Teoría de Colas

Jessica Nathaly Pulzara Mora
jessica.pulzara@udea.edu.co

Departamento de ingeniería de sistemas

UNIVERSIDAD DE ANTIOQUIA

Concepto

La teoría de colas es el estudio matemático del comportamiento de líneas de espera. Esta se presenta, cuando los “clientes” llegan a un “lugar” demandando un servicio a un “servidor”, el cual tiene una cierta capacidad de atención. Si el servidor no está disponible inmediatamente y el cliente decide esperar, entonces se forma la línea de espera.

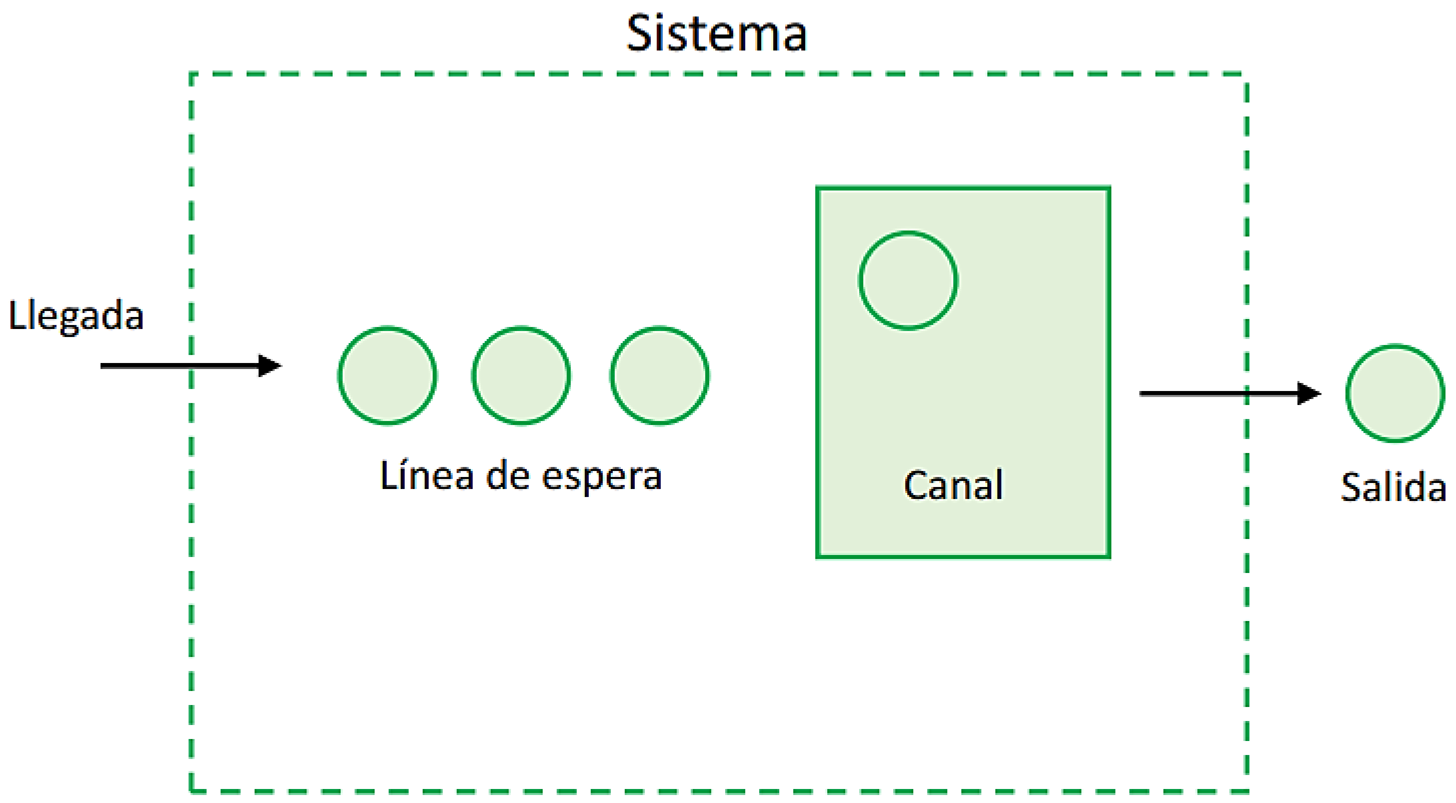
Ejemplos

Imagine situaciones donde se forman *colas* (líneas de espera):

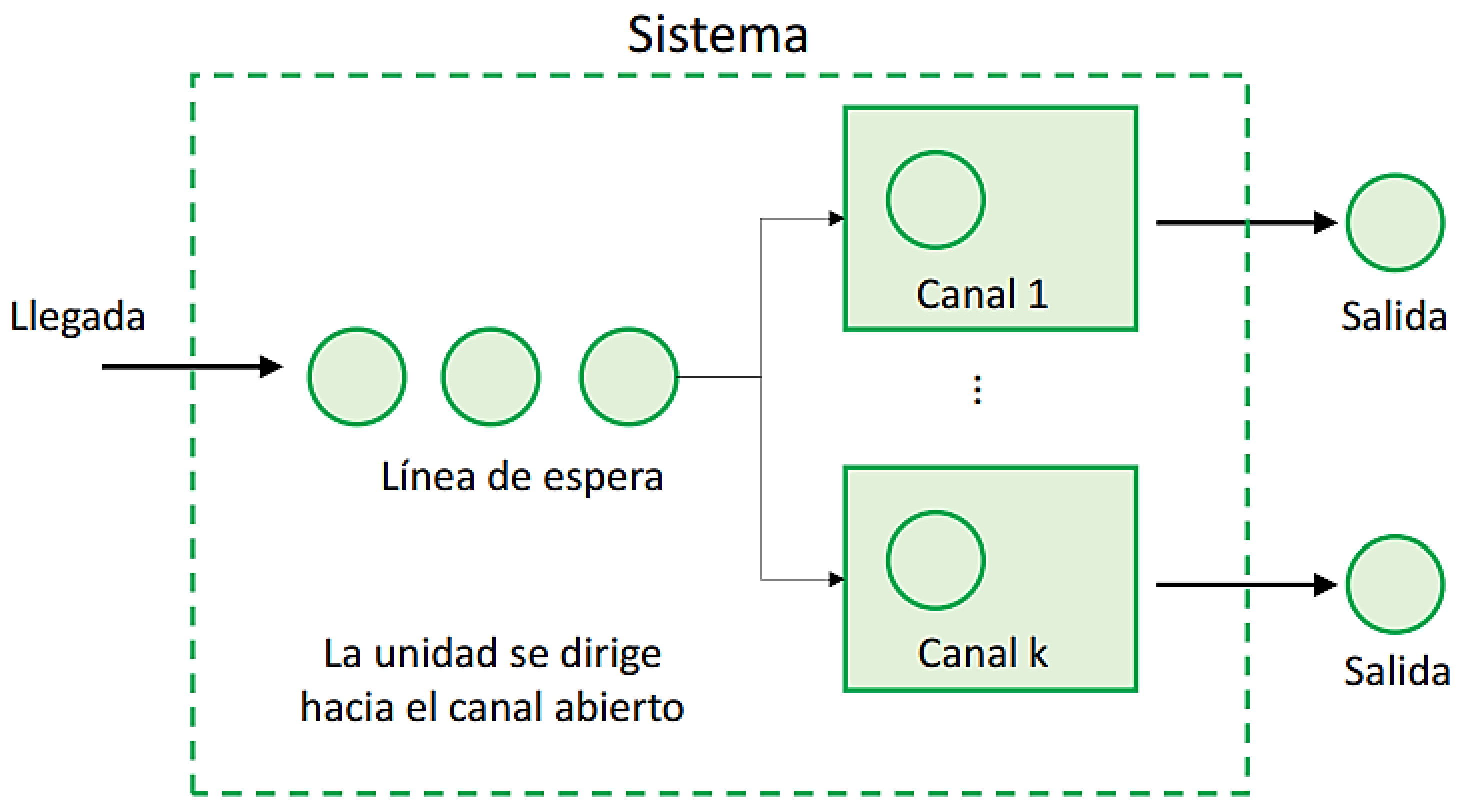
1. Realizar un pedido en un local de comida en centro comercial.
2. Ir al banco a realizar una consignación.
3. Un *call center*.
4. Un servidor, al que le llegan peticiones/consultas.

Estructura de un sistema de colas

Sistema de un canal



Sistema de varios canales



Líneas de espera

- El funcionamiento de las líneas de espera (o colas) es un proceso estocástico.
  - El número de llegadas es una variable aleatoria.
  - El tiempo entre llegadas es una variable aleatoria.
  - Los servicios se demoran un tiempo aleatorio.
- Para modelarlo, se requiere establecer un proceso en tiempo continuo.

Líneas de espera

Consideremos un proceso que evoluciona en tiempo continuo. Definamos lo siguiente:

$$P(x,t) = P(\text{hay exactamente } x \text{ llegadas durante un intervalo de tiempo } t)$$

Tiempo entre llegadas

Si consideramos un proceso Poisson, el número de llegadas en el tiempo t se distribuye así:

$$P(x, t) = e^{-\lambda t} \frac{(\lambda t)^x}{x!}$$

y el tiempo entre llegas en el intervalo $[0, t]$ tiene la siguiente probabilidad:

$$P(T \leq t) = 1 - e^{-\lambda t}$$

Tiempo de servicio

La probabilidad de que el tiempo de servicio sea menor que o igual a un tiempo de duración t es:

$$P(T \leq t) = 1 - e^{-\mu t}$$

donde μ es el número medio de unidades que pueden ser atendidas por periodo de tiempo.

Proceso Poisson

Proceso Poisson

Un proceso de llegadas es **Poisson**. con tasa λ si tienes estas propiedades:

a. **Homogeneidad:** la probabilidad de tener x llegadas es la misma para los intervalos de tiempo de tamaño τ.

b. **Independencia:** el número de llegadas en un intervalo de tiempo es independiente de la historia de llegadas previas a dicho intervalo.

Suma de procesos Poisson

- Considere las variables aleatorias $X_i \sim P(\lambda_i \tau_i),\ i = 1, 2, \ldots, m$. Entonces:

$$\sum_{i=1}^{m} X_i \sim P\left(\lambda \sum_{i=1}^{m} \tau_i\right)$$

Ejemplo

Considere la fila de un supermercado como un proceso Poisson con tasa de llegada $\lambda = 10$ clientes por minuto. Sean M el número de llegadas de 9:00 a 9:10 a.m. y N el número de llegadas de 9:30 a 9:35 a.m.

Se tiene,

$$\begin{aligned} \tau_1 &= 10 \qquad \lambda_1 = 100, \text{ y } M \sim P(\lambda_1), \\ \tau_2 &= 5 \qquad \lambda_2 = 50, \text{ y } N \sim P(\lambda_2), \end{aligned}$$

entonces:

$$M + N \sim P(\lambda_1 + \lambda_2), \text{ es decir, } M + N \sim P(150)$$

Notación de Kendall

Simbología para comunicar de forma resumida el modelo de colas que se esté trabajando:

$$a/b/c$$

- a: distribución de llegadas.
- b: distribución de las salidas (tiempo de servicio).
- c: cantidad de servidores/canales paralelos.

Notación de Kendall

$$a/b/c$$

Los símbolos a y b pueden tomar los siguientes valores:

- M: distribución markoviana (o de Poisson) de llegadas y salidas.
- D: tiempo constante (o determinístico).
- G: distribución genérica.

$c = 1, 2, \ldots, \infty$

Notación de Kendall extendida

$$(a/b/c) : (D/K/N/) \quad \text{o también} \quad a/b/c/K/N/D$$

- a, b, c tienen la misma función.
- D: disciplina en las colas.
- K: número máximo (finito o infinito) permitido en el sistema (haciendo cola o en servicio).
- N: tamaño de la fuente (finita o infinita).

Notación de Kendall extendida

$$(a/b/c) : (D/K/N/) \quad \text{o también} \quad a/b/c/K/N/D$$

D puede tomar los siguientes valores:

- *FIFO*: primero en llegar, primero en ser servido.
- *LIFO*: último en llegar, primero en ser servido.
- *SIRO*: servicio en orden aleatorio.
- *GD*: disciplina general (cualqueir tipo de disciplina).

$K = 1, 2, \ldots, \infty$, y $N = 1, 2, \ldots, \infty$.

Operación constante

- **Periodo transitorio**: cuando un sistema de colas comienza a funcionar, no hay unidades a la espera de servicio. La actividad comienza a aumentar gradualmente.
- **Periodo estacionario**: el sistema esta funcionando hace algún tiempo, y finalmente alcanza una *operación constante* o un *estado estable*.
- Los modelos que estudiaremos describen las características de una cola en estado estable.

Características de operación

Los modelos de lineas de espera se componen de fórmulas y relaciones matemáticas que pueden utilizarse para determinar las características de operación (medidas de desempeño) de una línea de espera. Las características de operación de interés incluyen:

Características de operación

- La probabilidad de que no haya unidades en el sistema P_0
- El número promedio de unidades en la línea de espera L_q
- El número promedio de unidades en el sistema (el número de unidades en la línea de espera más el número de unidades que están siendo atendidas) L_s
- El tiempo promedio que una unidad pasa en la línea de espera W_q
- El tiempo promedio que una unidad pasa en el sistema (el tiempo de espera más el tiempo para que atiendan) W_s
- La probabilidad de que una unidad que llega tenga que esperar para que la atiendan P_w

Parámetros de la cola

- λ: tasa de llegadas
- μ: tasa de servicios
- Factor de uso:

$$\frac{\lambda}{\mu}$$

- c, número de canales.

Expresiones generales para cola infinita

Fórmula de Little: Número de unidades en el sistema y en la cola, si $N \to \infty$

$$L_s = \lambda W_s$$

$$L_q = \lambda W_q$$

Tiempo en el sistema

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

Cantidad promedio de servidores ocupados:

$$\bar{c} = L_s - L_q$$

Modelos de colas con llegadas Poisson y tiempo de servicio exponencial

Modelo de un solo canal (M/M/1) : (GD/∞/∞)

Nota: estas ecuaciones se cumplen cuando $\mu > \lambda$.

1. Probabilidad de que no haya unidades en el sistema:

$$P_0 = 1 - \frac{\lambda}{\mu}$$

2. Número promedio de unidades en la línea de espera:

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

Modelo de un solo canal (M/M/1) : (GD/∞/∞)

3. Número promedio de unidades en el sistema:

$$L_s = L_q + \frac{\lambda}{\mu}$$

4. Tiempo promedio que la unidad pasa en la línea de espera:

$$W_q = \frac{L_q}{\lambda}$$

Modelo de un solo canal (M/M/1) : (GD/∞/∞)

5. Tiempo promedio que una unidad pasa en el sistema

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

6. Probabilidad de que haya n unidades en el sistema:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

7. Probabilidad de que una unidad que llega tenga que esperar a ser atendida:

$$P_W = \frac{\lambda}{\mu}$$

Ejemplo

Un pequeño negocio de comida rápida es capaz de atender a sus clientes a una tasa de servicio de 1 cliente por minuto. Los clientes que son atendidos llegan al negocio una tasa de 0.75 clientes por minuto. ¿Cuáles son las características de operación?

- $P_0 = 0.25$
- $L_q = 2.25$
- $L_s = 3$
- $W_q = 3$
- $W_s = 4$
- $P_W = 0.75$

Modelo de varios canales (M/M/c) : (GD/∞/∞)

Nota: estas ecuaciones se cumplen cuando $c\mu > \lambda$.

1. Probabilidad de que no haya unidades en el sistema:

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

2. Número promedio de unidades en la línea de espera:

$$L_q = \frac{(\lambda/\mu)^c \lambda\mu}{(c-1)!\,(c\mu - \lambda)^2} P_0$$

Modelo de varios canales (M/M/c) : (GD/∞/∞)

3. Número promedio de unidades en el sistema:

$$L_s = L_q + \frac{\lambda}{\mu}$$

4. Tiempo promedio que una unidad pasa en la línea de espera:

$$W_q = \frac{L_q}{\lambda}$$

Modelo de varios canales (M/M/c) : (GD/∞/∞)

5. Probabilidad de que una unidad que llega tenga que esperar a ser atendida:

$$P_W = \frac{1}{c!}\left(\frac{\lambda}{\mu}\right)^c \frac{c\mu}{c\mu - \lambda} P_0$$

6. Tiempo promedio que una unidad pasa en el sistema:

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

Modelo de varios canales (M/M/c) : (GD/∞/∞)

7. Probabilidad de que haya n unidades en el sistema:

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0, \quad \text{si } n \leq c$$

$$P_n = \frac{(\lambda/\mu)^n}{c!\, c^{n-c}} P_0, \quad \text{si } n > c$$

Ejemplo

El administrador del pequeño negocio de comida rápida también lo atiende. El observó la situación actual, y considera que el tiempo de espera de 3 minutos es muy alto para un cliente que va de afán y está hambriento. Por esa razón, consigue un ayudante con experiencia, que también es capaz de atender a sus clientes a una tasa de servicio de 1 cliente por minuto. ¿Cuáles son las características de operación?

- $P_0 = 0.4545$
- $L_q = 0.1227$
- $L_s = 0.8727$
- $W_q = 0.1636$
- $W_s = 1.1636$
- $P_W = 0.2045$

Comparación

	Valor de la medida	
Medida	Un despachador	Dos despachadores
W_s	4	1.1636
L_q	2.25	0.1227
W_q	3	0.1636
P_W	0.75	0.2045

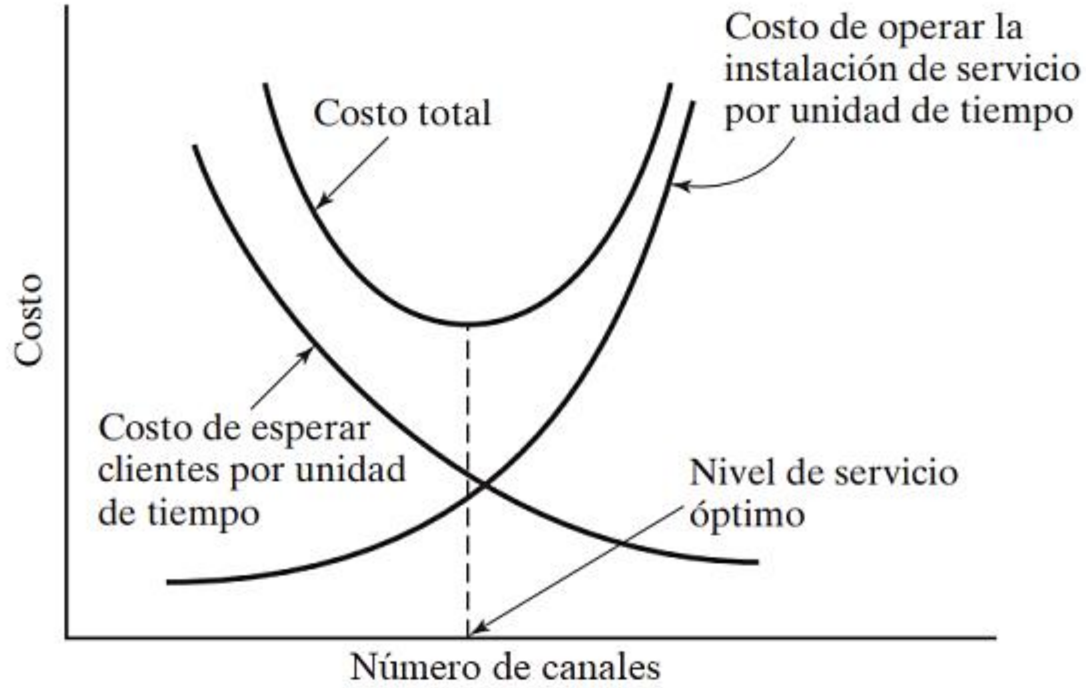
Comparación modelo de costos

¿Es siempre adecuado tener muchos canales?

$$\text{Costo Total} = k_T = k_W L_s + k_s c$$

- k_W costo de la espera de un cliente por unidad de tiempo.
- k_s costo del servicio por unidad de tiempo.

¿Es siempre adecuado tener muchos canales?



$$\text{Costo Total} = k_W L_s + k_s c$$

En general, k_W es un valor complejo de estimar.

En resumen, para graficar:

0. Inicio
1. Estimar el costo de la espera de un cliente por unidad
2. Elegir valores mínimo (c_{min}) y máximo (c_{Max}) para la cantidad de servidores c.
3. Fijar un valor para la cantidad de servidores c.
   a. Calcular L_s y fijar c.
   b. Calcular el costo con la fórmula del costo total.
   c. Graficar la pareja ordenadas (c, k_T).
   d. Si $c < c_{Max}$, volver al paso 3, de lo contrario, ir al paso 4.
4. Fin

Comparación nivel de aspiración

Considerar las siguientes medidas de desempeño:

1. Porcentaje de servidores ociosos:

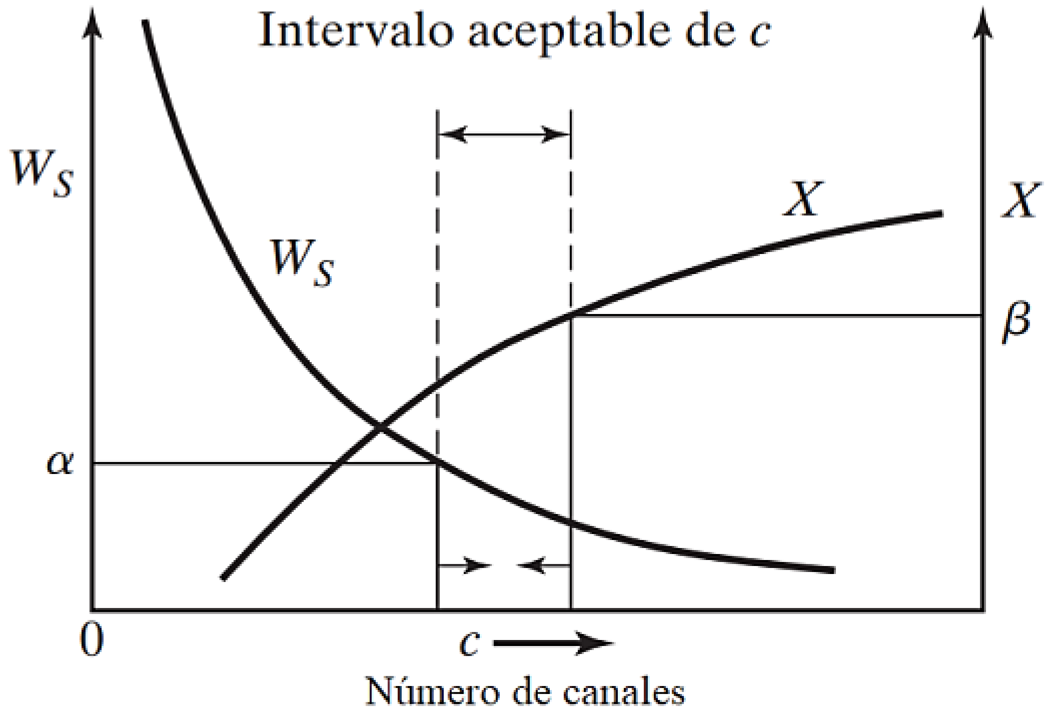
$$X = \frac{c - \bar{c}}{c} \times 100$$

2. Tiempo promedio en el sistema (W_s).

Queremos encontrar el número de servidores que cumpla lo siguiente:

$$W_s \leq \alpha \qquad X \leq \beta$$

Podemos graficar X y W_s contra c, para determinar un intervalo aceptable c^*



$$W_s \leq \alpha \qquad X \leq \beta$$

En resumen, para graficar:

0. Inicio
1. Fijar los valores de α y β.
2. Elegir valores mínimo (c_{min}) y máximo (c_{Max}) para la cantidad de servidores c.
3. Fijar un valor para la cantidad de servidores c.
   a. Calcular las características de operación (W_s, L_q, L_s)
   b. Calcular X.
   c. Graficar las parejas ordenadas (c, W_s) y (c, X)
   d. Si $c < c_{Max}$, volver al paso 3, de lo contrario, ir al paso 4.
4. Determinar graficamente el intervalo c^*.
5. Fin

Notas:

- Para el método de comparación por el modelo de costos, el costo del servicio por unidad de tiempo se puede obtener fijando/calculado el costo de operación del servidor (salario del recepcionista/tendero/operario, costo de operación de la máquina que hace el servicio, etc.)
- La selección del valor de c se puede hacer mediante *optimización*.
- Para el modelo de nivel de aspiración, los parámetros α y β son fijados por el ingeniero que organiza el sistema de colas.

Bibliografía

En el curso apenas *damos un pincelazo* a este tema. Si quiere ampliar el conocimiento, puede ver:

- Kendall's notation
  https://en.wikipedia.org/wiki/Kendall%27s_notation

- Un par de textos aun introductorios, pero donde pueden encontrar información adicional: