Sistemas M/G/1 y M/D/1

Grupo 4

Gastón Ugalde - Mariano Nozzi - Ivo Otero - Renato Corbellini - Juan Pablo Dominguez

Agenda

Repaso **notación de Kendall**

Características de cada sistema

Similitudes y **Diferencias**

Cálculo de:

- **Esperanza** de # en sistema
- Tiempo de **Permanencia**

Ejemplo Práctico (AWS **SQS**)

🤔 Recordando: ¿Qué era la notación de Kendall?

- Es una forma corta y estandarizada de describir cómo funciona un sistema de colas. Se usa para modelar, analizar y optimizar sistemas donde hay espera.

- Gracias a esta notación se pueden anticipar cuellos de botella, calcular una espera en promedio, calcular cuántos recursos faltantes y qué estrategia de atención resulta más eficiente.

Notación de Kendall clásica:

Es **la forma más común** de representar un **sistema de colas** y se escribe así:

A

Representa:
Tipo de distribución del tiempo entre llegadas. *(exponencial / aleatoria)*

B

Representa:
Tipo de distribución del tiempo de servicio.
(D = determinístico, G = general)

C

Representa:
Número de servidores *(o canales de atención).*

Notación de Kendall extendida:

Se usa cuando necesitamos **describir más detalles** del sistema y se agregan:

K

Representa:
Capacidad del sistema *(máximo de clientes en el sistema: en cola + en servicio).*

N

Representa:
Tamaño de la población *(total de clientes posibles).*

D

Representa:
Disciplina de la cola *(ej. FIFO, LIFO, prioridad, etc.)*

Sirve para entender y optimizar sistemas reales

Veamos tres casos donde estos modelos fueron clave para mejorar la experiencia de millones de personas

Netflix

¿Cuál era el problema? *Optimizar el tiempo de carga del contenido cuando hay alta demanda (por ejemplo, cuando se estrena una serie muy esperada)*

1. Se identificó que las llegadas de usuarios eran aleatorias (especialmente en horarios pico) y que el tiempo de servicio podría variar, según el contenido solicitado y la velocidad de red. Por eso se modeló como un sistema **M/G/1**.

2. Se organizó la atención en orden de llegada (**FIFO**) y se optimizó el envío de datos desde el servidor más cercano para evitar saturación. Así, se redujeron tiempos de espera y cargas excesivas.

3. Gracias a esto, se aplicaron estrategias como el caché local (guardar los contenidos más vistos en servidores regionales), lo que permitió acortar el tiempo de respuesta sin necesidad de aumentar el número de servidores reales.

Google Search:

¿Cuál era el problema? *Procesar búsquedas de miles de millones de usuarios rápidamente.*

1. Las búsquedas se modelaron como **llegadas aleatorias (M)**, con tiempos de procesamiento que varían ligeramente según la consulta **(G)**, y múltiples servidores distribuidos. El sistema puede representarse como un conjunto de colas por servidor.

2. Se aplicó una política **FIFO**, en la que cada búsqueda es atendida en el orden de llegada. Además, se envía la misma búsqueda a varios servidores (réplicas), y se toma la **respuesta más rápida** para reducir la espera del usuario.

3. El análisis ayudó a mejorar la **distribución de tareas**, a equilibrar las cargas entre servidores y a tomar decisiones más eficientes en la asignación de recursos, manteniendo la experiencia del usuario fluida incluso con alta demanda.

Disney:

¿Cuál era el problema? *Miles de personas haciendo fila para subirse a juegos casi en simultáneo.*

1. Las Se observó que las llegadas son aleatorias, pero el tiempo de cada juego es fijo. Por eso se utilizó el modelo **M/D/1**, ideal para sistemas con servicios regulares y alta concurrencia.

2. Se diseñaron colas organizadas con **FIFO**, asegurando que cada visitante fuera atendido en orden. También se estudió la capacidad de cada atracción para mantener un flujo constante.

3. Con el análisis matemático del sistema, se crearon colas virtuales, turnos anticipados y sistemas de reservas por app, que reducen el tiempo de espera física sin perder el orden ni la eficiencia.

Entonces:

Estos ejemplos muestran que, más allá de las fórmulas, los modelos matemáticos como los que describen la notación de Kendall ayudan a tomar decisiones concretas: permite organizar la atención, planificar mejor los recursos y mejorar la experiencia de los usuarios en sistemas de alta demanda

Sistema M/G/1: El flexible

M (Llegadas - Markovianas/Poisson):

- Llegadas aleatorias e independientes (Proceso de Poisson).
- Tasa promedio de llegadas: λ.
- Ej: Clientes llegando a una tienda en momentos impredecibles.

G (Servicio - General):

- Tiempos de servicio con distribución general o arbitraria.
- Conocemos el tiempo promedio de servicio: E[S] o μ.
- Conocemos la varianza del tiempo de servicio: σ^2.
- Ej: Un taller mecánico donde cada reparación toma un tiempo diferente.

1 (Servidor):

- Un único servidor disponible.

En resumen M/G/1: Ideal para llegadas aleatorias y tiempos de servicio variables.

Sistema M/D/1: El Predecible

M (Llegadas - Markovianas/Poisson):

- Llegadas aleatorias e independientes (Proceso de Poisson).
- Tasa promedio de llegadas: λ.
- Ej: Clientes llegando a una tienda en momentos impredecibles.

D (Servicio - Deterministico):

- Tiempos de servicio constantes y predecibles.
- Cada servicio toma exactamente $1/\mu$.
- Varianza del tiempo de servicio $\sigma^2 = 0$ (¡cero variabilidad!).
- Ej: Una máquina embotelladora que procesa cada botella en el mismo tiempo exacto.

1 (Servidor):

- Un único servidor disponible.

En resumen M/D/1: Perfecto para llegadas aleatorias pero con tareas de servicio estandarizadas y de duración fija.

Puntos Clave

- Ambos sistemas comparten llegadas Poisson (M) y un servidor (1).
- La diferencia principal radica en el tiempo de servicio:
  - M/G/1: General (flexible, con varianza).
  - M/D/1: Determinístico (constante, sin varianza).

Entender estas características es crucial para modelar y optimizar procesos de espera.

Similitudes y diferencias

Sistemas M/G/1 y M/D/1

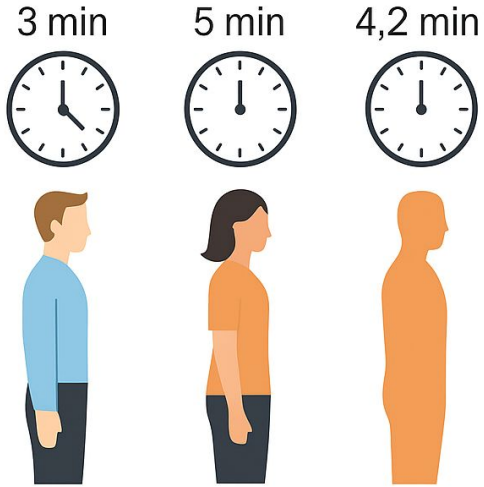
Similitudes

- Ambos modelos utilizan la notación de Kendall (A/B/c)
- Las llegadas siguen un proceso de Poisson
- Ambos tienen un solo servidor
- Cola infinita
- Disciplina de cola FIFO
- Sistema estable cuando $\lambda < \mu$

Diferencias

M/G/1		M/D/1
General	Distribución	Determinística
$\sigma^2 > 0$	Variabilidad del servicio	$\sigma^2 = 0$

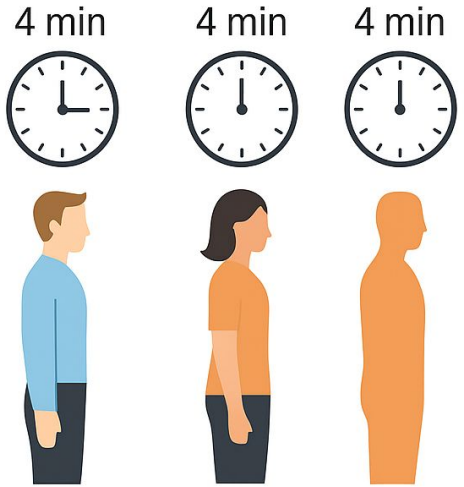
M/G/1



Tiempo de atención variable

→ General (G)

M/D/1

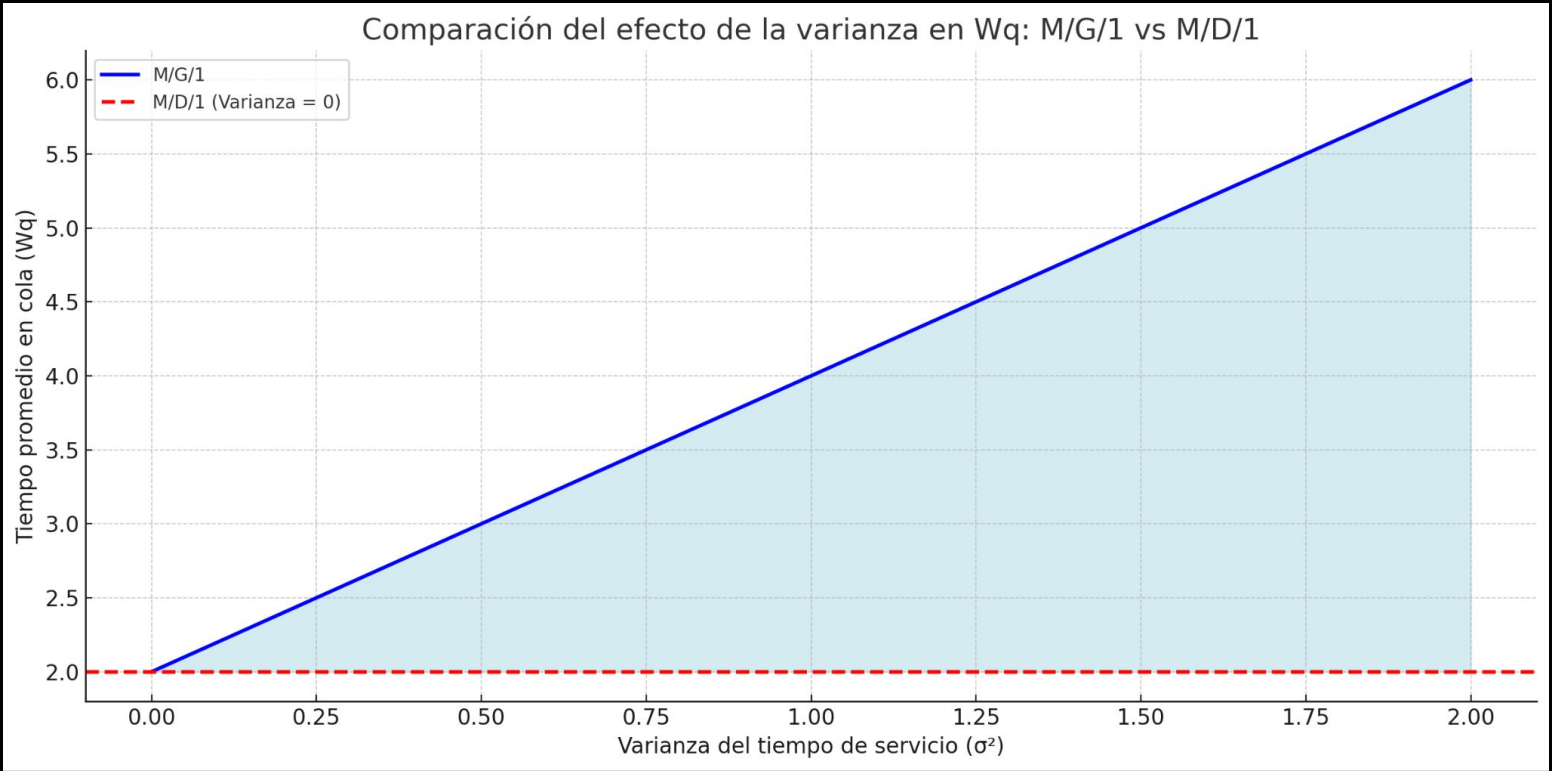


Tiempo de atención constante

→ Determinístico (D)

¿Por qué es importante esta diferencia?

En M/G/1, la varianza del tiempo de servicio **afecta directamente el rendimiento** del sistema.



Fórmula de espera en la cola (Wq)

M/G/1

$$W_q = \frac{\lambda . \sigma^2 + \rho^2}{2(1-\rho)}$$

M/D/1

$$\sigma^2 = 0$$

$$W_q = \frac{\lambda . \sigma^2 + \rho^2}{2(1-\rho)} \rightarrow W_q = \frac{\rho^2}{2(1-\rho)}$$

Esperanza de **# clientes en el sistema**

clientes en cola + clientes en servidor

Esperanza de **Tiempo de Permanencia**

tiempo de espera en cola + tiempo de servicio

Aclaración sobre **Esperanza**

La esperanza matemática se refiere a una métrica que **se espera que haya en el futuro**

- **No se basa en datos obtenidos estadísticamente**
- **No surge como resultado de un muestreo**

Número promedio de clientes en el sistema → Resultado de cálculos que se realizan sobre **algo que ya ocurrió**

—

Esperanza matemática → Valor medio que se calcula para algo que en el futuro **se espera que ocurra**

Ley de Little

$$E[n] = \lambda \cdot E[T]$$

E[n]: Esperanza del número de clientes en en el sistema en un instante cualquiera

λ: Tasa promedio de llegadas al sistema

E[T]: Esperanza del tiempo de permanencia en el sistema

Todos los valores de los tiempos de servicio tienen una dispersión del valor medio que se representa con el desvío estándar de los tiempos de servicio

$$\sigma$$

Recordamos que la varianza es el cuadrado del desvío estándar (Var = σ^2).

Sistemas M/G/1

$$\mathbb{E}[n] = \frac{\rho}{1-\rho} \cdot \left[1 - \frac{\rho}{2} \cdot (1 - \mu^2\sigma^2)\right]$$

$$\mathbb{E}[T] = \frac{1}{\mu(1-\rho)} \cdot \left[1 - \frac{\rho}{2} \cdot (1 - \mu^2\sigma^2)\right]$$

$\boldsymbol{\varrho = \lambda / \mu}$ (utilización del sistema)

$\boldsymbol{\mu}$: tasa de servicio

σ^2: varianza del tiempo de servicio

Ejemplo práctico

Una oficina de atención al cliente recibe en promedio 2 clientes por minuto. El tiempo que tarda un empleado en atender a un cliente varía, pero en promedio es de 0.4 minutos por cliente, con una varianza del tiempo de servicio de 0.01 minutos².

Se desea analizar el desempeño del sistema de atención al cliente suponiendo que sigue un modelo M/G/1, es decir, llegadas de Poisson, un solo servidor, y tiempos de servicio con distribución general.

Tasa de llegada → $\boldsymbol{\lambda} = 2$ clientes/minuto

Tiempo de servicio promedio → $E[S] = 0.4$ minutos

Varianza del **tiempo de servicio** → $\sigma^2 = 0.01$ minutos²

Ejemplo práctico

Tasa de llegada → $\lambda = 2$ clientes/min
T. de servicio prom → $E[S] = 0.4$ min
Var. del **t. de servicio** → $\sigma^2 = 0.01$ min²

Tasa de servicio μ

$$\mu = \frac{1}{\mathbb{E}[S]} = \frac{1}{0.4} = 2.5 \text{ clientes/minuto}$$

Utilización del sistema ρ

$$\rho = \lambda/\mu = 2/2.5 = 0.8$$

El servidor está ocupado el **80% del tiempo**

Ejemplo práctico

Tasa de llegada → $\lambda = 2$ clientes/min
T. de servicio prom → $E[S] = 0.4$ min
Var. del **t. de servicio** → $\sigma^2 = 0.01$ min²
Tasa de servicio $\mu = 2.5$ clientes/min
Utilización del sistema $\rho = 0.8$

E[T] (esperanza de tiempo en el sistema)

$$\mathbb{E}[T] = \frac{1}{\mu(1-\rho)} \cdot \left[1 - \frac{\rho}{2} \cdot (1 - \mu^2\sigma^2)\right]$$

$$\mathbb{E}[T] = \boxed{1.25 \text{ minutos}}$$

Ejemplo práctico

Tasa de llegada → $\lambda = 2$ clientes/min
T. de servicio prom → $E[S] = 0.4$ min
Var. del **t. de servicio** → $\sigma^2 = 0.01$ min²
Tasa de servicio $\mu = 2.5$ clientes/min
Utilización del sistema $\rho = 0.8$
Esp. t. en sistema $\mathbf{E[T]} = 1.25$ min

E[n] (número promedio de clientes)

$$\mathbb{E}[n] = \lambda \cdot \mathbb{E}[T] = 2 \cdot 0.45 = 0.9 \text{ clientes}$$

En promedio, hay <u>0.9 clientes en el sistema</u>

Sistemas M/D/1

Sistemas M/G/1

$$\mathbb{E}[n] = \frac{\rho}{1-\rho} \cdot \left[1 - \frac{\rho}{2} \cdot (1 - \mu^2\sigma^2)\right]$$

$\sigma^2 = 0$

$$\mathbb{E}[n] = \frac{\rho}{1-\rho} \cdot \left(1 - \frac{\rho}{2}\right)$$

$\boldsymbol{\varrho = \lambda / \mu}$ (utilización del sistema)

$\boldsymbol{\mu}$: tasa de servicio

Sistemas M/G/1

$$\mathbb{E}[T] = \frac{1}{\mu(1-\rho)} \cdot \left[1 - \frac{\rho}{2} \cdot (1 - \mu^2\sigma^2)\right]$$

$\sigma^2 = 0$

$$\mathbb{E}[T] = \frac{1}{\mu(1-\rho)} \cdot \left(1 - \frac{\rho}{2}\right)$$

Ejemplo práctico 2

Una cabina de peaje procesa vehículos que llegan al sistema a razón de 15 autos por hora. El tiempo que tarda en atender cada vehículo es constante, de exactamente 3 minutos por vehículo.

Se desea modelar el sistema como un M/D/1 (llegadas Poisson, tiempo de servicio determinístico, un servidor).

Tasa de llegada → $\boldsymbol{\lambda} = 15$ autos/hora

Tiempo de servicio → 3 minutos por vehículo

Varianza del **tiempo de servicio** → $\sigma^2 = 0$

Ejemplo práctico 2

Tasa de llegada → $\lambda = 15$ vehículos/hora
T. de servicio → 3 min/vehículo
Var. del **t. de servicio** → $\sigma^2 = 0$

Tasa de servicio μ

$$\mu = \frac{60}{3} = 20 \text{ autos por hora}$$

Utilización del sistema ϱ

$$\rho = \frac{\lambda}{\mu} = \frac{15}{20} = 0.75$$

El servidor está ocupado el **75% del tiempo**

Ejemplo práctico 2

Tasa de llegada → $\lambda = 15$ vehículos/hora
T. de servicio → 3 min/vehículo
Var. del **t. de servicio** → $\sigma^2 = 0$
Tasa de servicio $\mu = 20$ vehículos/hora
Utilización del sistema $\rho = 0.75$

E[T] (esperanza de tiempo en el sistema)

$$\mathbb{E}[T] = \frac{1}{\mu(1-\rho)} \cdot \left(1 - \frac{\rho}{2}\right)$$

$$\mathbb{E}[T] = \frac{1}{20(1-0.75)} \cdot \left(1 - \frac{0.75}{2}\right) = \frac{1}{20 \cdot 0.25} \cdot (1 - 0.375)$$

$$= \frac{1}{5} \cdot 0.625 = 0.125 \text{ horas} = 7.5 \text{ minutos}$$

Ejemplo práctico 2

Tasa de llegada → $\lambda = 15$ vehículos/hora
T. de servicio → 3 min/vehículo
Var. del **t. de servicio** → $\sigma^2 = 0$
Tasa de servicio $\mu = 20$ vehículos/hora
Utilización del sistema $\rho = 0.75$
Esp. t. en sistema **E[T]** $= 7.5$ min

E[n] (número promedio de clientes)

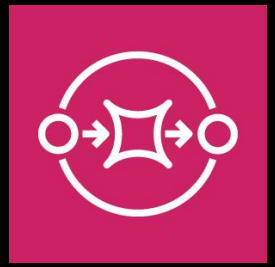
$$\mathbb{E}[n] = \lambda \cdot \mathbb{E}[T] = 15 \cdot 0.125 = 1.875 \text{ autos}$$

El sistema presenta una **utilización del 75%**

En promedio:

- Hay **menos de 2 vehículos** en el sistema en cualquier momento (**1.875**)
- Los vehículos **permanecen 7,5 minutos** en el sistema

Simple Queue Service



¿Preguntas?