobar que la función de distribución de la variable $\mathcal{W}_q$ viene dada por:

$$F_{\mathcal{W}_q}(t) = \begin{cases} 1-\rho, & \text{si } t = 0 \\ 1-\rho e^{-\mu(1-\rho)t}, & \text{si } t > 0. \end{cases}$$

Consideremos, a continuación, $\mathcal{W}$ la variable aleatoria continua que representa el tiempo que pasa un cliente en el sistema, y sea W su esperanza.

Al igual que en el caso anterior, calculemos W_q condicionando por $\mathcal{L}$.

$$\begin{aligned} W &= E\left(\mathcal{W}\right) = E\left(E\left(\mathcal{W} | \mathcal{L}\right)\right) && (6.8) \\ &= \sum_{n=0}^{\infty} E\left(\mathcal{W} | \mathcal{L} = n\right) P\left(\mathcal{L} = n\right) && (6.9) \end{aligned}$$

Con un razonamiento análogo al caso anterior, se demuestra que

$$E\left(\mathcal{W}|\mathcal{L}=n\right)=\frac{n+1}{\mu}.$$

Volviendo de nuevo a la expresión 6.8,

$$\begin{aligned}
W &= \sum_{n=0}^{\infty} E\left(\mathcal{W}_q|\mathcal{L}=n\right)P\left(\mathcal{L}=n\right) \\
&= \sum_{n=0}^{\infty} \frac{n+1}{\mu} p_n \\
&= \frac{1}{\mu}\left(\sum_{n=0}^{\infty} n\rho^n(1-\rho)+\sum_{n=0}^{\infty}\rho^n(1-\rho)\right) \\
&= \frac{1}{\mu}\left(\frac{\rho}{1-\rho}+1\right) \\
&= \frac{1}{\mu}\frac{1}{1-\rho} \\
&= \frac{1}{\mu-\lambda}
\end{aligned}$$

Aunque usualmente la característica de interés es la esperanza del tiempo de espera en el sistema, para el cálculo de otras medidas se puede comprobar que la función de densidad de la variable $\mathcal{W}$ viene dada por:

$$F_{\mathcal{W}}(t)=(\mu-\lambda)e^{-(\mu-\lambda)t}, \quad t>0.$$

6.3.2. Fórmulas de Little

Analizando los resultados obtenidos en los apartados anteriores se puede observar que se pueden encontrar relaciones entre las distintas medidas de eficiencia. Estas relaciones se conocen con el nombre de fórmulas de Little y en algunos casos se verifican de modo general y en otro para modelos más restrictivos.

- $W=W_q+\frac{1}{\mu}$. Esta relación es bastante intuitiva y se fundamenta en el siguiente razonamiento. Si Y es la variable aleatoria que representa el tiempo de servicio, entonces es claro que $\mathcal{W} = \mathcal{W}_q + Y$. Puesto que la esperanza de la suma es la suma de las esperanzas, se tiene que $E(\mathcal{W}) = E(\mathcal{W}_q) + E(Y)$, obteniéndose el resultado. Esta relación, por lo tanto, no sólo se verifica en el modelo $M/M/1$ sino en cualquier modelo general de colas.

- $L_q = \lambda W_q$. Supongamos que un cliente llega al sistema. En promedio entrará al servicio después de un tiempo W_q. Supongamos que justo cuando va a entrar al servicio se da la vuelta y cuenta los clientes que están en cola detrás de él; en promedio ese número sería L_q. Puesto que en promedio cada uno de los L_q que están en la cola han tardado en llegar $\frac{1}{\lambda}$ respecto del anterior, el tiempo que ha estado esperando nuestro cliente en cola ha sido $L_q\left(\frac{1}{\lambda}\right)$.

- $L = \lambda W$. Esta expresión se conoce comúnmente como la fórmula de Little, pues se debe a un trabajo de Little de 1961. Se puede demostrar que está condición y la anterior se siguen verificando para un modelo de colas de un único canal con llegadas exponenciales y disciplina FIFO, sin importar la distribución del tiempo de servicio. Incluso en un artículo de Jewel de 1967 se presenta un listado de condiciones menos restrictivas.

- $L = L_q + \frac{\lambda}{\mu}$. Es consecuencia inmediata de las anteriores. Se verifica en aquellos modelos en los que la fórmula de Little también lo haga.

- $W_q = \frac{1}{\mu}L$. Justo cuando un cliente llega al sistema espera encontrarse L clientes delante de él. Para empezar su servicio tendrá que esperar a que finalice el servicio de los L anteriores. Puesto que el tiempo de servicio promedio es $\frac{1}{\mu}$, el tiempo medio que espera en cola es $\frac{1}{\mu}L$ (observar que se está utilizando la propiedad de pérdida de memoria de la exponencial para asegurar que el cliente que está siendo servido cuando nuestro cliente se incorpora a la cola también tardará en finalizar su servicio un tiempo exponencial μ). Las condiciones para que se verifique esta relación son mucho más restrictivas que para la fórmula de Little, y se cumplen sólo en modelos del tipo $M/M/1$ u otros muy similares.

Proposición 6.9. *Si en un modelo de colas se verifica la fórmula de Little y que* $W_q = \frac{1}{\mu}L$*, entonces* $L = \frac{\rho}{1-\rho}$.

Demostración. Por hipótesis tenemos que $W_q = \frac{1}{\mu}L$. Puesto que siempre se cumple que $W = W_q + \frac{1}{\mu}$, se tiene que:

$$\frac{L}{\mu} = W - \frac{1}{\mu}$$

Como, por hipótesis, se verifica la fórmula de Little, $W = \frac{L}{\lambda}$, tenemos que

$$\frac{L}{\mu} = \frac{L}{\lambda} - \frac{1}{\mu},$$

de donde se sigue que

$$L\left(\frac{1}{\lambda} - \frac{1}{\mu}\right) = \frac{1}{\mu},$$

y finalmente,

$$L = \frac{\lambda}{\mu - \lambda} = \frac{\rho}{1-\rho}.$$

□

6.3.3. Modelo $M/M/1/K$

Analizaremos a continuación una modificación del modelo $M/M/1$ que se basa en suponer que la capacidad del sistema está limitada a K clientes.

Las ecuaciones de estado del sistema, $p_n(t)$, dadas para el modelo $M/M/1$, siguen siendo válidas si $n < K$. Analizaremos el caso $n = K$. Para que el sistema esté en el estado K en el instante $t + \Delta t$ pueden darse tres situaciones:

- Que el sistema esté en estado K en el instante t, y en el intervalo $(t, t + \Delta t]$ no se produzcan llegadas ni servicios. La probabilidad asociada a esta situación es

$$p_K(t)\left(1 - \lambda\Delta t + o(\Delta t)\right)\left(1 - \mu\Delta t + o(\Delta t)\right).$$

- Que el sistema esté en estado K en el instante t, y en el intervalo $(t, t + \Delta t]$ se produzca una llegada y ningún servicio. Como la capacidad máxima del sistema es K, el cliente que llega es rechazado y el sistema permanecería en estado K. La probabilidad asociada a esta situación es

$$p_K(t)\,(\lambda\Delta t + o(\Delta t))\,(1 - \mu\Delta t + o(\Delta t))\,.$$

- Que el sistema esté en estado $k-1$ en el instante t, y en el intervalo $(t, t+\Delta t]$ se produzca una llegada y ningún servicio. La probabilidad asociada a esta situación es

$$p_{K-1}(t)\,(\lambda\Delta t + o(\Delta t))\,(1 - \mu\Delta t + o(\Delta t))\,.$$

Por lo tanto,

$$\begin{aligned}
p_K(t + \Delta t) &= p_K(t)\,(1 - \lambda\Delta t + o(\Delta t))\,(1 - \mu\Delta t + o(\Delta t)) \\
&\quad + p_K(t)\,(\lambda\Delta t + o(\Delta t))\,(1 - \mu\Delta t + o(\Delta t)) \\
&\quad + p_{K-1}(t)\,(\lambda\Delta t + o(\Delta t))\,(1 - \mu\Delta t + o(\Delta t)) \\
&= p_k(t)\,(1 - \mu\Delta t) + p_{K-1}(t)\,(\lambda\Delta t))\,(1 - \mu\Delta t) + o(\Delta t)
\end{aligned}$$

de donde se obtiene la ecuación diferencial

$$\frac{dp_K(t)}{dt} = -\mu p_K(t) + \lambda p_{K-1}(t)$$

y tomando límites cuando $t \to \infty$, se obtiene para el caso estacionario

$$p_K = \frac{\lambda}{\mu} p_{K-1}.$$

Por lo tanto, el sistema de ecuaciones en diferencia para las probabilidades de estado en situación estacionaria del modelo $M/M/1/K$ es:

$$\begin{cases} p_1 = \frac{\lambda}{\mu} p_0 \\ p_{n+1} = \frac{\lambda+\mu}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1}, \quad 1 \leq n \leq K-1 \\ p_K = \frac{\lambda}{\mu} p_{K-1} \end{cases}$$

Por el razonamiento iterativo utilizado para el análisis del modelo $M/M/1$, sabemos que

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0, \qquad 0 \leq n \leq K-1.$$

De la expresión de p_K se deduce que la anterior relación también se verifica para $n = K$. Por lo tanto,

$$p_n = (\rho)^n p_0, \qquad 0 \leq n \leq K,$$

siendo $\rho = \frac{\lambda}{\mu}$. El valor de p_0 lo podemos obtener de la condición $\sum_{n=0}^{K} \rho^n p_0 = 1$, de donde se sigue que

$$p_0 = \frac{1}{\sum_{n=0}^{K} \rho^n}.$$

El denominador de la anterior expresión corresponde a la de una serie geométrica finita cuyo valor es

$$\sum_{n=0}^{K} \rho^n = \begin{cases} \frac{1-\rho^{K+1}}{1-\rho}, & \rho \neq 1 \\ K+1, & \rho = 1 \end{cases}$$

Por lo tanto,

$$p_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases}$$

de donde se sigue que la expresión final de las probabilidades de estado es:

$$p_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases} \tag{6.10}$$

Observese que en este caso la solución para el estado estacionario existe incluso si $\rho \geq 1$. Intuitivamente esto se debe a que la limitación en la capacidad del sistema provoca que éste no se desborde. Tambien se observa que si $K \to \infty$ y $\rho < 1$, entonces $p_n \to (1-\rho)\rho^n$, lo cual es consistente con los resultados obtenidos en el modelo $M/M/1$.

Medidas de Eficiencia

En primer lugar obtendremos el **número medio de clientes en el sistema**, L, mediante la expresión $L = \sum_{n=0}^{K} np_n$. Obsérvese que si $\rho = 1$, entonces de 6.10 se sigue que

$$L = \sum_{n=0}^{K} np_n = \sum_{n=0}^{K} n\frac{1}{K+1} = \frac{1}{K+1}\left(\sum_{n=0}^{K} n\right) = \frac{1}{K+1}\frac{K(K+1)}{2} = \frac{K}{2}.$$

Si $\rho \neq 1$, entonces

$$\begin{aligned}
L &= \sum_{n=0}^{K} np_n = \sum_{n=0}^{K} np_0\rho^n = p_0\rho\sum_{n=1}^{K} n\rho^{n-1} = p_0\rho\sum_{n=0}^{K} n\rho^{n-1} \\
&= p_0\rho\frac{d}{d\rho}\left(\sum_{n=0}^{K}\rho^n\right) = p_0\rho\frac{d}{d\rho}\left(\frac{1-\rho^{K+1}}{1-\rho}\right) \\
&= p_0\rho\frac{1-(K+1)\rho^K + K\rho^{K+1}}{(1-\rho)^2} \\
&= \frac{\rho\left[1-(K+1)\rho^K + K\rho^{K+1}\right]}{(1-\rho^{K+1})(1-\rho)}.
\end{aligned}$$

Para el **tamaño medio de la cola**, L_q, se obtiene que

$$\begin{aligned}
L_q &= 0p_0 + \sum_{n=1}^{K}(n-1)p_n \\
&= \sum_{n=1}^{K} np_n + \sum_{n=1}^{k} p_n \\
&= \sum_{n=0}^{K} np_n + (1-p_0) \\
&= L - (1-p_0),
\end{aligned}$$

de donde se sigue que:

$$L_q = L - (1 - p_0) = \begin{cases} L - \frac{\rho(1-\rho^K)}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)}, & \rho = 1 \end{cases}$$

Podemos calcular el **tiempo medio de estancia en el sistema** de un cliente, W, condicionado por el número de clientes en el sistema cuando el cliente se incorpora al mismo. Obsérvese que para que el cliente no sea rechazado tiene que haber a lo sumo $K-1$ clientes en el sistema, luego la variable por la que hay que condicionar no es $\mathcal{L}$ sino $\overline{\mathcal{L}} = (\mathcal{L}|\mathcal{L} \leq K-1)$, cuyas probabilidades puntuales son:

$$\overline{p_n} = P(\mathcal{L} = n|\mathcal{L} \leq K-1) = \frac{P(\mathcal{L} = n)}{P(\mathcal{L} \geq K-1)} = \frac{p_n}{1-p_K}, \quad n \leq K-1.$$

Entonces, se tiene que:

$$W = \sum_{n=0}^{K-1} E(W|\overline{\mathcal{L}} = n)\overline{p_n}.$$

Obsérvese que para n fijo, se tiene que $E(W|\overline{\mathcal{L}} = n) = \frac{n+1}{\mu}$, luego

$$\begin{aligned} W &= \sum_{n=0}^{K-1} E(W|\overline{\mathcal{L}} = n+1)\overline{p_n} \\ &= \sum_{n=0}^{K-1} \frac{n+1}{\mu} \frac{p_n}{1-p_K} \\ &= \frac{1}{\mu(1-p_K)} \left(\sum_{n=0}^{K-1} (n+1)p_n \right) \\ &= \frac{1}{\mu(1-p_K)} \left(\sum_{n=0}^{K-1} np_n + \sum_{n=0}^{k-1} p_n \right) \\ &= \frac{1}{\mu(1-p_K)} \left([L - Kp_K] + [1 - p_K] \right) \end{aligned}$$

Se comprueba fácilmente que $(L - Kp_K) + (1 - p_K) = L + 1 - (K+1)p_K = \frac{L}{\rho}$, luego

$$
\begin{aligned}
W &= \frac{1}{\mu(1-p_K)}\left([L - Kp_K] + [1 - p_K]\right) \\
&= \frac{1}{\mu(1-p_K)}\frac{L}{\rho} \\
&= \frac{L}{\lambda(1-p_K}.
\end{aligned}
$$

Para el cálculo del **tiempo medio de espera en cola**, W_q, podemos utilizar la relación $W = W_q + \frac{1}{\mu}$. Asimismo, se puede comprobar que $W_q = \frac{L_q}{\lambda(1-p_K)}$.

6.4. Procesos de Nacimiento y Muerte

La mayor parte de los modelos elementales de colas suponen que las entradas (llegadas) y las salidas (servicios) ocurren de acuerdo a un proceso de nacimiento y muerte, que es un tipo de proceso estocástico en tiempo continuo con espacio de estados discretos y que posee importantes aplicaciones en diversas área. En lo que se refiere a la teoría de colas, el término nacimiento se refiere a la llegada de un nuevo cliente y el término muerte se refiere a la finalización de un servicio y la consiguiente salida del cliente. El proceso de nacimiento y muerte describe en términos probabilísticos cómo cambia el número de clientes en el sistema, $N(t)$, con respecto al tiempo t. Las suposiciones del proceso de nacimiento y muerte son las siguientes:

1. En un intervalo suficientemente pequeño de tiempo, el estado del sistema sólo puede aumentar en uno (un nacimiento), disminuir en uno (una muerte) o permanecer como está.

2. Dado $N(t) = n$,

- la probabilidad de que no ocurra ningún nacimiento en el intervalo $(t, t+\Delta t]$ es $1 - \lambda_n \Delta t + o(\Delta t)$
- la probabilidad de que ocurra un nacimiento en el intervalo $(t, t+\Delta t]$ es $\lambda_n \Delta t + o(\Delta t)$
- la probabilidad de que ocurran dos o más nacimientos en el intervalo $(t, t+\Delta t]$ es $o(\Delta t)$

3. Dado $N(t) = n \geq 1$,
   - la probabilidad de que no ocurra ninguna muerte en el intervalo $(t, t+\Delta t]$ es $1 - \mu_n \Delta t + o(\Delta t)$
   - la probabilidad de que se produzca una muerte en el intervalo $(t, t+\Delta t]$ es $\mu_n \Delta t + o(\Delta t)$
   - la probabilidad de que ocurran dos o más muertes en el intervalo $(t, t+\Delta t]$ es $o(\Delta t)$
4. Los nacimientos y las muertes son independientes entre sí
5. El proceso verifica la propiedad de Markov, es decir si $t_1 < t_2 < \ldots < t_k < t$, entonces

$$\begin{aligned} &P\left(N(t) \in S \subset \mathbb{Z}_+ | N(t_k) = n_k, N(t_{k-1}) = n_{k-1}, \ldots, N(t_1) = k_1\right) = \\ &\quad = P\left(N(t) \in S \subset \mathbb{Z}_+ | N(t_k) = n_k\right). \end{aligned}$$

Al igual que en el análisis del modelo $M/M/1$, vamos a intentar obtener las probabilidades de estado $p_n(t)$, a través de un sistema de ecuaciones en diferencia. Para ello, vamos a analizar cómo podría alcanzar el sistema el estado n en el instante $t + \Delta t$ en función del estado en el instante t. Supongamos inicialmente que $n \geq 1$.

- Si el sistema está en el estado n en el instante t, entonces para que siguiera en el mismo estado en el instante $t + \Delta t$ pueden darse dos situaciones en el intervalo $(t, t + \Delta t]$:
  - o que no se dé ningún nacimiento ni ninguna muerte, lo cual tiene probabilidad $(1 - \lambda_n \Delta t + o(\Delta t))\,(1 - \mu_n \Delta t + o(\Delta t))$
  - o que ocurran un nacimiento y una muerte, lo cual tiene probabilidad $(\lambda_n \Delta t + o(\Delta t))\,(\mu_n \Delta t + o(\Delta t))$.
- Por otro lado, si el sistema está en el estado $n + 1$ en el instante t, entonces para que cambiase a estado n en el instante $t + \Delta t$, en el intervalo $(t, t + \Delta t]$ debería producirse una muerte y ningún nacimiento, lo cual tiene probabilidad $(1 - \lambda_{n+1} \Delta t + o(\Delta t))\,(\mu_{n+1} \Delta t + o(\Delta t))$.
- Finalmente, si el sistema está en el estado $n - 1$ en el instante t, entonces para que cambiase a estado n en el instante $t + \Delta t$, en el intervalo $(t, t + \Delta t]$ debería producirse un nacimiento y ninguna muerte, lo cual tiene probabilidad $(\lambda_{n+1} \Delta t + o(\Delta t))\,(1 - \mu_{n+1} \Delta t + o(\Delta t))$.

Por lo tanto, si $n \geq 1$, podemos escribir

$$\begin{aligned}
p_n(t + \Delta t) = {} & p_n(t)\,(1 - \lambda_n \Delta t + o(\Delta t))\,(1 - \mu_n \Delta t + o(\Delta t)) + \\
& + p_n(t)\,(\lambda_n \Delta t + o(\Delta t))\,(\mu_n \Delta t + o(\Delta t)) + \\
& + p_{n+1}(t)\,(1 - \lambda_{n+1} \Delta t + o(\Delta t))\,(\mu_{n+1} \Delta t + o(\Delta t)) + \\
& + p_{n-1}(t)\,(\lambda_{n+1} \Delta t + o(\Delta t))\,(1 - \mu_{n+1} \Delta t + o(\Delta t))\,.
\end{aligned}$$

Juntando los término $o(\Delta t)$, quedaría

$$p_n(t + \Delta t) = p_n(t)(1 - \lambda_n \Delta t - \mu_n \Delta t + p_{n+1}(t)\,(\mu_{n+1} \Delta t) + p_{n-1}(t)\,(\lambda_{n+1} \Delta t) + o(\Delta t). \tag{6.11}$$

Para el caso $n = 0$, el razonamiento sería el siguiente:

- Si el sistema está en el estado 0 en el instante t, entonces para que siguiera en el mismo estado en el instante $t + \Delta t$, en el intervalo $(t, t + \Delta t]$ no debe darse ningún nacimiento, lo cual tiene probabilidad $(1 - \lambda_0 \Delta t + o(\Delta t))$.
- Si el sistema está en el estado 1 en el instante t, entonces para que cambiase a estado 0 en el instante $t+\Delta t$, en el intervalo $(t, t+\Delta t]$ debería producirse una muerte y ningún nacimiento, lo cual tiene probabilidad $(1 - \lambda_1 \Delta t + o(\Delta t))\,(\mu_1 \Delta t + o(\Delta t))$.

Por lo tanto, para $n = 0$, podemos escribir:

$$\begin{aligned} p_0(t + \Delta t) = p_0(t)\,(1 - \lambda_0 \Delta t + o(\Delta t)) + \\ + p_1(t)\,(1 - \lambda_1 \Delta t + o(\Delta t))\,(\mu_1 \Delta t + o(\Delta t))\,. \end{aligned}$$

y juntando los términos $o(\Delta t)$, quedaría

$$p_0(t + \Delta t) = p_0(t)(1 - \lambda_0 \Delta t) + p_1(t)\,(\mu_1 \Delta t) + o(\Delta t). \tag{6.12}$$

Si en la ecuación 6.11 (respectivamente, 6.12) restamos en ambos términos $p_n(t)$ (respectivamente, $p_0(t)$), dividimos por Δt y tomamos límites cuando $\Delta t \to 0$, el sistema de ecuaciones diferenciales en diferencia que se obtiene para las probabilidades de estado del proceso de nacimiento y muerte sería

$$\begin{cases} \frac{dp_n(t)}{dt} = -(\lambda_n + \mu_n)p_n(t) + \mu_{n+1}p_{n+1}(t) + \lambda_{n-1}p_{n-1}(t), & n \geq 1 \\ \frac{dp_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t) \end{cases}$$

La probabilidades en el estado estacionario se obtienen igual que en el caso $M/M/1$, pues al tomar límites cuando $t \to \infty$, $p_n(t)$ se vuelve independiente de t. Por lo tanto, el sistema anterior quedaría:

$$\begin{cases} 0 = -(\lambda_n + \mu_n)p_n + \mu_{n+1}p_{n+1} + \lambda_{n-1}p_{n-1}, & n \geq 1 \\ 0 = -\lambda_0 p_0 + \mu_1 p_1 \end{cases}$$

o, equivalentemente,

$$\begin{cases} (\lambda_n + \mu_n)p_n = \mu_{n+1}p_{n+1} + \lambda_{n-1}p_{n-1}, \quad n \geq 1 \\ \lambda_0 p_0 = \mu_1 p_1 \end{cases} \tag{6.13}$$

El sistema de ecuaciones 6.13 recibe el nombre de ecuaciones de balance y representan una equivalencia entre la tasa media de entrada y la tasa media de salida de cada estado del sistema.

Podemos obtener la solución del sistema 6.13 por iteración.

$$p_2 = \frac{\lambda_1 + \mu_1}{\mu_2}p_1 - \frac{\lambda_0}{\mu_2}p_0 = \frac{\lambda_1 + \mu_1}{\mu_2}\frac{\lambda_0}{\mu_1}p_0 - \frac{\lambda_0}{\mu_2}p_0 = \frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0.$$

Igualmente,

$$p_3 = \frac{\lambda_2 + \mu_2}{\mu_3}p_2 - \frac{\lambda_1}{\mu_3}p_1 = \frac{\lambda_2 + \mu_2}{\mu_3}\frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0 - \frac{\lambda_1}{\mu_3}\frac{\lambda_0}{\mu_1}p_0 = \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}p_0.$$

En vista de los anteriores resultados, parece razonable conjeturar que para $n \geq 1$,

$$p_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_1\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_2\mu_1}p_0 = p_0\prod_{i=1}^{n}\frac{\lambda_{i-1}}{\mu_i}. \tag{6.14}$$

Demostraremos la relación 6.14 por inducción matemática. Sabemos que la fórmula es correcta para $n = 1, 2, 3$. Supongamos que es cierta para $n = 1, \ldots, k$, veamos que también lo es para $k + 1$.

$$\begin{aligned} p_{k+1} &= \frac{\lambda_k + \mu_k}{\mu_{k+1}}p_k - \frac{\lambda_{k-1}}{\mu_{k+1}}p_{k-1} \\ &= \frac{\lambda_k + \mu_k}{\mu_{k+1}}\left(p_0\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i}\right) - \frac{\lambda_{k-1}}{\mu_{k+1}}\left(p_0\prod_{i=1}^{k-1}\frac{\lambda_{i-1}}{\mu_i}\right) \\ &= p_0\frac{\lambda_k}{\mu_{k+1}}\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i} + p_0\frac{\mu_k}{\mu_{k+1}}\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i} - p_0\frac{\mu_k}{\mu_{k+1}}\prod_{i=1}^{k}\frac{\lambda_{i-1}}{\mu_i} \\ &= p_0\prod_{i=1}^{k+1}\frac{\lambda_{i-1}}{\mu_i}. \end{aligned}$$

Para obtener el valor de p_0 tenemos en cuenta que

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \left(1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right).$$

Por lo tanto, una condición necesaria y suficiente para que exista la solución del estado estacionario es que la serie

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \tag{6.15}$$

sea convergente.

Nota. Para el modelo $M/M/1$, se tiene que $\lambda_i = \lambda$ y $\mu_i = \mu$, por lo que la serie anterior queda $\sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n$ que es convergente si $\frac{\lambda}{\mu} < 1$.

Nota. Supongamos que $\lambda_n = \lambda$ y $\mu_n = n\mu$. Entonces, $p_n = p_0 \frac{\lambda^n}{n!\mu^n}$, y la serie 6.15 quedaría

$$\sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} = \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!},$$

que es convergente y vale $e^{-\frac{\lambda}{\mu}} - 1$. Por lo tanto, en este caso, se tiene que

$$p_0 = e^{-\frac{\lambda}{\mu}}, \qquad p_n = e^{-\frac{\lambda}{\mu}} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}, n \geq 1,$$

que es una distribución de Poisson de parámetro $\frac{\lambda}{\mu}$.

6.5. Modelos de colas exponenciales con varios servidores en paralelo

En esta sección consideramos un modelos de colas en el que las llegadas se producen según un proceso de Poisson de parámetro λ, hay c servidores que atienden a los clientes cuyos tiempos de servicio son independientes e idénticamente distribuidos, exponenciales de parámetro μ. Estas colas se ajustan a un modelo de nacimiento y muerte en el que $\lambda_n = \lambda$, para todo n. Por otro lado, recordemos que μ_n es el número medio de servicios que finalizan por unidad de tiempo cuando en el sistema hay n clientes. Si $n > c$, entonces los c servidores están ocupados y como cada uno de ellos sirve a μ clientes por unidad de tiempo, en total finalizarían $c\mu$ servicios por unidad de tiempo. Si $1 \leq n \leq c$, entonces sólo n de los c canales están ocupados, y por lo tanto la tasa de finalización de servicios sería $n\mu$. En definitiva,

$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c \\ c\mu, & n > c. \end{cases}$$

De la expresión 6.14 se tiene que

$$p_n = \begin{cases} \frac{\lambda^n}{n!\mu^n} p_0, & 1 \leq n \leq c \\ \frac{\lambda^n}{c!c^{n-c}\mu^n} p_0, & n \geq c. \end{cases} \tag{6.16}$$

Utilizando la condición de que $\sum_{n=0}^{\infty} p_n = 1$, podemos derivar el valor de p_0.

$$p_0 \left(\sum_{n=0}^{c-1} \frac{\lambda^n}{n!\mu^n} + \sum_{n=c}^{\infty} \frac{\lambda^n}{c!c^{n-c}\mu^n} \right) = 1.$$

Denotando $r = \frac{\lambda}{\mu}$, $\rho = \frac{\lambda}{c\mu} = \frac{r}{c}$, tenemos que la expresión anterior quedaría

$$p_0 \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \sum_{n=c}^{\infty} \frac{r^n}{c!c^{n-c}} \right) = 1.$$

Analicemos ahora la serie $\sum_{n=c}^{\infty} \frac{r^n}{c!c^{n-c}}$.

$$\sum_{n=c}^{\infty} \frac{r^n}{c!c^{n-c}} = \frac{r^c}{c!} \sum_{n=c}^{\infty} \left(\frac{r}{c}\right)^{n-c} = \frac{r^c}{c!} \sum_{m=0}^{\infty} \rho^m, \tag{6.17}$$

que converge a $\frac{r^c}{c!}\frac{1}{1-\rho}$ siempre que $\rho = \frac{r}{c} < 1$.

Por lo tanto, para el modelo $M/M/c$, si se verifica que $\rho = \frac{r}{c} = \frac{\lambda}{c\mu} < 1$ entonces la solución estacionaria existe y el valor de p_0 viene dado por

$$p_0 = \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{cr^c}{c!(c-r)} \right)^{-1} = \left(\sum_{n=0}^{c-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!} \left(\frac{\lambda}{\mu}\right)^c \left(\frac{c\mu}{c\mu - \lambda}\right) \right)^{-1}. \tag{6.18}$$

Nota. Obsérvese que si $c = 1$, las expresiones 6.16 y 6.18 se convierten en las correspondientes al caso $M/M/1$.

Medidas de eficiencia

Tamaño esperado de la cola, L_q

El tamaño de la cola $\mathcal{L}_q$ es cero si el número de clientes en el sistema, n, es menor o igual que c y $n - c$ en caso contrario. Por lo tanto,

$$L_q = E(\mathcal{L}_{\text{II}}) = \sum_{n=c}^{\infty}(n-c)p_n = \sum_{n=c}^{\infty} \frac{n}{c^{n-c}c!} r^n p_0 - \sum_{n=c}^{\infty} \frac{c}{c^{n-c}c!} r^n p_0 \tag{6.19}$$

Examinemos la primera serie del lado derecho de 6.19

$$\begin{aligned}
\frac{p_0}{c!}\sum_{n=c}^{\infty}\frac{n}{c^{n-c}}r^n &= \frac{p_0}{c!}\frac{r^{c+1}}{c}\left[\sum_{n=c}^{\infty}(n-c)\left(\frac{r}{c}\right)^{n-c-1} + \sum_{n=c}^{\infty} c\left(\frac{r}{c}\right)^{n-c-1}\right] \\
&= \frac{p_0}{c!}\frac{r^{c+1}}{c}\left[\sum_{n=0}^{\infty} n\left(\frac{r}{c}\right)^{n-1} + \sum_{n=0}^{\infty} c\left(\frac{r}{c}\right)^{n-1}\right] \\
&= \frac{p_0}{c!}\frac{r^{c+1}}{c}\left[\sum_{n=0}^{\infty} n\left(\frac{r}{c}\right)^{n-1} + c\frac{c}{r}\sum_{n=0}^{\infty} c\left(\frac{r}{c}\right)^{n}\right] \\
&= \frac{p_0}{c!}\frac{r^{c+1}}{c}\left[\frac{1}{(1-\frac{r}{c})^2} + \frac{\frac{c^2}{r}}{1-\frac{r}{c}}\right]
\end{aligned}$$

El valor de la segunda serie del lado derecho de 6.19, se obtiene fácilmente a partir de 6.17

$$\sum_{n=c}^{\infty}\frac{c}{c^{n-c}c!}r^n p_0 = p_0 c\sum_{n=c}^{\infty}\frac{r^n}{c^{n-c}c!} = \frac{p_0 c r^c}{c!\left(1-\frac{r}{c}\right)}.$$

Por lo tanto, la expresión 6.19 quedaría

$$L_q = p_0\frac{r^{c+1}}{c\cdot c!}\left[\frac{1}{\left(1-\frac{r}{c}\right)^2} + \frac{\frac{c^2}{r}}{1-\frac{r}{c}} + \frac{\frac{c^2}{r}}{1-\frac{r}{c}}\right]$$

$$p_0\frac{r^{c+1}}{c\cdot c!}\left[\frac{1}{\left(1-\frac{r}{c}\right)^2}\right]$$

En definitiva,

$$L_q = \left[\frac{\left(\frac{\lambda}{\mu}\right)^c \lambda\mu}{(c-1)!(c\mu - \lambda)^2} \right] p_0. \tag{6.20}$$

Tiempo medio de espera en cola

Como es habitual, en el modelo $M/M/1$ asumimos que la disciplina de servicio es FIFO, entendiendo por FIFO que los clientes entran al servicio por orden de llegada, aunque al haber varios canales es posible que un cliente que ha llegado después que otro salga del sistema antes que el primero. Haciendo uso de la fórmula de Little, obtenemos que:

$$W_q = \frac{L_q}{\lambda} = \left[\frac{\left(\frac{\lambda}{\mu}\right)^c \mu}{(c-1)!(c\mu - \lambda)^2} \right] p_0. \tag{6.21}$$

Tiempo medio de estancia en el sistema

Utilizando la relación general entre W y W_q y el resultado obtenido en 6.21, se tiene que:

$$W = W_q + \frac{1}{\mu} = \left[\frac{\left(\frac{\lambda}{\mu}\right)^c \mu}{(c-1)!(c\mu - \lambda)^2} \right] p_0 + \frac{1}{\mu} \tag{6.22}$$

Número medio de clientes en el sistema

Utilizando la fórmula de Little y la expresión 6.22, se obtiene que:

$$L = \lambda W = \left[\frac{\left(\frac{\lambda}{\mu}\right)^c \lambda\mu}{(c-1)!(c\mu - \lambda)^2} \right] p_0 + \frac{\lambda}{\mu} \tag{6.23}$$

Distribuciones de probabilidad de $\mathcal{W}_q$ y $\mathcal{W}$

Aunque, como hemos visto, las distribuciones de probabilidad de $\mathcal{W}_q$ y $\mathcal{W}$ no son necesarias para obtener W_q y W, las dejamos indicadas en este apartado para poder responder a cuestiones acerca de la probabilidad de que un determinado cliente espere una cierta cantidad de tiempo. En concreto, $\mathcal{W}_q$ es una variable mixta, con un punto aislado $t = 0$ en el que se alcanza probabilidad

$$P(\mathcal{W}_q = 0) = 1 - \frac{c\left(\frac{\lambda}{\mu}\right)^c}{c!\left(c - \frac{\lambda}{\mu}\right)} p_0$$

y una parte continua en $(0, +\infty)$ regida por la función de densidad

$$f_{\mathcal{W}_q}(t) = \frac{\left(\frac{\lambda}{\mu}\right)^c \mu e^{-(\mu c - \lambda)t}}{(c-1)!}, \quad t > 0$$

Asimismo, $\mathcal{W}$ es una variable continua, cuya función de densidad viene dada por:

$$f_{\mathcal{W}}(t) = \frac{\mu e^{-\mu t}\left[\lambda - c\mu + \mu \cdot P(\mathcal{W}_q = 0)\right] - \left[1 - P(\mathcal{W}_q = 0)\right]\left[\lambda - c\mu\right]\mu e^{-(c\mu - \lambda)t}}{\lambda - (c-1)\mu}, \quad t > 0.$$

6.6. Modelo $M/M/c$ con fuente de entrada finita

Consideramos una variación del modelo $M/M/c$ consistente en que la fuente de entrada es limitada; es decir, el tamaño de la población de potenciales clientes es finito. Sea N el tamaño de dicha población. De este modo, cuando en el sistema se encuentran n clientes, restan solo $N - n$ clientes potenciales en la fuente de entrada.

La aplicación más importante de este modelo es el problema de reparación de máquinas, en el que se asigna a uno o más técnicos la responsabilidad de mantener operativas un grupo de N máquinas. Cuando estas maquinas se estropean acuden al sistema de mantenimiento en espera de ser reparadas, y cuando están operativas están fuera del sistema.

En el modelo con población finita los clientes alternan entre estar dentro y fuera del sistema, así pues por analogía con el modelo $M/M/c$ se supone que el tiempo que pasa cada miembro fuera del sistema es una variable exponencial de parámetro λ. Cuando n miembros están dentro, $N-n$ están fuera, y por lo tanto la distribución de probabilidad del tiempo que falta para la próxima llegada al sistema es el mínimo de $N-n$ variables exponenciales independientes de parámetro λ. Se puede demostrar que esta distribución se ajusta a una exponencial de parámetro $\lambda(N-n)$. Así que el modelo es un caso especial del proceso de nacimiento y muerte en el cual

$$\lambda_n = \begin{cases} (N-n)\lambda, & 0 \le n \le N \\ 0, & n > N \end{cases}$$

y

$$\mu_n = \begin{cases} n\mu\lambda, & 1 \le n \le c \\ c\mu\lambda, & c \le n \le N \\ 0, & n > N \end{cases}$$

Medidas de eficiencia para el caso $M/M/1$ con población finita

$$p_0 = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

$$p_n = \frac{N!}{(N-n)!} \left(\frac{\lambda}{\mu}\right)^n p_0, \qquad 1 \le n \le N$$

$$L_q = N - \frac{\lambda+\mu}{\lambda}(1-p_0), \qquad L = N - \frac{\mu}{\lambda}(1-p_0)$$

$$W_q = \frac{L_q}{\overline{\lambda}}, \qquad W = \frac{L}{\overline{\lambda}}$$

donde

$$\overline{\lambda} = \lambda(N - L).$$

Medidas de eficiencia para el caso $M/M/c$ con población finita

$$p_0 = \left[\sum_{n=0}^{c-1} \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=c}^{N} \binom{N}{n} \frac{n!}{c! c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

$$p_n = \begin{cases} \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n p_0, & 0 \leq n \leq c \\ \binom{N}{n} \frac{n!}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n p_0, & c \leq n \leq N \end{cases}$$

$$L = p_0 \left[\sum_{n=0}^{c-1} n \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{c!} \sum_{n=c}^{N} n \binom{N}{n} \frac{n!}{c^{n-c}} \left(\frac{\lambda}{\mu}\right)^n \right].$$

$$L_q = L - c + p_0 \sum_{n=0}^{c-1} (c - n) \binom{N}{n} \left(\frac{\lambda}{\mu}\right)^n.$$

$$W_q = \frac{L_q}{\overline{\lambda}}, \qquad W = \frac{L}{\overline{\lambda}}$$

donde

$$\overline{\lambda} = \lambda(N - L).$$

6.7. Modelo de colas con tiempos de servicio no exponencial

Modelo $M/G/1$

En este modelo se supone que el sistema de colas tiene un servidor, las llegadas se producen según un proceso de Poisson de tasa λ y los clientes tienen tiempos de servicio independientes e idénticamente distribuidos de media $\frac{1}{\mu}$ y varianza σ^2.

Cualquier sistema de colas de este tipo alcanza en algún momento el estado estable si $\rho = \frac{\lambda}{\mu} < 1$. Las medidas de eficiencia para este modelo toman las siguientes expresiones (la referente a L_q recibe el nombre de fórmula de Pollaczek-Khintchine)

$$\begin{aligned} p_0 &= 1 - \rho \\ L_q &= \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)} \\ L &= L_q + \rho \\ W_q &= \frac{L_q}{\lambda} \\ W = \frac{L}{\lambda} &= W_q + \frac{1}{\mu} \end{aligned}$$

Obsérvese que las medidas de eficiencia incrementan su valor conforme σ^2 aumenta. Esto indica que el funcionamiento del servidor tiene gran transcendencia en la eficiencia global de la instalación. Asimismo, se observa que las fórmulas anteriores se reducen a las del modelo $M/M/1$ cuando los tiempos de servicio siguen una distribución exponencial.

Modelo $M/D/1$

Cuando el servicio consiste básicamente en la misma tarea rutinaria que el servidor realiza para todos los clientes, tiende a haber poca variación en el tiempo de servicio requerido. En el modelo $M/D/c$ se asume que el tiempo de servicio siempre es igual a una constante fija.

Cuando sólo se tiene un servidor, el modelo $M/D/1$ se reduce a un caso particular del modelo $M/G/1$ en el cual $\sigma^2 = 0$, con lo que la fórmula de Pollaczek-Khintchine se reduce a:

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

Modelo $M/E_k/1$

El modelo $M/D/c$ supone una variación cero en los tiempos de servicio, mientras que el modelo $M/M/1$ supone una gran variabilidad ($\sigma = \frac{1}{\mu}$). Entre estos dos casos extremos hay un gran intervalo ($0 < \sigma < \frac{1}{\mu}$) en el que caen muchas de las distribuciones de servicio reales. Una distribución teórica de tiempos de servicio que concuerda con este rango intermedio es la distribución de Erlang.

La distribución de Erlang de parámetros k y ν es la suma de k variables aleatorias independientes exponenciales de parámetro ν. Así pues, su media es $\frac{k}{\nu}$ y su varianza es $\sigma^2 = \frac{k}{\nu^2}$. Particularizando las expresiones del modelo $M/G/1$ a una distribución Erlang de media $\frac{1}{\mu}$ (es decir, tomando $\nu = k\mu$), se obtiene que las medidas de eficiencia del modelo $M/E_k/1$ vienen dadas por:

$$p_0 = 1 - \rho$$

$$L_q = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)}, \qquad L = L_q + \rho$$

$$W_q = \frac{L_q}{\lambda}, \qquad W = \frac{L}{\lambda} = W_q + \frac{1}{\mu}$$