Pablo Serrano Yáñez-Mingot,
José Alberto Hernández Gutiérrez

Una introducción amable a la Teoría de Colas

APUNTES DE TEORÍA DE REDES — CURSO 23/24

Departamento de Ingeniería Telemática
Universidad Carlos III de Madrid

Control de versiones

- (2024-01-17) Última versión compilada.
- (2023-12-07) Versión inicial curso 23/24.

Página web

La última versión de los apuntos, así como un boletín de problemas, siempre se encontrará disponible en la siguiente página web:
`https://www.it.uc3m.es/pablo/teoria-colas/`

Agradecimientos

A los lectores presenciales: José Luis Vázquez por los primeros ánimos, Ignacio Soto por mirar las primeras versiones de este documento, Jaime García por sugerir varias correcciones y Jorge Martín por detectar otros fallos.

Y a los lectores *remotos* que también han detectado errores y han tenido la paciencia y amabilidad de comunicarlos: Carlos Andreoli, Adrià Casmitjana.

Corrección de errores

Se agradece cualquier información sobre posible errores numéricos, conceptuales o de cualquier otro tipo. Para informar de ellos, basta con mandar un correo a:
`pablo~arroba~it~punto~uc3m~punto~es`

Índice general

Repaso de estadística y probabilidad

Se dice que el tratamiento moderno de la probabilidad nace cuando Antoine Gombaud, conocido como *Caballero de Méré* planteó a Blaise Pascal el siguiente problema: «¿Qué es más probable, sacar al menos un seis en cuatro tiradas de un dado, o sacar al menos un doble seis en veinticuatro tiradas de dos dados?» La motivación del problema es que De Méré era un gran aficionado a las apuestas, y dadas unas posibilidades para apostar, deseaba saber las ganancias esperables con cada una antes de escoger la mejor opción.[1]

[1] Su razonamiento (incorrecto) era el siguiente: en el primer caso, la probabilidad de un seis es de 1/6, por lo que en cuatro tiradas la probabilidad será de 4×1/6, esto es, 2/3. En el segundo caso, la probabilidad de un doble seis es de 1/36, por lo que en veinticuatro tiradas dicha probabilidad será 24×1/36, esto es, 2/3

En el contexto del problema planteado se pueden realizar (al menos) dos tipos de análisis

- Describir cuantitativamente el comportamiento de un dado, con objeto de «caracterizar» mediante una serie de análisis y variables su incertidumbre: de esto trata la *estadística descriptiva*.
- Predecir cómo de «esperable» es un determinado resultado, partiendo de un modelo del comportamiento de los dados: de esto trata la *teoría de la probabilidad*.

Por lo tanto, se podría decir que la estadística descriptiva parte de un conjunto de datos para estimar un modelo de comportamiento, mientras que la teoría de la probabilidad parte de unos modelos de comportamiento dados para predecir el futuro.

Estadística descriptiva

Sea una serie de n observaciones $O = \{o_i\}_{i=1}^{n}$ numéricas de un determinado suceso gobernado por el azar, esto es, de una *variable aleatoria*. A continuación se describen algunas de las principales herramientas que pueden emplearse para caracterizar dicha serie O.

Histograma

El histograma consiste en una representación gráfica en forma de barras, donde el eje horizontal (de abscisas) abarca el conjunto de posibles valores de O (al menos, entre su mínimo y su máximo), y la altura de cada barra indica bien la frecuencia relativa del conjunto de datos que abarca su base, bien el número total de muestras que quedan en dicha base.

Ejemplo 1.1. Veinte tiradas consecutivas de un dado (de seis caras) arrojan los siguientes resultados

$$X = \{3, 1, 6, 3, 4, 2, 4, 5, 2, 1, 1, 2, 2, 4, 6, 5, 4, 4, 3, 1\}.$$

Un histograma correspondiente a dicho conjunto X es el representado en la Figura 1.1 al margen, donde la altura de cada barra indica el número de muestras que caen en dicha base.

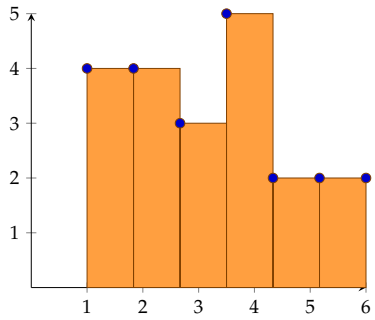


Figura 1.1: Histograma del conjunto de datos X del Ejemplo 1.1.

EL HISTOGRAMA ES UNO DE LOS ANÁLISIS que menos información descarta sobre un conjunto de datos, aunque descarta cualquier tipo de información sobre su orden: or ejemplo, los conjuntos de datos $X_1 = \{0, 0, 0, 1, 1, 1\}$ y $X_2 = \{0, 1, 0, 1, 0, 1\}$ tienen el mismo histograma. A continuación, se presentan descriptores numéricos del conjunto de datos, que proporcionan otras caracterizaciones parciales del mismo.

Moda, mediana y percentiles

A partir del histograma (aunque no necesariamente), se pueden obtener dos descriptores del conjunto de datos:

- La *moda*, que es el valor que más se repite.
- La *mediana*, que es el valor tal que la mitad del conjunto de datos queda por debajo del mismo (y, por lo tanto, la otra mitad queda por encima).

Para el conjunto de datos X del Ejemplo 1.1, se puede deducir fácilmente que la moda es 4, dado que se corresponde con la barra más alta según la Figura 1.1. Para obtener la mediana, se puede acudir al *histograma acumulado*: partiendo de la misma división del histograma original, en cada conjunto de datos se cuentan las muestras correspondientes a dicho conjunto y todas las muestras anteriores (por lo que el valor de la última barra será el número total de muestras n). El valor que coincida con la mitad del número de muestras (esto es, $n/2$) será la mediana, lo que para el caso del Ejemplo 1.1, según se aprecia en la Figura 1.2, es el 3.

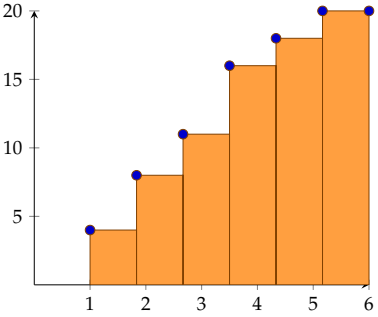


Figura 1.2: Histograma *acumulado* del conjunto de datos X del Ejemplo 1.1.

SI EL CONJUNTO DE DATOS DADO REPRESENTA una serie de valores, y es necesario realizar una «predicción» sobre el siguiente valor, la *moda* podría ser un buen candidato dado que parece ser el valor más frecuente; la *mediana* podría ser otro candidato razonable, dado que se queda «a la mitad» del conjunto de datos observado.

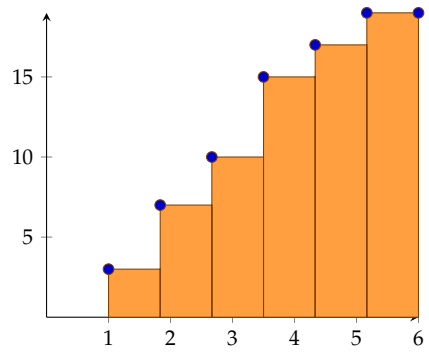
Como se ha visto, para calcular la mediana basta con ordenar el conjunto de datos de menor a mayor e identificar el elemento central del mismo. De forma análoga, se pueden definir diversos *percentiles*, en función del número de observaciones que queden por debajo del elemento: el percentil 25 (P_{25}) es aquel elemento mayor que el 25 % de las muestras, el percentil 50 coincide con la mediana, el percentil 75 es aquel valor mayor que el 75 % de las

observaciones, etc. A estos percentiles también se les conoce como primer, segundo y tercer cuartil, respectivamente, dado que dividen el conjunto de datos en cuatro partes proporcionalmente iguales, y se representan como Q_1, Q_2 y Q_3. Mediante el uso de los percentiles adecuados se puede tener una cierta idea de la dispersión de los datos. En concreto, el *rango intercuartil* (RQ) se define como la diferencia entre el tercer y el primer cuartil.

Ejemplo 1.2. Sea otra vez el caso del vector X del Ejemplo 1.1, de valor

$$X = \{3, 1, 6, 3, 4, 2, 4, 5, 2, 1, 1, 2, 2, 4, 6, 5, 4, 4, 3, 1\}.$$

El valor de Q_2, como se ha visto, corresponde con la mediana, y se puede obtener fácilmente con el histograma acumulado: basta con identificar el valor que divide por la mitad el conjunto ordenado de muestras. Dado que hay 20 muestras, dicha mitad es el número 10 y por lo tanto $Q_2 = 3$ (véase figura al margen). El primer cuartil Q_1 es a su vez la mediana de la primera mitad de valores, mientras que el tercer cuartil Q_3 es la mediana de la segunda mitad de valores. Por lo tanto, $Q_1 = 2$ y $Q_3 = 4$. El rango intercuartil sería, por lo tanto, $RQ = 2$.



Repetición de la Figura 1.2: Histograma acumulado de X.

De forma parecida al caso del histograma, aunque más severa, tanto la moda como la mediana no sirven para caracterizar inequívocamente un conjunto de datos: por ejemplo, los conjuntos

$$O_1 = \{5, 1, 5, 100\}$$
$$O_2 = \{4, 5, 5, 6\}$$

tienen el mismo valor de mediana y moda.

Media

Otro parámetro muy usado para describir el conjunto de datos es la *media* (o media aritmética), que se obtiene de calcular

$$\bar{o} \triangleq \frac{1}{n} \sum_{o_i \in O} o_i.$$

Si una determinada observación o_i se repite n_i veces, el cálculo se puede expresar como

$$\bar{o} = \sum_{i=1}^{n} \frac{n_i}{n} o_i,$$

donde n_i/n coincide con la *frecuencia relativa* de la observación o_i (esto es, número de veces que aparece o_i entre el total de observaciones).

Ejemplo 1.3. Sea un equipo de fútbol que, durante 10 jornadas, marca los siguientes goles:

$$G = \{3, 3, 4, 2, 0, 2, 4, 3, 0, 4\}$$

En estas 10 jornadas, la media del número de goles marcados es:

$$\overline{G} = \frac{3+3+4+2+2+4+3+4}{10} = 2.5.$$

que se representa en la Figura 1.3.

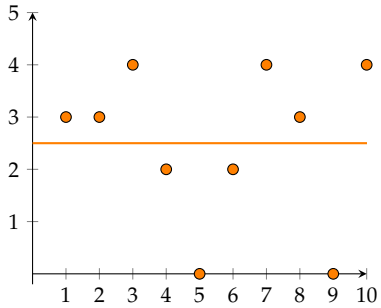


Figura 1.3: Goles y media del equipo del Ejemplo 1.3.

LA MEDIA DE UN SUBCONJUNTO DE LAS OBSERVACIONES, que viene definido por una *condición* expresada sobre el conjunto global, se denomina *media condicionada*. Si $O_C \subseteq O$ es el subconjunto de O que cumple con la condición C, con m elementos, entonces la media condicionada viene dada por

$$\overline{o_C} = \frac{1}{m} \sum_{o_i \in O_C} o_i.$$

Ejemplo 1.4. Sea el caso del Ejemplo 1.3. A partir del conjunto de goles G se puede hacer el subconjunto de los resultados en los que el equipo marcó al menos un gol, esto es

$$G_{g>0} = \{3,3,4,2,2,4,3,4\},$$

cuya media viene dada por

$$\overline{G_{g>0}} = \frac{3+3+4+2+2+4+3+4}{8} = 3.125.$$

A DIFERENCIA DE LA MEDIANA Y LA MODA, que sólo se basan en las frecuencias relativas de cada valor, y son un valor del conjunto de datos, la media multiplica cada valor por su frecuencia relativa, por lo que en general no coincide con un valor del conjunto (en el caso del Ejemplo 1.4 anterior, los goles sólo pueden ser números naturales). La media es el número que minimiza la distancia a los datos,[2] de ahí su utilidad para predecir el comportamiento de una variable aleatoria (como se ilustra en la Figura 1.3, donde los goles aparecen «alrededor» del valor indicado por la media).

[2] Esto es, la media coincide con

$$\bar{o} = \min_{x \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (o_i - x)^2$$

Varianza y desviación típica

De forma similar al rango intercuartil, las métricas de varianza y desviación típica aportan información sobre la *dispersión* del conjunto de datos observado. Para ello, realizan un cálculo a partir de las distancias entre cada muestra y valor medio del conjunto. La *varianza*, que se representa como σ^2, es la media del cuadrado de la distancia entre cada muestra o_i y la media $\bar{o}$,

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (o_i - \bar{o})^2,$$

por lo que tiene como unidad el cuadrado de la unidad del conjunto de muestras. La *desviación típica* σ emplea la misma unidad que

dicho conjunto, dado que se calcula como la raíz cuadrada de la varianza:

$$\sigma^2 = \sqrt{\sigma^2}.$$

Para el caso del Ejemplo 1.3, la desviación típica es $\sigma \approx 1{,}51$ goles y $\sigma^2 \approx 2{,}78$ goles$\times$ goles (cuesta encontrar una interpretación para la unidad «goles^2»).

Definición de probabilidad

Cuando se está tratando con un suceso aleatorio, la estadística descriptiva sirve para realizar un análisis *a posteriori* de los resultados obtenidos. Por contra, la teoría de la probabilidad se emplea para estimar *a priori* los resultados que se esperan obtener.

Desde el punto de vista *clásico*, la probabilidad de un suceso es una variable que representa su frecuencia relativa (en un experimento controlado): para el caso de una moneda, decir que la «probabilidad de que salga cara» sea $1/2$ significa que en, p.ej., mil lanzamientos, la cantidad de veces que saldrá cara será muy próxima a quinientos.

Ejemplo 1.5. Sea el caso de un equipo de fútbol (FCB) con las estadísticas indicadas en la Tabla 1.1 al margen, para un histórico de 1000 partidos. En dicha tabla se indica el número de partidos ganados, empatados y perdidos por dicho equipo, distinguiéndose también entre cuando se enfrenta a otro equipo (RMA) y contra el resto de equipos.

Resultado	Total	vs. RMA	Resto
Ganar	900	5	895
Empatar	65	10	55
Perder	35	15	20
Total	1000	30	970

Tabla 1.1: Resultados de un equipo FCB al jugar contra el resto de equipos y contra el RMA.

Considerando los resultados globales, el resultado «ganar» aparece en 900 de los 1000 experimentos. En este contexto, cuando se dice que el equipo FCB tiene una probabilidad de 0,9 (o del 90 %) de ganar un partido, se está suponiendo que cada partido es un «experimento» independiente y que la probabilidad del evento «ganar» se corresponde con la frecuencia relativa medida de dicho evento. De esta forma, se parte de la estadística descriptiva de los resultados del equipo de fútbol para construir un modelo que permita predecir (o explicar) su comportamiento, según la teoría de la probabilidad

Dado que la interpretación clásica de la probabilidad no se ajusta a todos los casos en que aparece dicho término, p.ej. cuando el experimento no es repetible, aparece una definición *subjetiva* de la misma: la probabilidad de un suceso es un número que se emplea para estimar la posibilidad de que dicho suceso ocurra, donde 0 representa su imposibilidad y 1 la absoluta certeza. Se define la probabilidad complementaria de un suceso A como la probabilidad de que dicho suceso no tenga lugar. Si $\Pr(A)$ representa la probabilidad del suceso A, su probabilidad complementaria es

$$\Pr(\neg A) \triangleq 1 - \Pr(A).$$

Variables aleatorias discretas

De momento, se considerará que el conjunto de posibles resultados es un espacio muestral contable, esto es, el caso de *variables aleatorias discretas*. A cada uno de los elementos de dicho conjunto $\mathbb{S}$ se le asigna una «probabilidad»,[3] representada como $\Pr(s_i)$, $s_i \in \mathbb{S}$, que debe cumplir las siguientes propiedades

$$0 \leq \Pr(s_i) \leq 1$$
$$\sum \Pr(s_i) = 1$$

[3] La S es por el nombre en inglés, *sample space*.

Ejemplo 1.6. Tirar un dado es un *suceso aleatorio*, y el número que salga es el resultado de un *experimento*. El resultado de una tirada de un dado es una *observación*. El conjunto de todas las posibles observaciones es el *espacio muestral*. Para el caso de un dado de seis caras, dicho espacio $\mathbb{S}$ es

$$\mathbb{S} = \{1, 2, 3, 4, 5, 6\}$$

Si el dado está debidamente equilibrado, se puede suponer que la probabilidad de cada resultado es la misma, esto es, $\Pr(s_i) = 1/6$ para todos los eventos. Con un número muy elevado de experimentos, la proporción de veces que se obtiene un determinado resultado se aproximará a dicha probabilidad.

La memoria en una variable aleatoria

Uno de los primeros «desencuentros» entre intuición y realidad suele estar relacionado con la memoria (más bien, su ausencia) en un experimento aleatorio, como p. ej. el lanzamiento del dado o una moneda. En este último caso, aún sabiendo que las probabilidades de cara y cruz son idénticas, si resulta que p. ej. cinco lanzamientos consecutivos son caras, la «intuición» lleva a pensar que el siguiente lanzamiento *tiene* que ser cruz. Sin embargo, la moneda no dispone de esta «memoria» del proceso aleatorio, por lo que la probabilidad de cara (o cruz) es la misma que en cada lanzamiento.

Sea X_n el resultado del lanzamiento enésimo, donde H indica cara y T indica cruz.[4] La probabilidad de que el sexto lanzamiento sea cara, sabiendo que los anteriores lanzamientos también ha sido cara,[5] se puede representar como

$$\Pr(X_6 = H \mid X_1 = H, X_2 = H, X_3 = H, X_4 = H, X_5 = H),$$

que se trata de una probabilidad *condicionada*, donde la barra $\mid$ separa el evento «de interés» (izquierda) de la condición dada (derecha).[6] El hecho de que los lanzamientos no tengan memoria supone que, independientemente del número y resultado de los lanzamientos anteriores, la probabilidad de obtener cara es siempre la misma, esto es

$$\Pr\left(X_6 = H \mid \{X_i = H_i\}_{i=1}^{5}\right) = \Pr(X_6 = H) = \frac{1}{2}$$

[4] Por sus nombres en inglés, esto es, *heads* and *tails*.

[5] No hay que confundir esta probabilidad con la probabilidad de obtener una secuencia de 6 caras, mucho menor.

[6] La probabilidad condicionada se define formalmente en la siguiente sección

lo que se cumplirá para cualquier lanzamiento m y conjunto de posibles resultados anteriores $\{R_i\}_{i=1}^{m-1}$.

No todos los procesos aleatorios carecen de memoria. Para el caso del «juego» de la *ruleta rusa*, si no se hace rodar el tambor en cada intento, la probabilidad de que al apretar el gatillo el arma dispare no sería constante a lo largo del tiempo, sino que va aumentando según el turno (compárense las Figuras 1.4 y 1.5). Si X_n es el resultado del turno n-ésimo, 0 representa que no hubo disparo y 1 que sí lo hubo, se tendrá que la probabilidad de que en el quinto turno haya un disparo (sabiendo que en los cuatro anteriores no lo hubo) es mayor que la probabilidad de un disparo en el primer intento:

$$\Pr(X_5 = 1 \mid X_1 = 0, X_2 = 0, X_3 = 0, X_4 = 0) > \Pr(X_1 = 1)$$

por lo que dicha variable aleatoria sí que posee memoria.

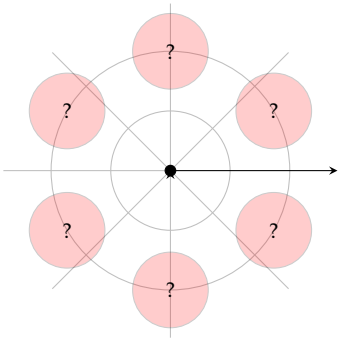


Figura 1.4: En un primer momento, la probabilidad de que la bala esté en cualquier hueco es de uno entre seis.

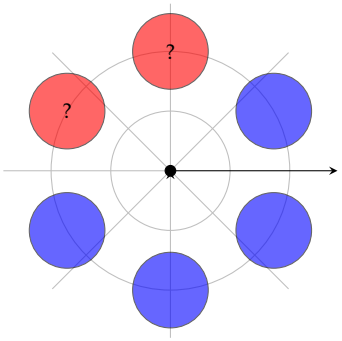


Figura 1.5: En esta variante del juego, tras cuatro disparos la probabilidad ha cambiado notablemente.

Probabilidad condicionada

Cuando se consideran varios eventos, para averiguar si existe relación entre ellos se puede analizar cómo se produce uno de los mismos, cuando «se fija» un valor particular en el otro: por ejemplo, para el caso de un equipo de fútbol (Ejemplo 1.1), se puede analizar con qué frecuencia se produce el evento «Ganar» en función de si se juega contra el equipo RMA o contra otro equipo; al lanzar dos dados, se puede analizar si el valor que se obtiene en uno de ellos cuando en el otro se obtiene un determinado número.

La probabilidad del evento A condicionada al evento B se define como

$$\Pr(A \mid B) \triangleq \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A, B)}{\Pr(B)},$$

donde $\Pr(A \cap B)$ representa la probabilidad de que sucedan A y B (el suceso conjunto). Por simplificar la notación, en general se sustituirá el símbolo de intersección ($\cap$) por una sencilla coma (como se ha realizado en la ecuación anterior), quedando esta probabilidad representada como $\Pr(A, B)$.

Ejemplo 1.7. Sea el caso de dos dados, A y B, que se tiran a la vez 1750 veces, con los resultados de la Tabla 1.2. La probabilidad de sacar un 4 con el dado A cuando sale un 3 en el dado B se expresaría como:

$$\Pr(A = 4 \mid B = 3) = \frac{\Pr(A = 4, B = 3)}{\Pr(B = 3)},$$

que precisa: (*i*) por una parte, calcular la frecuencia relativa del evento

$$\Pr(A = 4, B = 3),$$

que sucede en 40 de los 1750 casos (según se aprecia en la tabla), y (*ii*) por otra parte, calcular la frecuencia relativa del evento

$$\Pr(B = 3),$$

A	B					
	1	2	**3**	4	5	6
1	45	45	43	41	49	58
2	58	55	42	49	49	44
3	53	47	40	44	44	45
4	47	60	**40**	48	58	46
5	51	59	57	50	41	54
6	45	56	42	60	48	53

Tabla 1.2: Resultado de lanzar 1750 veces los dados A y B.

que sucede en 43+42+40+40+57+42=264 casos. Por lo tanto,

$$\Pr(A = 4 \mid B = 3) = \frac{40/1750}{264/1750} = \frac{40}{264} = \frac{10}{66} \approx 0{,}151,$$

que es similar a la probabilidad de que el dado A saque un 4:

$$\Pr(A = 4) = \frac{47 + 60 + 40 + 48 + 58 + 46}{1750} = \frac{299}{1750} \approx 0{,}166\ ,$$

lo que, como se verá más adelante, no es casualidad.

Ejemplo 1.8. Sea otra vez el caso del Ejemplo 1.5, con los 1000 partidos del equipo FCB resumidos al margen, donde se supone que las frecuencias relativas sirven para modelar el sistema con probabilidades.

Resultado	Total	vs. RMA	Resto
Ganar	900	5	895
Empatar	65	10	55
Perder	35	15	20
Total	1000	30	970

Repetición de la Tabla 1.1

La probabilidad de que el equipo FCB gane se estimó como

$$\Pr(G) = \frac{900}{1000} = \frac{9}{10}\ ,$$

bastante próxima a la unidad. También se puede calcular la probabilidad de que FCB juegue contra RMA, que se obtiene como

$$\Pr(RMA) = \frac{30}{1000} = \frac{3}{100}\ .$$

La probabilidad de que un partido sea una victoria contra RMA se obtiene como

$$\Pr(G, RMA) = \frac{5}{1000} = \frac{1}{200}\ ,$$

mientras que la probabilidad de ganar condicionada a jugar contra el RMA (o, lo que es lo mismo, la probabilidad de que un partido contra RMA sea una victoria) se calcula como

$$\Pr(G \mid RMA) = \frac{\Pr(G, RMA)}{\Pr(RMA)} = \frac{1/200}{3/100} = 1/6,$$

que es mucho menor que $\Pr(G)$, la probabilidad de ganar sin distinguir entre equipos.[7] Este resultado implica que la probabilidad de ganar depende del equipo contra el que se juegue.

[7] Otra forma de calcular $\Pr(G \mid RMA)$ consiste en considerar únicamente los resultados de la columna «vs. RMA» de la tabla, de lo que se obtiene

$$\Pr(G \mid RMA) = \frac{5}{30} = 1/6.$$

Distribución marginal

Como se ha visto en los ejemplos anteriores, a veces se dispone de información desglosada sobre dos tipos de eventos diferentes, pero se precisa únicamente caracterizar el comportamiento de uno de ellos. En una situación con dos variables aleatorias, X e Y, la *distribución marginal* de X se define como la distribución de dicha variable sin hacer referencia a la otra variable Y, y se obtiene directamente para cualquier valor x como

$$\Pr(X = x) = \sum_y \Pr(X = x, Y = y)$$

Ejemplo 1.9. Siguiendo con el caso de los dados del Ejemplo 1.7, el número de veces que con el dado A se sacó un 1, 2, ..., 6 es, respectivamente, 281, 297, 273, 299, 300 y 300, según se representa en la Tabla 1.3. Dividiendo estos valores por el número total de lanzamientos (1750), se puede estimar la distribución marginal de los resultados del dado A:

$$\Pr(A) \approx \{0.161, 0.169, 0.156, 0.171, 0.171, 0.171\} ,$$

todos ellos valores muy próximos a 1/6, como era de esperar.

A	B						
	1	2	3	4	5	6	
1	45	45	43	41	49	58	281
2	58	55	42	49	49	44	297
3	53	47	40	44	44	45	273
4	47	60	40	48	58	46	299
5	51	59	57	50	41	54	300
6	45	56	42	60	48	53	300
	299	322	264	293	298	300	1750

Tabla 1.3: Resultado de lanzar los dados (mismos valores que en la Tabla 1.2), incluyendo los valores totales para el dado A (última columna) y para el dado B (última fila).

Independencia

Como se ha indicado, la probabilidad condicionada permite identificar si los diferentes tipos de eventos en un conjunto de observaciones están relacionados entre sí: en el caso del Ejemplo 1.8 se ha comprobado que la probabilidad de ganar un partido depende del contrincante, mientras que en el Ejemplo 1.9 se aprecia que los resultados de un dado no dependen de los de otro dado. Este tipo de análisis permiten definir la independencia entre diferentes eventos, como se hace a continuación.

Dos sucesos son *independientes* si el hecho de que uno suceda no afecta a que se pueda producir el otro: si A y B son independientes, la probabilidad de que A se produzca no varía en función de que B se haya producido o no (esto es, de que se considere B como condición). Por lo tanto, la probabilidad condicionada permite comprobar si A y B son independientes, dado que en tal caso se tiene que cumplir que

$$\Pr(A \mid B) = \Pr(A) ,$$

lo que no ocurre en el caso del fútbol (Ejemplo 1.8), pero sí en el de los dados (Ejemplo 1.7).

No se debe confundir independencia con *exclusión*: dos eventos A y B son mutuamente excluyentes si resulta imposible que sucedan conjuntamente:

$$\Pr(A \cap B) = \Pr(A, B) = 0$$

de lo que se deduce que la probabilidad condicionada es

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)} = 0,$$

lo que no coincide con la definición de independencia.

Ejemplo 1.10. Los eventos «cara» y «cruz» de una moneda son mutuamente excluyentes, ya que el hecho de que se produzca uno implica que el otro no se ha producido (por lo que no son independientes). Un jugador que *nunca* marcase un gol al equipo RMA implicaría que el evento «marcar gol» no es independiente del evento «jugar contra RMA», lo que permite deducir, p. ej., que si el jugador ha marcado un gol es que *no* jugaba contra dicho equipo (lo que ilustra que son eventos dependientes).

Múltiples condiciones - regla de la cadena (*)

Se pueden aplicar las reglas de la probabilidad condicionada para desarrollar relaciones con varias condiciones. Sea el caso de tres eventos A, B y C. Suponiendo que la condición sea C, se puede escribir la probabilidad de A, B condicionada a C como

$$\Pr(A, B \mid C) = \frac{\Pr(A, B, C)}{\Pr(C)}$$

mientras que suponiendo la condición sea B, C, se puede escribir que

$$\Pr(A, B, C) = \Pr(A \mid B, C) \Pr(B, C)$$

Aplicando esta ecuación en el numerador de la expresión anterior, se tiene

$$\Pr(A, B \mid C) = \frac{\Pr(A \mid B, C) \Pr(B, C)}{\Pr(C)}$$

que por definición de probabilidad de B condicionada a C, resulta

$$\Pr(A, B \mid C) = \Pr(A \mid B, C) \Pr(B \mid C)$$

En un caso general, se puede aplicar la *regla de la cadena* o regla del producto general:

$$\begin{aligned}\Pr(X_1, X_2, X_3, X_4) &= \Pr(X_1 \mid X_2, X_3, X_4) \Pr(X_2, X_3, X_4) \\ &= \Pr(X_1 \mid X_2, X_3, X_4) \Pr(X_2 \mid X_3, X_4) \Pr(X_3, X_4) \\ &= \Pr(X_1 \mid X_2, X_3, X_4) \Pr(X_2 \mid X_3, X_4) \Pr(X_3 \mid X_4) \Pr(X_4)\end{aligned}$$

Propiedades de la probabilidad

Dados dos sucesos A y B, la probabilidad de que *alguno* de los dos suceda se obtiene como (Figura 1.6 al margen)

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A, B).$$

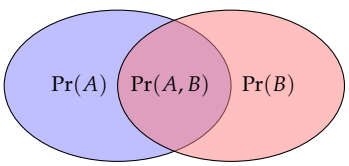


Figura 1.6: La probabilidad de A o B debe tener en cuenta el caso conjunto.

Si A y B son *excluyentes*, se ha visto que entonces

$$\Pr(A, B) = \Pr(A \mid B) \Pr(B) = 0\ ,$$

por lo que queda

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A, B) = \Pr(A) + \Pr(B).$$

Si A y B son *independientes*, la probabilidad del suceso conjunto viene dada por

$$\Pr(A \cap B) = \Pr(A, B) = \Pr(A \mid B) \Pr(B) = \Pr(A) \Pr(B).$$

Ejemplo 1.11. La probabilidad de que un dado saque un cinco o un número par, dado que son eventos excluyentes, viene dada por

$$\Pr(5 \cup \text{par}) = \Pr(5) + \Pr(\text{par}) = \frac{1}{6} + \frac{1}{2} = \frac{2}{3}\,.$$

La probabilidad de que salga un número par o un valor mayor que tres sería

$$\Pr(>3 \cup \text{par}) = \Pr(>3) + \Pr(\text{par}) - \Pr(>3 \cap \text{par}) = \frac{1}{2} + \frac{1}{2} - \frac{2}{6} = \frac{2}{3}\,.$$

Ley de la probabilidad total

Si $\{B_i\}$ es una *partición* del espacio de eventos,[8] la distribución marginal (página 12) se puede expresar en términos de la probabilidad condicionada

$$\Pr(A) = \sum_i \Pr(A, B_i) = \sum_i \Pr(A \mid B_i) \Pr(B_i)$$

Según cómo sea de compleja la partición, será recomendable comprobar que cumple todos los casos de interés, esto es

$$\sum_i \Pr(B_i) = 1$$

[8] Una partición de un conjunto A es una colección de subconjuntos de A que (*i*) su unión da lugar a A, y (*ii*) los subconjuntos son disjuntos dos a dos. Por ejemplo, dado el conjunto $S = \{1, 2, 3, 4\}$, Una partición sería los subconjuntos $S_a = \{1, 2\}$ y $S_b = \{3, 4\}$, mientras que los subconjuntos $S_x = \{1, 2, 3\}$ y $S_y = \{3, 4\}$ no serían una partición de S.

Ejemplo 1.12. En el caso del fútbol, el subconjunto «vs. RMA» y «Resto» constituyen una partición del espacio de eventos «Total». A partir de esto, la probabilidad de ganar se puede expresar como

Resultado	Total	vs. RMA	Resto
Ganar	900	5	895
Empatar	65	10	55
Perder	35	15	20
Total	1000	30	970

$$\Pr(G) = \Pr(G \mid \text{vs. RMA}) \Pr(\text{vs. RMA}) + \Pr(G \mid \text{Resto}) \Pr(\text{Resto})$$

que queda

$$\Pr(G) = \frac{5}{30}\frac{30}{1000} + \frac{895}{970}\frac{970}{1000} = \frac{900}{1000}\,,$$

como era de esperar.

La ley de la probabilidad total permite en varias ocasiones reducir la complejidad de un problema planteado, a base de realizar una partición de dicho espacio, y calcular resultados parciales sobre el mismo.

Ejemplo 1.13. Suponga que una máquina A genera un número entero al azar, entre 1 y 5, y que otra máquina B genera otro número entre 1 y 3. Para calcular la probabilidad de que el número generado por B (n_b) sea mayor o igual que el generado por A (n_a), se puede emplear el diagrama de la figura, en el que todos los casos tienen la misma probabilidad (suponiendo que las máquinas se comporten de forma independiente):

$$\Pr(n_a = x, n_b = y) = \Pr(n_a = x)\Pr(n_b = y) = \frac{1}{5}\cdot\frac{1}{3} = 1/15$$

Para calcular $\Pr(n_b \geq n_a)$, por la sencillez del ejemplo, resulta inmediato contar el número de casos «favorables» (círculos negros, seis) sobre el total de casos (quince), lo que se obtiene

$$\Pr(n_b \geq n_a) = 6/15 = 2/5\ .$$

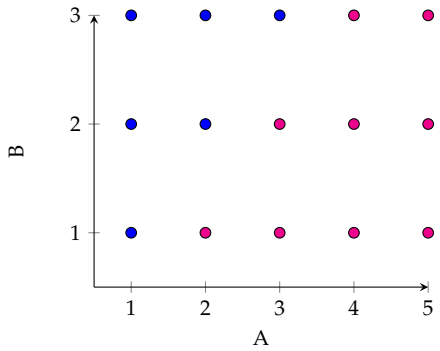


Figura 1.7: Probabilidad de que la máquina B genere un número mayor o igual que la máquina A.

Empleando la ley de la probabilidad total, se puede resolver realizando la partición sobre cada uno de los posibles resultados que arroja la máquina A, llegándose al mismo resultado:

$$\begin{aligned}\Pr(n_b \geq n_a) &= \sum_{x=1}^{5} \Pr(n_b \geq n_a | n_a = x)\Pr(n_a = x) \qquad (1.1)\\ &= \sum_{x=1}^{3} \Pr(n_b \geq n_a | n_a = x)\frac{1}{5} = (1 + 2/3 + 1/3)\cdot\frac{1}{5} = 2/5.\end{aligned}$$

Suma de variables aleatorias independientes

Gracias a la ley de la probabilidad total, se puede calcular la distribución de probabilidades de la *suma* de dos variables aleatorias independientes, a partir de la distribución de las mismas. Sean dos variables aleatorias independientes n_a y n_b. Lo que se persigue es calcular la probabilidad de que su suma tenga cualquier posible valor x, esto es,

$$\Pr(n_a + n_b = x), \quad \forall x \in [\min n_a + \min n_b, \max n_a + \max n_b]$$

Condicionando a los posibles valores que pueda tomar n_b, dicho cálculo se puede expresar como

$$\begin{aligned}\Pr(n_a + n_b = x) &= \sum_y \Pr(n_a + n_b = x \mid n_b = y)\Pr(n_b = y)\\ &= \sum_y \Pr(n_a = x - y)\Pr(n_b = y), \qquad (1.2)\end{aligned}$$

de lo que resulta que la distribución de probabilidades de la suma de dos variables aleatorias es la *convolución* de las distribuciones de probabilidad de las variables aleatorias.

Ejemplo 1.14. Para el caso de las variables n_a y n_b del Ejemplo 1.13, n_a puede tomar los siguientes valores, todos ellos con igual probabilidad

$$n_a = \{1, 2, 3, 4, 5\}$$

mientras que n_b puede tomar los siguientes valores, también equiprobables

$$n_b = \{1, 2, 3\} .$$

Por lo tanto, su suma tomará $n_a + n_b$ valores entre 2 y 8.

Resulta inmediato calcular la probabilidad de que su suma valga 2 (el mínimo), ya que al ser variables aleatorias independientes, se deduce que

$$\Pr(n_a + n_b = 2) = \Pr(n_a = 1)\Pr(n_b = 1) = \frac{1}{15} ,$$

Para calcular la probabilidad de que la suma valga p.ej. 4, el desarrollo de (1.2) resulta en

$$\begin{aligned}\Pr(n_a + n_b = 4) &= \Pr(n_a = 3)\Pr(n_b = 1) + \\ &= \Pr(n_a = 2)\Pr(n_b = 2) + \\ &= \Pr(n_a = 1)\Pr(n_b = 3) = \frac{1}{5}.\end{aligned}$$

Realizando el cálculo para todos los posibles valores de la suma se obtiene el resultado de la Figura 1.8, donde se puede comprobar que la distribución de $n_a + n_b$ se corresponde con la convolución de las distribuciones de n_a y n_b.

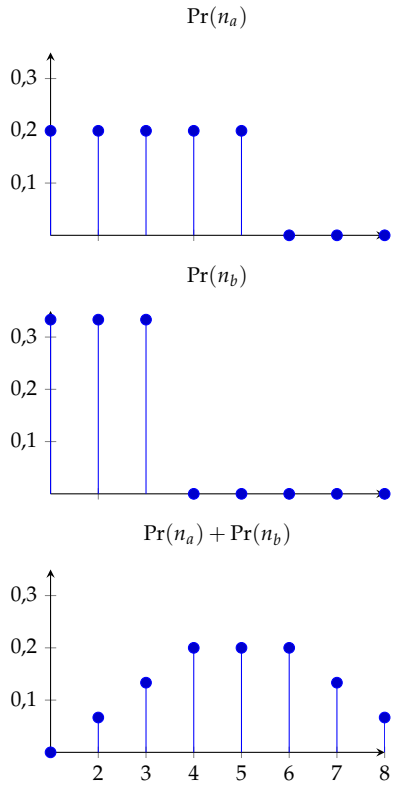


Figura 1.8: Funciones de masa de probabilidad de n_a (superior), n_b (medio) y de su suma (inferior).

Momentos y esperanzas

La media y la varianza muestrales, vistas anteriormente (páginas 7 y 8, respectivamente), son variables que persiguen caracterizar con una única cifra la *forma* del conjunto de datos. Un «momento» es una generalización de estas variables, consistente en un valor numérico que describe cómo una serie de valores se distribuye alrededor de uno que se toma como referencia. En general, el momento de orden n respecto al valor c se define como

$$\mu_n(c) \triangleq \sum (x_i - c)^n \ \Pr(x_i)$$

esto es, se realiza la suma de una «transformación» de cada posible valor x_i de la variable aleatoria X y se multiplica por su probabilidad $\Pr(x_i)$. Ponderar el valor de cada transformación por la probabilidad también se conoce como calcular la **esperanza** de dicha transformación, por lo que el momento de orden n respecto al valor c también se puede expresar como

$$\mathbb{E}[(X - c)^n] = \mu_n(c)$$

Si el valor de referencia es el 0, se trata de un momento estándar (o centrado respecto al origen). En función de los valores de n y c se pueden definir diferentes momentos:

- La **media** o **esperanza matemática** es el momento de primer orden respecto al origen, esto es,

$$\mu = \mathbb{E}[X] = \sum x_i \Pr(x_i),$$

y pretende representar la «tendencia» del conjunto de valores de la distribución.

- El momento de segundo orden se expresaría como,

$$\mu_2 = \mathbb{E}[X^2] = \sum x_i^2 \Pr(x_i),$$

y se trata de una suma de términos estrictamente positivos, proporcionando una idea de la «amplitud» de los valores de la variable aleatoria.[9]

[9] Puede interpretarse como la *energía* de una determinada variable.

- La **varianza** se define como el momento de segundo orden respecto a la media,

$$\sigma^2 = \mu_2(\mu) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum (x_i - \mu)^2 \Pr(x_i),$$

y mide también la «amplitud» de la separación entre cada posible valor y la media, proporcionando una medida de dispersión.[10] La desviación típica, como se ha visto anteriormente, consiste en la raíz cuadrada de la varianza.

[10] Que puede interpretarse como la *componente alterna* de la energía de una determinada variable.

Propiedades

Si X es una variable aleatoria y k una constante, a partir de las definiciones anteriores se pueden deducir las siguientes propiedades:

$$\begin{aligned} \mathbb{E}[kX] &= k\mathbb{E}[X] \\ \mathbb{E}[(kX)^2] &= k^2\mathbb{E}[X^2] \\ \mathbb{E}[X^2] &= \mathbb{E}[X]^2 + \sigma^2(X) \\ \sigma^2(k+X) &= \sigma^2(X) \end{aligned} \tag{1.3}$$

La expresión (1.3) es particularmente destacable, ya que relaciona el momento de segundo orden con el de primer orden y la varianza, permitiendo calcular uno en función de los otros.[11]

[11] Puede interpretarse, además, como que la energía total de una variable (momento de segundo orden) es igual a la suma de la energía de la componente continua (la esperanza al cuadrado) más la energía de la componente alterna (la varianza).

Si X e Y son dos variables aleatorias (independientes, o no), se tiene que la media de su suma coincide con la suma de las medias

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

y si además son variables aleatorias independientes, entonces la varianza de su suma es la suma de las varianzas

$$\sigma^2(X + Y) = \sigma^2(X) + \sigma^2(Y)$$

Esperanza condicionada

De forma análoga al caso de la media condicionada, visto en estadística descriptiva, se puede definir la esperanza de una variable aleatoria condicionada a que se cumpla un cierto requisito. De esta forma, la esperanza de una variable aleatoria X condicionada a que otra variable aleatoria Y valga y se puede expresar como

$$\mathbb{E}[X \mid Y = y] = \sum_X x \Pr(X = x \mid Y = y) = \sum \frac{x \cdot \Pr(X = x, Y = y)}{\Pr(Y = y)}$$

Ejemplo 1.15. Sea una variable aleatoria X que puede tomar los valores $\{1,2,3,4\}$ con probabilidades $\{1/5,1/5,1/5,2/5\}$, respectivamente. La esperanza de dicha variable aleatoria es

$$\mathbb{E}[X] = 1 \cdot \frac{1}{5} + 2 \cdot \frac{1}{5} + 3 \cdot \frac{1}{5} + 4 \cdot \frac{2}{5} = \frac{14}{5} \, .$$

La esperanza de X condicionada a que tome valores menores o iguales que 2, en cambio, es

$$\mathbb{E}[X \mid X \leq 2] = \frac{1 \cdot \Pr(X = 1) + 2 \cdot \Pr(X = 2)}{\Pr(X = 1) + \Pr(X = 2)} = \frac{3}{2}.$$

La esperanza condicionada, al igual que la probabilidad condicionada, permite resolver problemas al dividirlos en casos más sencillos de analizar, permitiendo en algunos casos aplicar una especie de «recursión», como se ilustra a continuación con un par de ejemplos.

Ejemplo 1.16. Una tarjeta inalámbrica transmite tramas de 1000 bits a 1 Mbps y 2 Mbps, seleccionando cada tasa de forma completamente aleatoria y retransmitiendo hasta que sea recibida con éxito. Las tramas a 1 Mbps se reciben con éxito el 50 % de las veces, mientras que las tramas a 2 Mbps nunca se transmiten con éxito.

Una forma de calcular el número medio de intentos N necesario hasta que se transmite con éxito es mediante la probabilidad condicionada, como se describe a continuación. La probabilidad de que el primer intento sea un éxito es

$$\Pr(\text{éxito intento 1}) = \Pr(\text{1 Mbps}) \ \Pr(\text{éxito a 1 Mbps}) = \frac{1}{4}$$

La esperanza pedida se puede expresar como

$$\begin{aligned}\mathbb{E}[N] =& \mathbb{E}[N \mid \text{éxito intento 1}] \Pr(\text{éxito intento 1}) + \\ & \mathbb{E}[N \mid \text{fallo intento 1}] \Pr(\text{fallo intento 1}) \, ,\end{aligned}$$

donde el valor de $\mathbb{E}[N \mid \text{éxito intento 1}]$ es directamente 1, mientras que

$$\Pr(\text{fallo intento 1}) = 1 - \Pr(\text{éxito intento 1}).$$

Queda, por último, obtener la expresión de $\mathbb{E}[N \mid \text{fallo intento 1}]$, esto es: el número medio de intentos si el primer intento ha sido un fallo. Dado que cada intento es independiente, y el proceso no tiene memoria alguna, tras el primer fallo el sistema se encuentra en las mismas condiciones que al principio, salvo que ya ha pasado un intento, por lo que se puede deducir que

$$\mathbb{E}[N \mid \text{fallo intento 1}] = 1 + \mathbb{E}[N] \, ,$$

Resolviendo la ecuación resultante, se obtiene que

$$\mathbb{E}[N] = 4 \, .$$

Ejemplo 1.17. Sea una grupo de excursionistas perdidos en una mina. En un momento dado llegan a una sala, con tres diferentes rutas. Una de ellas les lleva a la salida tras 10 h de camino, mientras que las otras dos les lleva de vuelta a la misma sala, tras 5 h y 20 h de camino. Suponga además que cada vez que llegan a dicha sala escogen una de las tres rutas al azar.

Sea el T el tiempo que tardan en salir de la mina una vez que han llegado a dicha sala. Condicionando a la elección de cada ruta (indicadas como A, B y C en la Figura 1.9), se tiene que la esperanza de dicho tiempo es

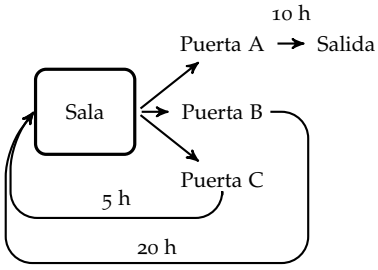


Figura 1.9: Sala con tres posibles caminos.

$$E[T] = E[T|A]\Pr(A) + E[T|B]\Pr(B) + E[T|C]\Pr(C).$$

Donde $E[T|A] = 10\ h$ es el tiempo que tardan en salir una vez que han acertado la ruta adecuada, y $\Pr(A) = \Pr(B) = \Pr(C) = 1/3$ por el mecanismo de elección de ruta (que se supone completamente aleatorio). En el caso de una ruta que no sea A, lo que se tiene es que se vuelve a la misma situación que al inicio, por lo que

$$E[T|B] = 20\ h + E[T]\ \ ,\ \ E[T|C] = 5\ h + E[T].$$

Resolviendo para $E[T]$, se tiene que $E[T] = 35$ h.

Distribuciones de probabilidad discretas

A continuación se presentan algunas de las distribuciones de probabilidad discretas más comunes. Estas distribuciones vienen definidas por la *función (de masa) de probabilidad*, que es la que relaciona cada posible resultado x_i con su correspondiente probabilidad $\Pr(x_i)$.

Distribución uniforme

Una variable aleatoria discreta uniforme (Figura 1.10) es aquella que puede tomar n posibles valores, cada uno de ellos con probabilidad $1/n$. Por lo tanto, su función de masa de probabilidad viene dada por

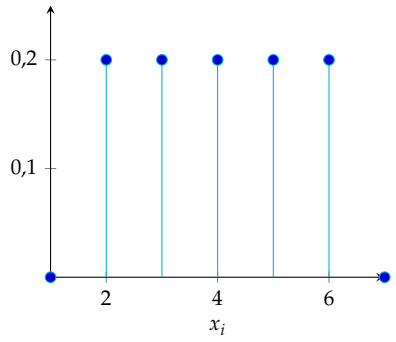


Figura 1.10: Variable aleatoria uniformemente distribuida entre 2 y 6.

$$\Pr(x_i) = \frac{1}{n}\ \forall\ i.$$

En general, los resultados de los experimentos suelen corresponderse con n números enteros consecutivos, desde un valor a hasta otro b, por lo que $n = b - a + 1$. Con esta variable se puede modelar, p.ej., el lanzamiento de un dado, o el resultado de rellenar una pregunta «tipo test» al azar. Su media y varianza vienen dados por

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{n^2-1}{12}.$$

Ensayo de Bernoulli

Un ensayo de Bernoulli (ilustrado en la Figura 1.11) se define como un experimento que se realiza una única vez y puede salir bien

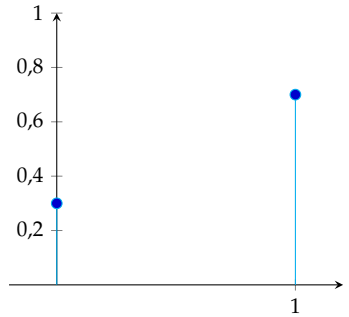


Figura 1.11: Ensayo de Bernouilli con $p = 3/4$.

o mal. Se trata de la variable aleatoria discreta más sencilla, que únicamente puede tomar dos valores, típicamente 0 y 1, donde el 1 representa un *éxito* y sucede con probabilidad p, y el 0 representa un *fracaso* y sucede con probabilidad $1 - p$. Su media y varianza vienen dados por

$$\mu = p \qquad \sigma^2 = p(1-p).$$

Distribución geométrica

En este caso, se repite el ensayo de Bernouilli hasta que se produzca un éxito. Ahora la variable aleatoria es el número de experimentos necesarios hasta que se produjo dicho éxito (Figura 1.12). Suponiendo que los ensayos se realizan de forma independiente, la probabilidad de necesitar k experimentos en total viene dada por

$$\Pr(k) = (1-p)^{k-1}p, \quad k = 0, 1, 2, \ldots$$

esto es, que se produzcan $k-1$ fallos consecutivos, cada uno con probabilidad $(1-p)$, y un éxito al final, con probabilidad p. La distribución geométrica tiene como media y varianza

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2}.$$

Nótese que el número de intentos necesarios para obtener un éxito no está acotado (de hecho, si el valor de p es muy próximo a cero, la media tiende a ser muy elevada), por lo que $\Pr(k)$ toma valores en el conjunto de los números naturales (en la Figura 1.12 se representan valores hasta $k = 10$).

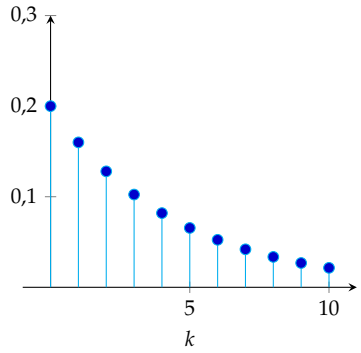


Figura 1.12: Variable aleatoria geométrica con $p = 1/5$.

Distribución binomial

La distribución binomial modela el caso de n ensayos de Bernouilli independientes, cada uno con una probabilidad de éxito p, siendo la variable aleatoria de interés el número total de éxitos. Las probabilidades de n éxitos y n fracasos resultan inmediatas de calcular (p^n y $(1-p)^n$, respectivamente), mientras que para el resto de probabilidades es preciso tener en cuenta que hay varias formas en las que pueden ocurrir los éxitos, por lo que es preciso emplear combinatoria. La probabilidad de tener k éxitos viene dada por

$$\Pr(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}, \tag{1.4}$$

y la media y varianza de la distribución binomial resultan ser

$$\mu = np \qquad \sigma^2 = np(1-p).$$

De hecho, considerando cada uno de los n experimentos como un ensayo de Bernouilli, de media p y varianza $p(1-p)$, la media y la varianza de la distribución binomial se pueden deducir a partir de las propiedades de la suma de variables aleatorias: la media es n veces la media de un ensayo (p) y la varianza es n veces la varianza de un ensayo ($p(1-p)$). Se representan en la Figura 1.13 dos variables aleatorias binomiales.

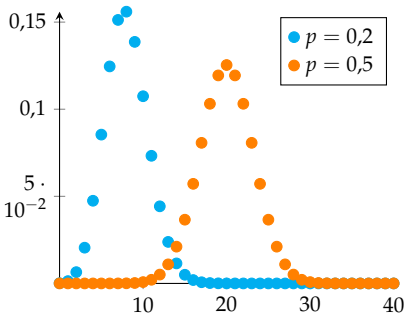


Figura 1.13: Variables binomiales con $n = 40$ y dos valores de p.

Distribución de Poisson

La distribución de Poisson se puede interpretar como una distribución binomial, donde el número de ensayos a realizar es *muy grande*, y la probabilidad de éxito *muy baja*. Sirve por tanto para modelar sucesos poco frecuentes,[12] con una media que viene dada por

$$\lambda \triangleq n \cdot p.$$

[12] En la binomial resulta complicado calcular probabilidades para valores elevados de n, dada la presencia del factorial.

Reemplazando $p = \lambda/n$ en la probabilidad de k éxitos en una Binomial (1.4), se tiene que

$$\Pr(k) = \frac{n(n-1)(n-2)\cdots(n-k)!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k},$$

que puede aproximarse, con $n \to \infty$ (pero manteniendo λ finito), como

$$\Pr(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

La media y varianza de la distribución de Poisson (ilustrada en la Figura 1.14 para diferences valores del párametro λ) resultan ser:

$$\mu = \lambda \qquad \sigma^2 = \lambda.$$

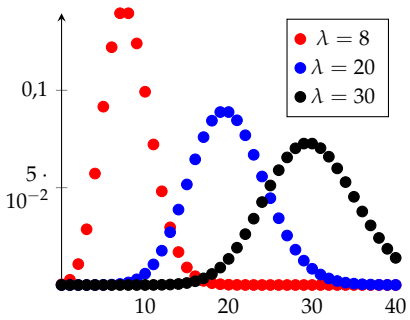


Figura 1.14: Variables aleatorias de Poisson.

Ejemplo 1.18. Sean X_1 y X_2 dos variables aleatorias independientes de Poisson, con media λ_1 y λ_2, respectivamente. La distribución de su suma $Z = X_1 + X_2$ se puede obtener como

$$\Pr(Z = n) = \sum_{x=0}^{n} \Pr(X_1 = x)\Pr(X_2 = n - x) = \sum_{x=0}^{n} \frac{\lambda_1^x}{x!}e^{-\lambda_1}\frac{\lambda_2^{n-x}}{(n-x)!}e^{-\lambda_2}$$

que, operando, queda[13]

$$\Pr(Z = n) = \frac{(\lambda_1 + \lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)},$$

esto es, *otra* variable aleatoria discreta de Poisson, de media $\lambda_1 + \lambda_2$.

[13] Es necesario usar el desarrollo

$$(a+b)^m = \sum_{n=0}^{m}\binom{m}{n}a^n b^{m-n}.$$

Variables aleatorias continuas

Las variables aleatorias discretas tienen un espacio muestral contable, por lo que resulta muy sencillo de interpretar la probabilidad de un determinado resultado $\Pr(x_i)$ como la frecuencia relativa del mismo. Sin embargo, cuando el espacio muestral ya no es un conjunto finito, esta interpretación resulta más difícil, dado que la «frecuencia relativa» de un determinado valor $x_i \in \mathbb{R}$ es 0: en cualquier segmento de la recta real hay un número infinito de valores, por lo que la probabilidad de cada uno será cero.

Ejemplo 1.19. Sea el caso del segundero de un reloj que se mueva de forma continua (como en la Figura 1.15). En una mirada «al azar» a dicho reloj (esto es, un *experimento*), el valor del segundero se encuentra el alguna posición entre 0 y 60 segundos, todas ellas con la misma probabilidad (Figura 1.16).

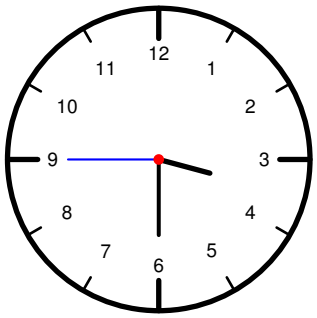


Figura 1.15: Si el segundero de un reloj avanza de forma continua, una mirada al azar lo encontrará en cualquier posición en el intervalo $[0, 60)$ segundos.

Si θ es la variable aleatoria que representa el valor del segundero, se puede hablar de la probabilidad de que su valor se encuentre en una determinada *región*, por ejemplo, $\Pr(\theta \in [0, 15]) = 1/4$, o $\Pr(\theta \in [0, 45]) = 3/4$. Pero a diferencia del caso de las variables aleatorias discretas, dado que en $[0, 60)$ hay un incontable número de valores, deja de tener sentido preguntarse por la probabilidad de que θ tome un valor dado.

Función de distribución

En el caso de una variable aleatoria continua, por lo tanto, resulta más sencillo tratar con la frecuencia relativa (o probabilidad) de un *conjunto* de valores. Sea X una variable aleatoria, que puede tomar un determinado conjunto de valores en el espacio continuo. Su función de distribución F_X determina la probabilidad de que un resultado de dicha variable aleatoria sea menor que un determinado valor x que se tome como referencia, esto es[14]

$$F_X(x) \triangleq \Pr(X \leq x).$$

[14] El signo $\leq$ es una convención para el caso continuo, dado que, por lo visto anteriormente, tendría el mismo valor el signo $<$.

En general, cuando queda claro (por el contexto) a qué variable aleatoria se está haciendo referencia, en muchas ocasiones se simplifica la notación y desaparece el subíndice de $F_X(x)$, esto es, se emplea únicamente $F(x)$. Por definición, se tiene que la función de distribución cumple que:

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 1,$$

A partir de la función de distribución se puede obtener la probabilidad de que una variable aleatoria se encuentre en un determinado intervalo, dado que

$$\Pr(a < X \leq b) = F_X(b) - F_X(a)$$

La función complementaria de la función de distribución F_X se define como

$$F_X^C(x) \triangleq 1 - F_X(x),$$

y por lo tanto representa la probabilidad de que la variable aleatoria X tome un valor superior a x, esto es,

$$F_X^C(x) = \Pr(X > x).$$

Cuando la variable aleatoria X represente un «tiempo de vida» o funcionamiento de un componente, sistema, etc., se emplea el término «función de supervivencia» (S_X) para referirse a F_X^C, dado que representa la probabilidad de que dicho componente esté operativo al menos hasta el valor proporcionado

$$S_X(t) \triangleq \Pr(X > t) = F_X^C(x)$$

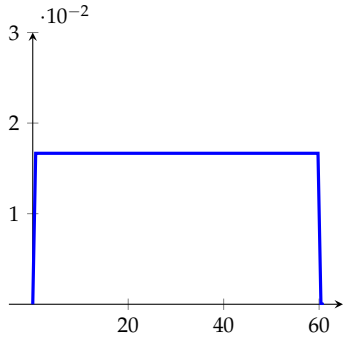


Figura 1.16: La posición del segundero puede ser cualquiera en el infinito de números existente en $[0, 60)$ segundos, lo que se representa con una función de densidad f constante en dicho intervalo.

Función de densidad

La función de densidad sirve para representar la forma en que una variable aleatoria continua se distribuye. Dada la variable aleatoria X, se define como aquella funcion no-negativa f_X que cumple que

$$\Pr(a < X \leq b) = \int_a^b f_X(x)dx,$$

por lo que también se puede interpretar como la derivada de la función de distribución F_X (si existe), esto es

$$f_X(x) = \frac{d}{dx}F_X(x).$$

Por lo tanto, se tiene que la función de distribución se puede expresar como la integral de la función de densidad,

$$F_X(x) = \int_{-\infty}^{x} f_X(y)dy.$$

Por lo descrito anteriormente sobre las variables aleatorias continuas, se tiene que el valor de la función de densidad evaluada en cualquier punto x es 0, dado que la probabilidad de un determinado valor en cualquier segmento de los números reales es nulo. Sí que tendría sentido calcular la probabilidad de que una variable aleatoria se encuentre en un rango ε alrededor de un valor x, esto es,

$$\Pr(X \in [x - \varepsilon/2, x + \varepsilon/2]) = \int_{x-\varepsilon/2}^{x+\varepsilon/2} f_X(\tau)d\tau \approx f_X(x)\varepsilon$$

por lo que se puede decir (aunque no sea muy formal) que aquellos valores de x donde f_X sea mayor son más probables, (teniendo en cuenta que la probabilidad de un determinado valor es cero).

Variable aleatoria uniformemente distribuida

Una variable aleatoria X se distribuye uniformemente entre a y b (lo que se puede representar como $X \sim U[a, b]$) si su función de densidad es *plana* y sólo existe en dicho intervalo, esto es

$$f_X(x) = \frac{1}{b-a},\ a \leq x \leq b.$$

Esta variable sirve, por tanto, para modelar aquellas situaciones en las que los posibles resultados de un experimento se encuentran acotados entre dos números, y no existe ningún valor «más probable» entre dichas cotas. Esto se correspondería, por ejemplo, con medir el ángulo de un minutero con respecto al cero en una mirada al azar (suponiendo un desplazamiento continuo del mismo), y serviría para modelar situaciones donde un evento puede suceder «en cualquier momento». Se ilustra (para dos casos particulares) en la Figura 1.17.

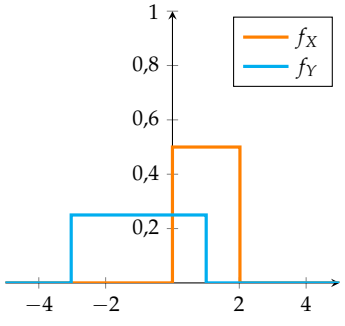


Figura 1.17: Función de densidad de las variables aleatorias $X \sim U(0, 2)$ e $Y \sim U(-3, 1)$

Correspondencias entre probabilidad discreta y continua

Momentos

Para el caso de las variables aleatorias continuas los momentos se definen de forma análoga al caso discreto, ponderando cada transformación de x por su valor de la función de densidad $f(x)$. Por lo tanto, el momento de orden n respecto al valor c se define como

$$\mu_n(c) = \int_{-\infty}^{\infty} (x-c)^n \, f(x)dx$$

- La **media** o esperanza se define como

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} x \, f(x)dx$$

- El momento de segundo orden es

$$\mu_2 = \mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 \, f(x)dx$$

- La **varianza** se calcula como:

$$\sigma^2 = \mu_2(\mu) = \int_{-\infty}^{\infty} (x-\mu)^2 \, f(x)dx$$

- La **esperanza condicionada** de la variable aleatoria X dada una condición Y se puede expresar como

$$\mathbb{E}[X \mid Y] = \frac{\int \tau \, f_{X,Y}(\tau)d\tau}{\Pr(Y)}$$

Ejemplo 1.20. Para el caso de una variable aleatoria $X \sim U[a,b]$, se puede calcular que la media y la varianza vienen dados por

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12} \ ,$$

La media de los valores mayores que la media se puede expresar mediante una esperanza condicionada

$$\mathbb{E}[X \mid X > (a+b)/2] = \frac{\int_{(a+b)/2}^{b} x \, f(x)dx}{\int_{(a+b)/2}^{b} f(x)dx}$$

Relación entre momentos

Se puede demostrar la misma relación entre los momentos que para el caso discreto, esto es

$$\mathbb{E}[X^2] = \mathbb{E}[X]^2 + \sigma_X^2 \ ,$$

y si X e Y son dos variables aleatorias, entonces

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \ ,$$

mientras que si, además, son independientes

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \ .$$

Suma de variables aleatorias

De forma similar al caso de variables aleatorias discretas, la distribución de la suma de dos variables aleatorias continuas se distribuye según su convolución. Esto es, si X e Y son dos variables aleatorias que toman valor sobre los números reales, con funciones de densidad f_X e f_Y, respectivamente, su suma $Z = X + Y$ tiene como función de densidad

$$f_Z(x) = (f_X * f_Y)(x) = \int_{-\infty}^{\infty} f_X(x-y) f_Y(y) dy \ ,$$

expresión que guarda bastante relación con la ya vista en el caso discreto.[15]

[15] La expresión (1.2), es decir:

$$\Pr(n_a + n_b = x) = \sum_y \Pr(n_a = x - y) \Pr(n_b = y)$$

Ejemplo 1.21. Sea una variable aleatoria exponencial X, con función de densidad

$$f_X(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

y sea otra variable aleatoria exponencial independiente Y, con función de densidad (donde $\lambda \neq \mu$)

$$f_Y(t) = \mu e^{-\mu t}, \quad t \geq 0$$

La función de densidad de su suma viene dada por

$$f_Z(x) = \int_{-\infty}^{\infty} f_X(x-y) f_Y(y) dy \ = \int_{-\infty}^{\infty} \lambda e^{-\lambda(x-y)} \mu e^{-\mu y} dy$$

El cálculo de la convolución lleva a (Figura 1.18)

$$\begin{aligned} f_Z(x) &= \int_{-\infty}^{\infty} \lambda e^{-\lambda(x-y)} \mu e^{-\mu y} dy \\ &= \int_0^x \lambda e^{-\lambda(x-y)} \mu e^{-\mu y} dy \\ &= \lambda\mu \int_0^x e^{-\lambda(x-y)} e^{-\mu y} dy \\ &= \lambda\mu e^{-\lambda x} \left(\frac{1}{\mu - \lambda} e^{-(\mu-\lambda)y} \Big|_x^0 \right) = \frac{\lambda\mu \left(e^{-\lambda x} - e^{-\mu x}\right)}{\mu - \lambda} \end{aligned}$$

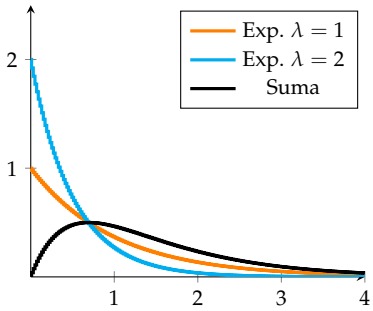


Figura 1.18: Función de densidad de dos v.a. exponencials y de su suma

Comparación de variables aleatorias

De forma análoga al caso discreto, para el caso continuo se puede calcular la probabilidad de que una determinada variable aleatoria X_1 sea mayor que otra variable aleatoria X_2 mediante la ley de la probabilidad total:

$$\Pr(X_1 > X_2) = \int_{X_2} \Pr(X_1 > X_2 \mid X_2 = \tau) f_{X_2}(\tau) d\tau \ ,$$

expresión que, de nuevo, resulta similar a la ya vista en (1.1) y que se repite a continuación:

$$\Pr(n_b \geq n_a) = \sum_x \Pr(n_b \geq n_a | n_a = x) \Pr(n_a = x)$$

Ejemplo 1.22. Sean $X_1 \sim U[0,1]$ y $X_2 \sim U[0,2]$, representadas en la Figura 1.19. La probabilidad de que un valor de X_1 sea menor que uno de X_2 se puede obtener como

$$\begin{aligned}\Pr(X_1 < X_2) &= \int_{X_2} \Pr(X_1 < X_2 \mid X_2 = \tau) f_{X_2}(\tau) d\tau \\ &= \int_{X_2} \Pr(X_1 < \tau) f_{X_2}(\tau) d\tau\end{aligned}$$

donde $\Pr(X_1 < \tau)$ es, por definición, F_{X_1}, y se trata de una función a trozos (Figura 1.20 al margen)

$$F_{X_1}(x) = \begin{cases} 0, & \text{si } x \leq 0 \\ x, & \text{si } 0 \leq x \leq 1 \\ 1, & \text{si } x \geq 1 \end{cases}$$

Para realizar realizar el cálculo, conviene dividir la integral como

$$\Pr(X_1 < X_2) = \int_0^1 F_{X_1}(\tau) \cdot f_{X_2} d\tau + \int_1^2 F_{X_1}(\tau) \cdot f_{X_2} d\tau \,,$$

de lo que resulta

$$\Pr(X_1 < X_2) = \int_0^1 \tau \cdot \frac{1}{2} d\tau + \int_1^2 1 \cdot \frac{1}{2} d\tau \; = \frac{3}{4} \,.$$

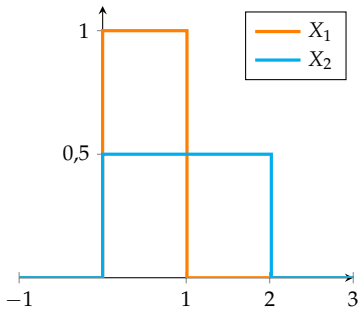


Figura 1.19: Función de densidad f de $X_1 \sim U[0,1]$ y $X_2 \sim U[0,2]$

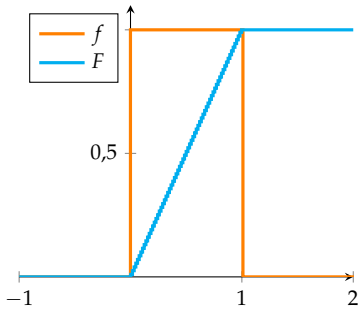


Figura 1.20: Funciones de densidad y distribución de $X_1 \sim U[0,1]$

Caracterizar el máximo (o mínimo) de varias variables aleatorias

Sea una variable aleatoria $X_{\text{máx}}$ definida como el máximo de un conjunto de N variables aleatorias independientes, esto es

$$X_{\text{máx}} = \text{máx}\{X_1, X_2, \ldots, X_N\},$$

que puede modelar, p.ej., el tiempo de ejecución de un programa compuesto por varios procesos en paralelo, o el tiempo necesario para recibir una trama cuando se ha dividido en varios fragmentos y cada uno es enviado por un camino diferente. Para caracterizar completamente la variable $X_{\text{máx}}$, es preciso obtener su función de densidad $f_{X_{\text{máx}}}$, o bien su función la de distribución $F_{X_{\text{máx}}}$, dado que a partir de una se puede obtener la otra.

En este caso, resulta relativamente sencillo obtener la función de distribución, ya que ésta se define como

$$F_{X_{\text{máx}}}(t) \triangleq \Pr(X_{\text{máx}} < t) \,,$$

y resulta que el cálculo de $\Pr(X_{\text{máx}} < t)$ se puede expresar en términos de las N variables aleatorias: para que un determinado valor t sea mayor que $X_{\text{máx}}$, tiene que ser mayor que las N variables;[16] dado que éstas son independientes, se tiene que

$$\Pr(X_{\text{máx}} < t) = \Pr(X_1 < t) \cdot \Pr(X_2 < t) \ldots \Pr(X_N < t) \,. \tag{1.5}$$

Una vez obtenida la expresión (1.5) para el caso de estudio, se puede obtener la esperanza, desviación típica, etc. Siguiendo un método similar, también puede caracterizarse el *mínimo* de variables aleatorias, si bien es preciso ser cauto con la comparación a realizar entre el valor t genérico y el conjunto de variables aleatorias.

[16] Por ejemplo: para que el jugador más alto de un equipo mida menos de 2 m se tiene que cumplir que todos los jugadores midan menos de 2 m.

Ejemplo 1.23. Sean dos variables aleatorias $X_1 \sim U[0,1]$ y $X_2 \sim U[0,1]$, y sea la variable aleatoria $X_{\text{máx}}$ definida por su máximo

$$X_{\text{máx}} = \text{máx}\{X_1, X_2\}$$

La función de distribución de $X_{\text{máx}}$ viene dada por

$$\Pr(X_{\text{máx}} < t) = \Pr(X_1 < t) \cdot \Pr(X_2 < t) = t \cdot t = t^2, \text{ para } t \in [0,1] ,$$

por lo que su función de densidad es

$$f_{X_{\text{máx}}}(t) = 2t, \ t \in [0,1] .$$

Obtenida la función de densidad, se puede calcular su esperanza

$$\mathbb{E}[X_{\text{máx}}] = \int_0^1 t f_{X_{\text{máx}}}(t) dt = \frac{2}{3}$$

que, como era de esperar, resulta mayor que la esperanza de cualquiera de ellas.

Otras distribuciones aleatorias continuas

Además de la distribución uniforme, obviamente existen otras variables aleatorias continuas, como la distribución normal o gaussiana (que también sirve para aproximar la distribución binomial para valores de n elevados), la distribución t de Student o la distribución exponencial. En el siguiente tema se analiza con detalle esta última.

La variable aleatoria exponencial

La variable aleatoria exponencial es una de las variables aleatorias continuas más sencillas de definir, dado que se caracteriza únicamente con un parámetro. También presenta ventajas desde el punto de vista analítico, ya que resulta relativamente «cómodo» realizar operaciones con la misma (p. ej., cálculo de esperanzas condicionadas).

Además de estas ventajas en cuanto al modelado analítico,[1] una variable aleatoria exponencial resulta adecuada para modelar sucesos entre eventos que tengan «poco que ver» entre sí: en el caso más representativo, el tiempo entre dos llamadas de teléfono consecutivas, realizada por una población «grande» de individuos, suele modelarse con una variable aleatoria exponencial. También resulta habitual suponer que el tiempo hasta el primer fallo de un componente fabricado en cadena se pueda modelar con una variable aleatoria exponencial.

[1] Que ya justificarían, hasta cierto punto, su uso en el análisis de sistemas.

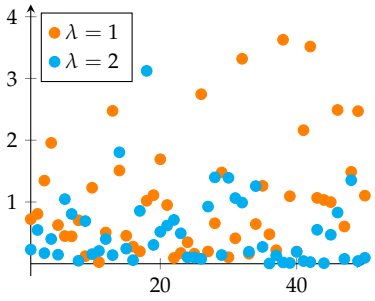


Figura 2.1: 50 muestras aleatorias de dos variables aleatorias exponenciales.

Definición y caracterización básica

Una variable aleatoria continua X sigue una distribución exponencial si su función de densidad f_X tiene la siguiente expresión:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

donde $\lambda > 0$ es el único parámetro que caracteriza la variable aleatoria.[2] A modo de ejemplo, se muestran en la Figura 2.1 cincuenta muestras de dos variables exponenciales con diferente valor de λ.

En la Figura 2.2 se representa la función de densidad de una variable aleatoria exponencial para distintos valores del parámetro. Se observa que la función de densidad toma el valor de λ en el eje de ordenadas ($x = 0$), por lo que un valor elevado en dicho punto supone que la variable tenderá a generar valores relativamente bajos, y viceversa.

[2] Hay autores que prefieren la definición $f(x) = \frac{1}{\beta} e^{-x/\beta}$.

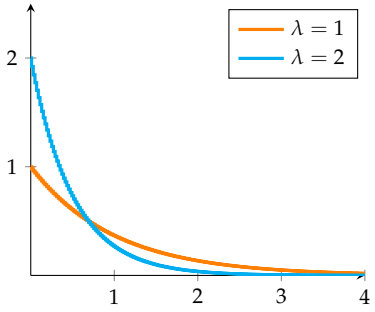


Figura 2.2: Función de densidad de probabilidad de una variable aleatoria exponencial

Esperanza y desviación típica

Para calcular la esperanza se aplica la definición de la misma para el caso continuo

$$\mathbb{E}[X] = \int_X x f(x) dx \, ,$$

lo que se traduce para el caso de la variable aleatoria exponencial en

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx \ .$$

Dicha integral se resuelve *por partes*,[3] lo que lleva a

$$\mathbb{E}[X] = \left| -xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x} \right|_0^\infty = \frac{1}{\lambda}.$$

[3] Recuérdese la regla de la integral por partes: $\int u \cdot dv = u \cdot v - \int v \cdot du$.

Ejemplo 2.1. Sea una población de usuarios donde el tiempo entre dos llamadas de teléfono T consecutivas se distribuye según una variable aleatoria exponencial, de media 5 minutos. Esto se puede representar como

$$T \sim \exp(\lambda = 5), \quad f_T(t) = \frac{1}{5} e^{-t/5} \ , \quad t \geq 0 \ .$$

Se ilustran en la Figura 2.3 cincuenta muestras de una variable aleatoria con dichas características, así como su valor medio.

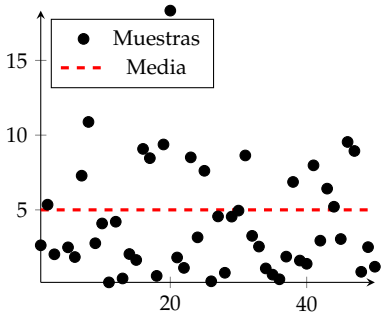


Figura 2.3: 50 muestras aleatorias una v.a.exponencial con $\lambda = 1/5$.

La desviación típica se puede obtener, bien por la definición de la misma

$$\sigma_X^2 = \int (x - \mathbb{E}[X])^2 f(x) dx \ ,$$

bien calculando el momento de segundo orden, y aplicando

$$\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \ .$$

En ambos casos el resultado es

$$\sigma_X^2 = \frac{1}{\lambda^2} \ ,$$

de donde se tiene que en una variable aleatoria exponencial la media y la desviación típica coinciden:

$$\mathbb{E}[X] = \sigma_X = \lambda^{-1}$$

Función de distribución y mediana

Como se ha visto en el capítulo anterior, la función de distribución F de una variable aleatoria X se define como

$$F(x) \triangleq \Pr(X \leq x) = \int_{-\infty}^{x} f_X(\tau) d\tau,$$

por lo tanto, para el caso de la variable aleatoria exponencial se tiene que

$$F(x) = \int_0^x \lambda e^{-\lambda\tau} d\tau = -e^{-\lambda\tau} \Big|_0^x = 1 - e^{-\lambda x}, \quad x \geq 0.$$

En la Figura 2.4 se representa la función de distribución para los mismos valores de λ que en la Figura 2.2. Como se aprecia en la figura, cuanto mayor es el valor de λ, más rápido crece la función de distribución, dado que hay mas probabilidad de los valores sean próximos a cero.

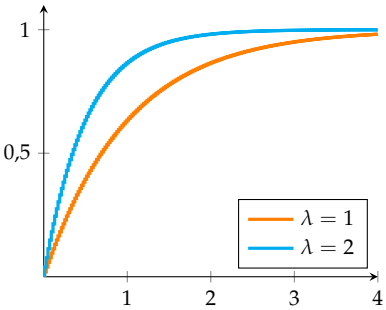


Figura 2.4: Función de distribución de una variable aleatoria exponencial.

Ejemplo 2.2. Suponga que el tiempo de vida de una bombilla se puede modelar como una variable aleatoria exponencial de media tres meses. Se pide:

1. ¿Cuál es la probabilidad de una bombilla dure menos de un mes?
2. De un lote de 100 bombillas, ¿cuántas durarán más de 63 días?

Si t es la variable aleatoria que representa el tiempo de vida de una bombilla, su media es $1/\lambda = 3$ meses, por lo que $\lambda = 1/3$. La primera pregunta es directamente $\Pr(t < 1 \text{ mes})$, lo que se tiene como

$$\Pr(t < 1) = F(1) = 1 - e^{-\lambda 1} = 1 - e^{-1/3} \approx 0.283$$

La segunda pregunta se puede interpretar en términos de la variable aleatoria binomial: hay n=100 bombillas, y cada una tiene una probabilidad p de sobrevivir más de 63 días (que hay que calcular). Teniendo en cuenta que 63 días equivalen a 2.1 meses, se tiene que

$$p \triangleq \Pr(t > 2.1) = e^{-\lambda 2.1} \approx e^{-0.7} \approx 0.5,$$

por lo que la probabilidad de que una bombilla supere los 63 días es $p = 1/2$. Por lo tanto, de un lote de 100 bombillas, el número esperado de «supervivientes» tras 63 días es $n \cdot p$, esto es, 50.

UNA VEZ OBTENIDA LA FUNCIÓN DE DISTRIBUCIÓN, el cálculo de la mediana x_M resulta inmediato, dado que se trata del percentil 0.5. Basta con resolver:

$$x_M : F(x_M) = 0.5,$$

de donde se obtiene que

$$x_M = \frac{\ln 2}{\lambda} \approx 0.7\frac{1}{\lambda}.$$

Por lo tanto, para el caso de una variable aleatoria exponencial se tiene que la mediana es un 70 % del valor de la media, por lo que la distribución se encuentra «escorada» hacia los valores próximos a cero (como se ilustra en la Figura 2.5).

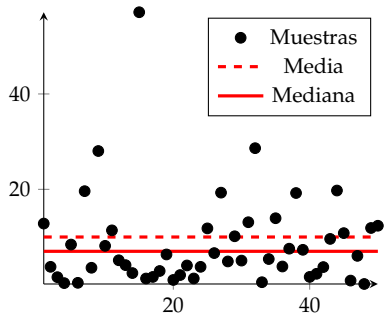


Figura 2.5: Muestras de una v.a.exp. con media 10.

La función de supervivencia

La función de supervivencia es otro nombre de la función complementaria F^C

$$F^C(x) \triangleq 1 - F(x).$$

Dicha función tiene una interpretación especial cuando la variable aleatoria sirve para modelar *tiempos de vida*, dado que si $F(t)$ pasa a ser la probabilidad de que dicho tiempo de vida sea menor o igual que t, la función complementaria indica la probabilidad de que se *sobreviva* más allá de dicho umbral t.

La función de supervivencia se suele representar por $S(t)$ y, para el caso de la variable aleatoria exponencial, su expresión es

$$S(t) = e^{-\lambda t} \quad .$$

Si se representa la función de supervivencia en escala logarítmica, se obtiene una recta de pendiente $-\lambda$ (Figura 2.6). Si se trata de un tiempo de vida o funcionamiento de un componente, λ representa la tasa media a la que se rompe. Por lo tanto, cuanto menor sea esta pendiente (esto es, $S(t)$ quede más plana), más fiable será dicho componente (y cuanto mayor sea la pendiente, más probable será que falle).

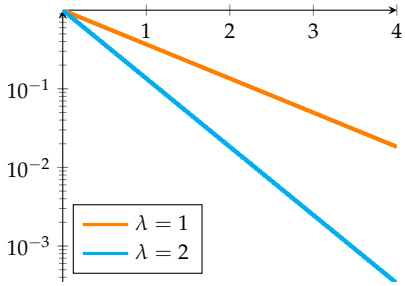


Figura 2.6: Función de supervivencia de una variable aleatoria exponencial (eje y en escala logarítmica).

La propiedad «sin memoria» de la variable aleatoria exponencial

Como ya se vio anteriormente,[4] en ocasiones la intuición y la realidad suelen entrar en contradicción: se suele pensar, por ejemplo, que una «mala racha» apostando tiene que acabar en algún momento, que ocurrirá tanto antes cuanto más larga sea la mala racha acumulada, o bien que los jugadores están «en racha» (p. ej. en baloncesto) cuando en realidad dichas rachas son ocurrencias propias del proceso aleatorio.

[4] La memoria en una variable aleatoria, página 10

A continuación, se presenta en primer lugar el caso de la variable aleatoria geométrica, donde dado un experimento de referencia, el número de intentos hasta tener un éxito no depende de los intentos anteriores a la referencia. Partiendo de este caso particular de variable discreta, luego se abordará el caso continuo.

La falta de memoria de la v.a. geométrica

Supóngase un dado de seis caras que se lanza repetidamente. Sea X_n la variable aleatoria definida como el resultado del lanzamiento n-ésimo. Si el dado está bien hecho, se tiene que

$$\Pr(X_n = 6) = 1/6 \; \forall n$$

y suponiendo que cada lanzamiento es independiente (es decir: lo normal), el resultado de un lanzamiento no afecta al siguiente lanzamiento, por lo que la probabilidad de obtener un seis habiendo obtenido un seis en el lanzamiento anterior es siempre la misma

$$\Pr(X_{n+1} = 6 \mid X_n = 6) = \Pr(X_n = 6) = 1/6$$

Sea ahora la variable N_6 definida como el número de lanzamientos hasta que aparece el primer seis (que puede ir de uno hasta infinito). Se puede identificar con una variable aleatoria geométrica con probabilidad de éxito $p = 1/6$, dado que las probabilidades de los primeros valores de N_6 se pueden obtener como

$$\Pr(N_6 = 1) = \Pr(X_1 = 6) = 1/6$$
$$\Pr(N_6 = 2) = \Pr(X_1 \neq 6)\Pr(X_2 = 6) = (5/6)(1/6)$$
$$\Pr(N_6 = 3) = \Pr(X_1 \neq 6)\Pr(X_2 \neq 6)\Pr(X_3 = 6) = (5/6)^2(1/6)$$

Si se toma el segundo lanzamiento como referencia, y no se ha tenido éxito hasta el momento, la probabilidad de que se obtenga un seis en el tercer lanzamiento (el siguiente lanzamiento) puede expresarse como

$$\Pr(N_6 = 3 \mid N_6 > 2)$$

donde $\Pr(N_6 > 2)$ representa la falta de éxitos en el primer y segundo lanzamiento, por lo que N_6 tiene que ser mayor que dos

$$\Pr(N_6 > 2) = 1 - \Pr(N_6 = 1) - \Pr(N_6 = 2) = 25/36$$

se puede desarrollar la probabilidad condicionada anterior como

$$\Pr(N_6 = 3 \mid N_6 > 2) = \frac{\Pr(N_6 = 3, N_6 > 2)}{\Pr(N_6 > 2)} = \frac{\Pr(N_6 = 3)}{\Pr(N_6 > 2)}$$

donde se ha simplificado el evento $N_6 > 2$ del numerador por redundante. Substituyendo los valores de las probabilidades en la ecuación anterior, queda

$$\Pr(N_6 = 3 \mid N_6 > 2) = \frac{(5/6)^2(1/6)}{25/36} = 1/6$$

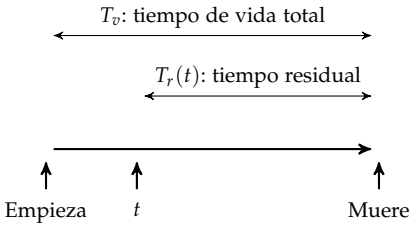
que coincide con la probabilidad de que en un lanzamiento se obtenga un seis.

La memoria en un tiempo de vida

El tiempo de vida de un elemento (p. ej., un componente electrónico) se define como el tiempo que pasa desde que dicho elemento es puesto a funcionar hasta que se estropea. Si su valor se conoce a priori y es T_v, según avance el tiempo t a dicho elemento le quedarán

$$T_r(t) = T_v - t$$

unidades de tiempo (Figura). A este tiempo que falta desde el instante actual de referencia hasta que el elemento «muera» se le llama tiempo restante o *residual* de vida. Por lo tanto, dado un tiempo de vida finito y conocido, según avance el tiempo su tiempo residual irá decayendo: se sabe exactamente cuándo dejará de funcionar.



En el caso de que el tiempo de vida sea una variable aleatoria, resulta interesante saber la probabilidad de que sobreviva más allá de un tiempo (s). Al poner a funcionar un elemento, dicha probabilidad se puede expresar en términos de su tiempo de vida (T_v):

$$\Pr(T_v > s)$$

Mientras que según avance el tiempo ($t > 0$), la probabilidad se expresa en términos del tiempo residual

$$\Pr(T_r(t) > s)$$

Sabiendo que el elemento sigue funcionando tras un tiempo t, esta última probabilidad se puede expresar en términos de T_v, teniendo en cuenta que:

1. Hay que restar de T_v el tiempo t que ya ha pasado (según lo visto anteriormente)
2. Que el elemento siga funcionando tras t nos da información sobre T_v (una condición)

Por lo tanto,

$$\Pr(T_r(t) > s) = \Pr(T_v - t > s \mid T_v > t)$$

Ejemplo 2.3. Supóngase que la duración de una película (su «tiempo de vida») se puede modelar con una variable aleatoria uniformemente distribuida entre 90 y 150 minutos (Figura 2.7). Un espectador que vaya al cine y sepa esta información, sabe que estará como mínimo 90 minutos sentado, dado que $\Pr(T_v > 90') = 1$. Durante el transcurso de la película, si se pregunta por la probabilidad de que «le quede» más de media hora ($\Pr(T_r(t) > 30')$), dicha probabilidad cambia con el tiempo:

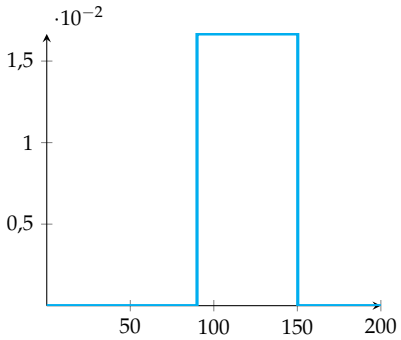


Figura 2.7: Duración uniformemente distribuida entre 90′ y 150′.

- Por una parte, durante la primera hora sabe que es seguro que le quede más de media hora ($\Pr(T_r(t) > 30') = 1$), dado que no hay películas que duren menos de $60 + 30$ minutos.
- Por otra parte, a partir de dos horas (120′) sabe que terminará en media hora o antes (ninguna película dura más de 150′). Esto se puede expresar como que $\Pr(T_r(t) > 30') = 0$.
- Entre estos dos extremos, $30' \leq t \leq 120'$, cuanto más avance el tiempo t, más probable es que la película acabe antes de media hora (dicho de otra forma: la probabilidad de que la película *sobreviva media hora más* va disminuyendo con t).

Por lo tanto, según avanza t el espectador puede ir *actualizando* la incertidumbre sobre lo que resta de película, teniéndose que $\Pr(T_r(t) > s)$ va decreciendo con t: cuanto más tiempo pasa, menos probable es que la película dure más allá de s.

En general, como en el ejemplo anterior, el tiempo que lleva sobreviviendo un elemento se puede emplear como una condición para intentar saber más sobre su tiempo residual de vida (esto es, lo que le queda). En el caso de la variable aleatoria exponencial, como se verá a continuación, esto no es así.

La propiedad «sin memoria»: definición

En términos intuitivos, que un tiempo de vida no tenga memoria significa que la probabilidad de que sobreviva más allá de una referencia es siempre la misma.[5] En términos formales, se dice que una variable aleatoria X no tiene memoria (posee la propiedad *sin memoria*) si cumple la siguiente propiedad

$$\Pr(X > t + s \mid X > t) = \Pr(X > s) \quad \forall s, t. \tag{2.1}$$

[5] Por ejemplo, que un disco duro no tenga memoria supone que la probabilidad de que dure más de año al comprarlo es igual a la probabilidad de que dure más de dos años cuando ya ha pasado un año.

Para el caso en que X represente un tiempo de vida, la ecuación (2.1) iguala dos expresiones (izquierda y derecha, respectivamente):

1. La probabilidad de que sobreviva un tiempo $t + s$, sabiendo que ya ha sobrevivido un tiempo t
2. La probabilidad de que sobreviva un tiempo s (partiendo desde 0, es decir, como si acabase de ser puesto en funcionamiento)

De lo que se deduce que la distribución del tiempo restante de vida no depende del tiempo t que lleve con vida.

Ejemplo 2.4. Sean dos bombillas: una acaba de encenderse y la otra lleva 40 años encendida. Si el tiempo de vida T_v de dichas bombillas es la misma variable aleatoria y no tiene memoria, la probabilidad de que la bombilla *dure un año más*:

$$\Pr(T_r > 1)$$

es la misma en los dos casos: en el primer caso coincide por definición con la probabilidad de que la bombilla dure más de un año (esto es, la expresión anterior), y en el segundo caso con la probabilidad de que dure más de 41 años, sabiendo que lleva 40 años encendida

$$\Pr(T_v > 41 \mid T_v > 40) \, .$$

La variable aleatoria exponencial no tiene memoria, lo que se puede demostrar desarrollando la expresión de la parte izquierda de (2.1), por definición de probabilidad condicional

$$\Pr(X > s + t \mid X > t) = \frac{\Pr(X > s + t, X > t)}{\Pr(X > t)}$$

En este cociente aparece, en el numerador, la probabilidad conjunta de que X sea mayor que $s + t$ y que X sea major que s, lo que se puede simplificar como

$$\Pr(X > s + t, X > t) = \Pr(X > s + t)$$

por lo que la parte izquierda de (2.1) queda como

$$\Pr(X > s + t \mid X > t) = \frac{\Pr(X > s + t)}{\Pr(X > t)}$$

Estas probabilidades se pueden expresar en términos de la función de supervivencia, lo que lleva a

$$\Pr(X > s + t \mid X > t) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} \tag{2.2}$$

que coincide con la parte derecha de (2.1):

$$\Pr(X > s) = e^{-\lambda s}$$

lo que demuestra que la exponencial tiene memora.

De hecho, se puede demostrar que la variable aleatoria exponencial es la *única* distribución de probabilidad continua sin memoria, lo que se debe a la forma en que se puede simplificar en (2.2) el cociente de funciones de distribución complementarias.

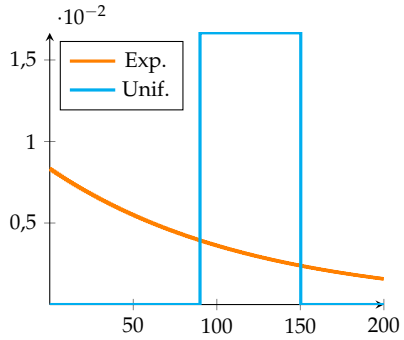


Figura 2.8: Variables aleatorias unif. y exp., ambas de media 120

Ejemplo 2.5. Siguiendo con el caso del Ejemplo 2.3, supóngase ahora que la duración de una película se puede modelar con una variable aleatoria exponencial, de media 120 minutos (Figura 2.8). En estas condiciones, la probabilidad de que una película acabe durante los próximos 30 minutos es siempre la misma, tanto si acaba de empezar como si lleva 3 horas. Si hay varias películas en reproducción en distintas salas, la probabilidad de que una de ellas acabe en los próximos 30 minutos es la misma, independientemente de cuándo hayan empezado.

VISUALMENTE, QUE LA VARIABLE ALEATORIA EXPONENCIAL NO tenga memoria se puede interpretar como se ilustra en la Figura 2.9: partiendo de la gráfica superior, al «dividir» el valor de la zona sombreada en oscuro ($\Pr(X > t+s)$) sobre toda la zona sombreada ($\Pr(X > t)$), se tiene el resultado de la gráfica inferior ($\Pr(X > s)$). Esto coincide con la división realizada anteriormente:

$$\frac{\Pr(X > s+t)}{\Pr(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s},$$

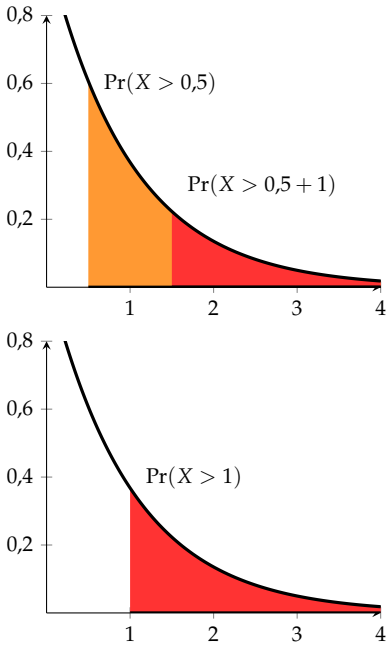


Figura 2.9: La probabilidad del futuro $s+t$ condicionada a (esto es, dividido por) una referencia t (figura superior) es igual a la incertidumbre a partir de dicha referencia s (figura inferior).

Ejemplo 2.6. Suponga que en una red Gigabit Ethernet el tiempo entre tramas se puede modelar como una variable aleatoria exponencial de media $\lambda^{-1} = 1$ ms. La probabilidad de que no se observe ninguna trama durante 2 ms es

$$\Pr(T > t)_{t=2\text{ ms}} = e^{-\lambda t} = e^{-2} \quad .$$

Por otra parte, la probabilidad de que no se observe trama alguna durante 4 ms, sabiendo que ya han pasado 2 ms sin tráfico en la red, resulta ser

$$\Pr(T > 4 \mid T > 2) = \Pr(T > 2) = e^{-2}.$$

Como era de esperar, la probabilidad de que no llegue ninguna trama durante 2 ms ($\Pr(T > 2)$) es la misma a lo largo del tiempo, independientemente de que justo acabe de llegar una trama o de que hayan transcurrido 2 ms desde la última llegada.

LA PROPIEDAD «SIN MEMORIA» de la variable aleatoria exponencial supone, por tanto, que la incertidumbre sobre lo que pueda pasar a futuro se mantiene constante: que haya pasado más o menos tiempo no afecta a la probabilidad de que un componente siga funcionando. Como se ilustra a continuación, esta propiedad no ocurre para otras variables aleatorias.

Ejemplo 2.7. Suponga un flujo de voz sobre IP, donde el tiempo entre tramas X se puede modelar con una variable aleatoria uniformemente distribuida entre 19 y 21 ms, esto es,

$$X \sim U[19, 21] \text{ ms}$$

Sea el instante $t = 0$ el momento en que se produce la primera llegada. Dado que la siguiente llegada ocurrirá, como mínimo, tras 19 ms, es seguro que se producirá tras 1 ms

$$\Pr(X > 1 \text{ ms}) = 1 \ .$$

A partir del instante $t = 19$ ms, es posible que se produzca una llegada, pero no más allá del instante $t = 21$ ms (Figura 2.10). Por una parte, se tiene p.ej. que

$$\Pr(X > 19 \text{ ms}) = 1 \ ,$$

mientras que 1 ms más tarde, la probabilidad de que llegue tras 20 ms es del 50 %

$$\Pr(X > 20 \text{ ms}) = 1/2 \ .$$

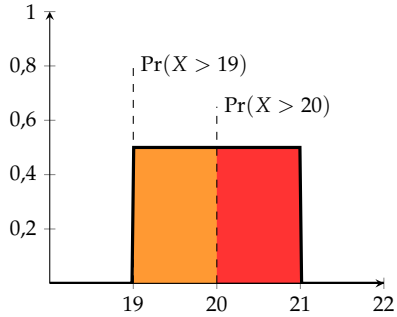


Figura 2.10: La variable aleatoria uniforme sí tiene memoria.

A partir de estos resultados, se puede comprobar que este proceso tiene memoria, ya no se cumple la igualdad

$$\Pr(X > t + s \mid X > t) = \Pr(X > s)$$

para cualquier valor de t y s: como se acaba de ver, con $s = 1$ ms y $t = 19$ ms, se tiene que

$$\Pr(X > 20 \text{ ms} \mid X > 19 \text{ ms}) = 1/2 \ ,$$

que no coincide con

$$\Pr(X > 1 \text{ ms}) = 1$$

de lo que se concluye que esta variable aleatoria uniformemente distribuida sí que tiene memoria.

Cálculo de la esperanza condicionada

Sea una variable aleatoria exponencial X, con función de densidad

$$f(t) = \lambda e^{-\lambda t} \ .$$

Como se ha visto, su esperanza viene dada por

$$\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x} dx \ = \frac{1}{\lambda} \ .$$

Supóngase ahora que interesa calcular la esperanza de aquellos valores mayores que una constante k, esto es,

$$\mathbb{E}[X \mid X > k] \ ,$$

lo que puede hacerse, en principio, aplicando la definición de esperanza condicionada

$$\mathbb{E}[X \mid X > k] = \frac{\int_k^\infty x\lambda e^{-\lambda x} dx}{\Pr(X > k)} = \frac{ke^{-\lambda k} + \frac{1}{\lambda}e^{-\lambda k}}{e^{-\lambda k}} = k + \frac{1}{\lambda} = k + \mathbb{E}[X]$$

Existe otra forma de realizar el cálculo, basándose en la propiedad sin memoria vista anteriormente: dicha propiedad ilustra que la incertidumbre «pasado un valor k» es la misma que la incertidumbre al principio:

$$\Pr(X > t + s \mid X > t) = \Pr(X > s)$$

Por lo tanto, la «aleatoriedad» de los valores mayores que $t + s$, fijando la referencia en t, es el mismo que el de los valores partiendo de 0 (Figura 2.11). De lo que se deduce que

$$\mathbb{E}[X \mid X > k] = k + \mathbb{E}[X]\,,$$

que es el mismo resultado obtenido anteriormente.

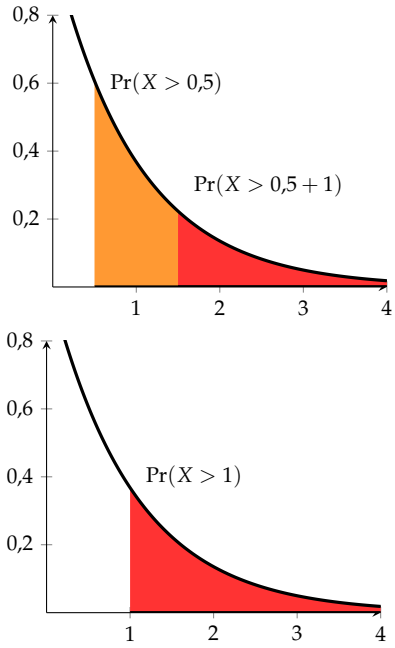


Figura 2.11: (repetida) El comportamiento a partir de $t + s$, fijando la referencia en t, es igual que el comportamiento a partir de s.

Análisis de múltiples variables aleatorias exponenciales

Hasta ahora se ha considerado el caso de una única variable aleatoria exponencial, pero en varias ocasiones la situación será más complicada: si, por ejemplo, el tiempo de vida de un componente hardware se puede modelar según una variable aleatoria exponencial, el tiempo de vida de un dispositivo compuesto por *varios* componentes vendrá determinado por aquel que se rompa antes.

Mínimo de variables aleatorias exponenciales

Sea el caso de dos variables aleatorias exponencialmente distribuidas, X_1 y X_2, independientes y de media $1/\lambda_1$ y $1/\lambda_2$, respectivamente. Se pretende caracterizar la variable definida por el mínimo de ambas

$$X_{\text{mín}} = \text{mín}\{X_1, X_2\},$$

lo cual se correspondería, por ejemplo, con un escenario con dos bombillas con tiempos de vida exponenciales, donde se quiere analizar cómo se distribuye el tiempo que pasa hasta que una de ellas deja de funcionar.

Dado que dicho mínimo también es una variable aleatoria, para analizarlo es preciso obtener su función de distribución F o su complementaria F^C. En este caso lo más sencillo resulta obtener esta última,

$$F^C_{X_{\text{mín}}}(t) \triangleq \Pr(X_{\text{mín}} > t),$$

dado que la probabilidad de que $X_{\text{mín}}$ sea mayor que un tiempo t coincide con la probabilidad de que las dos variables aleatorias independientes X_1 y X_2 sean mayores que dicho t (por ejemplo: la

probabilidad de que dos bombillas sigan encendidas tras un tiempo t es la probabilidad de que cada una de ellas siga encendida tras un tiempo t). Por lo tanto:

$$\Pr(X_{\text{mín}} > t) = \Pr(X_1 > t) \cdot \Pr(X_2 > t),$$

que se puede expresar como

$$\Pr(X_{\text{mín}} > t) = e^{-\lambda_1 t} e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}$$

De lo anterior, se tiene que la función de supervivencia de la variable aleatoria $X_{\text{mín}}$ queda como

$$F^C_{X_{\text{mín}}}(t) = e^{-(\lambda_1+\lambda_2)t},$$

que se comprueba, por simple inspección, que se corresponde con la función de supervivencia de una variable aleatoria exponencial, de media $1/(\lambda_1 + \lambda_2)$. Por lo tanto, el mínimo de dos variables aleatorias exponenciales independientes se distribuye según otra variable aleatoria exponencial, con la inversa de su media $\lambda_{\text{mín}}$ igual a la suma de las inversas de las medias.

Ejemplo 2.8. Sea un sistema con dos clientes realizando peticiones a un servidor. Si el tiempo entre peticiones se puede modelar con una variable aleatoria exponencial, de media $1/\lambda_1 = 2$ ms para el primer cliente y $1/\lambda_2 = 1$ ms para el segundo, el tiempo entre peticiones del agregado también se distribuirá según una variable aleatoria exponencial, de media $1/(\lambda_1 + \lambda_2) = 2/3 \approx 0{,}66$ ms.

El anterior resultado es generalizable al caso de múltiples variables aleatorias exponenciales independientes: sea $X_1, \ldots, X_n$ un conjunto de n variables, con medias $1/\lambda_1, \ldots, 1/\lambda_n$, respectivamente. La variable aleatoria definida por el mínimo de todas ellas se puede obtener como:

$$\begin{aligned}\Pr(\text{mín}\{X_1, \ldots, X_n\} > t) &= \prod_{i=1}^{n} \Pr(X_i > t) \\ &= \prod_{i=1}^{n} e^{-\lambda_i t} \\ &= e^{-(\sum_{i=1}^{n} \lambda_i)t},\end{aligned}$$

que es una variable aleatoria exponencial, con media $1/\sum_{i=1}^{n} \lambda_i$.

Ejemplo 2.9. Sea un router con tres componentes básicos: CPU, fuente de alimentación y memoria. El tiempo de vida de la CPU sigue una variable aleatoria exponencial de media 10 años; el de la fuente de alimentación sigue una exponencial de media 2 años y medio; y el de la memoria otra exponencial de media 2 años.

Según lo visto anteriormente, el tiempo de vida del router es una exponencial de media $1/\lambda$, donde dicho parámetro se calcula como

$$\lambda = \frac{1}{10} + \frac{1}{2.5} + \frac{1}{2} = \frac{1+4+5}{10} = 1 \text{ rotura/año}.$$

por lo que el tiempo de vida es de 1 año.

Componente	T^o vida (años)	Tasa roturas (# / año)
CPU	10	0.1
Memoria	2.5	0.4
Alimentación	2	0.5
Total	1	1

Tabla 2.4: Router con tres componentes. La tasa de rotura es la suma de las tasas de rotura.

COMO ILUSTRA EL EJEMPLO ANTERIOR, una forma de interpretar el mínimo de varias variables aleatorias exponenciales, para el caso en que dichas variables modelen tiempo de vida, es considerar que dado un tiempo medio de vida $1/\lambda$, existe «una tasa» de roturas por unidad de tiempo λ. De esta forma, se tiene que la tasa λ a la que se rompe un componente formado por varios elementos viene dada por la suma de todas las tasas, esto es

$$\lambda = \sum_i \lambda_i.$$

lo que se puede apreciar en la Tabla 2.4, donde la tasa de rotura del router del ejemplo anterior resulta la suma de las tasas de rotura de los diferentes componentes (y no de los tiempos de vida, obviamente).

Comparación de variables aleatorias exponenciales

En la sección anterior el objetivo es caracterizar el tiempo de vida de un sistema compuesto por diferentes elementos, cada uno con un tiempo de vida que se distribuye según una variable aleatoria exponencial. En esta sección se aborda un análisis parecido, pero sobre el valor de estos tiempos de vida en términos *relativos*, esto es, la probabilidad de que una variable aleatoria sea menor que otra.[6]

[6] Considerando el ejemplo de las dos bombillas, el objetivo sería calcular qué bombilla es más probable que se rompa antes.

Sean dos variables aleatorias exponenciales independientes X_1 y X_2, de medias λ_1^{-1} y λ_2^{-1}, respectivamente. La probabilidad de que una muestra de la primera variable sea inferior a una muestra de la segunda se puede calcular condicionando en una de ellas:

$$\begin{aligned} \Pr(X_1 < X_2) &= \int_0^\infty \Pr(X_2 > X_1 \mid X_1 = \tau) f_{X_1}(\tau) d\tau \qquad (2.3) \\ &= \int_0^\infty \Pr(X_2 > \tau) f_{X_1}(\tau) d\tau \,, \end{aligned}$$

donde $\Pr(X_2 > \tau)$ es la función de supervivencia de una exponencial de media $1/\lambda_2$ y $f_{X_1}(\tau)$ es la función de densidad de una exponencial de media $1/\lambda_1$. Por lo tanto,[7]

[7] Nótese que el cálculo de (2.3) se puede plantear de cuatro formas equivalentes, según se condicione al valor de X_1 o X_2, y calculando $\Pr(X_1 < X_2)$ o bien la probabilidad complementaria de $\Pr(X_2 < X_1)$.

$$\begin{aligned} \Pr(X_1 < X_2) &= \int_0^\infty e^{-\lambda_2 \tau} \lambda_1 e^{-\lambda_1 \tau} d\tau \\ &= \lambda_1 \int_0^\infty e^{-(\lambda_1+\lambda_2)\tau} d\tau \\ &= \frac{\lambda_1}{\lambda_1 + \lambda_2} \,, \end{aligned}$$

Se puede comprobar que $\Pr(X_1 < X_2) \to 1$ cuando $\lambda_1 \to \infty$, lo cual es lógico dado que en esos casos los valores de X_1 serán muy bajos (su media $\lambda_1^{-1} \to 0$), por lo que serán menores que los valores de X_2 con bastante probabilidad. También se obtiene el mismo resultado si $\lambda_2 \to 0$, por los motivos correspondientes.

Ejemplo 2.10. Sean dos variables aleatorias exponenciales independientes, X_1 y X_2, de media 5 y 10, respectivamente (Figura 2.12). La probabilidad de que un valor de X_1 sea menor que un valor de X_2 viene dada por

$$\Pr(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1/5}{1/5 + 1/10} = 2/3.$$

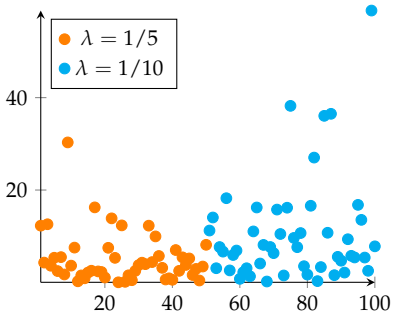


Figura 2.12: Cincuenta muestras de una v.a.exp. de media 5 y cincuenta muestras de otra v.a.exp. de media 10.

Este resultado se puede generalizar, esto es, calcular la probabilidad de que una variable aleatoria exponencial sea menor que otras variables aleatorias exponenciales. Para ello, suponiendo el caso de n variables aleatorias exponenciales, basta con considerar el mínimo de las $n-1$ variables aleatorias exponenciales contra la que se realiza la comparación.

Sean las variables aleatorias exponenciales independientes X_1, X_2 y X_3 de medias λ_1^{-1}, λ_2^{-1} y λ_3^{-1}. La probabilidad de que X_1 sea menor que las otras dos se puede calcular como

$$\begin{aligned}\Pr(X_1 < X_2, X_1 < X_3) &= \Pr(X_1 < \underbrace{\text{mín}\{X_2, X_3\}}_{\lambda=\lambda_2+\lambda_3}) \\ &= \frac{\lambda_1}{\lambda_1 + (\lambda_2 + \lambda_3)}\end{aligned}$$

La generalización es inmediata: sea $X_1, \ldots, X_n$ un conjunto de n variables aleatorias exponenciales independientes, con tasas $\lambda_1, \ldots, \lambda_n$, respectivamente. La probabilidad de que la variable aleatoria exponencial i arroje el menor valor del conjunto de todas ellas viene dada por:

$$\begin{aligned}\Pr(X_i = \min_j\{X_j\}) &= \Pr(X_i < \min_{j\neq i}\{X_j\}) \\ &= \frac{\lambda_i}{\sum_j \lambda_j}\end{aligned}$$

Ejemplo 2.11. Sea el mismo router con tres componentes básicos: CPU, fuente de alimentación, y memoria. El tiempo de vida de la CPU es exponencial de media 10 años; el de la fuente de alimentación es exponencial de media 2 años y medio; y el de la memoria es exponencial de media 2 años. La probabilidad de que la CPU sea la causante del primer fallo se puede obtener como

$$\begin{aligned}\Pr(X_{CPU} < \text{mín}\{X_{Pwr}, X_{Mem}\}) &= \frac{\lambda_{CPU}}{\lambda_{CPU} + (\lambda_{Pwr} + \lambda_{Mem})} \\ &= \frac{1/10}{1/10 + 2/5 + 1/2} = \frac{1}{10}.\end{aligned}$$

Como comentario adicional, se tiene que una vez reparado el router, la probabilidad de que el siguiente fallo vuelva a ser causado por la CPU *también* es 1/10, dado que todas las variables aleatorias exponenciales se «han renovado» tras la reparación (por la propiedad sin memoria).

Componente	T^o vida (años)	Tasa roturas (# / año)
CPU	10	0.1
Memoria	2.5	0.4
Alimentación	2	0.5
Total	1	1

Tabla 2.5: (repetida) Router con tres componentes. La tasa de rotura es la suma de las tasas de rotura.

Suma de variables aleatorias exponenciales (*)

Sean X e Y dos variables aleatorias exponenciales independientes con la misma media $1/\lambda$. La función de densidad de su suma $Z = X + Y$ se calcula como

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy = \int_0^z \lambda e^{-\lambda(z-y)} \lambda e^{-\lambda y} dy$$

El cálculo de la convolución lleva a

$$\begin{aligned} f_Z(x) &= \int_{-\infty}^{\infty} \lambda^2 e^{-\lambda x} dy \\ &= \int_0^x \lambda^2 e^{-\lambda x} dy \\ &= \lambda^2 x e^{-\lambda x} \end{aligned}$$

Que es una expresión diferente a la vista en el Ejemplo 1.21 (página 26), donde se desarrolló la suma de dos variables aleatorias exponenciales de diferentes medias $1/\lambda \neq 1/\mu$, cuya función de densidad es:

$$f_Z(x) = \frac{\lambda\mu \left(e^{-\lambda x} - e^{-\mu x}\right)}{\mu - \lambda}$$

Se ilustra en la Figura 2.13 la función de densidad para la suma de dos exponenciales cuando éstas tienen y no tienen la misma media, que dan lugar a variables aleatorias diferentes, como se puede apreciar. En un caso general, la suma de k variables aleatorias exponenciales independientes con la misma media $1/\lambda$ da lugar a la distribución de Erlang, cuya función de distribución es

$$F(x,k,\lambda) = 1 - \sum_{n=0}^{k-1} \frac{1}{n!} e^{-\lambda x} (\lambda x)^n.$$

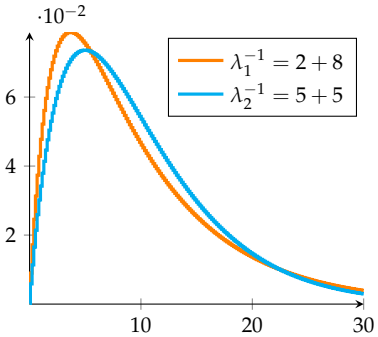


Figura 2.13: Función de densidad de la suma de dos exponenciales de media diferente y con la misma media

Resumen del tema

- Se dice que una variable aleatoria X no tiene memoria si cumple la siguiente propiedad

$$\Pr(X > t+s \mid X > t) = \Pr(X > s) \quad \forall s,t.$$

- La variable aleatoria exponencial es la única variable aleatoria sin memoria. Si sirve para modelar el tiempo de vida de un componente, esto implica que la probabilidad de que el tiempo restante de vida sea mayor que un determinado valor es siempre la misma (siempre que esté funcionando).
- Dado un conjunto de varias variables aleatorias exponenciales independientes, cada una con media λ_i^{-1}, se tiene que:
  - El mínimo de todas ellas es otra variable aleatoria exponencial, de media $(\sum \lambda_i)^{-1}$.
  - La probabilidad de que la variable aleatoria i-ésima sea menor que las otras viene dada por el cociente

$$\frac{\lambda_i}{\sum_j \lambda_j}$$

P. Serrano, J. A. Hernández
Universidad Carlos III de Madrid

Procesos de Poisson

Un proceso de conteo caracteriza la forma en la que se van sucediendo "eventos" en un sistema a lo largo de tiempo. Si estos eventos son *llegadas*, como pueden ser p.ej. clientes a una tienda, llamadas de teléfono a una central de conmutación, o tramas a un router en Internet, se habla de un *proceso de llegada*.

Un proceso de llegada *de Poisson* es aquel en el que, fijada una ventana de tiempo de duración T, el número de eventos que suceden en T sigue una variable aleatoria de Poisson. Este tipo de proceso de llegada suele emplearse para modelar situaciones en las que las llegadas no guardan relación entre sí, y –como se verá– tiene la característica de que el tiempo entre llegadas sigue una variable aleatoria exponencial. Esto supone que se trata de un proceso *muy impredecible*, dado que el tiempo que pasa entre llegadas no tiene memoria, por lo que es igual de probable que se produzca una llegada *ahora* que, por ejemplo, tras haber pasado diez minutos sin llegada alguna.[1]

[1] Si el tiempo entre llegadas siguiese una variable aleatoria uniformemente distribuida entre 90 y 120 minutos, tras 119 minutos sin una llegada, estaríamos *seguros* de que se producirá una llegada en el próximo minuto.

A continuación, primero se formaliza la definición de un proceso de conteo, para posteriormente presentar la *primera* definición de un proceso de llegada de Poisson (hay tres definiciones equivalentes).

Procesos de conteo

Un *proceso de conteo* es un proceso estocástico $\{N(t), t \geq 0\}$ que representa, en cada instante de tiempo, el número total de ocurrencias de un determinado evento hasta dicho instante de tiempo. Por lo tanto, dicho proceso debe tomar valores enteros positivos y debe ser creciente, lo que se puede expresar como:

1. $N(t) \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$.
2. Si $s < t$, entonces $N(s) \leq N(t)$.

Hay infinidad de ejemplos de procesos de conteo: número de veces que se reinicia un ordenador, *Champions League* que ha ganado un equipo de fútbol, visitas totales que tiene un video en *Youtube*, etc. El número de eventos que ocurren en el intervalo $(s, t]$ se representa como N_s^t y se puede obtener como

$$N_s^t = N(t) - N(s)$$

Ejemplo 3.1. Se el proceso de llegadas de la Figura 3.1, donde se tienen cuatro eventos en un intervalo de cinco unidades de tiempo. A la vista de la figura, se tiene que:

- $N_0^2 = N(2) - N(0) = 1$
- $N_{1,5}^{3,5} = 2$
- $N_0^{4,1} = 4$

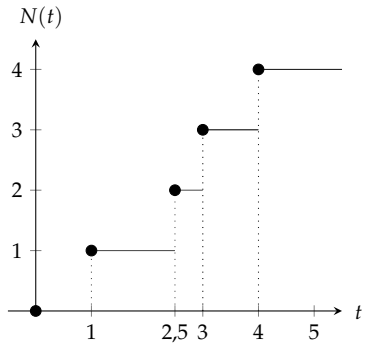


Figura 3.1: Ejemplo de un proceso con cuatro llegadas.

COMO SE HA MENCIONADO AL PRINCIPIO, un proceso de conteo tiene en cuenta únicamente *eventos* de un determinado tipo, y no otro tipo de variables como pudiera ser, p. ej., el número de usuarios en un sistema: en una localidad tanto los *nacimientos* como las *defunciones* pueden considerarse procesos de conteo, pero no así la variable *población* (esto es, nacimientos menos defunciones) a lo largo del tiempo.

Como se ilustra en la Figura 3.2 a continuación, normalmente se indica con S_i el instante de tiempo (absoluto) que se corresponde con el evento i, mientras que X_i representa el tiempo transcurrido desde el evento $(i-1)$ hasta el evento i.

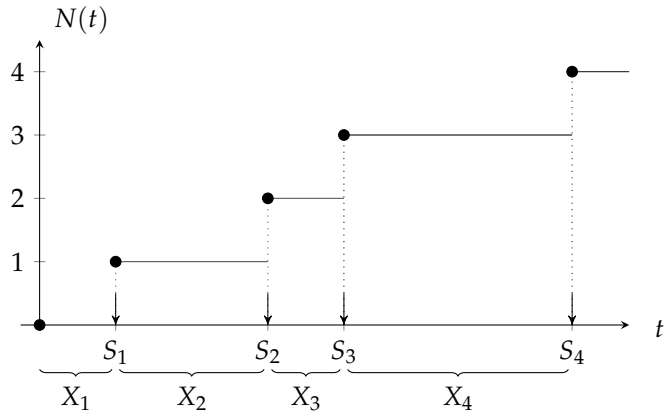


Figura 3.2: Ejemplo de un proceso de conteo

De esta forma, suponiendo que el proceso se inicia en $t = 0$, se tiene que

$$S_i = \sum_{j=1}^{i} X_j$$

Primera definición

Un proceso de llegadas (o de conteo) de Poisson es un proceso con incrementos independientes y estacionarios, que –como se ha mencionado anteriormente– guarda bastante relación con la variable aleatoria exponencial. Si bien hay tres formas equivalentes de definirlo, la definición más habitual (mencionada al principio del tema) parte de la variable aleatoria de Poisson:

Proceso de Poisson (1ª definición) Un proceso de Poisson $N(t)$ a tasa λ es un proceso de conteo con las siguientes propiedades:

(i) No hay eventos antes del instante inicial: $N(0) = 0$

(ii) El número de eventos entre 0 y un instante de tiempo t sigue una variable aleatoria de Poisson de media λt, esto es:

$$\Pr(N_0^t = n) = \Pr(N(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, \ldots, \forall\, t \tag{3.1}$$

(iii) El número de eventos en *cualquier intervalo* de longitud t sigue la misma distribución que en caso anterior, independientemente de cuándo empiece dicho intervalo, lo que puede expresarse como

$$\Pr(N_s^{s+t} = n) = \Pr(N_0^t = n)\ , \quad \forall s, t$$

Algunos sucesos que pueden modelarse con un proceso de Poisson son: los goles en un partido de fútbol, las peticiones a un servidor web, las mutaciones en una cadena de ADN, las llamadas de teléfono que recibe una central de conmutación, o las muertes por coz en la caballería prusiana entre 1875 y 1894.[2]

[2] Fue el economista y estadístico ruso Ladislaus Bortkiewicz el primero que empleó la variable aleatoria de Poisson en el análisis de fiabilidad de sistemas.

Ejemplo 3.2. Los goles en un partido de fútbol se pueden modelar como un proceso de Poisson. De esta forma, fijado un intervalo de tiempo T, el número de goles es una variable aleatoria discreta de Poisson. Supóngase un equipo donde los goles tienen una tasa de llegada de $\lambda = 3$ goles/partido. Así:

- La probabilidad de que no meta ningún gol en un partido entero ($T = 1$) es

$$\Pr(0 \text{ goles}) = \Pr(N_0^T = 0) = e^{-\lambda T} \approx 0{,}05$$

- La probabilidad de no meter ningún gol en la primera parte ($T = 1/2$) es

$$\Pr(N_0^{T/2} = 0) = e^{-\lambda T/2} \approx 0{,}22.$$

- La probabilidad de que meta tres goles en un partido:

$$\Pr(N_0^T = 3) = \frac{3^3}{3!} e^{-3} \approx 0{,}22$$

Estacionariedad e independencia

A raíz de la definición de proceso de Poisson se puede hablar de dos propiedades que pueden tener los procesos de llegadas (no sólo este tipo de procesos): estacionariedad e independencia. Por un lado, la primera hace referencia a que la probabilidad de n llegadas en un intervalo de tiempo t viene siempre dada por la misma variable aleatoria discreta de Poisson (de media λt), sin importar del instante inicial de dicho intervalo. Se puede definir como de la siguiente manera:

- Un proceso de llegadas tiene *incrementos estacionarios* (homogéneos en tiempo) si la variable aleatoria que caracteriza el número de eventos que suceden en un intervalo sólo depende de la longitud de dicho intervalo, y no de cuándo empiece. Esto es:

$$\Pr(N_{t_1}^{t_1+\tau} = n) = \Pr(N_{t_2}^{t_2+\tau} = n) \quad \forall t_1, t_2, n, \tau > 0$$

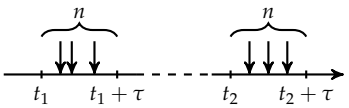


Figura 3.3: Incrementos estacionarios.

Que exista estacionariedad de los incrementos afecta a la mayor o menor incertidumbre de los mismos, dado que especifica si lo que pueda pasar en una ventana de tiempo τ depende del instante inicial de dicha ventana. Si los incrementos son estacionarios, la probabilidad de lo que pase durante un tiempo τ no depende del instante t_1 o t_2 a partir del cual se define dicho intervalo τ.

Ejemplo 3.3. En un libro (o unos apuntes de una asignatura), el número de errores tipográficos suele tener incrementos estacionarios: la cantidad de fallos en un bloque de texto depende sólo de su longitud (p.ej., en número de páginas), y no de si se encuentran p.ej. hacia el principio o el final del libro (la tasa de errores es constante).

En cambio, si se considera una granja de servidores que nunca se reparan, el proceso "servidor deja de funcionar" no tiene incrementos estacionarios: p.ej., si en un momento determinado todos los servidores dejan de funcionar, ya no volverán a producirse errores de funcionamiento.

POR OTRO LADO, independencia en los incrementos hace referencia a que lo que pueda pasar en un intervalo sólo viene determinado por la tasa λ y la longitud de dicho intervalo t, y no por lo que haya pasado en otros intervalos (siempre que los intervalos en consideración no se solapen). Esto puede definirse como:

- Un proceso de conteo tiene *incrementos independientes* si el número de eventos que suceden en intervalos disjuntos de tiempo no guardan relación entre sí. Esto es, para cuatro instantes de tiempo $s_1 < t_1 < s_2 < t_2$, se tiene que las probabilidades de que sucendan n eventos en el primer intervalo y m en el segundo son independientes (véase Figura 3.4):

$$\Pr(N_{s_1}^{t_1} = n, N_{s_2}^{t_2} = m) = \Pr(N_{s_1}^{t_1} = n) \Pr(N_{s_2}^{t_2} = m) \quad \forall n, m$$

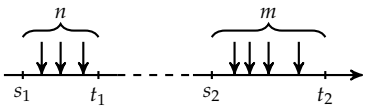


Figura 3.4: Incrementos independientes.

Que un proceso tenga incrementos independientes guarda bastante relación con su incertidumbre, dado que la independencia supone que los eventos que hayan sucedido en un intervalo anterior no afectan a lo que pueda suceder en un intervalo posterior.

Ejemplo 3.4. Sin considerar las "horas punta," un proceso con incrementos independientes podría ser la llegada de vehículos a un determinado cruce de una carretera en el que, *a priori*, el número de coches que pasan por un punto de la misma un martes de 2 a 3 AM no afecta al número de coches que pasan por dicho punto de 4 a 6 AM (aunque, siendo un intervalo más largo, puedan ser más).

Segunda definición

La primera definición de un proceso de Poisson puede entenderse como una visión *estática* del mismo, dado que en primer lugar se fija la ventana de tiempo t y posteriormente se analiza lo que ocurre en ella. A continuación se analiza el proceso desde un punto de vista *dinámico*, esto es, según suceden los eventos a lo largo del tiempo.

Tiempo entre llegadas

Sea X_1 el tiempo hasta el primer evento en un proceso de llegadas de Poisson, partiendo desde $t = 0$, según lo ilustrado en la Figura 3.5. La probabilidad de que dicho evento suceda más allá de un valor dado t

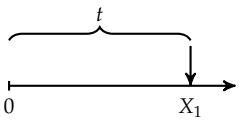


Figura 3.5: Que la primera llegada X_1 sea más allá de t coincide con que no haya llegadas durante un tiempo t.

$$\Pr(X_1 > t)$$

coincide con la probabilidad de que no suceda ninguna llegada en $(0, t)$, que viene dado por:

$$\Pr(N(t) = 0)$$

Igualando ambas expresiones, se tiene que

$$\Pr(X_1 > t) = \Pr(N(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t},$$

que es, por definición, la función de supervivencia de una variable aleatoria exponencial de media $1/\lambda$. Por lo tanto, se deduce que el tiempo hasta el primer evento X_1 se distribuye según una variable aleatoria exponencial.

Sea ahora el caso de la segunda llegada, que ocurre tras un tiempo X_2. Para analizar cómo se distribuye dicho tiempo, se supone que el tiempo X_1 de la primera llegada ocurre en un determinado instante s, y se analiza la probabilidad de que X_2 sea mayor que t condicionada a este valor de X_1, según se ilustra en la Figura 3.6:

$$\Pr(X_2 > t \mid X_1 = s) = \Pr(0 \text{ eventos en } (s, s+t) \mid X_1 = s)$$

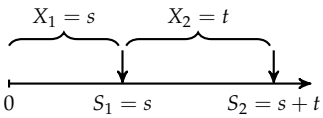


Figura 3.6: Caso de la segunda llegada.

Debido a la propiedad de incrementos estancionarios, esto se puede expresar como

$$\Pr(X_2 > t \mid X_1 = s) = \Pr(0 \text{ eventos en } (0, t)) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.$$

A la vista de lo anterior, se tiene que el tiempo entre la primera y la segunda llegada X_2 *también* se distribuye según una variable aleatoria exponencial, con el mismo parámetro λ. Este mismo desarrollo se puede repetir para la llegada i-ésima, obteniéndose que la variable X_i se distribuye según una variable aleatoria exponencial de media $1/\lambda$. De esto se deduce la siguiente definición de un proceso de Poisson:[3]

[3] El desarrollo presentado sirve para deducir la segunda definición partiendo de la primera. Para demostrar la equivalencia de ambas definiciones (que lo son) debería realizarse el razonamiento en el otro sentido, es decir, partir de la segunda definición y demostrar la primera.

Proceso de Poisson (2ª definición) Un proceso de Poisson $N(t)$ a tasa λ es un proceso de conteo en el que los tiempos entre eventos son independientes y se distribuyen según una variable aleatoria exponencial de media $1/\lambda$.

Ejemplo 3.5. Supóngase que el tiempo entre peticiones a un servidor puede modelarse como una variable aleatoria exponencial X de media 5 minutos. El número de peticiones por unidad de tiempo, por lo tanto, es un proceso de Poisson de media $\lambda = 1/5$ peticiones/minuto, esto es, 12 peticiones/hora. La probabilidad de no tener ninguna petición durante 30′ se puede calcular como

$$\Pr(X > 30') = \Pr(N_0^{30} = 0) = e^{-30/5} = e^{-6}$$

mientras que la probabilidad de no tener ninguna petición durante 5′ es

$$\Pr(X > 5') = \Pr(N_0^{5} = 0) = e^{-5/5} = e^{-1}$$

Propiedades de los procesos de Poisson

Una llegada de Poisson en un intervalo de tiempo dado puede ocurrir en *cualquier* momento, según lo visto en apartado anterior. A continuación se analizan tres propiedades fundamentales de los procesos de Poisson muy relacionadas con esta aleatoriedad.

Descomposición de procesos de Poisson

Si el apartado anterior trata sobre el agregado de procesos de Poisson, ahora se aborda el problema inverso: la descomposición de un proceso de Poisson en dos procesos. Sea un proceso de Poisson $\{N(t), \lambda\}$, que se descompone en dos procesos $N_1(t)$ y $N_2(t)$ de la siguiente forma: de forma independiente a cada llegada

- Con probabilidad p dicha llegada se asigna a N_1
- Con probabilidad $1 - p$ se asigna a N_2

tal y como se ilustra en la Figura 3.7.

En estas condiciones, como se demostrará a continuación, los procesos $N_1(t)$ y $N_2(t)$ también son procesos de llegada de Poisson e independientes, con tasas λp y $\lambda(1 - p)$, respectivamente.

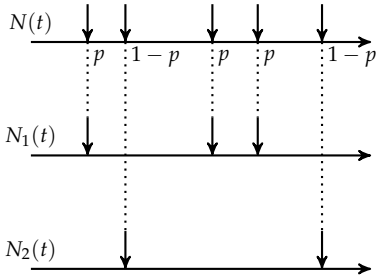


Figura 3.7: Descomposición de un proceso de Poisson en dos.

Ejemplo 3.6. En general, dividir un proceso de llegadas de Poisson no resulta en varios procesos de llegadas de Poisson. Sea un proceso de llegadas de Poisson $N(t)$ a tasa $\lambda = 10$ usuarios/hora, por tanto con la siguiente función de densidad del tiempo entre llegadas (suponiendo que t está en minutos):

$$f_X(t) = \frac{1}{6} e^{-t/6}, \quad t \geq 0$$

Supóngase que este proceso se divide en dos, según el tiempo entre una llegada y la siguiente X_i:

- $N_p(t)$: Llegadas pares.
- $N_i(t)$: Llegadas impares.

Ninguno de estos procesos es ya de Poisson, ya que el tiempo entre llegadas deja de ser exponencial, sino la *suma* de dos variables aleatorias exponenciales, según se ilustra en la Figura 3.8 (para los dos procesos de llegada). Como se vio en el capítulo anterior, dicha suma tiene una función de densidad

$$f_{N_i}(t) = f_{N_p}(t) = \frac{1}{6^2} t e^{-t/6}$$

que es diferente a la de una v.a.exponencial con la misma media

$$\frac{1}{12} e^{-t/12}, \quad t \geq 0$$

según se ilustra en la Figura 3.9

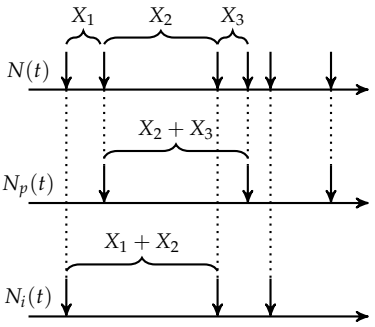


Figura 3.8: Descomposición de un proceso de Poisson en dos.

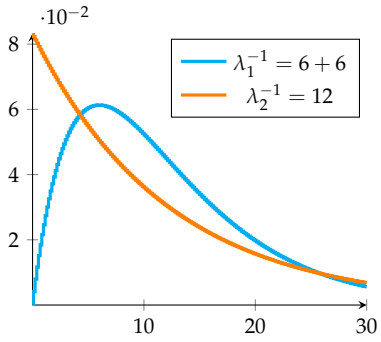


Figura 3.9: Función de densidad de una v.a.exponencial y de la suma de dos v.a.exp., con la misma media

A CONTINUACIÓN, SE DEMUESTRA QUE la descomposición de un proceso de Poisson en dos, seleccionando cada llegada con una probabilidad constante e independiente p, resulta en dos procesos de Poisson. Dicha demostración consiste en fijar un intervalo de tiempo t y analizar la probabilidad de tener n llegadas en el primer proceso y m llegadas en el segundo. Por simplicidad, no se hace referencia a la variable temporal t (puede ser cualquier intervalo de tiempo), por lo que $N(t)$, $N_1(t)$ y $N_2(t)$ pasan a expresarse como

$$N(t) \to N$$
$$N_1(t) \to N_1$$
$$N_2(t) \to N_2$$

La probabilidad (conjunta) de tener n llegadas en N_1 y m en N_2 se puede expresar de forma condicionada a tener $n + m$ llegadas en el proceso original N

$$\Pr(N_1 = n, N_2 = m) = \Pr(N_1 = n, N_2 = m \mid N = n+m) \Pr(N = n+m). \tag{3.2}$$

Obtener esta expresión requiere calcular dos probabilidades:

1. $\Pr(N = n + m)$, que es la probabilidad de tener $n + m$ llegadas en un proceso de Poisson, lo que resulta inmediato:

$$\Pr(N = n + m) = \frac{e^{-\lambda t}(\lambda t)^{n+m}}{(n+m)!} \quad . \tag{3.3}$$

2. $\Pr(N_1 = n, N_2 = m \mid N = n + m)$, que es la probabilidad de que, de un total de $n + m$ llegadas, n vayan al proceso N_1 y m vayan al proceso N_2.

Dado que toda llegada que no se asigne a N_1 se asignará a N_2, si el total de llegadas es $N = n + m$ no es necesario tener en cuenta tanto N_1 como N_2, basta con uno de ellos:[4]

$$\begin{aligned}\Pr(N_1 = n, N_2 = m \mid N = n + m) &= \Pr(N_1 = n \mid N = n + m) \\ &= \Pr(N_2 = m \mid N = n + m).\end{aligned}$$

[4] Por ejemplo, si $N = N_1 + N_2 = 3 + 7$, al suponer que $N_1 = 3$ ya se tiene implícito que $N_2 = 7$.

La probabilidad $\Pr(N_1 = n \mid N = n + m)$ representa la probabilidad de que de las $n + m$ llegadas, n se asignen al proceso N_1. Dado que para cada evento de llegada hay una probabilidad p de que esto ocurra, dicha probabilidad se puede calcular con la variable aleatoria binomial, interpretándose las $n + m$ llegadas como experimentos de Bernouilli que tienen éxito con probabilidad p. Por lo tanto, se tiene que[5]

$$\Pr(N_1 = n \mid N = n + m) = \binom{n+m}{n} p^n (1-p)^m. \tag{3.4}$$

[5] Se puede comprobar que se llegaría al mismo resultado calculando $\Pr(N_2 = m \mid N = n + m)$.

Substituyendo (3.3) y (3.4) en (3.2), y re-ordenando términos, se tiene que la probabilidad puede expresarse como

$$\Pr(N_1 = n, N_2 = m) = \frac{e^{-\lambda p t}(\lambda p t)^n}{n!} \cdot \frac{e^{-\lambda(1-p)t}(\lambda(1-p)t)^m}{m!}, \tag{3.5}$$

que resulta ser el producto de dos probabilidades independientes, esto es[6]

$$\Pr(N_1 = n, N_2 = m) = \Pr(N_1 = n) \cdot \Pr(N_2 = m)$$

donde N_1 es una variable aleatoria de Poisson de media $\lambda p t$ y N_2 es otra variable aleatoria de Poisson de media $\lambda(1-p)t$.

[6] Formalmente, para realizar este paso es preciso realizar el cálculo de las *distribuciones marginales* $\Pr(N_1 = n)$ y $\Pr(N_2 = m)$ a partir de (3.5), haciendo la suma sobre todos los valores de m y n, respectivamente.

Ejemplo 3.7. Sea un router al que llega un tráfico según un proceso de Poisson de tasa 100 tramas/ms y que, por un fallo en el *firmware*, descarta tramas con una probabilidad constante e independiente igual a $p = 1/100$. El proceso de salida de dicho router también es de Poisson, de tasa 99 tramas/ms.

Ejemplo 3.8. Sea un switch Ethernet con veinticuatro puertos de entrada y dos puertos de salida. El mecanismo de forwarding es tal que se puede suponer que el 70 % de todos los paquetes van al primer puerto de salida, y el 30 % restante al otro puerto (Figura 3.10). Si cada puerto de entrada recibe tramas según un proceso de Poisson a tasa $\lambda = 10$ paquetes/ms, el agregado de tramas que se reciben en el switch es otro proceso de Poison, de tasa

$$\lambda_T = 24 \times 10 = 240 \text{ paquetes/ms}.$$

Suponiendo que el mecanismo de *forwarding* se puede modelar como si cada trama es dirigida al puerto uno (dos) con una probabilidad del 70 % (30 %), entonces el proceso de salida del cada puerto también serán un proceso de Poisson, de tasas

$$\lambda_{o1} = 0{,}7 \times 240 = 168 \text{ paquetes/ms}$$
$$\lambda_{o2} = 0{,}3 \times 240 = 72 \text{ paquetes/ms}$$

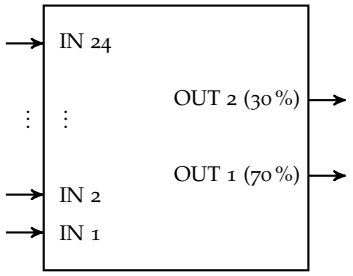


Figura 3.10: Switch con 24 puertos de entrada y 2 de salida.

Como ilustra el ejemplo anterior, las propiedades de agregado y descomposición de procesos de Poisson pueden simplificar en gran medida el modelado de sistemas, ya que los procesos resultantes seguirán manteniendo las propiedades de p.ej. "sin memoria," que resulta muy conveniente para su análisis. Sin embargo, en ocasiones surgirá el escollo de que los procesos de llegada no son procesos de Poisson: por ejemplo, un codificador de voz sobre IP con tiempo entre tramas prácticamente constante. En estas ocasiones, cuando haya *muchos* procesos (y, por lo tanto, el agregado sea *muy* impredecible), también se podrá aproximar el proceso global con un proceso de Poisson, como se verá más adelante.

Agregado de procesos de Poisson

Sean $\{N_1(t), \lambda_1\}$ y $\{N_2(t), \lambda_2\}$ dos procesos independientes de Poisson, cuyos tiempo entre llegadas son variables aleatorias exponenciales, de medias $1/\lambda_1$ y $1/\lambda_2$, respectivamente. Se persigue analizar el proceso de llegadas resultante de su agregado

$$N_T(t) = N_1(t) \cup N_2(t)$$

como se ilustra en la Figura 3.11.

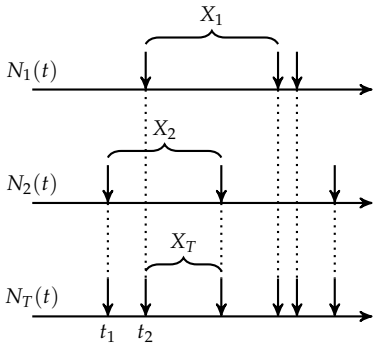


Figura 3.11: Agregado de dos procesos de llegada de Poisson.

La unión de dos procesos de llegada es otro proceso de llegadas, donde se unen los eventos ordenados en el tiempo de cada uno de los procesos. Por ejemplo, si las llegadas de cada proceso en los primeros 10 segundos fuesen

$$N_1 = \{3, 8, 9\}$$
$$N_2 = \{1, 4, 7, 10\}$$

el proceso agregado resultaría ser

$$N_T = \{1, 3, 4, 7, 8, 9, 10\}$$

Dicho proceso $N_T(t)$ se puede caracterizar de dos formas, según las dos primeras definiciones de un proceso de Poisson:

1. Analizando el tiempo entre llegadas X_T del proceso agregado.

Por definición, la construcción del proceso agregado consiste en, a cada instante, seleccionar la llegada de N_1 o N_2 que ocurra antes. Según se ilustra en la Figura 3.11, en la segunda llegada del proceso agregado (marcada como t_2), el tiempo (aleatorio) hasta la siguiente llegada X_T es el mínimo de los tiempos entre llegadas de cada proceso, X_1 y X_2 (por la propiedad sin memoria de X_2):

$$X_T = \text{mín}(X_1, X_2),$$

que es otra variable aleatoria exponencial,[7] de media $1/(\lambda_1 + \lambda_2)$. Dado que dicho tiempo entre llegadas es el inverso de la tasa del proceso, se tiene que el agregado es otro proceso de Poisson, pero de tasa $\lambda_1 + \lambda_2$.

[7] Mínimo de variables aleatorias exponenciales, página 38

2. Analizando el número de llegadas en una ventana de tiempo.

Dada una ventana de tiempo t, el número de llegadas total viene dado por la suma de las llegadas de cada uno de los procesos, que se corresponde con la suma de dos variables aleatorias discretas de Poisson, de media $\lambda_1 t$ y $\lambda_2 t$, respectivamente. Como ya se analizó,[8] dicha suma se trata de otra variable aleatoria de Poisson, de media $(\lambda_1 + \lambda_2)t$.

[8] Ejemplo 1.18, página 22

Dado que lo anterior se cumple para cualquier ventana de tiempo t, el agregado de dos procesos de Poisson es otro proceso de Poisson, cuya tasa es la suma de las tasas de aquellos.

INDEPENDIENTEMENTE DEL RAZONAMIENTO SEGUIDO, el resultado se puede generalizar a varios procesos: sea un conjunto de n procesos independientes de llegadas de Poisson, cada uno a una tasa λ_i. El agregado de dichos procesos es también un proceso de llegadas de Poisson, de tasa

$$\lambda = \sum_{1}^{n} \lambda_i \ .$$

Ejemplo 3.9. Un switch recibe tramas de tres ordenadores, con tasas de 10, 20 y 40 tramas por milisegundo. Suponiendo que dichos fujos son independientes y de Poisson, se tiene que el agregado de los dos primeros flujos es otro proceso de Poisson, de tasa 30 tramas/milisegundo, y que el agregado de éste con el tercer flujo es también un proceso de Poisson de tasa 70 tramas/milisegundo.

Teorema de Palm-Khintchine

En un proceso de llegadas de Poisson se tiene que en «cualquier instante de tiempo» existe la misma probabilidad de que se produzca una llegada. Como se ha visto, esto se puede deducir tanto de la segunda definición (la probabilidad de no tener una llegada durante t unidades de tiempo no varía con el tiempo) como de la tercera (la probabilidad de una llegada en un infinitesimal de tiempo Δt es siempre la misma). De hecho, es relativamente sencillo pensar en otros procesos de llegada que no sean de Poisson y que, por tanto, no cumplan estas características:

- Si el tiempo entre llegadas es constante e igual a T, pasado un breve intervalo de tiempo ϵ tras la última llegada es seguro que la siguiente llegada no sucederá hasta $T - \epsilon$.
- Si el tiempo entre llegadas se distribuye uniformemente en $[t_1, t_2]$, dado que esta variable aleatoria sí tiene memoria, la probabilidad de que no haya una llegada no es constante a lo largo del tiempo.

Estos casos ilustran que un proceso con el tiempo entre llegadas constante o uniformemente distribuido no se parecerá a un proceso

de llegadas de Poisson. Sin embargo, cuando se trate del agregado de *muchos* procesos de llegadas independientes (no necesariamente de Poisson), la situación puede que sea diferente: dado que en este proceso agregado será «muy complicado» predecir cuándo se producirá la siguiente llegada, su comportamiento será «parecido» al de un proceso de Poisson (y tanto más parecido cuantos más procesos compongan el agregado). Este resultado es el Teorema de Palm-Khintchine.[9]

[9] Obra de Conrad Palm (1907–1951), ingeniero y estadístico sueco, y Aleksandr Khinchine (1894–1959), matemático soviético.

Para una explicación intuitiva de que al agregar *muchos* procesos de llegadas se *tiende* a un proceso de Poisson, sea el agregado de varios procesos de llegadas independientes, cada uno con un tiempo entre llegadas que se distribuye uniformemente, como el ilustrado en la Figura 3.12 a continuación. Resulta evidente que, cada uno por separado, no son procesos de llegadas de Poisson, dado que –por ejemplo– el tiempo entre llegadas dista mucho de parecerse a una variable aleatoria exponencial.

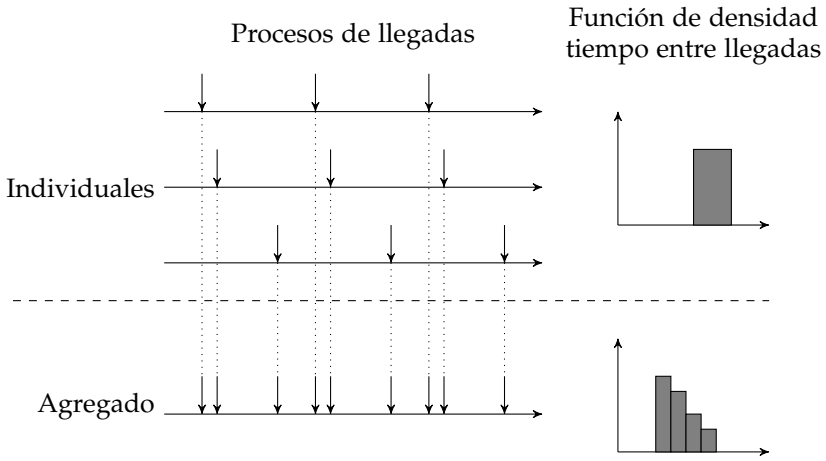


Figura 3.12: Ilustración del teorema de Palm-Khintchine: el agregado de procesos de llegadas independientes tiende a comportarse como un proceso de Poisson.

Sin embargo, si se considera el proceso resultante de agregar esos procesos, el tiempo entre llegadas se reduce: tras una llegada correspondiente a un proceso, es *bastante* probable que se produzca otra llegada correspondiente a cualquiera de los otros procesos. Por el mismo motivo, los tiempos entre llegadas elevados tienden a ser cada vez menos probables. Como se ilustra en la Figura 3.12, la función de densidad del proceso agregado tiende entonces a parecerse a la de una variable aleatoria exponencial.

El teorema de Palm-Khintchine formaliza este comportamiento, enunciando que, para un número *suficiente* de procesos de llegada, cada uno con una intensidad relativamente baja, dichos tiempos entre llegadas tienden a ser una variable aleatoria exponencial. Se puede interpretar este teorema como el teorema central del límite[10] para los procesos de llegadas.

[10] El teorema central del límite establece que la distribución de la suma de variables aleatorias independientes se aproxima bien a una distribución normal.

Teorema of Palm-Khintchine Sea un conjunto de n procesos de llegada independientes $\{N_i(t), t \geq 0\}$, cada uno de ellos a tasa λ_i,

donde $i = 1, 2, \ldots, n$. La superposición de todos los procesos

$$N(t) = \sum_{i=1}^{n} N_i(t), t \geq 0$$

tiende a ser un proceso de Poisson a tasa $\lambda = \sum \lambda_i$ según $n \to \infty$, siempre que se cumpla que:

1. La carga total λ sea finita.
2. Ningún proceso «domine» el agregado, es decir, $\lambda_i \ll \lambda$

La importancia del teorema de Palm-Khintchine es clara, dado que permite aproximar el comportamiento de procesos de llegada agregados mediante un proceso de Poisson, que resulta –por todo lo visto hasta ahora– sencillo de modelar. De esta forma, en multitud de situaciones si bien el proceso de eventos generados por *un* usuario no pueda considerarse de Poisson, el *agregado* de varios usuarios sí que se podrá considerar de Poisson.[11]

[11] De hecho, Erlang y Palm sólo encontraron argumentos heurísticos para modelar el número de llamadas a una centralita con una distribución de Poisson. Fueron Ososkov (1956) y Khintchine (1960) los que demostraron formalmente las condiciones necesarias y suficientes para esta aproximación.

Ejemplo 3.10. Sea un switch con sesenta y cuatro puertos de entrada, cada uno conectado a un teléfono VoIP. El codificador de voz genera una trama cada un tiempo aleatorio, que se distribuye según una variable aleatoria $\mathcal{U}(8, 12)$ ms. Por lo tanto, cada teléfono genera 1 trama cada 10 ms, pero no según un proceso de Poisson.

Sin embargo, en estas condiciones, teniendo en cuenta que el número de procesos es suficientemente elevado, se podría suponer que el agregado se puede modelar con un proceso de Poisson, de tasa

$$\lambda_{tot} = \sum_{i=1}^{64} \lambda_i = 64 \times \frac{1 \text{ trama}}{10 \text{ ms}} = 6.4 \text{ tramas/ms}.$$

Distribución condicionada de una llegada

Supóngase ahora que se conoce que ha ocurrido un evento (y sólo uno) en el intervalo $[0, t]$. Se pretende analizar la probabilidad de que dicho evento haya sucedido en un instante dado de dicho intervalo, con objeto de, por ejemplo, determinar si hay sub-intervalos donde es más probable que dicho evento haya ocurrido.

Ejemplo 3.11. Supóngase que los goles siguen una variable aleatoria discreta de Poisson y que un equipo ha metido un gol en la primera parte. Se plantea ahora la cuestión de estimar cuándo se produjo dicho gol, esto es, ¿en qué ventanas de tiempo es más probable que se haya producido dicho gol? ¿Hay *zonas calientes* donde el equipo marque con más probabilidad?

Una posible intuición sería suponer que el gol debe haberse producido hacia la mitad del tiempo considerado (como se ilustra en azul en la Figura 3.13). Otra opción sería pensar que, dado que el tiempo entre goles es exponencial, el gol debe tener una función

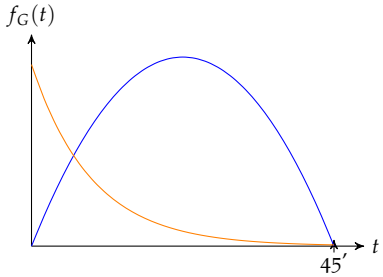


Figura 3.13: Sabiendo que un equipo ha marcado un único gol en la primera parte, se podría plantear si es más probable que hubiese ocurrido hacia la mitad del primer tiempo (azul) o hacia el principio del partido (naranja).

de densidad que se pareciese a la exponencial (en naranja en la Figura 3.13). Como se verá a continuación, ambas interpretaciones son incorrectas.

FORMALMENTE, LO QUE SE PERSIGUE es calcular qué valores de X_1 (esto es, la primera llegada) son más probables en un intervalo de longitud t, sabiendo que en dicho intervalo hubo una única llegada. Como se ha hecho anteriormente, para caracterizar una variable aleatoria se puede calcular su función de distribución F, por lo que el problema se puede formular como calcular la probabilidad

$$\Pr(X_1 < s \mid N(t) = 1)$$

para cualquier valor de $s \in [0, t]$. Esto se puede expresar, por definición de probabilidad condicionada, como

$$\Pr(X_1 < s | N(t) = 1) = \frac{\Pr(X_1 < s, N(t) = 1)}{\Pr(N(t) = 1)} \tag{3.6}$$

El numerador de la parte derecha (3.6) es la probabilidad de que haya una llegada en el intervalo $[0, s)$ y, a la vez, sólo una llegada en todo el intervalo $[0, t]$, por lo que se puede expresar como

$$\begin{aligned}\Pr(X_1 < s, N(t) = 1) &= \Pr(1 \text{ llegada en } (0, s)) \times \Pr(0 \text{ llegadas en } (s, t)) \\ &= \frac{(\lambda s)^1}{1!} e^{-\lambda s} \times \frac{(\lambda (t - s))^0}{0!} e^{-\lambda (t-s)}\end{aligned}$$

mientras que el denominador de la parte derecha (3.6) queda como

$$\begin{aligned}\Pr(N(t) = 1) &= \Pr(1 \text{ llegada en } (0, t)) \\ &= \frac{(\lambda t)^1}{1!} e^{-\lambda t}\end{aligned}$$

por lo que (3.6) queda como

$$\Pr(X_1 < s \mid N(t) = 1) = \frac{\frac{(\lambda s)^1}{1!} e^{-\lambda s} \times \frac{(\lambda (t-s))^0}{0!} e^{-\lambda (t-s)}}{\frac{(\lambda t)^1}{1!} e^{-\lambda t}} = \frac{s}{t} \ .$$

Según este resultado, la función de distribución del tiempo de la primera llegada coincide con la función de una variable aleatoria uniformemente distribuida en $[0, t]$ (ilustrada en la Figura 3.14), esto es

$$\Pr(X_1 < s \mid N(t) = 1) \sim U[0, t].$$

Dado que para la variable aleatoria uniforme la función de densidad es constante, se tiene que cualquier instante de tiempo s es igual de probable para que ocurra esa única llegada, por lo que no hay «zonas calientes» donde sea más probable que haya sucedido. En términos de incertidumbre, y refiriéndose al ejemplo anterior sobre un gol en un partido de fútbol, se tiene que dicho gol pudo haber ocurrido en cualquier momento de la primera parte: es tan probable un gol en el primer minuto como en el último.

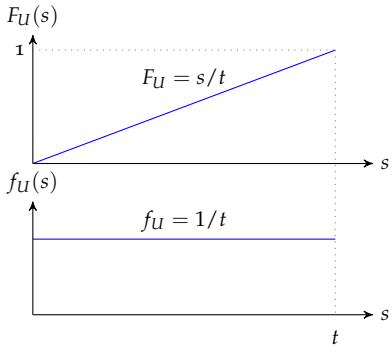


Figura 3.14: Función de distribución (arriba) y densidad (abajo) de una variable aleatoria uniformemente distribuida en $(0, t)$.

PASTA

El acrónimo PASTA proviene del inglés *Poisson Arrivals See Time Averages*, y podría traducirse como «un proceso de llegadas de Poisson *ve* medias temporales», siendo una de de las propiedades de mayor importancia en la teoría de colas.[12] Para definirla, es preciso distinguir entre dos «medias» en un proceso aleatorio:

[12] También se le conoce como la propiedad ROP, de *Random Observer Property*

- La media temporal del mismo, esto es, su valor medio a lo largo del tiempo.
- La media de los valores que obtenga un observador externo que «muestree» el proceso en algunos instantes de tiempo.

Ejemplo 3.12. Sea un teléfono VoIP que manda una trama de 80 B cada 10 ms a través de un enlace de 100 Kbps, con el siguiente patrón de tráfico La ocupación media del enlace puede obtenerse

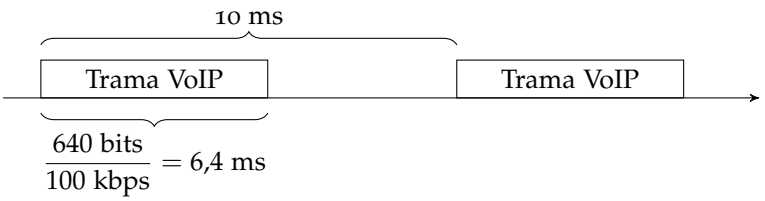


Figura 3.15: Patrón de tráfico de 80 B cada 10 ms en un enlace a 100 Kbps.

como el tiempo que se tarda en transmitir una trama dividido por el tiempo entre tramas, esto es

$$\rho = \frac{6{,}4 \text{ ms}}{10 \text{ ms}} = 0{,}64$$

por lo que se puede decir que el canal está ocupado el 64 % del tiempo. Sin embargo, un proceso de muestreo periódico cada 10 ms no medirá esta ocupación, ya que siempre «verá» lo mismo el 100 % del tiempo: bien el canal libre, bien el canal ocupado (según se ilustra en la figura a continuación).

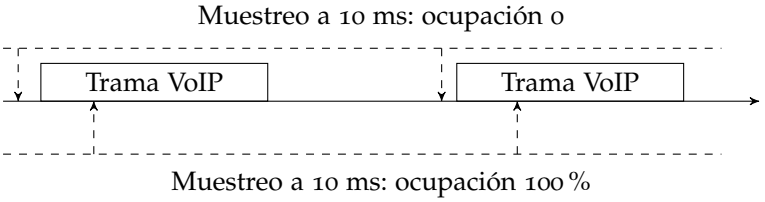


Figura 3.16: La ocupación media que estima un muestreo no suele coincidir con la ocupación a lo largo del tiempo.

Como se verá a continuación, cuando el muestreo sigue un proceso de Poisson, y por tanto tiene la adecuada naturaleza aleatoria, entonces lo que «ve» dicho muestreo coincide con la media a lo largo del tiempo del proceso.

SEA AHORA UN CASO GENERAL, como el proceso aleatorio $X(\tau)$ ilustrado en la Figura 3.17, que toma tres valores a lo largo de un tiempo t: A durante a unidades de tiempo, B durante b y C durante

c. Durante este intervalo total $t = a + b + c$, la media a lo largo de tiempo de X se puede calcular como:

$$\overline{X}(t) = \frac{1}{t}\int_0^t X(\tau)d\tau = \frac{A \cdot a + B \cdot b + C \cdot c}{t}$$

Se tiene que, en dicho intervalo t, se produce una llegada según un proceso de Poisson. Sea Y la variable aleatoria definida como el valor que tenga el proceso X cuando sucede dicha llegada, esto es

$$Y = \begin{cases} A & \text{si la llegada ocurre en } a \\ B & \text{si la llegada ocurre en } b \\ C & \text{si la llegada ocurre en } c \end{cases}$$

La esperanza de dicha variable Y se puede expresar como

$$\mathbb{E}[Y] = A \cdot \Pr(a) + B \cdot \Pr(b) + C \cdot \Pr(c), \tag{3.7}$$

donde $\Pr(x)$ representa la probabilidad de que la llegada de Poisson ocurra en el intervalo x. La probabilidad de que dicha llegada ocurra en el intervalo a se puede expresar como la probabilidad de que, dado que hay una llegada en t, dicha llegada ocurra en a y no en b o en c:

$$\Pr(a) = \frac{\Pr(1 \text{ llegada en } a)\Pr(0 \text{ llegadas en } b)\Pr(0 \text{ llegadas en } c)}{\Pr(1 \text{ llegada en } t)} \ ,$$

por lo que

$$\Pr(a) = \frac{\dfrac{(\lambda a)^1 e^{-\lambda a}}{1!} \cdot \dfrac{(\lambda b)^0 e^{-\lambda b}}{0!} \cdot \dfrac{(\lambda c)^0 e^{-\lambda c}}{0!}}{\dfrac{(\lambda t)^1 e^{-\lambda t}}{1!}} = \frac{a}{t} \ .$$

Siguiendo un razonamiento similar, se tiene que

$$\Pr(b) = \frac{b}{t} \quad \text{y} \quad \Pr(c) = \frac{c}{t} \ ,$$

es decir: la probabilidad de que la llegada se produzca en un determinado intervalo es igual a la longitud relativa de dicho intervalo con respecto al total.[13] Con esto, (3.7) queda como

$$\mathbb{E}[Y] = A \cdot \frac{a}{t} + B \cdot \frac{b}{t} + C \cdot \frac{c}{t} \ ,$$

que coincide con la expresión para $\overline{X}$. Por lo tanto, el valor esperado de lo que «ve» una llegada de Poisson ($\mathbb{E}[Y]$) coincide con la media temporal del proceso (X).

Ejemplo 3.13. Sea un aula que se llena con 40 alumnos de 9 a 12 de la mañana y el resto del tiempo permanece vacía. La ocupación media durante un día lectivo de dicha aula se puede calcular como

$$\overline{N} = \frac{3 \text{ horas}}{24 \text{ horas}} \times 40 \text{ alumnos} + \frac{21 \text{ horas}}{24 \text{ horas}} \times 0 = 5 \text{ alumnos.}$$

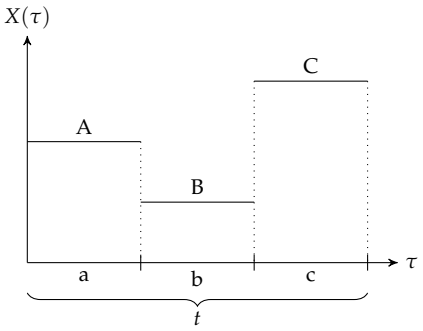


Figura 3.17: Proceso que pasa por tres estados a lo largo del tiempo.

[13] Este hecho también se podría haber deducido del análisis de la distribución condicionada de una llegada, visto en el apartado anterior.

Sin embargo, este número medio de alumnos no coincide ni con lo que «ve» un bedel que sólo llegue al aula para a las 8 de la mañana (es decir: 0 alumnos), ni con lo que «ve» un profesor que siempre llegue a clase 5 minutos tarde (40 alumnos), por lo que estos procesos de muestreo no son de Poisson.

Suponiendo que los guardias de seguridad tuviesen la orden de pasar por las aulas de forma «completamente aleatoria» (esto es, según un proceso de Poisson), sí se tendría que $1/8$ del tiempo un guardia vería 40 alumnos, y $7/8$ del tiempo el aula vacía: en este caso, la media de sus observaciones coincidirá con $N = 5$ alumnos por la propiedad PASTA.

Tercera definición

La primera definición de un proceso de Poisson se puede interpretar como una definición *estática*, dado que lo que caracteriza al proceso es que, fijada cualquier ventana de tiempo T_i a lo largo del mismo, el número de eventos que ocurren en dicho intervalo sigue la distribución de Poisson. Como se ilustra en la Figura 3.18, dada una tasa de llegadas λ se podrían ir definiendo las ventanas T_1, T_2, T_3, etc., (de un tamaño dado) y el resultado sería que el número de eventos en cada ventana ($N(T_i)$) sigue una variable aleatoria discreta de Poisson:

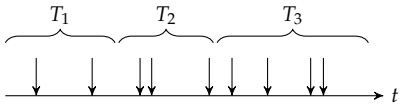


Figura 3.18: Ilustración de la primera definición de un proceso de Poisson.

$$N(T_i) \sim \text{Poisson}(\lambda T_i).$$

Por otro lado, la segunda definición tiene un carácter *dinámico*, que caracteriza lo que pasa entre una llegada y la siguiente. Como se ilustra en la Figura 3.19, en este caso la definición se basa en los tiempos entre llegadas X_1, X_2, X_3, ... que, para un proceso a tasa λ, se distribuyen según una variable aleatoria exponencial de media $1/\lambda$:

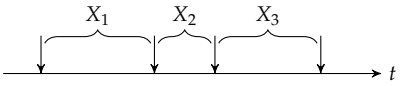


Figura 3.19: Ilustración de la segunda definición de un proceso de Poisson.

$$X_i \sim \text{Exp}(\lambda).$$

En este apartado se presenta la tercera (y última) definición del proceso de Poisson, que también tiene una interpretación dinámica pero en vez de analizar lo que ocurre entre una llegada y la siguiente, caracteriza lo que puede ocurrir «en cualquier momento» de tiempo: por lo tanto, se puede decir que tiene un carácter *instantáneo*. La definición se basa en intervalos tiempo muy cortos Δt y caracteriza lo que puede ocurrir en dichos intervalos. Como se ilustra en la Figura 3.20, en un determinado intervalo sólo podrán pasar dos cosas: que se produzca una llegada, o que no se produzca ninguna. Dado que esta definición trata con intervalos muy cortos de tiempo (esto es, infinitesimales), es preciso recordar la definición de una función $o(h)$.

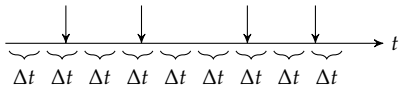


Figura 3.20: Ilustración de la tercera definición de un proceso de Poisson.

Funciones $o(h)$

Se dice que una función $f(x)$ es $o(h)$ si se cumple que

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

esto es, que la función $f(x)$ tiene a cero según x se hace menor de forma más rápida que si tuviese una caída lineal.

Por ejemplo, y tal y como se ilustra en la Figura 3.21, la función $f(x) = x^2$ es $o(h)$, dado que se cumple que

$$\lim_{h\to 0} \frac{h^2}{h} = \lim_{h\to 0} h = 0$$

La *intuición* detrás de este artificio de funciones $o(h)$ es que permiten realizar aproximaciones al manejar infinitesimales, ya que dichas funciones podrán «despreciarse» frente a componentes que crezcan (o se empequeñezcan) según una función lineal.

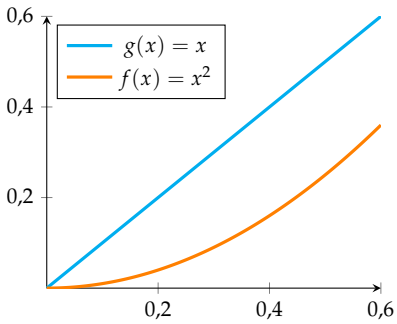


Figura 3.21: La función $f(x) = x^2$ es $o(h)$.

Resulta sencillo demostrar las siguientes propiedades

- Si las funciones $f(x)$ y $g(x)$ son $o(h)$, su suma $f(x) + g(x)$ también es $o(h)$.
- Si la función $f(x)$ es $o(h)$ y c es una constante, la función $c \cdot f(x)$ también es $o(h)$.

Tercera definición

En un proceso de llegadas a tasa λ, durante un intervalo de tiempo t deben producirse en media λt llegadas. Esta definición se basa en las funciones $o(h)$ para caracterizar lo que pasa cuando dicho intervalo t tiende a ser de tamaño mínimo.

Proceso de Poisson (3^a definición) Un proceso de Poisson $N(t)$ a tasa λ es un proceso de conteo que cumple las siguientes propiedades:

(i) $N(0) = 0$

(ii) $N(t)$ tiene incrementos independientes y estacionarios.

(iii) $\Pr(N(h) = 1) = \lambda h + o(h)$

(iv) $\Pr(N(h) \geq 2) = o(h)$

Dado que de (iii) y (iv) se puede deducir que

$$\Pr(N(h) = 0) = 1 - \lambda h + o(h),$$

se tiene que, sin considerar el término $o(h)$, en un intervalo de tiempo h suficientemente pequeño pueden pasar dos cosas:

- Que se produzca una llegada, con probabilidad λh
- Que no se produzca ninguna llegada, con probabilidad $1 - \lambda h$

Esta definición motiva que un proceso de Poisson sea considerado como un «proceso *puro* de llegadas», dado que en cada uno de los intervalos Δt se tiene que la probabilidad de que ocurra una llegada es siempre la misma, proporcional a λ, e independiente de los intervalos anteriores.

De la segunda definición a la tercera

Por la segunda definición de un proceso de llegadas de Poisson (el tiempo entre llegadas sigue una variable aleatoria exponencial), la probabilidad de que pase un tiempo t sin que se produzca una llegada viene dada por

$$\Pr(N(t) = 0) \triangleq \Pr(x > t) = e^{-\lambda t} \ ,$$

que, realizando la aproximación de la exponencial mediante una serie de Taylor centrada en el origen,[14] puede aproximarse como

$$\Pr(N(\Delta t) = 0) = 1 - \lambda \Delta t + o(\Delta t)$$

La probabilidad de tener al menos una llegada durante t es

$$\Pr(N(t) \geq 1) \triangleq \Pr(x < t) = 1 - e^{-\lambda t} \ ,$$

cuyo desarrollo es

$$\Pr(N(t) \geq 1) = \lambda \Delta t + o(\Delta t) \ .$$

Esta probabilidad incluye la posibilidad de que se produzca más de una llegada en un intervalo de tiempo t. Por lo tanto, la probabilidad de tener exactamente una llegada es la diferencia de la probabilidad de tener al menos una llegada y la probabilidad de tener al menos dos llegadas

$$\Pr(N(t) = 1) = \Pr(N(t) \geq 1) - \Pr(N(t) \geq 2)$$

Tener al menos dos llegadas en un intervalo de tiempo t coincide con la probabilidad de que la suma de dos variables aleatorias exponenciales sea menor que t. Según lo visto en el capítulo anterior[15] se obtiene que esta probabilidad es

$$\Pr(N(t) \geq 2) = 1 - e^{-\lambda t} - \lambda t e^{-\lambda t} \ ,$$

cuyo desarrollo es

$$\Pr(N(t) \geq 2) = (\Delta t)^2 + o((\Delta t)^2) \ ,$$

de lo que se deduce que $\Pr(N(t) \geq 2) = o(\Delta t)$ y, por lo tanto,

$$\Pr(N(t) = 1) = \lambda \Delta t + o(\Delta t)$$

A partir de estos resultados se deducen las propiedades (*iii*) y (*iv*) de la tercera definición. Dado que la variable aleatoria exponencial no tiene memoria estas probabilidades son constantes, por lo que los incrementos son independientes y estacionarios –propiedad (*ii*)–, mientras que la condición dada por la propiedad (*i*) es trivial.

Equivalencia entre las definiciones de un proceso de Poisson

Se han presentado tres definiciones de un proceso de llegadas de Poisson a tasa λ; de forma resumida, son:

[14] Serie de Mac Laurin, que para la exponencial es

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + o(x^n)$$

[15] Suma de variables aleatorias exponenciales (*), página 42

1. El número de llegadas en un intervalo de tiempo t sigue una distribución discreta de Poisson de media λt (definición «Poisson»).

2. El tiempo medio entre llegadas sigue una variable aleatoria exponencial de media $1/\lambda$ (definición «Exponencial»).

3. La probabilidad de una llegada en un intervalo h es $\lambda h + o(h)$ (definición «Infinitesimal»).

Además, se ha visto que de la primera definición se puede deducir la segunda (página 47, al analizar el tiempo entre llegadas de un proceso de Poisson) y que de la segunda se puede deducir la tercera (en el apartado anterior). Para demostrar que las tres definiciones son equivalentes, falta por ilustrar que de la tercera definición se puede deducir la primera, como se ilustra en la Figura 3.22.

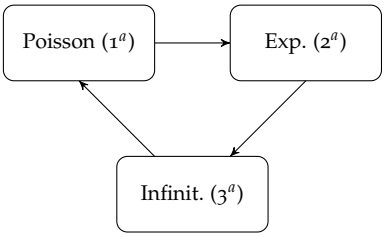


Figura 3.22: Equivalencia entre las definiciones de un proceso de llegadas de Poisson.

Ejemplo 3.14. Una manera *informal* de pasar de la tercera definición a la primera es la siguiente: sea un intervalo de tiempo t, como el ilustrado en la Figura 3.23, dividido en m sub-intervalos de longitud Δt (por lo que $m = t/\Delta t$). Considerando una llegada como un «éxito», el intervalo t se puede interpretar como una sucesión de m experimentos de Bernouilli, donde la probabilidad de éxito p es, por tanto,

$$p \triangleq \Pr(N(t/m) = 1) = \lambda \frac{t}{m} + o\left(\frac{t}{m}\right).$$

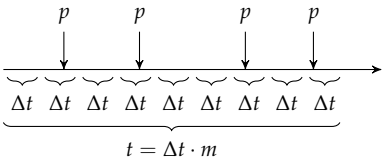


Figura 3.23: Llegadas de un proceso de Poisson vistas como experimentos de Bernouilli.

Como se vio en el primer tema (Distribución de Poisson, página 22), cuando m aumenta la distribución binomial se puede aproximar con una distribución de Poisson, de media mp. Cuando $\Delta t \to 0$ (esto es, $m \to \infty$) se tiene que

$$\lim_{m\to\infty} m \Pr(N(t/m) = 1) = \lim_{m\to\infty} m \left(\lambda \frac{t}{m} + o\left(\frac{t}{m}\right)\right),$$

que queda como

$$\lim_{m\to\infty} m \Pr(N(t/m) = 1) = \lambda t + \lim_{m\to\infty} m \cdot o\left(\frac{t}{m}\right).$$

Dado este resultado, queda la dificultad de calcular el límite de la expresión con una función $o(h)$. Para calcular este último límite, se multiplica y divide por t (que se trata como una constante), teniéndose que

$$\lim_{m\to\infty} m \cdot o\left(\frac{t}{m}\right) = \lim_{m\to\infty} t \cdot \frac{o\left(t/m\right)}{t/m} = 0,$$

por definición de función $o(h)$. Por lo tanto, la distribución de Poisson tiene de media

$$\lim_{m\to\infty} m \Pr(N(t/m) = 1) = \lambda t,$$

como era de esperar.

De la tercera definición a la primera (*)

La tercera definición establece lo que puede pasar en un infinitesimal de tiempo Δt, basándose en las funciones $o(h)$, mientras que la primera definición establece la probabilidad de n llegadas en un intervalo de tiempo t. Para simplificar, se empleará la siguiente notación:

$$\Pr(N(t) = n) = P_n(t).$$

A continuación se muestra cómo a partir de la tercera definición es posible deducir la expresión para $P_n(t)$, esto es, la variable aleatoria discreta de Poisson. Para ello, se parte de la tercera definición para analizar cómo varía $P_n(t)$ al pasar de t a $t + \Delta t$. A partir de esta diferencia, se obtiene la *derivada* de $P_n(t)$ mediante división por Δt, calculando el límite cuando $\Delta t \to 0$ (lo que permitirá simplificar la parte correspondiente a $o(h)$).

- Caso de $P_0(t)$

La probabilidad de no tener ninguna llegada en un instante de tiempo $t + h$ se puede expresar, por la propiedad de los incrementos independientes, como el producto de la probabilidad de no tener ninguna llegada en t por la probabilidad de no tener ninguna llegada en h:

$$P_0(t + h) = P_0(t) \cdot P_0(h) \tag{3.8}$$

Ya se ha visto que $P_0(h)$, según los puntos (*iii*) y (*iv*) de la definición, se puede expresar como

$$P_0(h) = 1 - P_1(h) - P_{\geq 2}(h) = 1 - \lambda h + o(h),$$

por lo que (3.8) queda como

$$P_0(t + h) = P_0(t) \cdot (1 - \lambda h + o(h)),$$

que se puede reordenar de la siguiente forma

$$\frac{P_0(t + h) - P_0(t)}{h} = -\lambda P_0(t) + \frac{o(h)}{h} P_0(t). \tag{3.9}$$

Calculando el límite cuando $h \to 0$, el término $\frac{o(h)}{h}$ desaparece (dado que $P_0(t)$ es una constante), por lo queda

$$\lim_{h \to 0} \frac{P_0(t + h) - P_0(t)}{h} = -\lambda P_0(t). \tag{3.10}$$

La parte izquierda de la anterior expresión corresponde con la definición de la derivada de $P_0(t)$. Expresando ésta como $P_0'(t)$, se tiene por lo tanto que para calcular $P_0(t)$ hay que resolver

$$P_0'(t) = -\lambda P_0(t),$$

que es una ecuación diferencial que tiene una solución de tipo

$$P_0(t) = Ke^{-\lambda t},$$

donde K es una constante que hay que determinar. De la propiedad (i) de la definición se tiene que $P_0(0) = 1$, por lo que $K = 1$ y por lo tanto la expresión para $P_0(t)$ es:

$$P_0(t) = e^{-\lambda t}.$$

De esta forma, queda demostrado que siguiendo la tercera definición de un proceso de Poisson, la probabilidad de no tener llegadas en un intervalo de tiempo t tiene la misma expresión que la dada por la primera definición.

- Caso de $P_1(t)$

Tener una única llegada durante un tiempo $t + h$ puede ocurrir de dos formas (Figura 3.24)

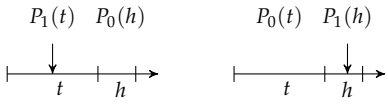


Figura 3.24: Por los incrementos independientes, $P_1(t + h)$ se puede calcular considerando los intervalos t y h por separado.

1. Teniendo una llegada en t y ninguna en h.
2. No teniendo llegadas en t y una llegada en h.

De nuevo, por la propiedad de los incrementos independientes, $P_1(t + h)$ se pude expresar como la suma del producto de dos probabilidades

$$P_1(t + h) = P_1(t)P_0(h) + P_0(t)P_1(h),$$

que resulta en la siguiente expresión:

$$P_1(t + h) = P_1(t)(1 - \lambda h + o(h)) + P_0(t)(\lambda h + o(h)),$$

Siguiendo un procedimiento similar al caso anterior, se obtiene la siguiente ecuación a resolver para calcular $P_1(t)$

$$P_1'(t) = -\lambda P_1(t) + P_0(t),$$

que tiene como solución:

$$P_1(t) = \lambda t e^{-\lambda t}.$$

Por lo tanto, siguiendo la tercera definición, se demuestra que la probabilidad de tener una llegada en un intervalo de tiempo t tiene la misma expresión que la dada por la primera definición.

- Caso de $P_n(t)$

De forma similar al caso anterior, la probabilidad de n llegadas en un intervalo $t + h$ de tiempo se puede expresar como

$$P_n(t + h) = P_n(t)(1 - \lambda h + o(h)) + P_{n-1}(t)(\lambda h + o(h)),$$

que da lugar a otra ecuación diferencial. Se puede demostrar, por inducción, que la solución general para la probabilidad de n llegadas es

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, \ldots$$

que equivale a la probabilidad de n llegadas según la primera definición de un proceso de llegadas de Poisson.

Resumen del tema

- Un proceso de Poisson a tasa λ es un proceso de llegadas con incrementos independientes y estacionarios.
- Se puede definir de tres formas equivalentes:
  1. El número de llegadas en un intervalo de tiempo t sigue una distribución discreta de Poisson de media λt.
  2. El tiempo medio entre llegadas sigue una variable aleatoria exponencial de media $1/\lambda$.
  3. La probabilidad de una llegada en un intervalo h es $\lambda h + o(h)$.
- Un proceso de llegadas de Poisson ve medias temporales (propiedad PASTA).
- El agregado de procesos de Poisson es otro proceso de Poisson.
- La descomposición de procesos de Poisson, de forma independiente a cada llegada, da lugar a otros procesos de Poisson.
- El agregado de procesos de llegada independientes (de cualquier tipo) tiende a comportarse como un proceso de Poisson (teorema de Palm-Khintchine).

Teoría de colas: fundamentos

Una cola es un sistema con uno o más recursos disponibles para una población de usuarios, que puede que tengan que esperar para poder acceder a los mismos por estar ocupados por otros usuarios. La «teoría de colas» se dedica al estudio de dichos sistemas, que sirven para modelar, por ejemplo, una central de conmutación de llamadas donde puede producirse saturación en las líneas (y se ruega volver a llamar dentro de unos minutos), o un centro de atención a usuarios, donde las llamadas pueden quedar a la espera de que un operador quede disponible.

En el campo de las comunicaciones, la teoría de colas se ha empleado tradicionalmente para modelar sistemas con conversaciones de voz, como los ejemplos ya mencionados. De manera más reciente, también se puede aplicar para modelar escenarios de conmutación de paquetes, si bien es preciso tener cierta cautela en las suposiciones realizadas a la hora de modelar el sistema.

Definición

En general, una cola es un sistema con una serie de recursos y una *línea de espera*, en la que las peticiones de una población de usuarios aguardan a que alguno de los recursos quede disponible para ser atendidos. Esto se ilustra en la siguiente figura, donde se aprecia que los principales elementos son:

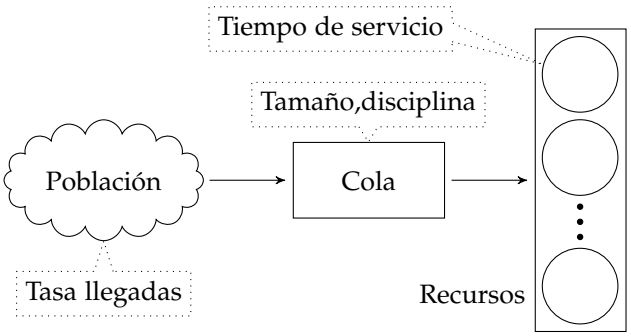


Figura 4.1: Caso general de una cola.

- Una **población** que genera peticiones al sistema. En un caso general, la *tasa* de generación de peticiones dependerá, entre otros factores, de si la población es *finita* o *infinita*. Si la población es finita, que un usuario haya generado una petición puede afectar a

la cantidad de peticiones que se puedan generar a continuación. Si la población es infinita, la tasa de generación no dependerá, en principio, del número de peticiones en el sistema.

- Una **cola** en la que las peticiones generadas por la población aguardan a ser atendidas. Dicha cola podrá tener un tamaño limitado o ilimitado. En el primer caso, el sistema *rechazará* las peticiones que no quepan, bien rechazando la nueva petición, o expulsado alguna que ya existiese.[1] Además de esta política de rechazo, la cola también debe definir el tipo de políticas para atender las peticiones (p.ej., *First In, First Out*), si hay peticiones con mayor prioridad que otras, o si las peticiones se atienden una a una o en bloque.[2]

- Una serie de **recursos** en paralelo, a los que los usuarios acceden tras pasar por la cola. Un usuario accede a un recurso, y –en general– dichos recursos serán idénticos, siendo t_s el tiempo medio de servicio en cualquiera de ellos.[3]

[1] En el caso extremo, la cola tendrá un tamaño nulo, por lo que las peticiones o acceden directamente a los recursos, o no accederán al sistema.

[2] P.ej., la técnica de *packet coalescing* consiste en no iniciar un servicio hasta que hay un número suficiente de tramas para ser atendidas.

[3] Por simplificar la notación, en general se empleará t_s (en vez de $\mathbb{E}[t_s]$) para referirse a un tiempo medio, pero la distinción estará clara por el contexto.

Ejemplo 4.1. Sea una línea de transmisión a 100 Mbps a la que llega un flujo de datos a 50 Mbps compuesto por tramas de 1500 B. En este caso, las peticiones a la cola son las *tramas* de dicho flujo, que llegan a una tasa de

$$\lambda = \frac{100 \text{ Mbps}}{1500 \text{ B/trama}} = 4{,}16 \text{ trama/ms}$$

Por otra parte, el recurso es la línea de transmisión, por lo que el tiempo de servicio se corresponde con el tiempo de transmisión:

$$t_s = \frac{1500 \text{ B}}{100 \text{ Mbps}} = 120 \ \mu s$$

lo que también puede interpretarse como que la línea tiene una capacidad de servicio de

$$\mu = \frac{1}{t_s} = 8{,}33 \text{ trama/ms}$$

Suponiendo que haya *ráfagas* en el proceso de llegada, es de suponer que algunas tramas lleguen mientras otra está siendo transmitida, por lo que aquellas tendrán que esperar en una cola (de salida) antes de ser enviadas. Por otra parte, nótese que la tasa de entrada λ es menor que la capacidad de la línea de transmisión μ. Dado que no se crean nuevas tramas dentro del sistema, la línea no estará siempre ocupada transmitiendo.

Variables de interés

Las principales variables que se suelen estudiar en una cola son las siguientes:

- Tiempo medio de estancia en el sistema (T): es el tiempo que transcurre desde que una petición llega al sistema (es generada por la población) hasta que ha sido finalmente atendida por un recurso (abandona el sistema). Dicho tiempo es en general una variable aleatoria con función de distribución $F_T(t)$.

- Tiempo medio de estancia en la cola (W): es el tiempo que transcurre desde que una petición llega al sistema hasta que empieza a ser atendida por uno de los recursos. De forma análoga, se puede definir la función de distribución de dicho tiempo $F_W(t)$.

Se trata de retardos medios para todos los usuarios. Si $T(k)$ representa el tiempo que pasó en el sistema el usuario k, el tiempo medio de estancia en el sistema puede definirse como

$$T = \lim_{k\to\infty} \frac{1}{k} \sum_{1}^{k} T(k)$$

Resulta claro que el tiempo medio total de estancia en el sistema está relacionado con el tiempo medio de espera en cola y el tiempo medio de servicio:

$$T = W + t_s.$$

Dado un tiempo medio de servicio t_s también se puede definir una tasa (máxima) de servicio $\mu = 1/t_s$, que se corresponde con el ritmo al que los usuarios «salen» de un recurso si éste siempre se encontrase ocupado. Por lo tanto, la expresión anterior también puede escribirse como

$$T = W + \frac{1}{\mu}.$$

- Número medio de usuarios en el sistema (N): que es la media temporal del número de usuarios (o peticiones) que se encuentran en el sistema, ya sea en la cola o siendo atendidos por un recurso, a lo largo del tiempo.

- Número medio de usuarios en la cola (Q): variable similar a la anterior, pero en este caso en vez de considerarse el sistema completo, sólo se considera la cola.

En ambos casos se trata de medias a lo largo del tiempo. Por lo tanto, si $N(\tau)$ representa el número de usuarios en el instante τ, el número medio de usuarios en el sistema durante un tiempo t se define como

$$N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(\tau) d\tau.$$

Ejemplo 4.2. Sea un sistema donde el número de usuarios a lo largo del tiempo es el indicado por la Figura 4.2: al principio hay dos usuarios, en $t = 10$ uno lo abandona y luego en $t = 15$ aparecen dos usuarios más, hasta $t = 25$.

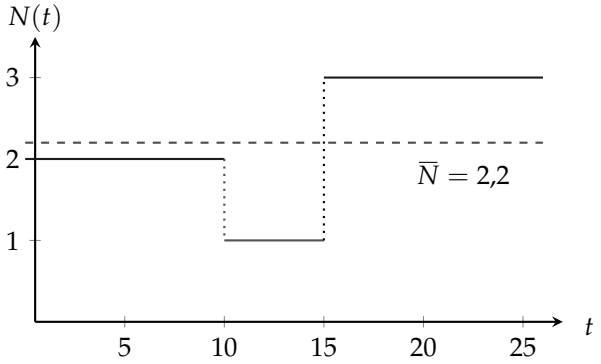


Figura 4.2: Número medio de usuarios en un sistema.

El número medio de usuarios en el sistema durante $t = 25$ se puede calcular como

$$N = \frac{2 \text{ usuarios } \cdot (10-0) + 1 \text{ usuario } \times (15-10) + 3 \text{ usuarios } \times (25-15)}{25-0} = 2.2 \text{ usuarios}$$

En general, se puede definir la «ocupación» de un recurso ρ como la proporción de tiempo que dicho recurso está atendiendo a un usuario. Con dicho parámetro, el número medio de usuarios en un sistema será igual al número medio de usuarios en la cola más el número medio de usuarios en los recursos, esto es:

$$N = Q + n \cdot \rho.$$

Cuando el tamaño de la cola sea finito y, por lo tanto, exista *rechazo* de usuarios, tiene interés tratar con otras dos variables que dependerán de la longitud de la cola:

- La probabilidad de pérdida, o probabilidad de bloqueo (P_B): probabilidad de que una determinada petición no pueda acceder al sistema, porque se encuentra lleno (todos los recursos y la cola están completamente ocupados).
- La tasa efectiva de tráfico cursado: si la población realiza peticiones a una tasa λ, pero existe dicha probabilidad de bloqueo P_B, entonces el tráfico que se cursa en dicho sistema no es el total del ofrecido, sino una fracción inferior, esto es:

$$\overline{\lambda} = \lambda \cdot (1 - P_B)$$

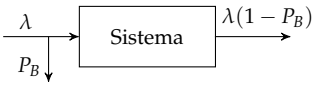


Figura 4.3: Tasa cursada en un sistema con bloqueo.

Notación de Kendall

Con objeto de estandarizar la forma en que definir los posibles sistemas de espera se aparece la *notación de Kendall*,[4] que emplea tres parámetros para especificar las características de una cola:

$$A/B/m/K$$

estos parámetros representan:

[4] Originariamente propuesta por el inglés David G. Kendall en 1953 con los tres primeros parámetros.

- A: especifica cómo se distribuye el tiempo entre peticiones, es decir, la variable aleatoria entre una petición que llega al sistema y la siguiente.
- B: especifica cómo es la variable aleatoria del tiempo de servicio, esto es, el tiempo que pasa desde que una petición accede a un recurso hasta que es atendida. Tanto A como B suelen servir para indicar alguna de las siguientes distribuciones de tiempo:
  - M (por Markov, o «sin Memoria»): en este caso se trata de la variable aleatoria exponencial.
  - D: si se trata de un caso en que dicho tiempo es una constante (esto es, se trata de una variable determinista).
  - G: si se analiza un caso genérico, sin tener que especificar cómo se distribuye la variable aleatoria que determina alguno de los tiempos.
- m: número de recursos idénticos en paralelo.
- K: capacidad máxima de todo el sistema, esto es, número máximo de peticiones que caben a la vez (en la cola y en los servidores). Si $Q_{\text{máx}}$ representa el número máximo de peticiones que caben en la cola, se cumple entonces que

$$K = Q_{\text{máx}} + m.$$

Ejemplo 4.3. Un sistema D/G/2/4 es una cola en la que:

1. El tiempo entre una petición y la siguiente es constante (D).
2. El tiempo de servicio para cada petición puede ser cualquier variable aleatoria (G).
3. Hay dos recursos en paralelo (2).
4. La cola tiene una capacidad máxima de dos usuarios, dado que el sistema puede albergar cuatro usuarios en total (4).

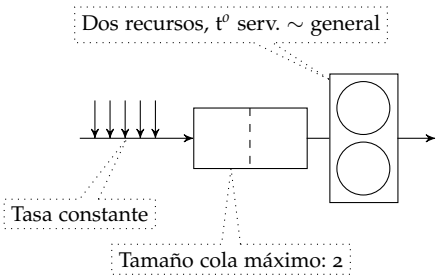


Figura 4.4: Sistema D/G/2/4

Ejemplo 4.4. En una cola $M/D/3/\infty$

1. El que el tiempo entre llegadas es exponencial (M).
2. El tiempo de servicio es constante para todos los usuarios (D).
3. Hay tres recursos en paralelo (3).
4. La longitud máxima de la cola puede suponerse ilimitada, dado que la capacidad del sistema es infinita (∞).

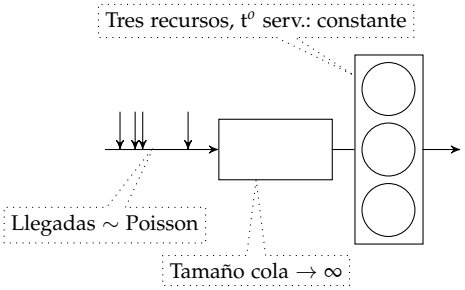


Figura 4.5: Sistema M/D/3/∞

Además de estos parámetros de la notación de Kendall, en ocasiones se emplean dos variables más, por lo que en el caso más completo la representación sería

$$A/B/m/K/N/Z$$

El significado de los dos últimos parámetros es el siguiente:

- N: hace referencia al tamaño (finito o no) de la población, lo que determina cómo varía la tasa de peticiones en función de las que se encuentran en el sistema (por ejemplo, si un usuario no realiza una petición hasta que la anterior ha sido atendida).
- Z: especifica la disciplina de la cola, por ejemplo, si en vez de FIFO se emplea algún tipo de mecanismo que da prioridad a unos usuarios frente a otros.

Los tres primeros parámetros de la notación de Kendall aparecerán siempre, mientras que los tres últimos sólo suelen hacerse explícitos cuando no se corresponden con los valores por defecto: si la capacidad del sistema K y la población N son infinitos, y si la disciplina de la cola es FIFO, no se suelen indicar.

Teorema de Little

Se trata de uno de los resultados fundamentales de la teoría de colas, que establece una relación muy sencilla entre dos variables relativamente sencillas de calcular: la tasa efectiva de entrada de usuarios en un sistema $\bar{\lambda}$ y el número medio de usuarios en dicho sistema N, y una variable que en general resulta algo más complicada de obtener: el tiempo medio total que pasan en el mismo T. La relación es:

$$N = \bar{\lambda} \cdot T$$

Parte de su importancia radica, además, en que se puede aplicar a casi cualquier sistema.

Ejemplo 4.5. Sea el caso de la Figura 4.6 con dos llegadas a un sistema, en el que la primera permanece dos unidades de tiempo, y la segunda tres unidades de tiempo. Dado que el tiempo total son cinco unidades de tiempo, la tasa media de llegadas al sistema es

$$\bar{\lambda} = \frac{2}{5},$$

mientras que el tiempo medio de estancia en el sistema es

$$T = \frac{1}{2}(2+3) = 5/2.$$

Se puede comprobar se cumple el teorema de Little, dado que el número medio de usuarios en el sistema es

$$N = \frac{0 \cdot 1 + 1 \cdot 1 + 2 \cdot 1 + 1 \cdot 2}{5} = 1,$$

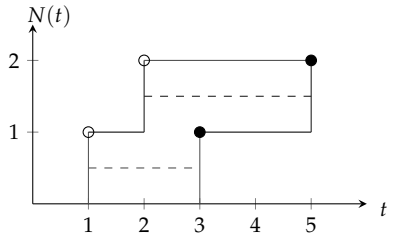


Figura 4.6: Ejemplo de un sistema con dos llegadas

que coincide con

$$\bar{\lambda} \cdot T = \frac{2}{5} \cdot \frac{5}{2} = 1.$$

PARA DEMOSTRAR EL TEOREMA DE LITTLE, sea un caso general como el ilustrado en la Figura 4.7, al que llegan usuarios según un proceso $\alpha(t)$ y del que salen según otro proceso $\beta(t)$. Dado que el tiempo total que pasa un usuario i en el sistema es la diferencia entre el instante en que sale y aquel en que entró, este tiempo T_i se corresponde con la separación en horizontal entre los procesos $\beta(t)$ y $\alpha(t)$.

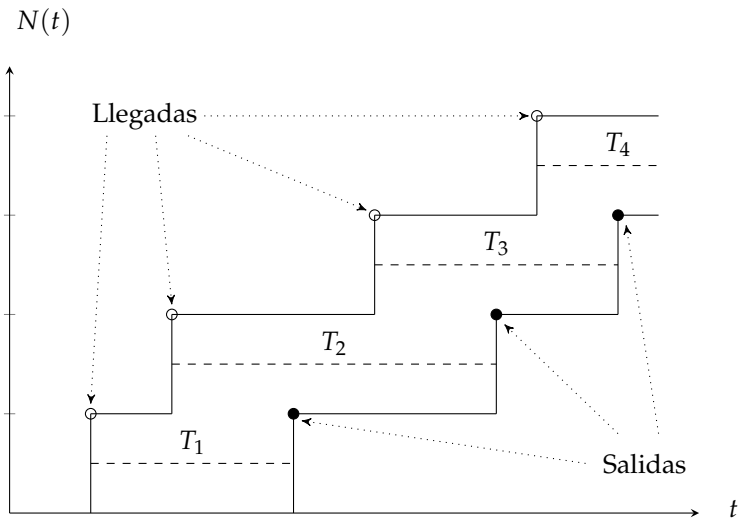


Figura 4.7: Ilustración gráfica del teorema de Little.

Por otro parte, el número de usuarios en el sistema en un instante de tiempo t es la diferencia (en vertical) entre el proceso de llegadas y el de salidas, esto es

$$N(t) = \alpha(t) - \beta(t).$$

Como se ha visto anteriormente, el número medio de usuarios en el sistema se puede calcular como

$$N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(\tau) d\tau = \lim_{t\to\infty} \frac{1}{t} \int_0^t (\alpha(\tau) - \beta(\tau))\, d\tau \qquad (4.1)$$

Nótese que la integral de (4.1) se corresponde con el cálculo del área entre $\alpha(t)$ y $\beta(t)$. Dado que cada «escalón» de estos procesos tiene de altura 1, dicha área también se puede calcular como la suma de los «rectángulos» para cada llegada en el sistema, de altura 1 y anchura T_i. Como en el instante t se han producido $\alpha(t)$ llegadas, se puede aproximar[5]

$$\int_0^t (\alpha(\tau) - \beta(\tau))\, d\tau \approx \sum_{i=1}^{\alpha(t)} T_i$$

[5] Esta demostración no pretende ser rigurosa.

Sustituyendo esta expresión en (4.1), queda

$$N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(\tau) d\tau = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{\alpha(t)} T_i$$

Multiplicando la parte derecha por $\alpha(t)/\alpha(t)$, reordenando términos y cálculos de los límites, resulta

$$N = \lim_{t\to\infty} \frac{\alpha(t)}{t} \cdot \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)} = \lim_{t\to\infty} \frac{\alpha(t)}{t} \cdot \lim_{t\to\infty} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)} \tag{4.2}$$

De lo que se obtiene que el número medio de usuarios en el sistema N es un producto de dos términos:

- El primer término es (el límite de) el cociente del número total de llegadas al sistema entre el tiempo. Por lo tanto, dicho cociente (si existe el límite) es la tasa media de llegadas al sistema:

$$\bar{\lambda} \triangleq \lim_{t\to\infty} \frac{\alpha(t)}{t}$$

- El segundo término es la suma de todos los tiempos de estancia en el sistema T_i, dividida por el número total de llegadas al sistema. Por lo tanto, este cociente es el tiempo medio de estancia en el sistema:

$$T = \lim_{t\to\infty} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

Por lo anterior, el teorema de Little se puede expresar como

$$N = \bar{\lambda} \cdot T$$

UN DETALLE IMPORTANTE a la hora de aplicar el teorema de Little es que $\bar{\lambda}$ es la tasa media de usuarios que *pasan* por el sistema, no la tasa *ofrecida* al sistema. Por lo tanto, en los sistemas donde haya usuarios rechazados[6] habrá que tener en cuenta esta distinción al aplicar el teorema.

[6] Siguiendo la notación de Kendall, todos aquellos que sean de tipo -/-/-/K, con K finito

Ejemplo: Aplicación de Little para un sistema con un recurso

Como se ha dicho anteriormente, la versatilidad del teorema de Little es que resulta valido en cualquier situación,[7] independientemente del sistema considerado. Un ejemplo inmediato es el caso de la Figura 4.8, donde se tiene un sistema con un único recurso.

[7] Para todos los casos que nos ocuparán, el teorema de Little *siempre* será válido, si está aplicado sobre las variables correctas.

Si se aplica el teorema a todo el sistema, se tiene la ya conocida relación

$$N = \bar{\lambda} \cdot T,$$

pero, además, también se puede considerar la cola y el recurso como dos sistemas diferentes, lo que permite establecer dos relaciones adicionales:

- Por un lado, el número de usuarios en la cola Q viene dado por la tasa de entrada a la cola (la misma que al sistema) y el tiempo que pasan en la misma, esto es

$$Q = \bar{\lambda} \cdot W$$

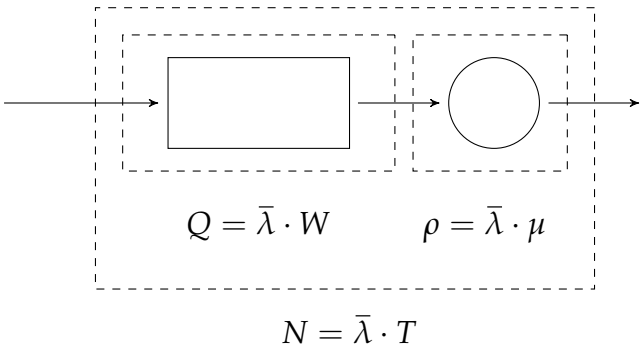


Figura 4.8: Little aplicado a un sistema con un recurso.

- Por otro lado, la ocupación media del recurso ρ viene dada por la tasa de entrada y el tiempo medio de servicio t_s (o su inversa μ), esto es

$$\rho = \overline{\lambda} \cdot t_s = \frac{\overline{\lambda}}{\mu}$$

De hecho, partiendo de la conocida expresión $T = W + t_s$, si se multiplica por $\overline{\lambda}$, se obtiene

$$\overline{\lambda} \cdot T = \overline{\lambda} \cdot W + \overline{\lambda} \cdot t_s$$

que resulta en la también conocida relación entre el número medio de usuarios en el sistema, en la cola y en el recurso:

$$N = Q + \rho.$$

Si en vez de un único recurso hubiese m recursos idénticos en paralelo, se tendría que

$$\overline{\lambda} \cdot t_s = m \cdot \rho$$

y, por lo tanto,

$$N = Q + m \cdot \rho \ .$$

Mediante el teorema de Little se establecen las relaciones fundamentales entre las variables de interés en una cola.

Resumen del tema

- El tiempo de estancia en el sistema es la suma de tiempo de espera en cola y el tiempo de servicio

$$T = W + t_s$$

- El teorema de Little establece la relación entre la tasa de llegadas, el tiempo de estancia y el número medio de usuarios

$$N = \overline{\lambda} \cdot T$$

Cadenas de Markov de tiempo discreto

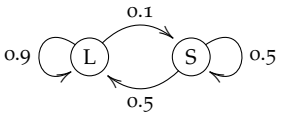
La propiedad sin memoria de la variable aleatoria exponencial implica poder descartar la historia del proceso: si un un proceso de llegadas es de Poisson, la probabilidad de tener una legada dentro de los próximos cinco minutos no depende de si se acaba de ocurrir una o no. Sin embargo, en varias ocasiones es preciso tener en cuenta el estado en que se encuentra un sistema, que se puede ver alterado por una nueva llegada: por ejemplo, en un cajero automático, la probabilidad de tener que esperar dependerá de si está ocupado (esto es, alguien llegó antes) o si está libre; además, si hay mucha gente esperando, es posible que nadie más quiera hacer cola (no se produzcan más llegadas).

Las cadenas de Markov son una herramienta analítica que permiten modelar sistemas en los que una serie de *estados* determinan completamente el posible comportamiento a futuro del sistema: en un cajero vacío, lo único que puede pasar en un futuro es que llegue un nuevo cliente, mientras que si el cajero está ocupado, además de que llegue un nuevo cliente, también cabe la posibilidad de que termine el cliente que está utilizándolo.

Ejemplo 5.1. Sea un equipo de fútbol donde hay dos jugadores encargados de lanzar los penalties, Lucas y Sergio. Lucas tiene un porcentaje de acierto del 90 %, mientras que Sergio acierta el 50 % de las ocasiones. La regla que siguen para repartirse los lanzamientos es: si uno falla, el siguiente penalty lo tira el otro, mientras que si mete gol, repite el siguiente. Para este sistema, la probabilidad de meter gol claramente depende del lanzador (el estado). Además, dado que el lanzador con más acierto tirará más penalties, tampoco se puede suponer que la probabilidad de que el equipo meta un penalty es la media de ambos

$$\Pr(\text{gol}) \neq \frac{(0.9 + 0.5)}{2} = 0.7,$$

Con dos posibles lanzadores de penalties, se puede considerar que el equipo tiene dos *estados*, L o S, según sea el lanzador. Se puede representar la regla que siguen para repartirse los lanzamientos con un diagrama como el indicado al margen, donde los círculos indican los posibles lanzadores y las flechas las probabilidades de que en el siguiente lanzamiento se repita lanzador (si



la flecha vuelve al estado: el 90 % de las veces para Lucas, el 50 % para Sergio) o se cambie (si pasa al siguiente estado: el 10 % de las veces para Lucas y el 50 % para Sergio). Además, dado el lanzador, no sólo está dada la probabilidad de éxito de *este* penalty, sino que también se podrá calcular la del *siguiente* penalty, sin necesidad de conocer lo sucedido en los lanzamientos anteriores.

UNA CADENA DE MARKOV[1] es un proceso aleatorio que pasa por una serie de estados y que cumple una propiedad: una vez conocido el estado en el que se encuentra el proceso, no es necesario conocer la historia pasada del mismo de cara a poder calcular lo que pueda suceder en el futuro.

[1] Por el matemático ruso Andréi Márkov (1856-1922).

Definición

Sea un proceso aleatorio $\{X_n\}$ que toma valores en un espacio finito $S = \{s_1, s_2, \ldots, s_k\}$, como el ilustrado en la Figura 5.1.

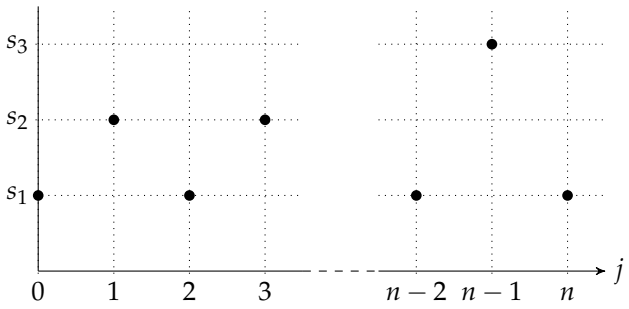


Figura 5.1: Cadena de Markov de tiempo discreto.

Según se observa en la figura, los valores que toma el proceso desde el instante de tiempo $j = 0$ hasta $j = n$ son

$$X_k = \{s_1, s_2, s_1, s_2, \ldots, s_1, s_3, s_1\}$$

Sea un proceso que se inicie en un instante 0 y que discurra hasta un tiempo indefinido. Dado instante de referencia n, el proceso se divide en dos partes:

- El «futuro»: los valores X_{n+1}, X_{n+2}, etc.
- El «pasado»: todos los valores desde X_0 hasta X_{n-1}.

El modelado analítico persigue predecir el comportamiento de un sistema, esto es, calcular las diferentes probabilidades de los diferentes eventos que puedan suceder: si el instante de referencia es n, se pretende calcular la probabilidad de que la cadena pase a cualquiera de los valores de S en el instante $n + 1$ (posteriormente se abordarían los instantes $n + 2$, $n + 3$, etc.):

$$\Pr(X_{n+1} = s_j \mid \text{desde el instante } n), \quad \forall s_j \in S,$$

para lo cual dispone de todo lo sucedido hasta n (incluyendo dicho instante), esto es,

$$\Pr(X_{n+1} = s_j \mid X_n = s_{i_n}, X_{n-1} = s_{i_{n-1}}, \ldots, X_0 = s_{i_0}), \quad \forall s_j \in S, \quad (5.1)$$

donde la secuencia $s_{i_0}, s_{i_1}, s_{i_2}, \ldots, s_{i_n}$ representa los estados por los que ha pasado el proceso desde el instante o hasta el instante n.

Ejemplo 5.2. Siguiendo con el ejemplo de los penalties, supóngase que se han lanzado 6 penalties, y los lanzadores han sido los siguientes

$$X = \{L, L, L, L, S, L\}$$

la probabilidad de que el siguiente penalty lo tire Sergio, según la expresión (5.1), se puede expresar como:

$$\Pr(X_7 = S \mid X_6 = L, X_5 = S, X_4 = L, X_3 = L, X_2 = L, X_1 = L),$$

mientras que la probabilidad de que el siguiente penalty lo tire Lucas sería:

$$\Pr(X_7 = L \mid X_6 = L, X_5 = S, X_4 = L, X_3 = L, X_2 = L, X_1 = L),$$

En este caso, según las reglas para repartirse los penalties, para calcular estas probabilidades no hace falta saber toda la historia del proceso, dado que el lanzador del siguiente penalty sólo depende del lanzador actual, esto es:

$$\begin{aligned}&\Pr(X_7 = L \mid X_6 = L, X_5 = S, X_4 = L, X_3 = L, X_2 = L, X_1 = L)\\ =&\Pr(X_7 = L \mid X_6 = L)\end{aligned}$$

cuando esto sucede, es decir, cuando el pasado no afecta a la predicción, sino únicamente el presente, se dice que el proceso aleatorio es una cadena de Markov.

Propiedad de Markov y homogeneidad

A continuación se formaliza esta propiedad sobre la falta de dependencia del futuro con el pasado que caracteriza a las cadenas de Markov.

Cadena de Markov Un proceso aleatorio $\{X_n\}$ en un espacio finito de valores $S = \{s_1, s_2, \ldots, s_k\}$ es una cadena de Markov de tiempo discreto si cumple la siguiente *propiedad de Markov*:

$$\begin{aligned}&\Pr(X_{n+1} = s_j \mid X_n = s_{i_n}, X_{n-1} = s_{i_{n-1}}, \ldots, X_1 = s_{i_1}, X_0 = s_{i_0})\\ =&\Pr(X_{n+1} = s_j \mid X_n = s_{i_n}) \quad \forall i, j, \{s_{i_k}\}\end{aligned}$$

Esta propiedad indica que, aunque se conozca toda la historia del proceso hasta el instante n (esto es, todos los estados $\{s_{i_0}, s_{i_1}, \ldots, s_{i_n}\}$ por los que ha pasado), la probabilidad de que el proceso pase al estado $X_{n+1} = s_j$ sólo depende del estado en que se encuentre en X_n, y no de dicha historia, para cualquier posible combinación de estados s_i, s_j e historia $\{s_{i_k}\}$.

Ejemplo 5.3. Sea una red multi-salto que emplea un algoritmo de *forwarding* aleatorio como la ilustrada en la Figura 5.3. Hay un mensaje a transmitir desde un nodo 'Origen' a un nodo 'Destino', y el algoritmo de retransmisión a cada nodo consiste en escoger un destino de entre los vecinos al azar. Este caso se podría modelar con una cadena de Markov, dado que la probabilidad de que un mensaje llegue a un nodo sólo depende del nodo en que se encuentre y no de todos los nodos que haya visitado.

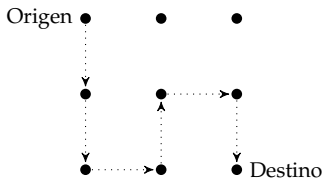


Figura 5.2: Ejemplo de reenvío aleatorio en una red multisalto.

La propiedad de Markov define que la probabilidad de pasar desde un estado s_i en el instante n a un estado s_j en el instante $n+1$ no depende de lo que haya sucedido en instantes anteriores a n. Si, además, la probabilidad de pasar de un estado al otro no depende del valor específico de n (esto es, no varía a lo largo del tiempo) la cadena de Markov es *homogénea*:

$$\Pr(X_n = s_j | X_{n-1} = s_i) = \Pr(X_k = s_j | X_{k-1} = s_i), \quad \forall n, k, s_i, s_j$$

Ejemplo 5.4. Según la Wikipedia,[2] algunos de los estados civiles más habituales son: soltero, comprometido, casado, divorciado, viudo. En este caso, aún sin tener en cuenta la historia del proceso, la probabilidad de pasar de un estado dependerá de la edad (esto es, del valor de n). Se trataría de un sistema inhomogéneo.

[2] https://es.wikipedia.org/wiki/Estado_civil

Únicamente se considerarán de ahora en adelante cadenas de Markov homogéneas.

Matriz de transiciones y diagrama de estados

Dado que en las cadenas homogéneas la probabilidad de transición de un estado al otro no depende del instante de tiempo considerado, tiene sentido definir dicha probabilidad para todos los estados. Ello lleva a la aparición de los siguientes dos conceptos fundamentales:

- La *probabilidad de transición* (p_{ij}): es la probabilidad de que el proceso pase al estado s_j en el instante n estando en el estado s_i en el instante $n-1$

$$p_{ij} \triangleq \Pr(X_n = s_j | X_{n-1} = s_i)$$

- La *matriz de transiciones* (P): es la matriz constituida por las probabilidades de transición, donde la componente p_{ij} representa la probabilidad de pasar desde el estado s_i al estado s_j. Por tanto, la fila de la matriz indica el estado actual, y la columna el posible estado siguiente.

Ejemplo 5.5. Siguiendo con el ejemplo de Lucas y Sergio (con probabilidades de acierto del 90 % y 50 %, respectivamente), la probabilidad del lanzador futuro se ilustra en la Tabla 5.6 al margen. La matriz de transiciones de esta cadena de Markov sería la siguiente:

	Siguiente	
Ahora	Lucas	Sergio
Lucas	0.9	0.1
Sergio	0.5	0.5

Tabla 5.6: Probabilidad del siguiente lanzador en función del actual.

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{pmatrix}$$

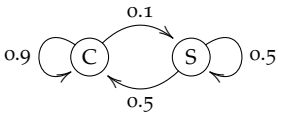
Esta matriz de transiciones, dado que representa las probabilidades de que del estado asociado a la fila i se pase al estado asociado a la columna j (incluyéndose el caso de que se repita estado), debe cumplir las siguientes propiedades:

$$p_{ij} \geq 0, \ \forall i, j$$
$$\sum_{j=1}^{k} p_{ij} = 1, \ \forall i$$

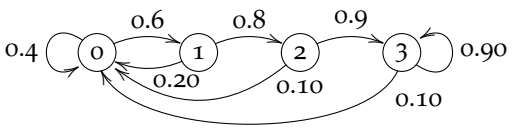
Además de la matriz de transiciones[3] se puede emplear el *diagrama de estados* de la cadena de Markov para representarla. Se trata de un diagrama como el visto con el ejemplo de los penalties, donde cada estado se representa con un círculo, y las transiciones entre estados con flechas acompañadas por la correspondiente probabilidad de transición, salvo que dicha probabilidad sea nula (en este caso no se dibuja la flecha).

[3] La matriz de transiciones también se denomina matriz de probabilidad, matriz de Markov o matriz estocástica.

Ejemplo 5.6. Para el caso de los penalties, el diagrama de estados sería el que aparece en la figura al margen. Se puede apreciar que la suma de todas las probabilidades asociadas a las flechas que salen de un mismo estado es igual a la unidad.



Ejemplo 5.7. Sea un jugador de baloncesto al que *se le calienta* la mano, esto es, que cuando encesta un tiro es más probable que enceste el siguiente tiro: su primer lanzamiento lo acierta el 60 % de las veces, si ya ha encestado un tiro el siguiente acierta el 80 % de los intentos, y con dos, tres o más tiros consecutivos encestados anota el 90 % de las veces. Este comportamiento se puede modelar con el siguiente diagrama

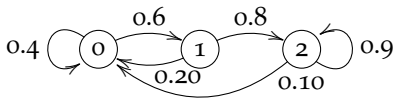


a partir del que se deduce la matriz de transiciones P_1 indicada al margen.

$$P_1 = \begin{pmatrix} 0{,}4 & 0{,}6 & 0 & 0 \\ 0{,}2 & 0 & 0{,}8 & 0 \\ 0{,}1 & 0 & 0 & 0{,}9 \\ 0{,}1 & 0 & 0 & 0{,}9 \end{pmatrix}$$

Este modelado del sistema no es incorrecto; sin embargo, tras una lectura detallada del comportamiento del jugador se puede comprobar que «sobra» un estado, dado que a partir del segundo

lanzamiento anotado la probabilidad de éxito es siempre la misma, por lo que no hace falta distinguir entre 2 y 3. El diagrama de estados pasa entonces a ser



con la correspondiente matriz P_2 al margen (más sencilla).

$$P_2 = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.2 & 0 & 0.8 \\ 0.1 & 0 & 0.9 \end{pmatrix}$$

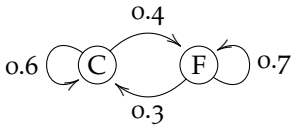
COMO ILUSTRA EL EJEMPLO ANTERIOR, en general será sencillo añadir estados «innecesarios» que aumenten la complejidad de la cadena. Sin embargo, también puede ocurrir que existan varias formas de modelar un sistema, en función de lo que se desee analizar, y que no resulte sencillo determinar cuál es la más adecuada. Es por ello que a la hora de modelar un sistema resulta crítico dedicar tiempo a determinar el número y significado de cada estado, teniendo en cuenta que siempre se ha de cumplir la propiedad de Markov: un estado debe definir una situación con toda la información necesaria para que la historia pasada del proceso no afecte a lo que pueda suceder en el futuro.

Ejemplo 5.8. Sea un modelo simple para predecir el tiempo, que a partir de si hace calor o frío hoy estima con qué probabilidad hará calor o frío mañana, según la Tabla 5.7.

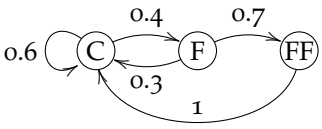
	Mañana	
Hoy	Pr(Calor)	Pr(Frío)
Calor	0.60	0.40
Frío	0.30	0.70

Tabla 5.7: Modelo para el tiempo con una memoria de un día.

Dado que sólo es preciso saber el tiempo hoy para predecir el de mañana, este sistema se puede modelar con una cadena de Markov de dos estados, con el siguiente diagrama



Supóngase que hay que extender el modelo para tener en cuenta que nunca hay más de dos días fríos seguidos. Dado que un estado contiene toda la información necesaria para predecir el futuro, hay que extender la información contenida en los mismos para «recordar» el número de días que hace frío de forma consecutiva. Un diagrama que modelase esta nueva situación sería el siguiente



Por lo tanto, si hace falta considerar la «historia» del proceso que se quiere modelar, hay que definir los estados de modo tal que tengan en cuenta dicha historia.

Tiempo de permanencia en un estado

Los elementos de la diagonal de P (esto es, p_{ii}) representan la probabilidad de que, llegada la cadena $\{X_n\}$ al estado s_i, perma-

nezca en dicho estado en la siguiente iteración:

$$p_{ii} \triangleq \Pr(X_{n+1} = s_i \mid X_n = s_i).$$

Por lo tanto, el número de iteraciones que una cadena permanece en un estado s_i se puede modelar con una distribución geométrica, donde la probabilidad de «éxito» es la probabilidad de que la cadena abandone dicho estado $(1 - p_{ii})$.

Sea D_i el número de veces que la cadena repite de forma consecutiva el estado s_i. Dicho número aleatorio es una variable aleatoria geométrica, donde la probabilidad de exactamente k repeticiones es

$$\Pr(D_i = k) = p_{ii}^k(1 - p_{ii})$$

mientras que su media es

$$\mathbb{E}[D_i] = \frac{1}{1 - p_{ii}}.$$

Ejemplo 5.9. Para el caso del tiempo de la Tabla 5.7 (repetida al margen), el número medio de días de calor consecutivos es de $1/(1 - 0{,}6) = 2{,}5$, mientras que el número medio de días consecutivos de frío será $1/(1 - 0{,}7) = 3{,}33$.

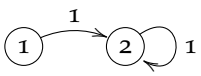
Hoy	Mañana Pr(Calor)	Pr(Frío)
Calor	0.60	0.40
Frío	0.30	0.70

Comunicación entre estados

A continuación se formaliza las posibles relaciones entre los estados de una cadena de Markov, en términos tanto de la probabilidad de «llegar» a un estado partiendo de otro, como de la «frecuencia» con la que puede ocurrir.

Irreducibilidad

En los ejemplos vistos hasta ahora, se puede comprobar que, partiendo desde cualquier estado, es posible «alcanzar» cualquier otro estado.[4] Sin embargo, esto no es necesariamente cierto en todas las cadenas de Markov, como se ilustra en el siguiente caso:



Dicha figura representa también una cadena de Markov, pero con las transiciones entre estados mucho más limitadas: en cuanto el proceso llegue al estado 2, permanecerá en el mismo para siempre (como se verá más adelante, se trata de un estado absorbente). De hecho, otro caso más «extremo» de cadena de Markov donde los estados no son alcanzables entre sí sería el siguiente:

en el que el futuro está completamente determinado por el *estado inicial* del sistema.

[4] Esto es, siempre hay un «camino» para llegar desde cualquier estado a cualquier otro estado.

En una cadena con «mucha» conexión entre estados se tendrá que el proceso no se queda «atrapado» en un estado en particular, mientras que en aquellas cadenas como las anteriores, será más frecuente que el proceso alcance un estado y no lo abandone. Para caracterizar una cadena de Markov siguiendo estas ideas se define la *comunicación entre estados*:

- Un estado s_j es *accesible* desde un estado s_i si la probabilidad de alcanzar el estado s_j desde el estado s_i es estrictamente mayor que cero, y se denota como:

  $$s_i \to s_j$$

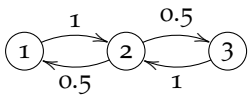
  Es decir, que es posible (no es imposible) que en algún momento se pueda alcanzar el estado s_j partiendo desde s_i:

  $$\Pr(X_{n+m} = s_j | X_m = s_i) > 0, \text{ para algún } n \geq 0$$

Esta propiedad no implica que $p_{ij} > 0$ (aunque sea una condición suficiente), sino que desde el estado s_i no es imposible alcanzar el estado s_j. Por definición, un estado es siempre accesible desde él mismo (dado que la condición es $n \geq 0$ y no $n > 0$).

Ejemplo 5.10. Sea la siguiente cadena, con la correspondiente matriz P asociada indicada en el margen.

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0{,}5 & 0 & 0{,}5 \\ 0 & 1 & 0 \end{pmatrix}$$



Se puede comprobar que 3 es accesible desde 1 ($1 \to 3$) a pesar de que $p_{13} = 0$, ya que se cumple que

$$\Pr(X_{m+2} = 3 | X_m = 1) = 0{,}5 \ \forall m.$$

LA PROPIEDAD DE ACCESIBILIDAD ES TRANSITIVA, dado que si el estado s_j es accesible desde el estado s_i, esto es,

$$s_i \to s_j$$

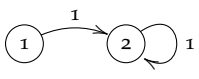
y el estado s_k es accesible desde el estado s_j,

$$s_j \to s_k$$

entonces se tiene que el estado s_k es accesible desde el estado s_i

$$s_i \to s_k$$

Sin embargo, la propiedad de accesibilidad no es conmutativa, dado que las flechas son unidireccionales: por ejemplo, en la cadena representada al margen se cumple que $1 \to 2$ pero no que $2 \to 1$.



Si $s_i \to s_j$ y $s_i \leftarrow s_j$, se dice que los estados s_i y s_j *se comunican (entre sí)*, y se representa como

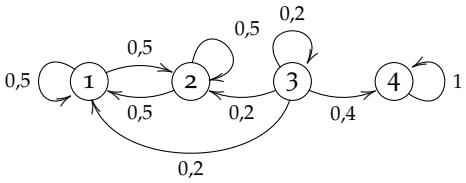
$$s_i \leftrightarrow s_j$$

Dos estados que se comunican entre sí se dice que están en la *misma clase*,, por lo que todos los miembros de la misma clase se comunican entre sí.

Ejemplo 5.11. La cadena de Markov dada por la siguiente matriz

$$P = \begin{pmatrix} 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}2 & 0{,}2 & 0{,}2 & 0{,}4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

con el siguiente diagrama de estados



Se puede comprobar que dicha cadena tiene tres clases: $\{1, 2\}$, $\{3\}$ y $\{4\}$ (por la definición de accesibilidad, un estado siempre se comunica consigo mismo).

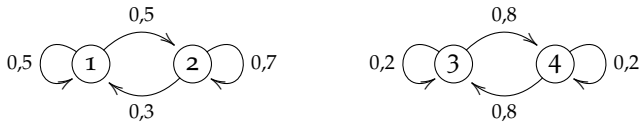
A PARTIR DE LA PROPIEDAD de comunicación entre estados se define una de las principales características de una cadena de Markov: la irreducibilidad.

Cadena de Markov irreducible: Una cadena de Markov $\{X_n\}$ con espacio de estados $S = \{s_1, \ldots, s_K\}$ y matriz de transición P es *irreducible* si, para todos sus estados $s_i, s_j \in S$ se cumple que $s_i \leftrightarrow s_j$. Si no es así, se dice que la cadena de Markov es *reducible*.

Ejemplo 5.12. Los ejemplos de los lanzamientos de penalties (Ejemplo 5.1) o predicción del tiempo (Ejemplo 5.8) son cadenas de Markov irreducibles, dado que desde cualquier estado se puede llegar a cualquier otro estado. Una ejemplo de cadena de Markov reducible sería la del ejemplo anterior (con tres clases), o la dada por la siguiente matriz de transición

$$P = \begin{pmatrix} 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}3 & 0{,}7 & 0 & 0 \\ 0 & 0 & 0{,}2 & 0{,}8 \\ 0 & 0 & 0{,}8 & 0{,}2 \end{pmatrix},$$

como se puede ver al representar su diagrama de estados



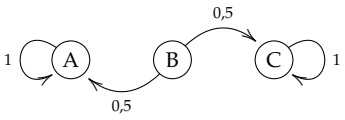
donde se aprecia que dicha matriz representa dos cadenas «independientes», por lo que el análisis del comportamiento de la cadena se podrá simplificar (esto es, *reducir*) a casos más sencillos, como se verá más adelante.

Recurrencia

Partiendo de un estado en una cadena de Markov, puede ocurrir que: (*i*) nunca se abandone, (*ii*) se vuelva a visitar pasado un tiempo, o (*iii*) se abandone para no volver nunca. Según el caso, se pueden definir diferentes tipos de estado.

- Un estado s_j es *absorbente* si $p_{jj} = 1$. Es un caso sencillo de identificar: en el diagrama de estados aparece un «bucle» con $p = 1$, y en la matriz P la posición diagonal en la fila o columna tiene el valor 1.

Cuando en una cadena el proceso alcanza un estado absorbente, ya nunca saldrá de dicho estado. Una cadena puede tener varios estados absorbentes, como se muestra en la figura al margen (estados A y B).

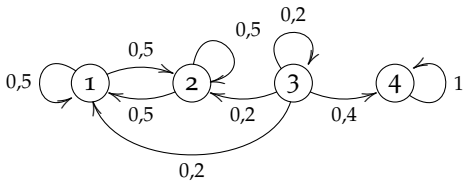


- Un estado s_j es *recurrente* (o persistente) si, partiendo de dicho estado, la cadena vuelve a él con total probabilidad:

$$\Pr(\text{volver a } s_j \text{ en algún momento} \mid X_0 = s_j) = 1$$

Ejemplo 5.13. Sea otra vez el caso de la cadena del Ejemplo 5.12, con el siguiente diagrama de estados y matriz P al margen.

$$P = \begin{pmatrix} 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 & 0 \\ 0{,}2 & 0{,}2 & 0{,}2 & 0{,}4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



Se puede deducir que:

- El estado 4 es recurrente, y se trata de un estado absorbente.
- El estado 3 no es recurrente: la probabilidad de volver a dicho estado es de 0.2, mientras que con probabilidad 0.8 la cadena abandona dicho estado para no volver jamás.
- Los estados 1 y 2 son recurrentes. Para que p.ej. 1 no fuese recurrente, la cadena debería abandonar dicho estado (lo que hace con probabilidad 1/2) y permanecer en 2 para siempre, pero la probabilidad de permanecer en 2 durante n iteraciones es

$$\left(\frac{1}{2}\right)^n \to 0,$$

por lo que se vuelve a 1 en algún momento.

La propiedad de recurrencia en un estado implica que en algún momento el proceso volverá a dicho estado. Se trata de una propiedad de clase: si un estado es recurrente, entonces todos los estados de su clase son recurrentes.

- Finalmente, un estado s_j que no es recurrente se llama *transitorio*: existe la posibilidad de que el proceso nunca vuelva a s_j una vez que lo abandone.

En una cadena irreducible (esto es, que sólo tiene una clase), o todos los estados son recurrentes, o todos son transitorios.

Evolución en el tiempo de una cadena

Como se ha visto, la matriz de transiciones P define la probabilidad de que la cadena pase de un estado a otro en una unidad de tiempo

$$p_{ij} \triangleq \Pr(X_{n+1} = s_j \mid X_n = s_i)$$

por lo tanto, si se sabe dónde se puede encontrar la cadena en un estado n (esto es, las probabilidades $\Pr(X_n = s_k)$ para todo s_k), se puede calcular la probabilidad de que esté en cualquier estado en $n+1$, ya que

$$\Pr(X_{n+1} = s_j) = \sum_{s_k} \Pr(X_{n+1} = s_j \mid X_n = s_k) \Pr(X_n = s_k), \forall s_j \quad (5.2)$$

Ejemplo 5.14. Sea una cadena con dos estados $\{s_1, s_2\}$ como la ilustrada en la figura al margen, que parte del estado s_1 en el instante 0, esto es

$$\Pr(X_0 = s_1) = 1$$

De aplicar la expresión (5.2) para el caso de X_1 resulta

$$\Pr(X_1 = s_1) = \Pr(X_1 = s_1 \mid X_0 = s_1) \Pr(X_0 = s_1) = p_{11}$$
$$\Pr(X_1 = s_2) = \Pr(X_1 = s_2 \mid X_0 = s_1) \Pr(X_0 = s_1) = p_{12}$$

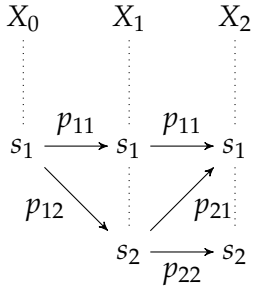


Figura 5.3: Cadena para el Ejemplo 5.14

Si se aplica ahora la expresión (5.2) para calcular $\Pr(X_2 = s_1)$, se obtiene

$$\begin{aligned}\Pr(X_2 = s_1) = &\Pr(X_2 = s_1 \mid X_1 = s_1) \Pr(X_1 = s_1) + \\ &\Pr(X_2 = s_1 \mid X_1 = s_2) \Pr(X_1 = s_2)\end{aligned}$$

que, a partir de los valores calculados anteriormente para $\Pr(X_1 = s_1)$ y $\Pr(X_1 = s_2)$, puede expresarse como

$$\begin{aligned}\Pr(X_2 = s_1) = &p_{11}p_{11} + \\ &p_{21}p_{12}\end{aligned}$$

lo que tiene una interpretación muy directa: si se parte de s_1 en $n = 0$, hay dos «caminos» para estar en s_1 en $n = 2$:

- Bien no se abandona dicho estado en dos unidades de tiempo (lo que sucede con probabilidad $p_{11}p_{11}$)
- Bien se abandona dicho estado para volver inmediatamente (lo que sucede con probabilidad $p_{12}p_{21}$)

GRACIAS A LA PROPIEDAD DE MARKOV, realizar predicciones sobre el comportamiento de un sistema es razonablemente sencillo, ya que basta con calcular probabilidades condicionadas entre un instante n y el siguiente $n+1$: por ejemplo, en una cadena de Markov, la probabilidad de estar en s_i en $n+2$ si se parte de s_j en n se puede expresar como

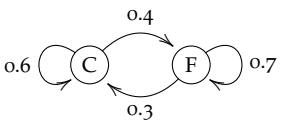
$$\Pr(X_{n+2} = s_i \mid X_n = s_j) = \sum_{\forall s_k} \Pr(X_{n+2} = s_i \mid X_{n+1} = s_k) \Pr(X_{n+1} = s_k \mid X_n = s_j)$$

mientras que si el sistema no fuese de Markov, habría que tener en cuenta la historia a lo largo del proceso:

$$\Pr(X_{n+2} = s_i \mid X_n = s_j) = \sum_{\forall s_k} \Pr(X_{n+2} = s_i \mid X_{n+1} = s_k, X_n = s_j) \Pr(X_{n+1} = s_k \mid X_n = s_j)$$

Ejemplo 5.15. Sea otra vez el modelo del tiempo con las probabilidades de transición en la tabla al margen. Según este modelo, sabiendo lo que sucede *hoy*, se puede calcular la probabilidad de lo que pueda suceder *mañana*. Por lo tanto, si hoy hace calor (c), mañana también hará calor con una probabilidad de p_{cc} =0.6, y frío (f) con una probabilidad de p_{cf} =0.4.

	Mañana	
Hoy	Pr(Calor)	Pr(Frío)
Calor	0.60	0.40
Frío	0.30	0.70



Una vez que se ha estimado lo que puede pasar mañana, se puede calcular la probabilidad de que haga calor *pasado mañana*

- Si mañana hace calor, pasado mañana hará calor con probabilidad p_{cc}. Por lo tanto, la probabilidad de dos días más con calor (CC) es
$$p_{cc}p_{cc} = 0.6 \times 0.6 = 0.36.$$
- Si mañana hace frío, pasado mañana hará calor con probabilidad 0.3. Por lo tanto, la probabilidad de primero frío y luego calor (FC)
$$p_{cf}p_{fc} = 0.4 \times 0.3 = 0.12.$$

Por lo tanto, si hoy hace calor, pasado mañana hará calor con una probabilidad de

$$\Pr(\text{Pasado mañana calor}) = 0.36 + 0.12 = 0.48.$$

La probabilidad de que haga frío se puede obtener con un razonamiento similar,[5] o bien directamente como

$$\Pr(\text{Pasado mañana frío}) = 1 - \Pr(\text{Pasado mañana calor}) = 0.52$$

[5] Que se obtendría como 0.6×0.4+0.4×0.7=0.52

Los anteriores ejemplos ilustran que realizar predicciones en una cadena de Markov es sencillo pero laborioso, dado que «basta» con calcular las probabilidades de todos los posibles caminos entre un estado inicial y otro final. A continuación se introduce un elemento clave (el vector de probabilidades de estado) que permite simplificar este cálculo.

Vector de probabilidades de estado

Dada una cadena de Markov de K estados, la probabilidad con la que dicha cadena se encuentra en el instante n en cada uno de esos estados se puede representar con el siguiente vector *fila*:

$$\pi^{(n)} \triangleq (\Pr(X_n = s_1), \Pr(X_n = s_2), \ldots, \Pr(X_n = s_k))$$

La componente i-ésima de dicho vector se representa como

$$\pi_i^{(n)} \triangleq \Pr(X_n = s_i),$$

y éstas siempre deben sumar uno:

$$\sum_{i=1}^{K} \pi_i^{(n)} = 1.$$

Ejemplo 5.16. Siguiendo con el ejemplo del tiempo, sea *hoy* el día $n = 0$ y hace calor. Esto se representa como

$$\pi^{(0)} = (\Pr(X_0 = C), \Pr(X_0 = F)) = (1, 0),$$

mientras que los vectores $\pi^{(1)}$ y $\pi^{(2)}$ son, respectivamente:

$$\pi^{(1)} = (0.6, 0.4)$$
$$\pi^{(2)} = (0.48, 0.52)$$

Dado un valor de las probabilidades de calor o frío en el día n (esto es, $\pi^{(n)}$), y la matriz de transiciones P, resulta sencillo calcular el valor de $\pi^{(n+1)}$, dado que por la ley de la probabilidad total

$$\Pr(X_{n+1} = C) = \Pr(X_{n+1} = C \mid X_n = C) \Pr(X_n = C) + \Pr(X_{n+1} = C \mid X_n = F) \Pr(X_n = F)$$
$$\Pr(X_{n+1} = F) = \Pr(X_{n+1} = F \mid X_n = C) \Pr(X_n = C) + \Pr(X_{n+1} = F \mid X_n = F) \Pr(X_n = F)$$

que puede escribirse como

$$\pi_1^{(n+1)} = p_{11}\pi_1^{(n)} + p_{21}\pi_2^{(n)}$$
$$\pi_2^{(n+1)} = p_{12}\pi_1^{(n)} + p_{22}\pi_2^{(n)}$$

y permite calcular la probabilidad de calor o frío de forma iterativa, como se realiza a continuación en la Tabla 5.8, donde la matriz de transiciones es

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}$$

Como se aprecia en la tabla, el cálculo de $\pi^{(n+1)}$ a partir de $\pi^{(n)}$ resulta mecánico, pues basta con multiplicar cada componente por el valor correspondiente de p_{ij} y sumar. A continuación se formaliza esta relación.

n	$\pi_1^{(n)}$ (Calor)	$\pi_2^{(n)}$ (Frío)
0	1.00	0
1	0.60 (=1 × 0.6 + 0 × 0.3)	0.40 (= 1 × 0.4 + 0 × 0.7)
2	0.48 (=0.6 × 0.6 + 0.4 × 0.3)	0.52 (= 0.6 × 0.4 + 0.4 × 0.7)
3	0.4440 (=0.48 × 0.6 + 0.52 × 0.3)	0.5560 (= 0.48 × 0.4 + 0.52 × 0.7)
4	0.4332 (=0.444 × 0.6 + 0.546 × 0.3)	0.5668 (= 0.444 × 0.4 + 0.556 × 0.7)

Tabla 5.8: Evolución de las probabilidades de calor y frío para el ejemplo del tiempo.

Ecuaciones de Chapman-Kolmogorov

Como se ha visto en el ejemplo anterior, por la ley de la probabilidad total en una cadena de Markov se cumple que

$$\Pr(X_n = s_i) = \sum_j \Pr(X_n = s_i \mid X_{n-1} = s_j) \Pr(X_{n-1} = s_j).$$

Mediante el vector de probabilidades de estado $\pi^{(n)}$ y la matriz de transiciones P se puede expresar esta relación para todos los estados s_i de forma compacta

$$\pi^{(n)} = \pi^{(n-1)} P, \quad n > 1$$

Ejemplo 5.17. Como se ha visto en el ejemplo anterior, en el caso de una cadena de dos estados las ecuaciones que permiten calcular el vector $\pi^{(n)}$ a cada instante son

$$\pi_1^{(n)} = p_{11}\pi_1^{(n-1)} + p_{21}\pi_2^{(n-1)}$$
$$\pi_2^{(n)} = p_{12}\pi_1^{(n-1)} + p_{22}\pi_2^{(n-1)}$$

que se puede comprobar que coinciden con la siguiente multiplicación

$$\left(\pi_1^{(n)}, \pi_2^{(n)}\right) = \left(\pi_1^{(n-1)}, \pi_2^{(n-1)}\right) \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

DADO QUE LA EXPRESIÓN $\pi^{(n)} = \pi^{(n-1)} P$ permite calcular la distribución de probabilidades $\pi^{(n)}$ a partir de $\pi^{(n-1)}$ para cualquier n, también se puede expresar $\pi^{(n-1)}$ en función de $\pi^{(n-2)}$:

$$\pi^{(n-1)} = \pi^{(n-2)} P$$

por lo que se puede calcular $\pi^{(n)}$ en función de $\pi^{(n-2)}$:

$$\pi^{(n)} = \pi^{(n-1)} P = \left(\pi^{(n-2)} P\right) P,$$

esto es,

$$\pi^{(n)} = \pi^{(n-2)} P^2.$$

Este resultado se puede generalizar para $\pi^{(n-3)}$, $\pi^{(n-4)}$, etc., lo que permite calcular cualquier valor de $\pi^{(n)}$ a partir de las potencias de la matriz P y un vector inicial $\pi^{(0)}$. Este resultado se conoce como las ecuaciones de Chapman-Kolmogorov.

Ecuaciones de Chapman-Kolmogorov: Dada una cadena de Markov $\{X_n\}$ con espacio de estados $S = \{s_1, \ldots, s_K\}$, matriz de transiciones P y vector inicial de probabilidades de estado $\pi^{(0)}$, para cualquier $\pi^{(n)}$ se cumple que

$$\pi^{(n)} = \pi^{(0)} P^n \tag{5.3}$$

Ejemplo 5.18. Sea el ejemplo del tiempo, con las siguientes potencias de la matriz P:

$$P = \begin{pmatrix} 0.6 & 0.4 \\ 0.3 & 0.7 \end{pmatrix}, \quad P^2 = \begin{pmatrix} 0.48 & 0.52 \\ 0.39 & 0.61 \end{pmatrix}, \quad P^3 = \begin{pmatrix} 0.444 & 0.556 \\ 0.417 & 0.583 \end{pmatrix}$$

Se puede comprobar que si $\pi^{(0)} = (1, 0)$, entonces las probabilidades de calor y frío en el día n (recopiladas en la tabla al margen) coinciden con $\pi^{(0)} P^n$.

	Probabilidad de	
n	Calor	Frío
0	1.00	0
1	0.60	0.40
2	0.48	0.52
3	0.444	0.556
4	0.4332	0.5668

Tabla 5.9: Probabilidades de calor y frío según la Tabla 5.8.

Significado de las filas la matriz P^n

Dado un vector inicial $\pi^{(0)}$, las potencias de P determinan el valor del vector de distribución de probabilidades $\pi^{(n)}$. Para analizar el significado de la matriz P^n, sea una cadena de Markov genérica de dos estados $\{s_1, s_2\}$ con la siguiente matriz de transiciones:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} .$$

El cuadrado de la matriz P es:

$$P^2 = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} p_{11}^2 + p_{12}p_{21} & p_{11}p_{12} + p_{12}p_{22} \\ p_{21}p_{11} + p_{22}p_{21} & p_{21}p_{12} + p_{22}^2 \end{pmatrix} .$$

El primer elemento de la primera fila de P^2 (es decir, P_{11}^2) se compone de dos términos:

1. p_{11}^2: probabilidad de que, estando en s_1, se siga en s_1 tras dos iteraciones.
2. $p_{12}p_{21}$: probabilidad de pasar de s_1 a s_2, y luego pasar de s_2 a s_1.

Por lo tanto, P_{11}^2 es la probabilidad de que, partiendo de s_1, en dos iteraciones el estado sea s_1 otra vez (pasando o no por s_2).

Por otra parte, el segundo elemento de la primera fila (P_{12}^2) se compone de:

1. $p_{11}p_{12}$: probabilidad de permanecer en s_1 y luego pasar a s_2.
2. $p_{12}p_{22}$: probabilidad de pasar a s_2 y luego permanecer en s_2.

por lo que se corresponde con la probabilidad de terminar en el estado s_2 tras dos iteraciones partiendo del estado s_1.

Ejemplo 5.19. En el ejemplo del tiempo, la potencia cuarta de la matriz P es

$$P^4 = P^2 \cdot P^2 = \begin{pmatrix} 0.48 & 0.52 \\ 0.39 & 0.61 \end{pmatrix} \begin{pmatrix} 0.48 & 0.52 \\ 0.39 & 0.61 \end{pmatrix} = \begin{pmatrix} 0.4332 & 0.5668 \\ 0.4251 & 0.5749 \end{pmatrix},$$

en la que su primera fila

$$P^4_{1*} = (0.4332, 0.5668)\ ,$$

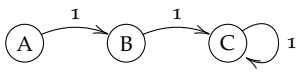
se corresponde con las probabilidades de calor y frío, respectivamente, en el cuarto día, suponiendo que $\pi^{(0)} = (1, 0)$.

SE PUEDE DEMOSTRAR QUE, EN UN CASO GENERAL, la fila i de la matriz P^n se corresponde con el vector de probabilidades de estado $\pi^{(n)}$ si el estado inicial en $n = 0$ es s_i. Esto es, dicha fila representa probabilidad de estar en cada uno de los estados $\{s_1, s_2, \ldots, s_k\}$ tras n iteraciones partiendo del estado s_i. Por ello, P^n_{ij} representa la probabilidad de llegar al estado s_j en n transiciones partiendo del estado s_i:

$$P^n_{ij} \triangleq \Pr(X_{m+n} = s_j \mid X_m = s_i)$$

y, por lo tanto, un estado s_j es accesible desde s_i si existe algún n tal que $P^n_{ij} > 0$.

Ejemplo 5.20. Sea la cadena dada por la siguiente matriz de transiciones y representada al margen



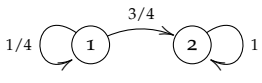
$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Resulta sencillo comprobar que el valor de P^2 resulta ser:

$$P^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Lo que significa que, independientemente del estado inicial, en $n = 2$ el sistema acabará en s_C.

Ejemplo 5.21. Sea la cadena con la siguiente matriz P



$$P = \begin{pmatrix} 1/4 & 3/4 \\ 0 & 1 \end{pmatrix}.$$

Se puede obtener que la potencia P^2 es igual a

$$P^2 = \begin{pmatrix} 1/16 & 15/16 \\ 0 & 1 \end{pmatrix}\ ,$$

mientras que P^4 resulta ser[6]

[6] Dado que, por definición, las filas de la matriz P^n tienen que sumar 1, realizar la potencia resulta razonablemente sencillo.

$$P^4 = \begin{pmatrix} 1/256 & 255/256 \\ 0 & 1 \end{pmatrix},$$

de lo que se puede intuir que el valor de P^n será igual a

$$P^n = \begin{pmatrix} (1/4)^n & 1-(1/4)^n \\ 0 & 1 \end{pmatrix} .$$

Otra forma de realizar el cálculo de P^n puede ser por inspección del comportamiento de la cadena. Para ello, se puede partir de la interpretación de las filas de P^n:

- Si la cadena empieza en el estado 2, permanecerá en él para siempre, al ser un estado absorbente. Por lo tanto, se tiene que

$$P^n_{2,*} = (0,1)$$

- Si la cadena empieza en estado 1, existe una probabilidad de $(1/4)^n$ de que permanezca en dicho estado pasadas n iteraciones. Dado que la suma de las probabilidades de cada fila tiene que ser 1, se tiene que

$$P^n_{1,*} = ((1/4)^n, 1-(1/4)^n)$$

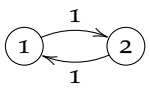
Por lo tanto:

$$P^n = \begin{pmatrix} (1/4)^n & 1-(1/4)^n \\ 0 & 1 \end{pmatrix} .$$

Periodicidad

Los componentes de P^n determinan la probabilidad de encontrarse en el estado s_j partiendo del estado s_i y, por lo tanto, también determinan la probabilidad de *no* encontrarse en cualquier estado en un momento dado. Sea el caso de la cadena de Markov dada por la matriz y el diagrama de transiciones a continuación:

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$



Se puede comprobar que las potencias P^n tienen la siguiente forma para el caso de n impar y par, respectivamente

$$P^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \; P^{2n} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Dado que la primera fila de P^{2n+1} es $(0,1)$, si el estado inicial es s_1, resulta *imposible* que la cadena que pase por dicho estado los valores de n impares.[7] En función de que exista o no esta imposibilidad de volver a un estado, dicho estado será periódico o aperiódico.

[7] Aunque a lo largo de las iteraciones la cadena pase la mitad de tiempo en s_1 y la otra mitad en s_2, no se puede decir p.ej. que la probabilidad de que esté en s_1 es de $1/2$ para cualquier valor de n.

Periodo de un estado: dada una cadena de Markov $\{X_n\}$ con espacio de estados S y matriz de transición P, el periodo de un estado s_i se define como el máximo común divisor (mcd) de todos los instantes n en los que la cadena podría volver a dicho estado:

$$d(s_i) \triangleq \text{mcd}\{n \geq 1 : P_{ii}^n > 0\}.$$

Ejemplo 5.22. En la cadena de dos estados del ejemplo anterior se cumple que

$$P^2 = P^4 = P^8 = \ldots = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

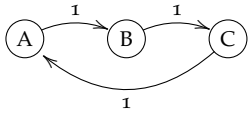
Por lo tanto, $d(s_1) = d(s_2) = 2$.

EN FUNCIÓN DEL VALOR QUE TOME $d(s_i)$ se define si un estado es periódico o no:

- Si $d(s_i) = 1$, el estado es aperiódico.
- Si $d(s_i) > 1$, el estado es periódico, con periodo $d(s_i)$.

Una cadena de Markov en la que todos sus estados son aperiódicos es una cadena *aperiódica*, mientras que en caso contrario, la cadena es *periódica*. Si dos estados pertenecen a la misma clase, tienen el mismo periodo, por lo que para identificar si una cadena irreducible es periódica o aperiódica basta con identificar el periodo de uno de sus estados.

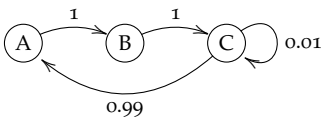
Ejemplo 5.23. Sea la siguiente cadena de Markov



$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

que se puede comprobar que es periódica con periodo $d = 3$.

Sin embargo, en la siguiente cadena, que resulta muy parecida" a la anterior



$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0{,}99 & 0 & 0{,}01 \end{pmatrix}$$

el estado C es aperiodico (dado que p_{CC} =0.01). Puesto que es una cadena irreducible, todos los demás estados, y la cadena, son aperiódicos.

Ejemplo 5.24. Sea la cadena definida por

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

se puede comprobar que el valor de P^8 es

$$P = \begin{pmatrix} 71/256 & 57/256 & 1/4 & 9/64 & 7/64 \\ 57/256 & 39/128 & 29/256 & 21/64 & 1/32 \\ 1/4 & 129/256 & 49/128 & 9/256 & 7/32 \\ 9/64 & 21/64 & 9/256 & 63/128 & 1/256 \\ 7/32 & 1/16 & 7/16 & 1/128 & 35/128 \end{pmatrix}.$$

Dado que $P^8_{ij} > 0$ para todo i y j, se cumple que $P^{8+n}_{ij} > 0$ también para todo valor de $n \geq 0$. Por lo tanto, todos los elementos de la diagonal serán estrictamente positivos a partir de P^8 y siguientes potencias, por lo que la cadena es aperiódica.

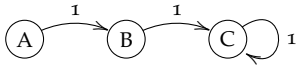
Distribuciones de estado estacionarias

El comportamiento de una cadena de Markov $\{X_n\}$ según avanza n se rige, como se ha visto, por las ecuaciones de Chapman-Kolmogorov:

$$\pi^{(n+1)} = \pi^{(n)} P.$$

A continuación se analiza el comportamiento de una cadena cuando $n \to \infty$, con objeto de poder obtener si existe un comportamiento «estable» de la misma cuando haya pasado «mucho tiempo». Esto serviría, p.ej., para calcular la probabilidad media de que un equipo acierte un penalty, o el % de días que hace frío o calor en un determinado pueblo.

Ejemplo 5.25. En el caso del Ejemplo 5.20 se tenía la cadena dada por la siguiente matriz de transiciones y representada al margen



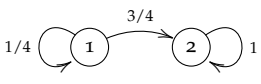
$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

De la que se puede se puede deducir que a partir de $n \geq 2$

$$P^n = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

lo que significa que, a partir de un valor de n, el sistema acababa en el estado s_C, esto es, $\pi^{(n)} = (0, 0, 1)$.

Para el caso del Ejemplo 5.21 se tenía la cadena con la siguiente matriz P



$$P = \begin{pmatrix} 1/4 & 3/4 \\ 0 & 1 \end{pmatrix}.$$

En el que

$$P^n = \begin{pmatrix} (1/4)^n & 1 - (1/4)^n \\ 0 & 1 \end{pmatrix}.$$

A partir de esta expresión, se tiene que según $n \to \infty$ la matriz P^n tenderá a

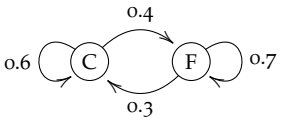
$$P^n \to \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix},$$

lo que significa que el sistema tenderá al vector $(0,1)$.

En los dos ejemplos anteriores se tiene que, a partir de un valor de n, la cadena permanece siempre en el mismo estado, por lo que $\pi^{(n)}$ tiene una componente con valor a 1 y el resto a 0. Sin embargo, como se verá en el ejemplo a continuación, también podrá pasar que el vector $\pi^{(n)}$ no cambie según avance n, aunque ninguna de sus componentes tenga valor igual a la unidad: es la *incertidumbre* sobre el estado en que puede estar la cadena la que no cambia.

Ejemplo 5.26. Sea el ejemplo del tiempo, con la matriz de transiciones y diagrama de estados ya conocidos

$$P = \begin{pmatrix} 0{,}6 & 0{,}4 \\ 0{,}3 & 0{,}7 \end{pmatrix}$$



Calculando las primeras potencias 2^k de P^n, se obtiene

$$P^2 = \begin{pmatrix} 0{,}48 & 0{,}52 \\ 0{,}39 & 0{,}61 \end{pmatrix}, \quad P^4 = \begin{pmatrix} 0{,}4332 & 0{,}5668 \\ 0{,}4251 & 0{,}5749 \end{pmatrix}, \quad P^8 = \begin{pmatrix} 0{,}4286 & 0{,}5714 \\ 0{,}4281 & 0{,}5715 \end{pmatrix},$$

por lo que las filas de la matriz P^n resultan muy similares. De hecho, a partir de la potencia novena las probabilidades entre filas son iguales hasta el cuarto decimal

$$P^9 = \begin{pmatrix} 0{,}4286 & 0{,}5714 \\ 0{,}4286 & 0{,}5714 \end{pmatrix} \approx P^{10} \approx P^{11} \approx \dots$$

Por lo tanto, el vector de distribución de probabilidades $\pi^{(n)}$ no cambia significativamente a partir de $n \geq 9$, independientemente del estado inicial. Esto supone que a partir de dicho valor de n

- La probabilidad de que haga calor (o frío) no depende del tiempo que hiciese al principio (el primer día).
- El 43 % (aprox.) de los días serán calurosos, y el 57 % fríos.

Que el vector de distribución de probabilidades no «cambie» a partir de un valor de n no supone que el sistema se haya vuelto *estático*, ni que –siguiendo con el ejemplo anterior– la probabilidad de que haga frío o calor un día no dependa del día anterior.[8] Lo que esto supone es que, teniendo en cuenta todas las posibles secuencias de estados que puede tener $\{X_n\}$, las probabilidades de las mismas ponderadas por las transiciones entre estados *convergen* (para $n \gg 1$) a un vector de distribución de probabilidades. A continuación se formaliza la definición de este vector.

[8] Si p.ej. se observase el día $n = 12$ y éste fuese caluroso, la probabilidad de que $n = 13$ fuese caluroso sería del 0.6, y no de 0.4286.

Definición de π

Sea una cadena de Markov con matriz de transiciones P donde, a partir de un valor de n, el vector de probabilidades de estado $\pi^{(n)}$

no varía. En estas condiciones, se cumple que

$$\pi^{(n)} = \pi^{(n+1)}, \quad n \gg 1$$

Aplicando las ecuaciones de Chapman-Kolmogorov a ambos lados de la equación, resulta

$$\pi^{(0)} P^n = \pi^{(0)} P^{n+1}$$

que puede expresarse como

$$\pi^{(0)} P^n = \left(\pi^{(0)} P^n\right) P$$

esto es

$$\pi^{(n)} = \pi^{(n)} P. \tag{5.4}$$

Por lo tanto, si un vector $\pi^{(n)}$ no cambia a partir de un valor de n, debe cumplir la relación dada por (5.4).[9] Este vector recibe el nombre de *distribución estacionaria* de la cadena de Markov y se representa sin superíndice

[9] De la expresión (5.4) también se tiene que dicho vector es un autovector de la matriz P, de autovalor la unidad.

$$\pi = \lim_{n\to\infty} \pi^{(n)}$$

Vector de distribución estacionaria: Sea $\{X_n\}$ una cadena de Markov con matriz de transiciones P. El vector de distribución de probabilidades $\pi = (\pi_1, \ldots, \pi_N)$ es una *distribución estacionaria* de la cadena si cumple que:

$$\pi P = \pi \tag{5.5}$$

Además, dado que es un vector de distribución de probabilidades, también debe cumplir que

$$\pi_i \geq 0 \ \forall i, \quad \sum \pi_i = 1. \tag{5.6}$$

Ejemplo 5.27. Para el caso del modelo del tiempo, con la siguiente matriz P

$$P = \begin{pmatrix} 0{,}6 & 0{,}4 \\ 0{,}3 & 0{,}7 \end{pmatrix}$$

Se puede comprobar que el siguiente vector

$$\pi = (3/7, 4/7).$$

cumple las condiciones dadas por (5.5) y (5.6), por lo que es una distribución de estados estacionaria (más adelante se abordará el cálculo del vector π).

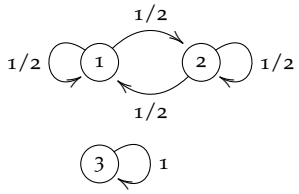
OTROS NOMBRES QUE SE SUELEN UTILIZAR para referirse a π son: distribución invariante, distribución de equilibrio o probabilidades de estado estacionario.

Convergencia a un único π

A partir de la definición de π, en una cadena puede ocurrir tanto que existan varios vectores que cumplan dicha definición, como que no exista ninguno. Además, también se ha visto que $\pi^{(n)}$ puede aproximarse a dicho π según $n \to \infty$, pero no en qué condiciones: la unicidad, existencia y convergencia a π dependen de la irreducibilidad y periodicidad de la cadena.

- Si una cadena es *reducible*, no todos los estados se comunican entre sí. En estas condiciones, el valor de $\pi^{(n)}$ cuando $n \to \infty$ puede ser muy diferente según sea el estado inicial.

Ejemplo 5.28. Sea el caso de la siguiente cadena reducible (representada al margen):



$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Se puede comprobar que hay al menos dos vectores de distribución estacionaria válidos:

$$\pi_a = (1/2, 1/2, 0) \text{ y}$$
$$\pi_b = (0, 0, 1),$$

dado que ambos cumplen con la definición.[10] Que el comportamiento de la cadena cuando $n \to \infty$ se corresponda con π_a o π_b dependerá del estado inicial y determinará que, en cualquier caso, se puede *reducir* al análisis de una cadena de un menor número de estados.

[10] Se puede demostrar que para cualquier $p \in (0,1)$, el vector $p\pi_a + (1-p)\pi_b$ cumple con las condiciones de distribución estacionaria.

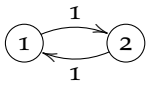
COMO ILUSTRA EL EJEMPLO ANTERIOR, en una cadena reducible no puede garantizarse que exista un único π, ya que al haber más de una clase el estado inicial puede determinar el comportamiento a lo largo de n. En una cadena irreducible, en cambio, dado que todos los estados se comunican, siempre existirá la posibilidad de visitar cualquier estado independientemente del estado inicial

- Si una cadena es *periódica*, si el estado inicial es periódico (con periodo $d > 1$), la cadena lo volverá a visitar en los múltiplos de dicho periodo $k \cdot d$, y sólo en dichos múltiplos: la probabilidad de que lo visite en cualquier otro momento es nula. Por lo tanto, es imposible que el vector de distribución de probabilidades de estado «converja» a un valor fijo π:

$$\nexists \lim_{n \to \infty} \pi^{(n)}$$

Ejemplo 5.29. Sea el caso de la cadena dada por

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$



La cadena alterna de forma estricta entre los estados 1 y 2, por lo que pasa la mitad del tiempo (aproximadamente) en cada estado. Sin embargo, si la cadena empieza en el estado 1, a lo largo de n el vector $\pi^{(n)}$ no converge a ningún valor, al no existir el límite, tal y como se ilustra en la Figura 5.4.

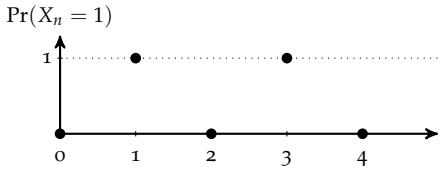


Figura 5.4: Probabilidad de estar en el estado 1 para el Ejemplo 5.29.

Si la condición de reducibilidad de una cadena permitía la existencia de varios vectores de distribución estacionaria, la periodicidad de una cadena impide su existencia. En una cadena de Markov irreducible y aperiódica se tiene que, además de que π exista y sea único, la cadena converge a dicho vector.

Cadenas de Markov irreducibles y aperiódicas: En una cadena de Markov irreducible y aperiódica se cumple que: (*i*) existe un vector de distribución estacionaria π, (*ii*) dicho vector es único, y (*iii*) independientemente del estado inicial, la cadena converge a dicho vector:

$$\lim_{n\to\infty} \pi^{(n)} = \pi.$$

Ejemplo 5.30. En el caso del modelo del tiempo, independientemente de que hoy haga calor o frío, la probabilidad de que dentro de 12 días haga calor o frío vendrá dada por

$$\pi = (3/7, 4/7),$$

dado que dicho vector es único y la cadena tiende al mismo.

Las condiciones de irreducibilidad y aperiodicidad son condiciones suficientes para garantizar la existencia, unicidad y convergencia a π, pero no son condiciones necesarias: existen cadenas que, sin cumplir dichas condiciones, pueden tener un único π y converger al mismo (como p.ej. las cadenas de los ejemplos 5.20 y 5.21).

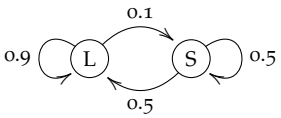
Cálculo de π en cadenas irreducibles aperiódicas

Dada una cadena y un vector de distribución de probabilidades «candidato» π, resulta sencillo comprobar si se trata de un vector de distribución estacionario: basta con comprobar si cumplen las ecuaciones (5.5) y (5.6), indicadas de nuevo al margen. Sin embargo, en un problema de modelado en general se querrá calcular dicho vector, con objeto de analizar alguna variable del sistema, como se ilustra en el ejemplo a continuación.

Un vector de distribución estacionaria π debe cumplir que:

$$\pi P = \pi$$
$$\sum \pi_i = 1$$

Ejemplo 5.31. Para el ejemplo de los lanzadores de penalties, una posible variable a analizar es la probabilidad media de que el equipo meta un penalty, que estará en un punto intermedio entre la del mejor lanzador y la del peor. Para calcular esta probabilidad de meter un penalty $\Pr(G)$, dado que se conocen las probabilidades de que S y L acierten (esto es, las probabilidades condicionadas):



$$\Pr(G \mid L) = 0.9$$
$$\Pr(G \mid S) = 0.5$$

se puede aplicar la ley de la probabilidad total, condicionando sobre el estado en que se encuentre la cadena

$$\Pr(G) = \Pr(G \mid L)\Pr(L) + \Pr(G \mid S)\Pr(S).$$

Para lo que es necesario calcular el vector π.

SI LA CADENA ES IRREDUCIBLE Y APERIÓDICA, el vector π existe y es único, por lo que sólo falta calcularlo. En una cadena de K estados, aplicar la expresión (5.5)

$$\pi P = \pi$$

resulta en un sistema de K ecuaciones, que *no* permite resolver las K incógnitas (esto es, las probabilidades de estar en cada uno de los estados), por las propiedades de la matriz de transiciones.[11] Para poder resolver el sistema, se descarta una de las K ecuaciones y se añade la condición (5.6)

$$\sum \pi_i = 1.$$

[11] Se trata de un sistema linearmente dependiente, dado que todas las filas suman 1: una columna se puede calcular en función del resto de columnas.

Ejemplo 5.32. Continuando con el caso de los lanzadores de penalties, la expresión $\pi = \pi P$ resulta en

$$\pi_L = 0.9\pi_L + 0.5\pi_S \tag{5.7}$$
$$\pi_S = 0.1\pi_L + 0.5\pi_S \tag{5.8}$$

donde π_L y π_S se corresponden con $\Pr(L)$ y $\Pr(S)$, esto es, las probabilidades de que el lanzador de un penalty sea Lucas y Sergio, respectivamente, cuando $n \to \infty$. Dicho sistema de ecuaciones no permite calcular π_L y π_S, sino que es preciso añadir la ecuación

$$\pi_L + \pi_S = 1. \tag{5.9}$$

A partir de la expresión (5.7) o (5.8) se obtiene

$$\pi_L = 5\pi_S\ ,$$

que sustituyendo en (5.9) resulta en

$$\pi_S = 1/6$$

y, por lo tanto

$$\pi_L = 5/6.$$

La probabilidad de acierto se calcula mediante la expresión ya vista

$$\Pr(G) = \Pr(G \mid L)\Pr(L) + \Pr(G \mid S)\Pr(S) \ ,$$

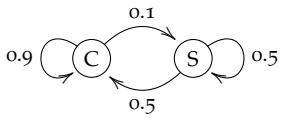
esto es,

$$\Pr(G) = \frac{9}{10} \cdot \frac{5}{6} + \frac{1}{2} \cdot \frac{1}{6} = \frac{5}{6}.$$

Tiempo medio de visita entre estados

En una cadena de Markov irreducible y aperiódica con matriz de transiciones P y número de estados K, sea n_{ij} la variable aleatoria discreta que representa el número de iteraciones en n que pasan desde que la cadena está en el estado s_i hasta que llega al estado s_j por primera vez (independientemente de los estados intermedios por los que pase).

Ejemplo 5.33. En el caso de los lanzamientos de penalties, el tiempo medio de visita desde L hasta S (n_{LS}) es una variable aleatoria geométrica, dado que si la cadena está en L, el número de iteraciones necesarias para alcanzar S es



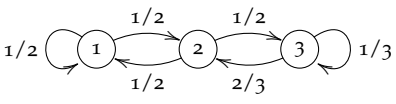
$$n_{LS} = \begin{cases} 1 & \text{Con probabilidad } p{=}0.1 \\ 2 & \text{Con probabilidad } p{=}0.9 \cdot 0.1 \\ \dots & \\ 1+k & \text{Con probabilidad } p{=}(0.9)^{\mathrm{k}} \cdot 0.1 \end{cases}$$

que tiene de media

$$\mathbb{E}[n_{LS}] = \frac{1}{1-0.9} = 10 \cdot$$

En un caso general, a diferencia del ejemplo anterior, no será posible caracterizar completamente el tiempo de visita entre dos estados cualquiera, dado que para ir s_i a s_j se podrá pasar por un número de estados intermedios $\{s_m\}$. Por lo tanto, el tiempo de estancia será una variable aleatoria geométrica, pero no será tan inmediato caracterizar el tiempo de visita entre estados.

Ejemplo 5.34. Para la cadena de la figura al margen, el tiempo de estancia en los estados s_1 y s_3 es una variable aleatoria geométrica, que coinciden con los tiempos de visita n_{12} y n_{32}, respectivamente. Sin embargo, el tiempo de visita n_{13} no es una variable aleatoria geométrica, dado que una vez abandonado el estado s_1 y llegado a s_2, la cadena puede volver a s_1 con probabilidad $1/2$. Por motivos similares, el tiempo n_{31} tampoco es una variable aleatoria geométrica.



Si bien la variable n_{ij} no se podrá caracterizar completamente en un caso general, sí que se puede calcular su esperanza, aplicando la ley de la probabilidad total. Para ello, se considera tanto el «salto directo» como todos los posibles caminos:

$$\begin{aligned}\mathbb{E}[n_{ij}] = &\Pr(\text{Pasar de } s_i \text{ a } s_j \text{ directamente}) \cdot 1 \\ &+ \sum_{s_k \neq s_j} \Pr(\text{Pasar de } s_i \text{ a } s_k) \cdot \mathbb{E}[n_{ij} \mid \text{Pasando por } s_k]\end{aligned} \tag{5.10}$$

esto es, se tiene en cuenta tanto el paso directo como todos los demás posibles *siguientes estados* tras abandonar s_i, y se calcula el valor de $\mathbb{E}[n_{ij}]$ dada esta condición.

La probabilidad de que partiendo de s_i el siguiente salto sea s_k es, por definición, p_{ik}. El número medio de saltos para llegar desde s_i a s_j si el siguiente salto es un estado intermedio s_k se puede expresar como la suma de dos componentes:

- La iteración que acaba de ocurrir, correspondiente a pasar de s_i a s_k (por lo que tiene como valor 1).
- El número medio de iteraciones para pasar de s_k a s_j que, por definición, es $\mathbb{E}[n_{kj}]$.

Por lo que queda

$$\mathbb{E}[n_{ij} \mid \text{Pasando por } s_k] = 1 + \mathbb{E}[n_{kj}]$$

De esta forma, la expresión (5.10) queda como

$$\mathbb{E}[n_{ij}] = p_{ij} + \sum_{s_k \neq s_j} p_{ik}(1 + \mathbb{E}[n_{kj}]),$$

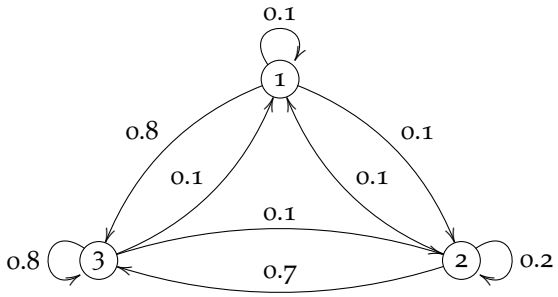
que, teniendo en cuenta que $\mathbb{E}[n_{jj}] = 0$, también puede escribirse como

$$\mathbb{E}[n_{ij}] = \sum_{s_k} p_{ik} \cdot (1 + \mathbb{E}[n_{kj}]). \tag{5.11}$$

o bien

$$\mathbb{E}[n_{ij}] = 1 + \sum_{s_k} p_{ik} \cdot \mathbb{E}[n_{kj}]. \tag{5.12}$$

Ejemplo 5.35. Sea la cadena dada por



de la que se quiere calcular la esperanza del número de iteraciones para alcanzar el estado 3 partiendo del estado 1. Siguiendo (5.12), se puede expresar como:[12]

[12] Por simplificar notación se emplea n_{13} en vez de $\mathbb{E}[n_{13}]$.

$$n_{13} = 1 + p_{11}n_{13} + p_{12}n_{23},$$

ecuación en la que, además del valor n_{13} que se desea calcular, también aparece n_{23}. Para ésta también se puede plantear una expresión siguiendo (5.12):

$$n_{23} = 1 + p_{21}n_{13} + p_{22}n_{23},$$

por lo que queda un sistema de dos ecuaciones con dos incógnitas, del que se obtiene

$$n_{13} = 9/7,1 \approx 1{,}267 \text{ iteraciones.}$$

Resumen del tema

- Un proceso aleatorio $\{X_n\}$ en un espacio S es una cadena de Markov si para cualquier n la probabilidad de que el proceso pase al estado $X_n = s_j$ sólo depende del estado en que se encuentre en $n-1$.
- Si estas probabilidades son constantes, la cadena de Markov es homogénea y se puede caracterizar con una matriz de probabilidades de transición P que no depende del instante de tiempo n.
- El vector de probabilidades de estado $\pi^{(n)}$ representa la probabilidad de que en el instante n la cadena se encuentre en cada uno de los posibles estados, y cumple que

$$\pi^{(n+1)} = \pi^{(n)}P = \pi^{(0)}P^n$$

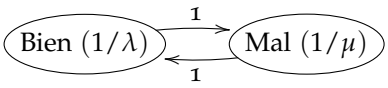
- Los estados que se comunican entre sí constituyen una clase. Si en una cadena todos los estados se comunican, la cadena es irreducible.
- Un estado es periódico si fuera de determinados instantes de tiempo, múltiplos del periodo, resulta imposible que la cadena pase por dicho estado. Si el periodo es 1, el estado es aperiódico.
- Un vector de distribuciones de estado es estacionario si cumple que $\pi = \pi P$.
- Una cadena irreducible y aperiódica tiene un único vector π al que converge según $n \to \infty$.
- El número medio de iteraciones para llegar desde el estado s_i al estado s_j se puede expresar como

$$n_{ij} = 1 + \sum_{k \neq j} p_{ik} \cdot n_{kj}$$

Cadenas de Markov de tiempo continuo

Las cadenas de Markov de tiempo discreto resultan adecuadas para sistemas en los que las transiciones tienen lugar en instantes de tiempo deterministas: el caso de un Aloha ranurado-donde hay transmisiones o colisiones en instantes de tiempo bien definidos, un *robot* que va navegando aleatoriamente por páginas en Internet, o si el tiempo (caluroso o frío) en un día se puede predecir partiendo del tiempo en los días anteriores. Sin embargo, cuando las transiciones no ocurren de forma discreta, sino tras un tiempo aleatorio y continuo, es necesario otro modelo.

Ejemplo 6.1. Sea el caso de un ordenador que se resetea cada un tiempo aleatorio, que se puede modelar según una variable aleatoria exponencial de media $1/\lambda$. El tiempo que tarda en arrancar también es aleatorio, con otro tiempo exponencial de media $1/\mu$. Dicho sistema se puede representar con una cadena como la siguiente



y, de hecho, guarda bastante relación con una cadena discreta periódica de dos estados. Sin embargo, dado que los tiempos de permanencia en cada estado son variables aleatorias continuas, resulta complicado determinar qué representaría una variable discreta n que modelase el avance del tiempo, así como un vector de distribución de probabilidades $\pi^{(n)}$.

Una cadena de Markov de tiempo continuo se puede interpretar como una "extensión" de las cadenas de tiempo discreto, donde el cambio de estado no sucede en instantes de tiempo bien definidos, sino tras instantes aleatorios que siguen una variable aleatoria exponencial. Debido a que la variable aleatoria exponencial no tiene memoria, se seguirá manteniendo la *propiedad de Markov*: lo que pueda pasar en el futuro vendrá determinado únicamente por el estado en que se encuentre la cadena, y no por el tiempo que lleve en él, ni por la historia pasada.

Definición

Una cadena de Markov de tiempo continuo es un proceso aleatorio en un espacio de estados contable, donde se cumple la propiedad de Markov: fijado un instante de tiempo s, la probabilidad de estar en un futuro $s+t$ en un estado j sólo dependerá del estado i en el que se encuentre el proceso (en el instante s), y no de la *historia* antes de s.

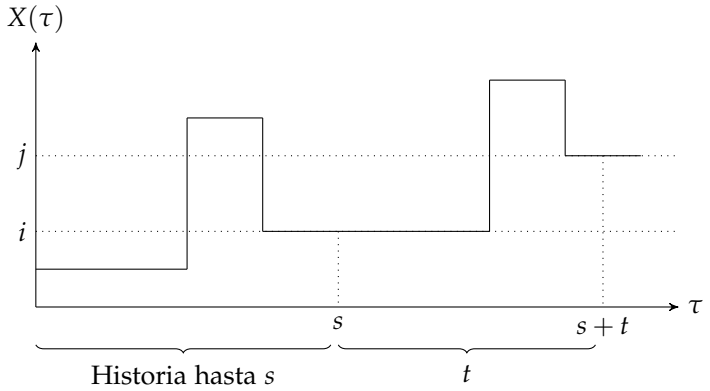


Figura 6.1: Cadena de Markov de tiempo continuo. La probabilidad de estar en el estado j en un tiempo $s+t$ ($\Pr(X(s+t)=j)$ sabiendo el valor de $X(s)$ sólo depende de t, y no de lo que haya pasado hasta s.

Cadena de Markov de tiempo continuo (1ª definición) Un proceso estocástico $\{X(t):\ t \geq 0\}$ en un espacio de estados S contable es una cadena de Markov de tiempo continuo si cumple la propiedad de Markov:

$$\Pr(X(t+s)=j \mid X(s)=i, X(t_{n-1})=i_{n-1}, \ldots, X(t_1)=i_1) \tag{6.1}$$
$$= \Pr(X(t+s)=j \mid X(s)=i).$$

para cualquier posible secuencia de instantes de tiempo

$$0 \leq t_1 \leq t_2 \leq \ldots t_{n-1} \leq s \leq t$$

y posible secuencia de números de S (el espacio de estados)[1]

$$i_1, i_2, \ldots, i_{n-1}, i, j \in S$$

[1] Los instantes de tiempo y secuencias de números sirven para definir cualquier posible historia antes de s en la expresión (6.1).

Si la probabilidad de pasar a j estando en i no depende del instante de tiempo s de referencia,

$$\Pr(X(t+s)=j \mid X(s)=i) = \Pr(X(t+u)=j \mid X(u)=i) \quad \forall s,u \geq 0$$

se trata de una *cadena homogénea* y se puede escribir

$$\Pr(X(t+s)=j \mid X(s)=i) = P_{ij}(t).$$

Tiempo de permanencia en un estado

Como se ha mencionado al inicio del tema, en una cadena de Markov de tiempo continuo el tiempo de estancia en un estado

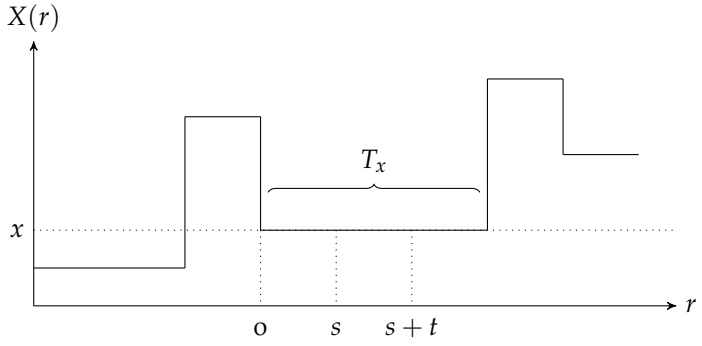


Figura 6.2: Tiempo de estancia en el estado x.

será una variable aleatoria exponencial. Esto se puede deducir a partir de la definición de la propiedad de Markov, como se realiza a continuación.

Sea T_x la variable aleatoria definida como el tiempo de estancia en el estado x (véase Figura 6.2). La probabilidad de que la cadena permanezca en x más de un tiempo $s+t$, suponiendo que permanece más de un tiempo s, se puede expresar como

$$\Pr(T_x > s+t \mid T_x > s) = \Pr\left(X(r) = x, \text{para } r \in [s, s+t] \mid X(r) = x, \text{para } r \in [0, s]\right)$$

La propiedad de Markov determina que, dado un punto de referencia s, la historia pasada no afecta al futuro. Por lo tanto,

$$\Pr(T_x > s+t \mid T_x > s) = \Pr\left(X(r) = x, \text{para } r \in [s, s+t] \mid X(s) = x\right),$$

Si la cadena es homogénea, la probabilidad no depende del valor de s, por lo que la parte derecha de la igualdad anterior se puede escribir como

$$\Pr(T_x > s+t \mid T_x > s) = \Pr\left(X(r) = x, \text{para } r \in [0, t] \mid X(0) = x\right),$$

donde esta última probabilidad resulta ser, por definición, la probabilidad de que el tiempo de estancia en el estado x sea mayor que t, esto es $\Pr(T_x > t)$.

Por todo lo anterior, se tiene que

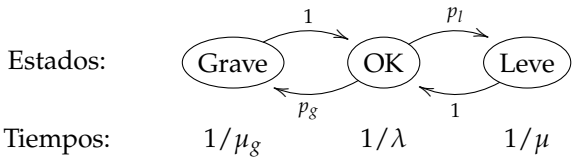
$$\Pr(T_x > s+t \mid T_x > s) = \Pr(T_x > t)$$

lo que define que la variable aleatoria T_x no tiene memoria, por lo que se trata de una variable aleatoria exponencial.

Ejemplo 6.2. El tiempo entre fallos de un servidor web se puede modelar con una variable aleatoria exponencial de media $1/\lambda$. Puede sufrir dos tipos de fallos: graves, con probabilidad p_g, que requieren un tiempo de reparación también exponencial de media $1/\mu_g$, y leves, con probabilidad $p_l = 1 - p_g$, que requieren otro tiempo de reparación exponencial de media $1/\mu_l$.

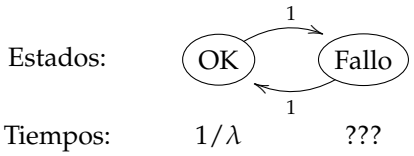
Este sistema se puede modelar con la siguiente cadena de Mar-

kov de tres estados



ya que en todo momento se cumple la propiedad de Markov: si p.ej. *ahora* no está funcionando, la probabilidad de que en los próximos 5' siga sin funcionar no depende del tiempo que lleve estropeado, sino únicamente de si el fallo fue grave o leve (esto es, del estado en que se encuentre).

Nótese que modelar este sistema con una cadena de Markov de dos estados plantea el problema del tiempo que permanece cuando el servidor ha fallado, que no podría modelarse como una variable exponencial de media $1/\mu_g$ o una variable exponencial de media $1/\mu_l$ (véase Figura 6.3):



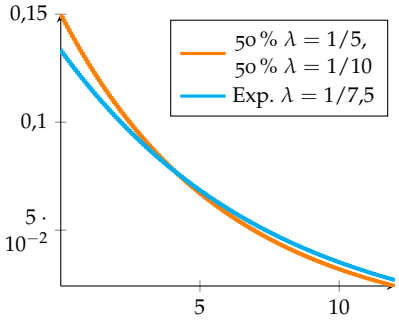


Figura 6.3: Función de densidad de una composición de v.a. exponenciales y de una v.a. exponencial

Ejemplo 6.3. Sea una partícula con un tiempo de vida exponencial, de media $1/\lambda$. Pasado dicho tiempo, la partícula se divide en dos partículas, con idénticas propiedades a la primera (véase la Figura 6.4 al margen).

En la situación inicial, con una única partícula, el tiempo hasta el siguiente *evento* será una variable aleatoria exponencial, de media $1/\lambda$. Con dos partículas, el siguiente evento vendrá dado por el *mínimo* de dos tiempos de vida: en cuanto una de ellas se divida, se pasará al siguiente estado con *tres* partículas, por lo que el tiempo hasta el siguiente evento se distribuirá según una variable aleatoria exponencial de media $1/(2\lambda)$.[2]

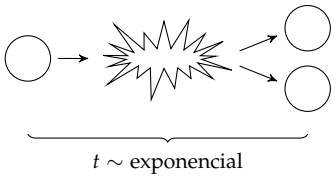
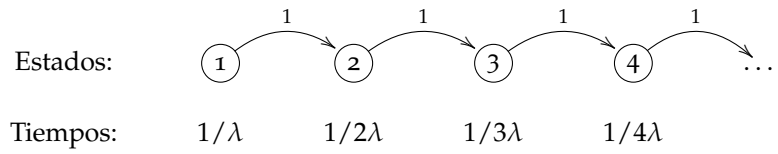


Figura 6.4: Partícula que se divide en dos con idénticas propiedades.

[2] Dado que son variables aleatorias continuas, la probabilidad de que las dos se dividan en el mismo instante de tiempo (que es continuo) es cero.

Siguiendo este razonamiento, el sistema se puede modelar con la siguiente cadena



Donde en el estado n hay n partículas, cada una con un tiempo medio de vida exponencial de media $1/\lambda$, y el tiempo medio de estancia en dicho estado es por tanto $1/(n\lambda)$.

Definición alternativa

Como se ha visto en los ejemplos anteriores, una cadena de Markov de tiempo continuo se puede relacionar con un diagrama de

estados que guarda una notable relación con las cadenas de Markov de tiempo discreto. A continuación se formaliza esta relación:

Cadena de Markov de tiempo continuo (2ª definición) Un proceso estocástico $\{X(t), t \geq 0\}$ en un espacio de estados contable S es una cadena de Markov de tiempo continuo si cumple que:

1. El tiempo de estancia en cada estado i se distribuye según una variable aleatoria exponencial independiente de media $1/v_i$.
2. El proceso abandona el estado i hacia el estado j con una probabilidad p_{ij}, cumpliéndose que $p_{ii} = 0$ y $\sum_j p_{ij} = 1, \forall i, j$.

Por lo tanto, una cadena de Markov de tiempo continuo es un proceso aleatorio que se *mueve* de un estado a otro según una cadena de Markov de tiempo discreto, pero en el que los tiempos de estancia en cada estado se distribuyen según una variable aleatoria exponencial. Por la propiedad de Markov, estos tiempos son variables aleatorias independientes, dado que de lo contrario sería preciso tener en cuenta la historia pasada del proceso. Un hecho importante de la cadena discreta asociada es que $p_{ii} = 0$, esto es, que nunca se "repite" el estado una vez que ha pasado el tiempo de permanencia correspondiente.

Ejemplo 6.4. Sea una gasolinera con un surtidor y una capacidad máxima para dos coches (incluyendo al que está repostando). Los coches llegan a repostar según un proceso de Poisson a tasa λ (por lo que el tiempo entre llegadas t_v se distribuye según una variable aleatoria exponencial de media $1/\lambda$) y el tiempo que permanecen ocupando el surtidor t_s se puede modelar con una variable aleatoria exponencial de media $1/\mu$ (por lo que, cuando el surtidor está ocupado, los coches "salen" según un proceso de Poisson a tasa μ).

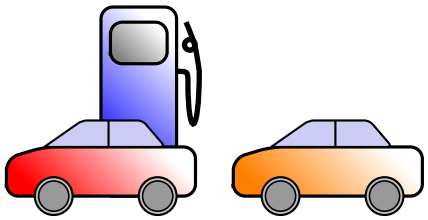
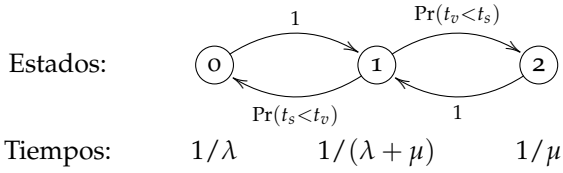


Figura 6.5: Gasolinera con un surtidor y sitio para dos coches.

Este sistema se puede modelar con la siguiente cadena



donde el estado representa el número de vehículos en la gasolinera:

- En el estado 0, el único evento que puede suceder es la llegada de un vehículo, a tasa λ. Por lo tanto, el tiempo de permanencia en este estado es exponencial de media $1/\lambda$ y, una vez pasado dicho tiempo, se pasa con probabilidad 1 al estado 1.
- En el estado 1, con un vehículo respostando, pueden suceder dos eventos: (*i*) llegada de otro vehículo, o (*ii*) fin del repostaje. Ambas cosas suceden tras una variable aleatoria exponencial, de media $1/\lambda$ y $1/\mu$, respectivamente, por lo que lo primero que suceda determinará el siguiente estado: si llega antes otro vehículo ($\Pr(t_v < t_s)$) se pasa al estado 2, mientras que si termina de repostar antes ($\Pr(t_s < t_v)$) se pasa al estado 0.

- En el estado 2, dado que no caben más vehículos, el único evento que puede suceder es el fin del repostaje, tras el que se pasa al estado 1.

El tiempo de permanencia en el estado 1 viene dado por el mínimo de dos variables aleatorias exponenciales, por lo que es otra exponencial, de media $1/(\lambda+\mu)$.[3] La probabilidad de que una llegada se produzca antes del fin del repostaje (o viceversa) viene dada por[4]

$$\Pr(t_v < t_s) = \frac{\lambda}{\lambda+\mu} \qquad \Pr(t_s < t_v) = \frac{\mu}{\lambda+\mu}$$

Además, ya que el tiempo de estancia es una variable aleatoria exponencial, se puede interpretar que la cadena "sale" del estado 1 según un proceso de Poisson a tasa $\lambda+\mu$, y se descompone en dos procesos: con probabilidad $\Pr(t_s < t_v)$ hacia el estado 0 y con probabilidad $\Pr(t_v < t_s)$ hacia el estado 2. Por lo tanto, las tasas de salida son

$$\text{Tasa de 1 a 2: } (\lambda+\mu)\cdot\frac{\lambda}{\lambda+\mu} = \lambda \text{ y tasa de 1 a 0: } (\lambda+\mu)\cdot\frac{\mu}{\lambda+\mu} = \mu,$$

con una interpretación muy inmediata: la gasolinera se "llena" a la tasa de llegada de vehículos, y se "vacía" a la tasa de repostaje.

[3] Mínimo de variables aleatorias exponenciales, página 38

[4] Comparación de variables aleatorias exponenciales, página 40

Evolución en el tiempo de la cadena

En las cadenas discretas la evolución del sistema sucede en instantes de tiempo bien delimitados (de n a $n+1$) según las ecuaciones de Chapman-Kolmogorov:

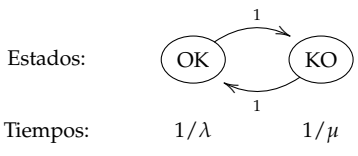
$$\pi^{(n+1)} = \pi^{(n)} P.$$

En el caso de una cadena de tiempo continuo, analizar la evolución de la cadena en el tiempo supone ser capaz de calcular la probabilidad de que la cadena esté en el estado i en un instante de tiempo t, esto es,

$$p_i(t) \triangleq \Pr(X(t) = i) \ ,$$

lo que resulta más complicado dado que hay un continuo de instantes de tiempo.

Ejemplo 6.5. Para el caso de la máquina que se estropea (Ejemplo 6.1), con el diagrama de estados indicado al margen, resulta claro que el sistema alterna entre dos estados, y que el tiempo de permanencia en cada uno de ellos se distribuye según una variable aleatoria exponencial, de distinta media. Una posible realización de dicha cadena sería la ilustrada en la siguiente figura:



Esta figura ilustra la dificultad del cálculo de la probabilidad de que la cadena esté en un determinado estado para un instante de tiempo, particularmente si se compara con el caso de las cadenas de Markov de tiempo discreto, como se discute a continuación.

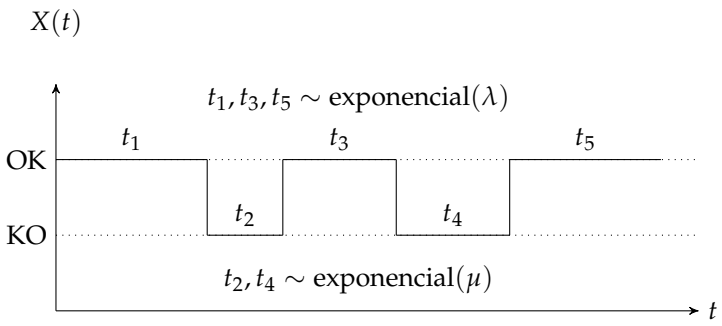


Figura 6.6: Evolución del estado (OK o KO) de una máquina a lo largo del tiempo.

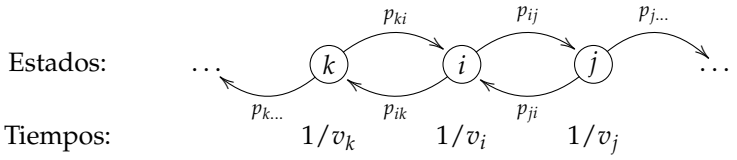
Dado un estado inicial i en $t = 0$ (esto es, unas *condiciones iniciales*), es sencillo calcular el tiempo que la cadena permanece en dicho estado: por lo visto anteriormente,[5] se trata de una variable aleatoria exponencial de media $1/v_i$, por lo que la probabilidad de que la cadena permanezca en dicho estado coincide con el tiempo de supervivencia:

$$\Pr(X(s) = i, \text{para } s \in [0,t] \mid X(0) = i) = e^{-v_i t} .$$

[5] Tiempo de permanencia en un estado, página 104

Sin embargo, resulta más complicado calcular la probabilidad de que, dado un estado inicial, la cadena se encuentre en dicho estado en cualquier instante de tiempo t, ya que es necesario considerar si ha permanecido en dicho estado, si ha visitado otros estados una o más veces, e incluso cuándo se han producido dichas visitas. Dada esta infinidad de posibilidades en la evolución de $p_i(t)$, para su cálculo se analizará lo que pueda ocurrir en un infinitesimal del tiempo Δt.

Para calcular la probabilidad de que la cadena se encuentre en el estado i en el instante de tiempo t (esto es, $p_i(t)$), sea una cadena genérica como la ilustrada a continuación



Dado que el tiempo de estancia en un estado es una variable aleatoria exponencial de media $1/v_i$, se puede interpretar que la cadena "sale" de un estado según un proceso de Poisson a tasa v_i. Con esto, se puede emplear la tercera definición de un proceso de Poisson para calcular lo que puede suceder tras un instante de tiempo Δt, y calcular la derivada de $p_i(t)$, de forma similar a como se demuestra que la tercera definición lleva a la primera.[6]

En un intervalo infinitesimal de tiempo Δt la probabilidad de que el sistema abandone el estado i es

$$\Pr(\text{salir de } i \text{ en } \Delta t) = v_i \cdot \Delta t + o(\Delta t).$$

ya que $1/v_i$ es el tiempo medio de estancia en i. Fijado un instante de tiempo t, la probabilidad de que en $t + \Delta t$ el sistema se

[6] Según la tercera definición, un proceso de Poisson $N(t)$ a tasa λ es un proceso de conteo que cumple las siguientes propiedades:

(i) $N(0) = 0$

(ii) $N(t)$ tiene incrementos independientes y estacionarios.

(iii) $\Pr(N(h) = 1) = \lambda h + o(h)$

(iv) $\Pr(N(h) \geq 2) = o(h)$

encuentre en el estado i se puede expresar en función de dónde se encontrase en el instante t, pudiendo ocurrir que:

- El sistema ya se encontraba en i y no lo abandona, o que
- El sistema se encontraba en cualquiera de los otros estados $j \neq i$ (lo que ocurre con probabilidad $p_j(t)$ para cada estado j), y lo abandona para moverse al estado i.

Por lo tanto, el cálculo de $p_i(t + \Delta t)$ se puede expresar como

$$\begin{aligned} p_i(t+\Delta t) =& p_i(t)(1 - \Pr(\text{salir de } i \text{ en } \Delta t)) \\ &+ \sum_{j \neq i} p_j(t) \Pr(\text{salir de } j \text{ en } \Delta t) \Pr(\text{ir de } j \text{ a } i) \end{aligned} \tag{6.2}$$

La probabilidad de que no lo abandone es

$$1 - \Pr(\text{salir de } i \text{ en } \Delta t) = 1 - (v_i \cdot \Delta t + o(\Delta t)),$$

mientras que la probabilidad de salir del estado j para llegar a i es

$$\Pr(\text{salir de } j \text{ en } \Delta t) \Pr(\text{ir de } j \text{ a } i) = (v_j \Delta t + o(\Delta t)) p_{ji}$$

donde p_{ji} es la probabilidad de pasar de j a i una vez se sale del estado j, como se ha visto anteriormente.

Por simplificar la notación, no se tendrán en cuenta las $o(\Delta t)$ en las siguientes expresiones, dado que desaparecerán cuando se realice el cálculo de la derivada de $p_i(t)$ (esto es, cuando $\Delta t \to 0$). A partir de las expresiones anteriores, (6.2) queda como:

$$\begin{aligned} p_i(t+\Delta t) =& p_i(t)(1 - v_i \Delta t) \\ &+ \sum_{j \neq i} p_j(t) v_j \Delta t p_{ji} \end{aligned}$$

Reordenando los elementos de la anterior ecuación, se obtiene

$$\frac{p_i(t+\Delta t) - p_i(t)}{\Delta t} = -p_i(t) v_i + \sum_{j \neq i} p_j(t) \cdot v_j p_{ji} \ ,$$

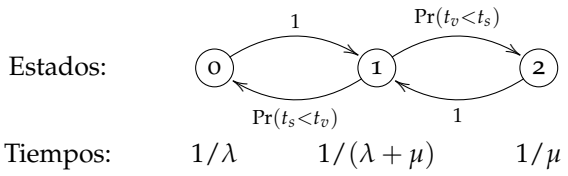
que cuando $\Delta t \to 0$ permite obtener la derivada:

$$\frac{dp_i(t)}{dt} \triangleq p_i'(t) = -p_i(t) v_i + \sum_{j \neq i} p_j(t) \cdot v_j p_{ji}, \quad i = 0, 1, \ldots, K \tag{6.3}$$

expresión que permite interpretar la derivada de $p_i(t)$ (esto es, el crecimiento de la probabilidad de que la cadena se encuentre en el estado i en el instante t), en términos de dos componentes:

- Un término negativo, que se corresponde con la probabilidad de que se encontrase en dicho estado i en t (esto es, $p_i(t)$) multiplicado por la tasa de salida desde el estado, esto es, v_i.
- Un término positivo, que es la suma de la probabilidad de que se encontrase en cualquier otro estado j ($p_j(t)$) por la *tasa de transición* desde j a i. Dicha tasa de transición es el producto de la tasa de salida de j (v_j) por la probabilidad de ir desde j a i (p_{ji}).

Ejemplo 6.6. En el ejemplo de la gasolinera, modelado con la siguiente cadena



Las tasas de salida desde cada estado son

$$v_0 = \lambda$$
$$v_1 = \lambda + \mu$$
$$v_2 = v_2$$

mientras que las tasas de transición entre estados son

$$\text{De 0 a 1} = \lambda \qquad \text{De 1 a 0} = \mu$$
$$\text{De 1 a 2} = \lambda \qquad \text{De 1 a 2} = \mu$$

Como es de esperar, la suma de las tasas de transición desde un estado coincide con la tasa de salida de dicho estado (p.ej., la suma de las tasas de 1 a 2 y de 1 a 0 coincide con v_1).

Matriz de tasas de transición

Las tasas de salida y entrada se pueden representar con la *matriz de tasas de transición* Q, también llamada *generador infinitesimal*, donde

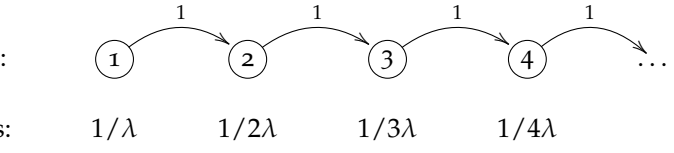
- El elemento de la fila i y columna j es la tasa de transición desde el estado i al estado j
$$q_{ij} = v_i \cdot p_{ij}$$
- Los elementos de la diagonal representan las tasas de salida desde los estados de la cadena, esto es
$$q_{ii} = -v_i$$

Dado que cuando una cadena sale de un estado debe ir a otro diferente, la tasa de salida del estado i se corresponde con la suma de las tasas de llegada desde el estado i a todos los demás estados. Por lo tanto, [7]

$$-q_{ii} = v_i = \sum_{j \neq i} q_{ij}$$

[7] En las cadenas de tiempo discreto la suma de las filas de la matriz P era igual a 1, lo que resultaba en que p.ej. una columna se podía obtener a partir del valor de todas las demás. En este caso, en las cadenas de tiempo continuo la suma de las filas de la matriz Q es igual a 0 (y también se puede deducir el valor de una columna a partir del resto).

Ejemplo 6.7. El caso de las partículas que se dividen se podía modelar con la siguiente cadena



que se corresponde con la siguiente matriz Q:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & 0 & \dots \\ 0 & -2\lambda & 2\lambda & 0 & 0 & \dots \\ 0 & 0 & -3\lambda & 3\lambda & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \dots \end{pmatrix}$$

Ejemplo 6.8. Para el caso de la gasolinera, la matriz de tasas de transición o generador infinitesimal resulta ser:

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & \mu & -\mu \end{pmatrix}$$

LA MATRIZ DE TASAS DE TRANSICIÓN permite expresar la evolución de una cadena continua de forma similar a como lo hacen las ecuaciones de Chapman-Kolmogorov para el caso discreto. Sea una cadena de K estados con las probabilidades de estado representadas en forma vectorial

$$\boldsymbol{p(t)} = [p_0(t), p_1(t), p_2(t), \dots, p_K(t)],$$

y las derivadas de dichas probabilidades, también en forma vectorial

$$\boldsymbol{p'(t)} = [p'_0(t), p'_1(t), p'_2(t), \dots, p'_K(t)].$$

A partir de estas definiciones, la expresión (6.3) para todos los estados queda como[8]

$$\boldsymbol{p'(t)} = \boldsymbol{p(t)}Q. \tag{6.4}$$

[8] Se repite aquí (6.3):

$$p'_i(t) = -p_i(t)v_i + \sum_{j \neq i} p_j(t) \cdot v_j p_{ji}$$

Ejemplo 6.9. De nuevo en el caso de la gasolinera, la ecuación (6.4) quedaría como

$$[p'_0(t), p'_1(t), p'_2(t)] = [p_0(t), p_1(t), p_2(t)] \begin{pmatrix} -\lambda & \lambda & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & \mu & -\mu \end{pmatrix},$$

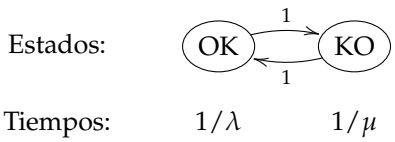
lo que daría lugar a un sistema de ecuaciones diferenciales.

ANALIZAR EL COMPORTAMIENTO EN TIEMPO de una cadena de Markov de tiempo continuo resulta más difícil que hacer lo mismo en el caso de las de tiempo discreto, dado que hay que tratar con ecuaciones diferenciales.[9] Para ilustrar esta complejidad, a continuación se analiza el sistema más sencillo.

[9] Al sistema de ecuaciones diferenciales habría que añadir la condición de normalización $\sum_i p_i(t) = 1$.

Ejemplo: análisis en el tiempo de una máquina con dos estados

Sea el caso del inicio del tema: una máquina con un tiempo de vida que se distribuye según una variable aleatoria exponencial de media $1/\lambda$ y un tiempo de reparación que se distribuye también según otra variable aleatoria exponencial de media $1/\mu$:



Se quiere calcular la probabilidad de que la máquina se encuentre funcionando en un instante de tiempo t, esto es, $p_{OK}(t)$. Según lo visto anteriormente, la derivada de esta probabilidad se puede expresar como

$$p'_{OK}(t) = -p_{OK}(t)\lambda + p_{KO}(t)\mu \ .$$

Con la condición de normalización $p_{OK}(t) + p_{KO}(t) = 1$ resulta la siguiente ecuación diferencial de primer orden

$$p'_{OK}(t) = \mu - p_{OK}(t)(\lambda + \mu),$$

que tiene una solución de la forma

$$p_{OK}(t) = \frac{\mu}{\lambda + \mu} + Ke^{-(\lambda+\mu)t}$$

donde K es una constante que depende de las condiciones iniciales del sistema:

- Si la máquina funciona en $t = 0$, entonces $p_{OK}(0) = 1$. De esto se deduce que $K = \lambda/(\lambda + \mu)$, por lo que

$$p_{OK}(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu}e^{-(\lambda+\mu)t}$$

- Si la máquina está estropeada en $t = 0$, entonces $p_{OK}(0) = 0$, de lo que se deduce que $K = -\mu/(\lambda + \mu)$, por lo que

$$p_{OK}(t) = \frac{\mu}{\lambda + \mu} - \frac{\mu}{\lambda + \mu}e^{-(\lambda+\mu)t}$$

Se ilustran ambas soluciones de $p_{OK}(t)$ en la Figura 6.7, para un caso donde $\mu = 2$ y $\lambda = 1$. Según se aprecia en la figura, independientemente de las condiciones iniciales, a partir de aproximadamente 1.5 unidades de tiempo el efecto de éstas desaparece, siendo la probabilidad de encontrar la máquina funcionando constante e igual a

$$p_{OK}(t) = \frac{\mu}{\lambda + \mu} \approx 0.66$$

Por lo tanto, y como sucedía en algunas cadenas de Markov de tiempo discreto, tras un tiempo suficientemente largo la probabilidad de que la máquina se encuentre o no funcionando no depende de las condiciones iniciales, sino de las características de la máquina: 2/3 del tiempo estará operativa y 1/3 del tiempo estará estropeada.

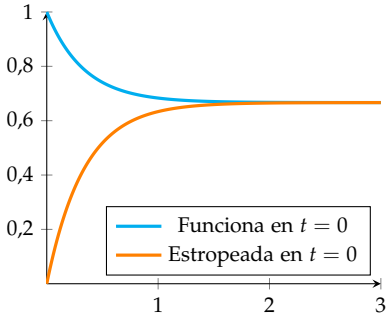


Figura 6.7: Evolución de $p_{OK}(t)$ según la máquina funcione o esté estropeada en $t = 0$, con $\mu = 2$ y $\lambda = 1$

Representación del diagrama de estados

Hasta ahora, las cadenas de Markov se han representado con dos tipos de información:

- Por una parte, la media de los tiempos exponenciales de permanencia en cada estado.
- Por otra parte, una cadena de Markov discreta con las probabilidades de transición entre estados[10]

También se ha visto que la ecuación que regula el comportamiento de la cadena para cada estado i

$$p_i'(t) = -p_i(t)v_i + \sum_{j\neq i} p_j(t) \cdot \underbrace{v_j p_{ji}}_{q_{ji}}, \quad i = 0, 1, \ldots, K$$

viene determinada por la tasas de salida v_i y de transición entre estados q_{ij}, que están relacionadas.

De hecho, a partir de las tasas de transición es posible deducir los tiempos medios de estancia y probabilidades de transición:[11] para calcular el tiempo medio de estancia en un estado basta con saber las tasas de transición desde dicho estado

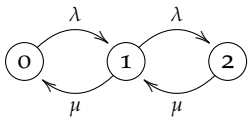
$$1/v_i = \frac{1}{\sum_{j\neq i} q_{ij}},$$

mientras que las probabilidades de transición se pueden obtener como

$$p_{ij} = \frac{q_{ij}}{v_i}.$$

Por lo tanto, basta con las tasas de transición entre estados q_{ij} para caracterizar completamente el comportamiento de una cadena de Markov, lo que simplifica la representación gráfica de las mismas.

Ejemplo 6.10. El caso de la gasolinera con un surtidor y hueco para dos coches queda completamente caracterizado con la siguiente cadena de Markov

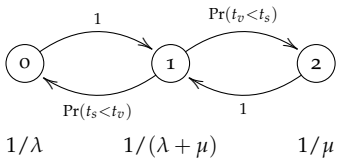


donde las flechas entre estados no representan las probabilidades de transición p_{ij} sino las tasas de transición q_{ij}, calculadas anteriormente y que constituyen el generador infinitesimal

$$Q = \begin{pmatrix} -\lambda & \lambda & 0 \\ \mu & -(\lambda+\mu) & \lambda \\ 0 & \mu & -\mu \end{pmatrix}$$

DE AHORA EN ADELANTE una cadena de Markov de tiempo continuo se representará con un diagrama como el anterior, donde cada estado se indique con un círculo y los números que acompañan a las flechas entre estados representan las tasas de transición entre estados q_{ij}.

Cadena de Markov para la gasolinera con un surtidor y hueco para dos coches:



[10] Que se conoce como cadena *embebida* y donde se cumple $p_{ii} = 0$.

[11] El procedimiento inverso se ha realizado anteriormente

Cálculo de la distribución de estado estacionaria

Al igual que en el caso de las cadenas de Markov de tiempo discreto, en las de tiempo continuo se puede deducir si la cadena tenderá a una distribución estacionaria. Son condiciones necesarias para que esto ocurra que la cadena sea *irreducible* y *recurrente*, esto es, que desde cualquier estado se pueda alcanzar cualquier otro estado, y que la esperanza del tiempo en volver a un estado (una vez que se abandona) sea finita.[12]

[12] En el caso de las cadenas de Markov de tiempo continuo no tiene sentido hablar de periodicidad.

Si una cadena de Markov de tiempo continuo es irreducible y recurrente,[13] existe una única distribución de probabilidades estacionaria. Dado que esta probabilidad es constante y no depende del instante de tiempo t, se debe cumplir que su derivada sea 0:

[13] Lo que se podría analizar a partir de la cadena de Markov embebida.

$$\frac{dp_i(t)}{dt} = 0, \text{ cuando } t \to \infty.$$

Además, dada esta independencia del tiempo, en la notación del vector de probabilidades de estado

$$\boldsymbol{p(t)} \triangleq [p_1(t), p_2(t), \ldots, p_K(t)]$$

se le puede suprimir la dependencia con t:

$$\boldsymbol{p} \triangleq [p_1, p_2, \ldots, p_K].$$

El cálculo de la distribución de probabilidades estacionaria se basa en este valor constante de $\boldsymbol{p}$. Según lo visto anteriormente, la ecuación (6.3) que rige el comportamiento de la probabilidad de estar en un estado i es la siguiente

$$p_i'(t) = -p_i(t)v_i + \sum_{j \neq i} p_j(t) \cdot \underbrace{v_j p_{ji}}_{q_{ji}}.$$

A partir de esta ecuación, igualando la derivada a 0 y obviando la dependencia con el tiempo, se obtiene que

$$p_i v_i = \sum_{j \neq i} p_j q_{ji}, \quad i = 0, 1, \ldots, K \tag{6.5}$$

lo que determina una propiedad de la cadena de Markov en el estado estacionario para todos los estados: la probabilidad de que la cadena esté en un estado multiplicada por su tasa de salida:

$$p_i v_i$$

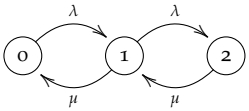
es igual a la suma de las probabilidades de que la cadena esté en los otros estados vecinos multiplicadas por las correspondientes tasas de llegadas:

$$\sum_{j \neq i} p_j q_{ji}$$

Se puede interpretar que se iguala el flujo de salida desde un estado con el flujo de entrada desde los demás estados, motivo por el que son conocidas como *ecuaciones de balance* o *ecuaciones de equilibrio*. De forma matricial (esto es, para todos los estados), las ecuaciones de balance (6.5) pueden escribirse como

$$0 = \boldsymbol{p} \cdot Q\,.$$

Ejemplo 6.11. En el caso de la gasolinera, aplicar las ecuaciones de balance resulta en el siguiente sistema de ecuaciones



$$\begin{aligned} p_0\lambda &= p_1\mu \\ p_1(\mu+\lambda) &= p_0\lambda + p_2\mu \\ p_2\mu &= p_1\lambda \end{aligned}$$

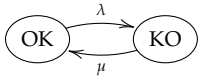
que resulta ser un sistema ecuaciones dependiente (recuérdese que las filas de la matriz Q suman 0).

DADO QUE APLICAR (6.5) NO BASTA para obtener un sistema que permita calcular las probabilidades estacionarias, se añade la condición de que la suma de las probabilidades debe valer 1

$$\sum_i p_i = 1. \tag{6.6}$$

De esta forma, en una cadena de Markov de K estados aplicando (6.5) se puede obtener un sistema de $K-1$ ecuaciones independientes, que se completa con (6.6) para obtener $\boldsymbol{p}$.

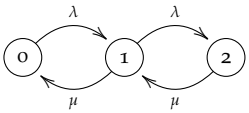
Ejemplo 6.12. Sea la máquina con tasa de rotura λ y de reparación μ. Aplicando (6.5) y (6.6) resulta el sistema de ecuaciones



$$\begin{aligned} 0 &= -p_{OK}\lambda + p_{KO}\mu \\ p_{OK} + p_{KO} &= 1 \end{aligned}$$

Con solución $p_{OK} = \frac{\mu}{\lambda+\mu}$ y, por lo tanto, $p_{KO} = \frac{\lambda}{\lambda+\mu}$.

Ejemplo 6.13. Para la gasolinera, se puede plantear el siguiente sistema de ecuaciones



$$\begin{aligned} p_0\lambda &= p_1\mu \\ p_2\mu &= p_1\lambda \\ p_0 + p_1 + p_2 &= 1 \end{aligned}$$

De las dos primeras ecuaciones se deduce que

$$p_1 = \frac{\lambda}{\mu}p_0, \quad p_2 = \frac{\lambda}{\mu}p_1 = \left(\frac{\lambda}{\mu}\right)^2 p_0$$

Aplicando estas igualdades en la tercera ecuación, resulta

$$p_0\left(1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2\right) = 1,$$

lo que permite calcular el valor del vector de probabilidades $\boldsymbol{p}$:

$$p_0 = \frac{1}{1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2} \quad p_1 = \frac{\frac{\lambda}{\mu}}{1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2} \quad p_2 = \frac{\left(\frac{\lambda}{\mu}\right)^2}{1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2}$$

Aplicación del Teorema de Little para calcular retardos

Una vez obtenido el vector de probabilidades en estado estacionario, en general resultará sencillo calcular los parámetros de interés del problema que se quiera resolver. Para el caso de redes de comunicaciones, en muchos casos uno de dichos parámetros será el tiempo medio de estancia en el sistema (a veces denominado retardo). Como se mencionó al presentar el Teorema de Little (página 70), dicho teorema permite relacionar dos variables relativamente sencillas de calcular (N y λ) con otra algo más difícil de obtener (T). Se ilustra su uso a continuación con un ejemplo.

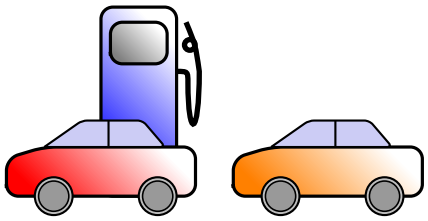
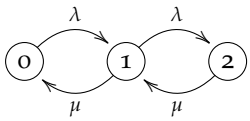


Figura 6.8: Gasolinera con un surtidor y sitio para dos coches.

Ejemplo 6.14. Sea una gasolinera con un surtidor y hasta dos coches, donde los coches llegan según un proceso de Poisson a tasa λ y el tiempo que permanecen ocupando el surtidor t_s sigue una variable aleatoria exponencial de media $1/\mu$. Como ya se ha visto, el sistema se puede modelar con la siguiente cadena, donde cada estado indica el número de coches en la gasolinera.



Si se denomina ρ al cociente entre λ/μ, se ha visto (Ejemplo 6.13) que la probabilidad de cada estado es:

$$p_0 = \frac{1}{1+\rho+\rho^2} \qquad p_1 = \frac{\rho}{1+\rho+\rho^2} \qquad p_2 = \frac{\rho^2}{1+\rho+\rho^2}$$

- Cálculo del retardo condicionando a cada estado

Dado que los vehículos llegan según un proceso de Poisson, la probabilidad de que una llegada vea la gasolinera llena (probabilidad condicionada a una llegada) coincide con la probabilidad de que el sistema se encuentre en p_2, por la propiedad PASTA (página 56). Considerando todas las opciones, un vehículo puede:

- Llegar con el sistema vacío (p_0) y ser atendido inmediatamente.
- Llegar cuando hay un vehículo siendo atendido (p_1), por lo que tiene que esperar.
- Llegar con la gasolinera llena (p_2), por lo que no entra en el sistema.

Sea T la variable aleatoria "retardo total que pasa un vehículo en la gasolinera," que sólo tiene sentido definir en los dos primeros casos. El cálculo de su esperanza se puede expresar como:

$$E[T] = E[T \mid \text{no esperar}]\,\Pr(\text{no esperar}) + E[T \mid \text{esperar}]\,\Pr(\text{esperar})$$

Donde el valor de las esperanzas condicionadas se puede deducir como:

- $E[T \mid \text{no esperar}]$ se trata, por definición, de $1/\mu$
- $E[T \mid \text{esperar}]$ en este caso, primero hay que esperar a que el coche que ya estaba termine su servicio ($1/\mu$) y añadir otro tiempo de servicio, por lo que sería $2/\mu$

Por otro lado, la probabilidad de que un coche no espere está implícitamente *condicionada* a que haya entrado en la gasolinera, dado que sólo en este caso tiene sentido definir T. Por lo tanto, las probabilidades de no esperar y de esperar son, respectivamente

$$\Pr(\text{no esperar}) = \frac{p_0}{p_0 + p_1} \qquad \Pr(\text{esperar}) = \frac{p_1}{p_0 + p_1}$$

Con todo lo anterior, se tiene que

$$E[T] = \frac{1}{\mu}\left(1 + \frac{\rho}{1+\rho}\right)$$

Por lo tanto, calcular el tiempo medio de estancia en un sistema $E[T]$ aún teniendo el vector de distribución de probabilidades (p_0, p_1, p_2) resulta algo complicado, dado que es preciso tener en cuenta cómo se encontraba el sistema en cada llegada.

- Cálculo del retardo mediante Little

Calcular el número medio de coches en el sistema es razonablemente sencillo:

$$E[N] = 0p_0 + 1p_1 + 2p_2 = \frac{\rho(1+2\rho)}{1+\rho+\rho^2}$$

Mientras que la tasa media de usuarios que entra al sistema también resulta sencillo, dado que en p_2 no se producen llegadas:

$$\hat{\lambda} = \lambda(p_0 + p_1) = \frac{\lambda(1+\rho)}{1+\rho+\rho^2}$$

De lo anterior, aplicando Little, el retardo se obtiene como

$$E[T] = \frac{E[N]}{\hat{\lambda}} = \frac{\rho(1+2\rho)}{\lambda(1+\rho)} = \frac{1}{\mu}\left(1 + \frac{\rho}{1+\rho}\right)$$

que coincide con el resultado anterior.

Tiempo medio de visita entre estados

De forma análoga al caso de las cadenas de Markov de tiempo discreto,[14] para calcular la esperanza del tiempo que pasa desde que la cadena está en un estado i hasta que alcanza un estado j se hace uso de la esperanza condicional.[15]

Sea m_{ij} la esperanza de dicho tiempo. Dado que la cadena permanece en el estado i en media un tiempo $1/v_i$ y pasa al estado j con probabilidad p_{ij}, por la ley de la probabilidad total se tiene que

$$m_{ij} = \frac{1}{v_i} + \sum_{k \neq i,j} p_{ik} \cdot m_{kj},$$

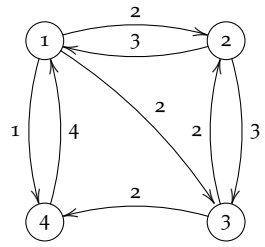
A partir de cualquiera de estas expresiones, para calcular la esperanza de los tiempos de visita entre estados es necesario plantear un sistema de ecuaciones, como se ilustra en el ejemplo a continuación.

[14] Tiempo medio de visita entre estados, página 99

[15] En el caso discreto, el número medio de iteraciones desde un estado s_i hasta otro s_j se podía expresar como

$$n_{ij} = 1 + \sum_{k \neq j} P_{ik} n_{ik}.$$

Ejemplo 6.15. Sea el caso de la cadena representada al margen, de la que se puede obtener la siguiente matriz de tasas de transición entre estados (recuérdese que $v_i = \sum_j q_{ij}$):



$$Q = \begin{pmatrix} -5 & 2 & 2 & 1 \\ 3 & -6 & 3 & 0 \\ 0 & 2 & -4 & 2 \\ 4 & 0 & 0 & -4 \end{pmatrix}$$

Con estos valores, para calcular m_{14} se puede plantear el siguiente sistema de ecuaciones

$$\begin{aligned} m_{14} &= \frac{1}{5} + \frac{2}{5}m_{24} + \frac{2}{5}m_{34} \\ m_{24} &= \frac{1}{6} + \frac{3}{6}m_{14} + \frac{3}{6}m_{34} \\ m_{34} &= \frac{1}{4} + \frac{2}{4}m_{24} \end{aligned}$$

del que se obtiene que $m_{14} = 8/9$.

Resumen del tema

- Una cadena de Markov de tiempo continuo homogénea es un proceso aleatorio $\{X(t)\}$ en un espacio de estados S, donde para cualquier instante de tiempo t e intervalo de tiempo s la probabilidad de estar en un estado en $t + s$ sólo depende del estado en que se encuentre en t.
- También se puede definir como un proceso aleatorio en el que el tiempo de permanencia en el estado i es exponencial y donde las probabilidades de transición entre estados son las dadas por una cadena de Markov de tiempo discreto (donde $p_{ii} = 0$).
- La evolución en el tiempo de la probabilidad de estar en un estado i viene dada por

$$p_i'(t) = -p_i(t)v_i + \sum_{j \neq i} p_j(t) \cdot v_j p_{ji},$$

- La distribución de probabilidades de estado estacionaria se calcula aplicando las expresiones

$$p_i v_i = \sum_{j \neq i} p_j q_{ji}, \quad \sum p_i = 1.$$

- El tiempo medio de visita entre estados se puede expresar como

$$m_{ij} = \frac{1}{v_i} + \sum_{k \neq i,j} p_{ik} \cdot m_{kj}.$$

Teoría de colas: sistemas básicos

Como se definió en un capítulo anterior, una cola es un sistema con uno o más recursos a compartir entre una población que realiza peticiones. Si una petición llega y el sistema tiene algún recurso disponible, será inmediatamente atendida (esto es, no tendrá que esperar). Si una petición llega y todos los recursos están ocupados, entonces no será atendida inmediatamente, y bien tendrá que esperar o será rechazada (en función de si hay espacio para esperar o no, respectivamente). Por lo tanto, el número de peticiones existentes será una variable crítica a la hora de modelar un sistema de este tipo.

En este capítulo se analiza un caso particular de este tipo de sistemas en el que el tiempo entre llegadas de peticiones sigue una variable aleatoria exponencial (esto es, el proceso de llegada de peticiones es de Poisson) y el tiempo de servicio de cada petición sigue otra variable aleatoria exponencial.[1] Por la propiedad "sin memoria" de la exponencial, estos sistemas se pueden modelar mediante una cadena de Markov de tiempo continuo donde el estado será el número de usuarios en el sistema. En cada estado sólo existirán a lo sumo dos posibles transiciones: bien se produce una nueva llegada, bien una petición ha sido atendida y abandona el sistema (este tipo de sistemas reciben el nombre de *procesos de nacimiento y muerte*).

[1] Según la Notación de Kendall $A/B/m/K$ (página 68) se trata de los sistemas de tipo $M/M/-/-$.

El sistema M/M/1

Se trata de un sistema en el que el tiempo entre llegadas es exponencial, de media $1/\lambda$, el tiempo de servicio también es exponencial, de media $1/\mu$, hay un único recurso para atender las peticiones, y la longitud de la cola no está limitada. El M/M/1 es el sistema más básico y puede servir para modelar sistemas sencillos con un único recurso atendiendo a una población variada, como, por ejemplo, un punto de acceso WiFi que transmita los datos de varios usuarios.

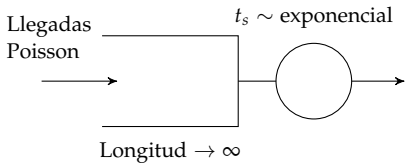
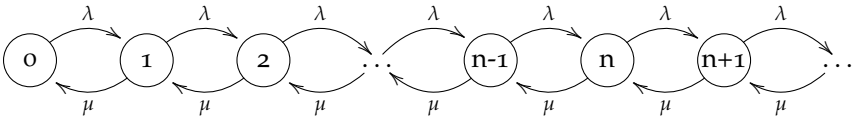


Figura 7.1: Sistema $M/M/1$.

Cálculo de la distribución de estado estacionario

El análisis de este sistema es parecido al caso de la estación de servicio con un único surtidor visto en el tema anterior, si bien ex-

tendido para cualquier número de usuarios. Se emplea el estado para indicar el número de usuarios en el sistema: en 0 el sistema se encuentra vacío, en 1 hay un usuario atendido en el recurso, en 2 además de un usuario atendido en el recurso hay otro esperando en la cola, etc. Sin embargo, dado que el número máximo de usuarios en el sistema no está limitado, la cadena de Markov tiene un número de estados infinito:



Para obtener la distribución de probabilidades de estado estacionaria de este sistema se aplican las expresiones vistas en el tema anterior (indicadas al margen), si bien con una estrategia diferente dado que el número de estados es infinito.

Ecuaciones para el cálculo de la distribución de estado estacionaria en una cadena de Markov:

$$p_i \cdot v_i = \sum_{j \neq i} p_j \cdot (v_j p_{ji})$$

$$\sum_i p_i = 1$$

Sea la ecuación de balance para el estado 0:

$$p_0 \lambda = p_1 \mu, \tag{7.1}$$

de la que se puede deducir que

$$p_1 = \frac{\lambda}{\mu} p_0 \tag{7.2}$$

Sea ahora la ecuación de balance para el estado 1:

$$p_1(\lambda + \mu) = p_0 \lambda + p_2 \mu \tag{7.3}$$

substituyendo (7.1) en la parte derecha de (7.3), se tiene que

$$p_1(\lambda + \mu) = p_1 \mu + p_2 \mu$$

de la que se puede despejar p_2:

$$p_2 = \frac{\lambda}{\mu} p_1,$$

y, aplicando (7.2), resulta

$$p_2 = \left(\frac{\lambda}{\mu}\right)^2 p_0$$

Se puede proceder de forma similar para el estado 2, de lo que se obtendría

$$p_3 = \frac{\lambda}{\mu} p_2 = \left(\frac{\lambda}{\mu}\right)^3 p_0$$

y, en general, se puede deducir que

$$p_{n-1} \lambda = p_n \mu, \tag{7.4}$$

lo que puede interpretarse como que la cadena está "balanceada" entre cada par de estados (Figura 7.2, al margen) dado que la tasa efectiva con la que se incrementa el número de usuarios ($p_n \lambda$) es igual a la tasa con la que disminuye ($p_{n+1} \mu$).

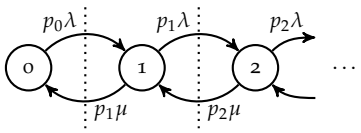


Figura 7.2: Equilibrio entre las tasas de incremento y disminución de estados.

A partir de esta expresión, se deduce:

$$p_n = \frac{\lambda}{\mu} p_{n-1} = \left(\frac{\lambda}{\mu}\right)^n p_0,$$

que puede expresarse en función del cociente $\rho = \frac{\lambda}{\mu}$ como

$$p_n = \rho^n p_0.$$

Para calcular la distribución de probabilidades en el estado estacionario, se aplica la anterior relación para todo n en la ecuación de normalización

$$\sum_{n=0}^{\infty} p_n = 1,$$

lo que da lugar a la siguiente ecuación en p_0:

$$\sum_{n=0}^{\infty} \rho^n p_0 = 1,$$

con solución

$$p_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1}.$$

Dado que se trata de una serie geométrica, para que p_0 tenga solución es preciso que $\rho < 1$. De esta forma, se tiene que[2]

$$p_n = \begin{cases} 1-\rho & \text{Para } n = 0 \\ \rho^n(1-\rho) & \text{Para } n \geq 0 \end{cases}$$

donde se hace explícito que la probabilidad de que el sistema esté vacío p_0 es igual a $1-\rho$, esto es, que la ocupación media del recurso ρ coincide con la probabilidad de que haya al menos un usuario en el sistema.

[2] Expresión para la suma infinita:

$$\sum_{i=0}^{\infty} \rho^i = \frac{1}{1-\rho}$$

Ejemplo 7.1. Sea un cajero al que llegan clientes según una tasa de 10 clientes/hora. El tiempo medio que un cliente pasa en el cajero se puede modelar como una variable aleatoria exponencial de media cuatro minutos. Se pide: calcular la probabilidad de que el cajero esté vacío, y de que haya más de un cliente esperando.

Según los datos, se tiene que:

$$\rho = \frac{\text{10 clientes/60 minutos}}{\text{1 cliente/4 minutos}} = \frac{10}{15} = 2/3.$$

La probabilidad de que el cajero esté vacío se puede calcular como

$$p_0 = 1 - \rho = 1/3$$

mientras que la probabilidad de que haya más de un cliente esperando se corresponde con la probabilidad de que haya más de dos usuarios en total. Dicha probabilidad se puede calcular como

$$\begin{aligned} \Pr(\text{Más de un cliente esperando}) &= \sum_{n>2}^{\infty} p_n = \\ &= 1 - p_0 - p_1 - p_2 \\ &= 1 - (1-\rho) - (1-\rho)\rho - (1-\rho)\rho^2 = \frac{8}{27} \end{aligned}$$

Sobre el significado de ρ

El parámetro ρ se define como el cociente entre la tasa de llegadas al sistema y la tasa máxima de salida que proporciona el servidor

$$\rho = \frac{\lambda}{\mu} = \frac{\text{tasa llegada}}{\text{tasa salida}},$$

lo que también puede interpretarse como el tiempo medio de servicio entre el tiempo medio entre llegadas

$$\rho = \frac{1/\mu}{1/\lambda} = \frac{\text{tiempo servicio}}{\text{tiempo entre llegadas}}.$$

Por lo tanto, el hecho de que $\rho < 1$ para que el sistema tenga una solución resulta muy coherente: si $\rho > 1$, se tendría que la tasa media de llegada al sistema es mayor que la tasa máxima de salida (o que el tiempo medio para servir un usuario es mayor que el tiempo medio entre llegadas), por lo que el sistema no sería capaz de cursar la carga de trabajo que se le presenta. Dado que se trata de una cola $M/M/1/\infty$, la longitud de la cola crecería de forma indefinida, por lo que el número de usuarios en el sistema tendería a infinito.

De esta forma, en un sistema sin rechazo donde el número máximo de usuarios no está ilimitado, si la tasa de llegadas es mayor que la tasa agregada de salida, el sistema se encontrará *congestionado* por lo que no tiene sentido realizar un modelado de sus prestaciones. Cuando el sistema sí rechaza usuarios (por ejemplo, el caso de la gasolinera del Ejemplo 6.4, página 107, que admitía un número máximo de coches) la condición $\rho < 1$ no resulta necesaria, dado que el número de estados de la cadena de Markov es finito.

Análisis de prestaciones básicas del sistema

Una vez obtenida la distribución de probabilidades del sistema, se puede obtener cualquiera de las diferentes variables de interés en una cola. Para el M/M/1 resulta sencillo calcular, por ejemplo, el número medio de usuarios en el sistema mediante[3]

$$N = \sum_{n=0}^{\infty} n p_n = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = \frac{\rho}{1-\rho}$$

[3] Expresión para la suma infinita

$$\sum_{n=0}^{\infty} n\rho^{n-1} = \frac{1}{(1-\rho)^2}$$

Una vez calculado N, el teorema de Little permite obtener

$$T = \frac{N}{\lambda} = \frac{1/\mu}{1-\rho}$$

Ejemplo 7.2. Sea el caso del cajero del Ejemplo 7.1, donde se tenía que $\rho = 2/3$. El número medio de usuarios en el sistema se puede calcular como

$$N = \frac{2/3}{1-2/3} = 2 \text{ usuarios.}$$

Mientras que el tiempo medio de estancia en el sistema es

$$T = \frac{t_s}{1-\rho} = \frac{4}{1/3} = 12 \text{ minutos.}$$

Resulta interesante analizar la *sensibilidad* de T respecto a cambios en ρ, esto es, cómo varía relativamente el retardo medio en el sistema respecto a cambios relativos en la carga ofrecida. Si la tasa de usuarios aumenta un 20 % (esto es, λ pasa de 10 a 12 clientes/hora), se tiene que la nueva carga es

$$\rho = 12/15 = 4/5,$$

por lo que el tiempo medio de estancia en el sistema pasa a ser

$$T = \frac{t_s}{1-\rho} = \frac{4}{1/5} = 20 \text{ minutos.}$$

Por lo tanto, un 20 % de aumento en la carga ha supuesto un incremento del 20/12≈1.66, es decir, del 66 % en el retardo en el sistema.

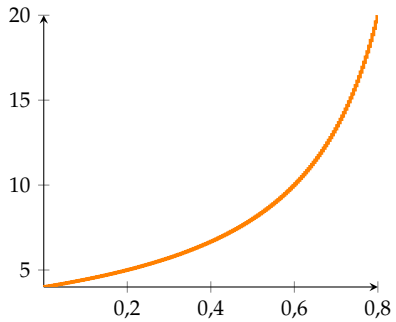


Figura 7.3: Retardo medio T vs. ρ en un sistema M/M/1 con $t_s = 4$.

ESTA ALTA SENSIBILIDAD DEL RETARDO a variaciones de la carga se debe a la forma de la expresión para T:

$$T = \frac{t_s}{1-\rho},$$

que presenta un comportamiento asintótico cuando $\rho \to 1$. De esta forma, cuando $\rho \to 0$ se tiene que $T \approx t_s$, dado que el sistema está casi siempre vacío, mientras que conforme aumenta ρ, en mayor medida aumenta el retardo en el sistema. Esto se debe a la "realimentación" en el mismo: cuando aumenta la carga, aumenta la probabilidad de que un usuario tenga que esperar en el sistema, lo que a su vez aumenta la probabilidad de que otro usuario llegue mientras aquél se encuentra esperando.

El cálculo del número medio de usuarios en la cola Q puede hacerse de forma análoga al del número medio de usuarios en el sistema, basta con tener en cuenta que, por ejemplo, con tres usuarios en el sistema, son dos los usuarios que están en la cola. Por lo tanto, Q se puede obtener como

$$Q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} n\cdot p_n - \sum_{n=1}^{\infty} p_n = N - (1-p_0) = N - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

A partir de Q, se puede obtener W mediante Little como

$$W = \frac{Q}{\lambda} = \frac{\rho 1/\mu}{1-\rho}$$

Otra forma de calcular W (y Q) pasa por resolver primero

$$W = T - t_s = \frac{t_s}{1-\rho} - t_s,$$

y a partir de este resultado, calcular $Q = \lambda \cdot W$.

Ejemplo 7.3. Sea una línea de transmisión a 100 Mbps, a la que llegan tramas de una longitud que puede modelarse con una variable aleatoria exponencial de media 1500 octetos. El tiempo entre tramas es otra variable aleatoria exponencial de media 0.18 ms. Bajo estas

suposiciones, se quiere calcular el retardo medio por trama y el tiempo medio de espera en cola.

Dado que las tramas tienen una longitud exponencial[4] y la velocidad de la línea es constante, el tiempo de transmisión de cada trama será también una variable aleatoria exponencial, de media

$$t_s = 1500\ B\ \cdot 8/100\ \text{Mbps} = 120\ \mu s,$$

y el sistema se puede modelar como un M/M/1. La ocupación del mismo se puede obtener como

$$\rho = \frac{t_s}{1/\lambda} = \frac{120}{180} = 2/3,$$

siendo el retardo medio

$$T = \frac{t_s}{1-\rho} = 360\ \mu s.$$

Dado que el tiempo de servicio es 120 μs, el tiempo medio de espera en cola será

$$W = 360 - 120 = 240\ \mu s.$$

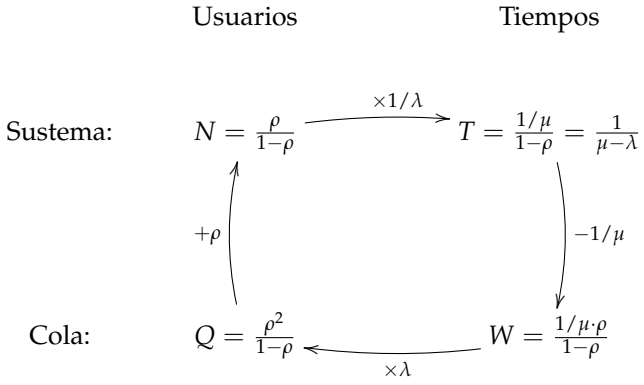
Dicho tiempo medio de espera en cola también se puede obtener como:

$$W = \frac{\rho\ t_s}{1-\rho} = \frac{(2/3)120\ \mu s}{1-2/3} = 240\ \mu s\ \ .$$

[4] Suponer que la longitud de una trama (un número entero) se puede modelar con una variable aleatoria continua será menos exacto a medida que la longitud de las tramas sea menor y, por tanto, mayor sea el error al redondear al siguiente entero.

Relación entre N, T, W y Q

A continuación se presenta un resumen de las expresiones para las prestaciones básicas de un sistema M/M/1. Dada una de ellas, resulta inmediato obtener cualquiera de las otras, bien aplicando el teorema de Little, bien por las expresiones básicas que relacionan las variables en una cola.



Funciones de distribución del retardo F_W y F_T (*)

Para el caso de un $M/M/1$ no resulta demasiado complicado calcular la función de distribución del retardo en cola $F_W(t)$, esto

es, la probabilidad de que el tiempo que tenga que esperar una petición hasta acceder al recurso sea menor o igual que t:

$$F_W(t) \triangleq \Pr(W \leq t).$$

Para realizar este cálculo, se aplica el teorema de la probabilidad total sobre el número de usuarios que se encontraban en el sistema al llegar una petición:

$$F_W(t) = \sum_{n=0}^{\infty} \Pr(W \leq t | N = n) \Pr(N = n). \tag{7.5}$$

En la expresión (7.5), $\Pr(W \leq t|N = n)$ es la probabilidad de que el tiempo de espera en cola W sea menor que t cuando hay n usuarios al llegar. En estas condiciones, W es la suma de los n tiempos de servicio de los usuarios por delante, por lo que su función de densidad es la convolución de n variables aleatorias exponenciales de media $1/\mu$ (una variable aleatoria de Erlang). Substituyendo el valor de $\Pr(N = n)$ y separando el caso $n = 0$, se tiene que (7.5) se puede desarrollar como[5]

[5] Cuando hay o usuarios al llegar al sistema, $W = 0$ por lo que $\Pr(W \leq t)$, esto es, $F_W(t)$, es igual a 1.

$$F_W(t) = (1 - \rho) + (1 - \rho) \sum_{n=1}^{\infty} \rho^n \int_0^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x} dx,$$

que se puede simplificar como

$$F_W(t) = (1 - \rho) + \rho \int_0^t \mu(1 - \rho) e^{-\mu(1-\rho)x} dx. \tag{7.6}$$

En esta expresión, el término de la integral se puede identificar con una variable aleatoria exponencial, de media $(\mu(1 - \rho))^{-1}$. Por lo tanto, (7.6) queda como

$$F_W(t) = 1 - \rho e^{-(\mu-\lambda)t}.$$

Siguiendo un razonamiento similar, se puede obtener que la función de distribución del tiempo total en un sistema $M/M/1$ viene dada por

$$F_T(t) = 1 - e^{-(\mu-\lambda)t} \ ,$$

esto es, una variable aleatoria exponencial de media $(\mu - \lambda)^{-1}$.

Ejemplo 7.4. Volviendo al ejemplo 7.1 del cajero, donde los usuarios llegan a una tasa de $\lambda = 10$ clientes/hora y el tiempo medio de servicio es $t_s = 4$ minutos. La carga del sistema y el retardo medio son, respectivamente,

$$\rho = \frac{2}{3} \quad \text{y T=12 minutos} \ ,$$

por lo que el tiempo medio de espera en cola es W=12-4=8 minutos. La probabilidad de que un usuario esté más de t unidades de tiempo esperando a usar el cajero vendrá dada por

$$1 - F_W(t) = \rho e^{-(\mu-\lambda)t} \ ,$$

por lo que la probabilidad de estar esperando más de 10 minutos será

$$\frac{2}{3} \cdot e^{-(\frac{1}{4}-\frac{1}{6})10} = \frac{2}{3} \cdot e^{-\frac{5}{6}} \approx 0,289 \ .$$

De forma análoga, la probabilidad de estar en el cajero más de 10 minutos se puede calcular como

$$1 - F_T(t) = e^{-(\mu-\lambda)t} = e^{-(\frac{1}{4}-\frac{1}{6})10} = e^{-\frac{5}{6}} \approx 0,435 \ .$$

Proceso de salida de un M/M/1 (*)

Sea un sistema M/M/1, con los tiempos entre llegadas y de servicio distribuidos según variables aleatorias exponenciales, de medias λ^{-1} y μ^{-1}, respectivamente, y –por lo tanto– con una probabilidad de ocupación $\rho = \lambda/\mu$. A continuación se analiza el proceso de salida de los usuarios de dicho sistema (una vez que han sido servidos).

En un instante t al azar el sistema M/M/1 puede estar libre, con probabilidad $p_0 = 1 - \rho$, u ocupado, con probabilidad ρ. Gracias a la ley de la probabilidad total, la función de distribución del tiempo hasta la siguiente salida se puede expresar como

$$F(t) = (1-\rho)F_{1-\rho}(t) + \rho F_\rho(t)$$

donde $F_{1-\rho}(t)$ es la función de distribución del tiempo hasta la siguiente salida cuando el sistema está vacío, y $F_p(t)$ la función de distribución del tiempo hasta la siguiente salida cuando el sistema está ocupado (es decir, hay al menos un usuario en el recurso).

Cuando el sistema está ocupado, el tiempo hasta la siguiente salida es una variable aleatoria exponencial de media $1/\mu$, por lo que la función de distribución es

$$F_\rho(t) = 1 - e^{-\mu t}.$$

Cuando el sistema está libre, el tiempo hasta la siguiente salida es la suma de dos variables aleatorias exponenciales: el tiempo hasta la siguiente llegada más su tiempo de servicio. La función de distribución de dicha suma es[6]

$$F_{1-\rho}(t) = 1 - \frac{\mu}{\mu-\lambda}e^{-\lambda t} + \frac{\lambda}{\mu-\lambda}e^{-\mu t}$$

[6] La función de densidad de la suma de dos variables aleatorias independientes $f(x)$ y $g(x)$ es su convolución $(f * g)(z) = \int_{-\infty}^{\infty} f(z-y)g(y)dy$.

Por lo tanto, la función de distribución del tiempo entre salidas en el sistema resulta ser

$$F(t) = (1-\rho)\left(1 - \frac{\mu}{\mu-\lambda}e^{-\lambda t} + \frac{\lambda}{\mu-\lambda}e^{-\mu t}\right) + \rho\left(1 - e^{-\mu t}\right),$$

que, tras simplificar, queda como

$$F(t) = 1 - e^{-\lambda t}$$

Por lo tanto, el tiempo entre salidas en un sistema M/M/1 estable ($\rho < 1$) se distribuye según una variable aleatoria exponencial

de media λ^{-1}, al igual que el tiempo entre llegadas.[7] El hecho de que el proceso de salida sea a la misma tasa que el de entrada no debe resultar sorprendente, dado que el sistema ni crea ni rechaza usuarios, sólo los retarda. Sí que resulta destacable que dicho tiempo sea también una variable aleatoria exponencial, por lo que el proceso de salida será de nuevo de Poisson.

[7] Nótese que no se ha demostrado que el tiempo entre una salida y la siguiente sean variables aleatorias independientes.

Ejemplo 7.5. Si el tiempo de servicio no es exponencial el proceso de salida no será de Poisson. Para ilustrar esto, sea un sistema como el de la Figura 7.4 a continuación, con un tiempo de servicio constante t_s. Dado que las llegadas son de Poisson, el tiempo entre llegadas será una variable aleatoria exponencial distribuida en $[0, \infty)$. Sin embargo, tras su paso por el sistema, el tiempo entre salidas pasará a estar distribuido en $[t_s, \infty)$, por lo que es seguro que dicho proceso de salida no será de Poisson (aunque sea un proceso a la misma tasa que el de entrada).

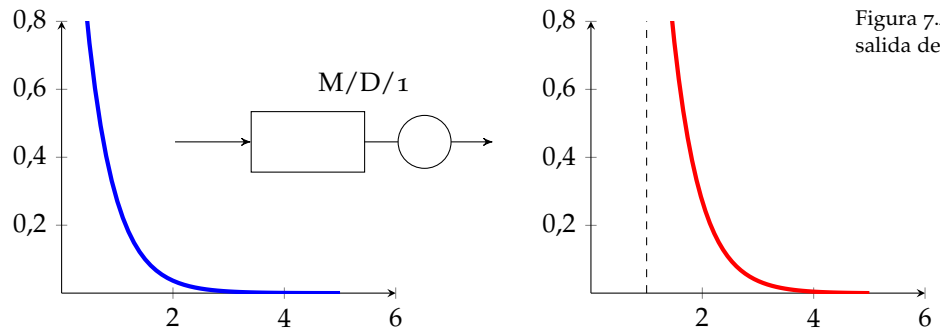


Figura 7.4: Tiempo entre llegadas y a la salida de un sistema $M/D/1$

El sistema M/M/m

Al igual que el M/M/1, se trata de un sistema con llegadas de Poisson y tiempos de servicio que siguen una variable aleatoria exponencial, pero en este caso hay m servidores idénticos en paralelo. Si hay varios servidores disponibles, un usuario que llegue al sistema será atendido por uno de ellos (no importa el que sea), mientras que si todos están ocupados, el usuario tendrá que esperar en cola hasta que uno quede libre. Se trata de un sistema que podría emplearse para modelar, por ejemplo, centros de atención telefónica (*call centers*), una granja de servidores que resuelven peticiones web, o un supermercado con varias cajas disponibles y una única línea de espera.

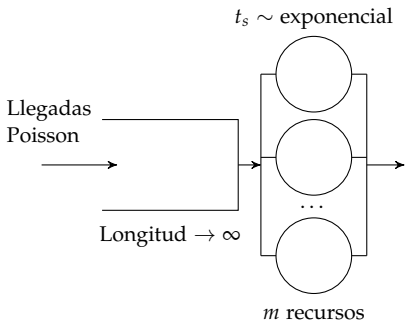


Figura 7.5: Sistema M/M/m.

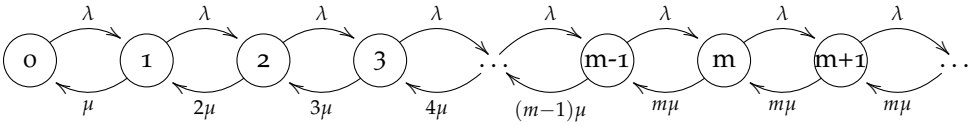
La cadena de Markov que modela el comportamiento será parecida a la del M/M/1, con la tasa de transición desde el estado n al $n+1$ siempre igual a λ. Sin embargo, la tasa de transición cuando se reduce el número de usuarios n en el sistema es diferente y depende de si n es mayor o menor que m:

- Cuando $n \leq m$, no hay usuarios en la cola sino que todos están

siendo atendidos. En estas condiciones, la tasa de transición al estado $n-1$ es $n \cdot \mu$.

- Cuando $n > m$, hay usuarios esperando en cola porque todos los recursos están ocupados, por lo que la tasa de transición será $m \cdot \mu$.

La cadena queda, por tanto:



La forma en la que se calcula la distribución de probabilidades estacionaria es muy parecida al caso anterior, aunque más laboriosa. Por una parte, se tiene que para los estados $n \leq m$, las relaciones son de tipo

$$p_1 = \frac{\lambda}{\mu} p_0, \quad p_2 = \frac{\lambda}{2\mu} p_1, \quad p_3 = \frac{\lambda}{3\mu} p_2, \quad p_4 = \frac{\lambda}{4\mu} p_3, \quad \ldots$$

por lo que la expresión para dichos estados queda

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \cdot p_0, \quad n < m \tag{7.7}$$

Por otra parte, para los estados $n > m$, las relaciones son

$$p_{m+1} = \frac{\lambda}{m\mu} p_m, \quad p_{m+2} = \frac{\lambda}{m\mu} p_{m+1}, \quad \ldots$$

por lo que la expresión en este caso es

$$p_n = \left(\frac{\lambda}{m\mu}\right)^{n-m} p_m, \quad n \geq m$$

que, sustituyendo p_m por su valor según (7.7), queda como

$$p_n = \left(\frac{\lambda}{m\mu}\right)^{n-m} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} p_0 = \left(\frac{\lambda}{m\mu}\right)^n \frac{m^m}{m!} p_0, \quad n \geq m \tag{7.8}$$

Antes de continuar con el análisis de la Cadena de Markov, conviene detenerse en los dos siguientes términos que aparecen en (7.7) y (7.8):

$$\rho \triangleq \frac{\lambda}{m\mu}, \quad I \triangleq \frac{\lambda}{\mu}$$

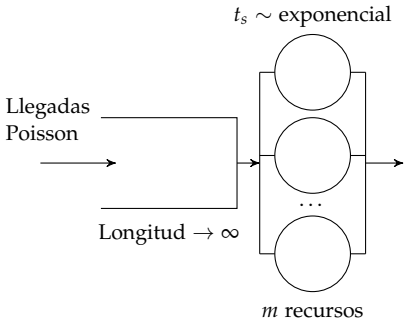
que, obviamente, se relacionan mediante la expresión $I = m\rho$.

Los términos ρ e I

Por una parte, el término ρ se define como el cociente entre la tasa de entrada λ y la tasa máxima de salida con todos los servidores ocupados $m\mu$. Por lo tanto, dado que la capacidad de la cola es infinita, para que se pueda obtener una solución del sistema será necesario que se cumpla (de forma similar al sistema M/M/1)

$$\lambda < m\mu \rightarrow \rho < 1 \, .$$

Además, el término ρ admite la misma interpretación que en el caso del M/M/1, como valor de ocupación de un recurso. Esto puede verse interpretando los m recursos en paralelo como un sistema al que acceden usuarios a una tasa λ y pasan un tiempo $t_s = 1/\mu$. Aplicando Little, el número medio de usuarios en el sistema será



$$N = \lambda t_s = \lambda/\mu = m\rho \ ,$$

por lo que ρ coincide con el porcentaje de tiempo que un servidor está ocupado.

De hecho, este número medio de servidores ocupados $m\rho$ coincide con la expresión para el término I. Por lo tanto, I podría interpretarse como el mínimo número de servidores que serían necesarios para atender a una demanda de tráfico λ, si el tiempo medio de servicio es $1/\mu$.

Cálculo de p_n y probabilidad de espera en cola

Como se ha visto anteriormente, las probabilidades de cada estado se pueden expresar en función de p_0:

$$p_n = \begin{cases} \dfrac{I^n}{n!} \cdot p_0 & \text{Si } n \leq m \\ \rho^n \dfrac{m^m}{m!} \cdot p_0 & \text{Si } n \geq m \end{cases}$$

Para calcular p_0 se aplica la ecuación $\sum p_n = 1$, de lo que resulta

$$p_0 \left(\sum_{n=0}^{m-1} \frac{I^n}{n!} + \sum_{n=m}^{\infty} \rho^n \frac{m^m}{m!} \right) = 1$$

lo que lleva al siguiente valor de la probabilidad de que el sistema se encuentre vacío:

$$p_0 = \left(\left(\sum_{n=0}^{m-1} \frac{I^n}{n!} \right) + \frac{I^m}{m!} \cdot \frac{1}{1-\rho} \right)^{-1} .$$

A partir de las anteriores expresiones es se puede calcular la probabilidad de que un usuario tenga que esperar, esto es, la probabilidad condicionada de que, al llegar al sistema, haya m o más usuarios: por la propiedad PASTA (página 56), esta probabilidad coincide con la probabilidad de que el sistema se encuentre en los correspondientes estados p_m, p_{m+1}, p_{m+2}, etc.

Por lo tanto, la probabilidad de esperar en cola P_Q se puede expresar como[8]

$$P_Q = \sum_{n=m}^{\infty} p_m$$

y puede obtenerse como

$$P_Q = \sum_{n=m}^{\infty} \frac{\rho^n}{m!} m^m p_0 = \frac{m^m}{m!} p_0 \sum_{n=m}^{\infty} \rho^n = \frac{I^m}{m!} \frac{1}{1-\rho} p_0$$

[8] Para el caso de p.ej. un *call center*, dicha probabilidad se trata de una métrica de "satisfacción" (o su ausencia) de los usuarios.

La expresión para calcular P_Q recibe el nombre de *Erlang-C*:

$$P_Q = E_c(m, I) = \frac{\frac{I^m}{m!}\frac{1}{1-\rho}}{\left(\sum_{n=0}^{m-1}\frac{I^n}{n!}\right) + \frac{I^m}{m!}\cdot\frac{1}{1-\rho}} = \frac{\frac{I^m}{m!}}{(1-\rho)\left(\sum_{n=0}^{m-1}\frac{I^n}{n!}\right) + \frac{I^m}{m!}} \tag{7.9}$$

Ejemplo 7.6. Suponga un servicio telefónico de atención al cliente con dos operarios. Las llamadas tienen una duración que se puede modelar con una variable aleatoria exponencial de media 5 minutos y se estima que hay una consulta cada 10 minutos, que se produce según un proceso de Poisson. En estas condiciones,

$$I = \frac{\lambda}{\mu} = \frac{1/10}{1/5} = 1/2 \quad \text{y} \quad \rho = \frac{\lambda}{m\mu} = \frac{1/10}{2/5} = 1/4 \ .$$

La probabilidad de que un usuario tenga que esperar para ser atendido viene dada por la expresión (7.9):

$$E_c(2, 1/2) = \frac{\frac{(1/2)^2}{2!}\frac{1}{1-1/4}}{\left(1 + \frac{1}{2} + \frac{(1/2)^2}{2!}\cdot\frac{1}{1-(1/4)}\right)} = \frac{1/8\cdot 4/3}{5/3} = \frac{1}{10} = 0,1 \ .$$

Por lo tanto, el 10 % de las llamadas tendrán que esperar antes de ser atendida. Si se considera que dicha proporción es muy alta, será preciso aumentar el número de operarios m.

Con $m = 3$ se tiene la misma I pero

$$\rho = \frac{\lambda}{m\mu} = \frac{1/10}{3/5} = 1/6 \ .$$

La probabilidad de esperar pasa a ser

$$E_c(3, 1/2) = \frac{\frac{(1/2)^3}{3!}\frac{1}{1-1/6}}{\left(1 + \frac{1}{2} + \frac{(1/2)^2}{2!} + \frac{(1/2)^3}{3!}\cdot\frac{1}{1-(1/6)}\right)} = \frac{1}{66} \approx 0,01515 \ ,$$

esto es, con un operario más la proporción de llamadas que tienen que esperar baja al 1,5 %.

Nótese que la Erlang-C, como se ilustra en el ejemplo anterior, es una ecuación *de análisis pero no de diseño*: fijado un escenario en términos de λ, μ y m dicha expresión permite calcular p_m, pero no es posible despejar el mínimo valor de m que garantiza unas prestaciones. Para ello, es preciso emplear un método de "prueba y error" o acudir a tablas precalculadas u otras herramientas numéricas.

Análisis de prestaciones básicas del sistema

En el caso del M/M/m, para obtener el rendimiento del sistema (a través de los parámetros N, T, W y Q) resulta más sencillo calcular en primer lugar el número medio de usuarios en cola Q.[9] Para realizar dicho cálculo, hay que tener en cuenta que hasta el estado $m+1$ no hay usuarios en cola. Por lo tanto,

$$Q = \sum_{n=m}^{\infty}(n-m)p_n = \sum_{n=m}^{\infty}(n-m)\rho^n\frac{m^m}{m!}p_0 = \rho^m\frac{m^m}{m!}p_0\sum_{n=m}^{\infty}(n-m)\rho^{n-m}$$

[9] En el caso del M/M/1, el análisis comienza con el número medio de usuarios en el sistema:

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

lo que lleva a

$$Q = \frac{I^m}{m!}\frac{\rho}{(1-\rho)^2}p_0 \tag{7.10}$$

que puede expresarse como (sustituyendo el término p_0)[10]

$$Q = \frac{\frac{I^m}{m!}\frac{\rho}{(1-\rho)^2}}{\left(\sum_{n=0}^{m-1}\frac{I^n}{n!}\right) + \frac{I^m}{m!}\cdot\frac{1}{1-\rho}}$$

[10] El valor de p_0 viene dado por:

$$p_0 = \left(\left(\sum_{n=0}^{m-1}\frac{I^n}{n!}\right) + \frac{I^m}{m!}\cdot\frac{1}{1-\rho}\right)^{-1}.$$

Una vez obtenido Q, el tiempo medio de espera en cola se calcula mediante Little

$$W = \frac{Q}{\lambda},$$

mientras que el tiempo medio de estancia en el sistema y número medio de usuarios en el sistema se obtiene de

$$T = W + \frac{1}{\mu}, \quad N = \lambda \cdot T = Q + m\rho$$

Ejemplo 7.7. Sea un sistema M/M/2 con

$$\begin{aligned}\lambda &= 2 \text{ usuarios/segundo}\\ \mu &= 2 \text{ usuarios/segundo}\end{aligned}$$

Con estos valores,

$$I = \frac{\lambda}{\mu} = 1, \quad \rho = \frac{\rho}{m} = \frac{1}{2}$$

Para calcular Q con la expresión (7.10), primero se calcula p_0 como

$$p_0 = \left(\left(\sum_{n=0}^{m-1}\frac{I^n}{n!}\right) + \frac{I^m}{m!}\cdot\frac{1}{1-\rho}\right)^{-1} = \left(1+1+\frac{1}{2}\cdot\frac{1}{1-1/2}\right)^{-1} = 1/3$$

El número medio de usuarios en cola se obtiene como

$$Q = \frac{I^m\cdot\rho}{m!(1-\rho)^2}p_0 = \frac{1\cdot 1/2}{2!(1-1/2)^2}\cdot\frac{1}{3} = 1/3\ ,$$

y, a partir de este valor,

$$W = \frac{Q}{\lambda} = \frac{1}{6} \text{ segundos}, \quad \text{y } T = W + \frac{1}{\mu} = \frac{2}{3} \text{ segundos}\ .$$

Por último, la probabilidad de que un usuario no tenga que esperar (1-P_Q) coincide con la probabilidad de que al menos haya un recurso disponible al llegar, esto es:

$$\Pr(\text{No esperar}) = \sum_{n=0}^{m-1} = p_0 + p_1 = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}\ .$$

Relación entre P_Q y Q (*)

Resulta destacable que las expresiones para P_Q y Q, dadas por (7.9) y (7.10), guardan la siguiente relación[11]

$$Q = \frac{\rho}{1-\rho} P_Q$$

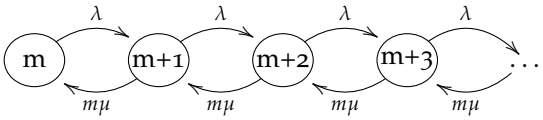
Dicha relación puede deducirse aplicando la esperanza condicionada para el cálculo de Q, según se tenga que esperar o no al sistema:

$$Q = \mathbb{E}[Q] = \mathbb{E}[Q \mid \text{Esperar}] P_Q + \mathbb{E}[Q \mid \text{No esperar}](1 - P_Q)$$

donde, por definición, $\mathbb{E}[Q \mid \text{No esperar}] = 0$ por lo que queda

$$\mathbb{E}[Q] = \mathbb{E}[Q \mid \text{Esperar}] P_Q$$

El cálculo de $\mathbb{E}[Q \mid \text{Esperar}]$ se corresponde con el número medio de usuarios en cola que "ve" un usuario que llega al sistema y tiene que esperar, esto es, que llega al estado m o superior. Considerando únicamente estos estados, la cadena de Markov resultante es



lo que se correspondería con un sistema M/M/1 (donde los estados empiezan a contar en m), con tasa de llegada λ y tasa de servicio $m\mu$. El número medio de usuarios en cola para el M/M/m, por lo tanto, se corresponde con el número medio de usuarios en el sistema M/M/1

$$\mathbb{E}[Q \mid \text{Esperar}] = N_{\text{M/M/1}} = \frac{\rho}{1-\rho}$$

de lo que se deduce la relación entre Q y P_Q.

Prestaciones de un M/M/m vs. un M/M/1 con la misma capacidad

Resulta interesante comparar las prestaciones de un sistema M/M/m donde cada servidor tiene una capacidad μ con las de un sistema M/M/1 con un servidor con capacidad $m \cdot \mu$, dado que en ambos casos la capacidad máxima de procesado es la misma, pero en el primero ésta se encuentra dividida entre varios servidores mientras que en el segundo se encuentra agregada en un único servidor.

Ejemplo 7.8. Sea ahora un sistema M/M/1 con $\lambda = 2$ usuarios/segundo y $1/\mu = 1/4$ segundo, esto es, con la misma tasa de entrada que en el ejemplo 7.7 (un sistema sistema M/M/2) pero la mitad del tiempo de servicio. Puede interpretarse que ambos sistemas tienen la misma "capacidad" dado que la μ del M/M/1 es el doble que la μ de cada uno de los servidores del M/M/2.

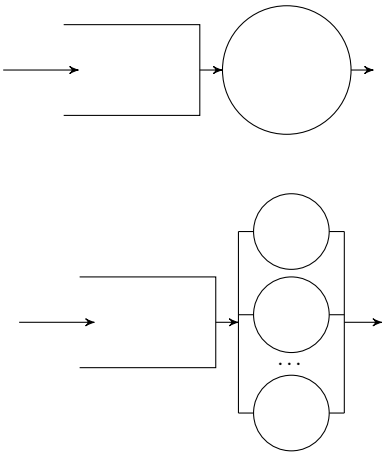


Figura 7.6: Comparación M/M/1 con M/M/m de la misma capacidad.

[11] Por comodidad, se repiten las expresiones aquí:

$$P_Q = \frac{I^m}{m!} \frac{1}{1-\rho} p_0$$

$$Q = \frac{I^m}{m!} p_0 \frac{\rho}{(1-\rho)^2}$$

La ocupación del M/M/1 viene dada por

$$\rho = \frac{\lambda}{\mu} = \frac{1}{2},$$

por lo que el retardo total en el sistema resulta

$$T = \frac{1/\mu}{1-\rho} = 1/2 \text{ segundo},$$

y, a partir de este retardo

$$W = T - t_s = \frac{1}{4} \text{ segundo}.$$

Se resumen los resultados de retardo para este sistema y el anterior en la Tabla 7.10, donde se aprecia que si bien en el M/M/2 el tiempo medio de espera en cola es menor que en el M/M/1, tanto el tiempo medio de servicio como el tiempo medio total de estancia en el sistema son mayores.

Parámetro	Sistema M/M/1	Sistema M/M/2
W (s)	1/4	1/6
t_s (s)	1/4	1/2
T (s)	1/2	2/3

Tabla 7.10: Retardos medios en cola, de servicio y total para un M/M/1 y M/M/2 con $\lambda = 2$ y $\rho = 1/2$.

A LA VISTA DE ESTOS RESULTADOS se tiene que, al comparar las prestaciones de un sistema $M/M/1$ con otro $M/M/m$, ambos con la misma capacidad agregada:

- Considerando las prestaciones globales del sistema (el retardo medio total T), resulta mejor agrupar recursos en un único servidor que distribuirlos en dos servidores menos potentes.
- A cambio, el tiempo medio de espera en cola W para el M/M/1 resulta ligeramente mayor (además de otras consideraciones, como p.ej. que un servidor muy potente pueda ser más caro que varios servidores con la misma potencia agregada).

Agregar la capacidad de procesado en un único servidor mejora las prestaciones globales del sistema, siendo éste uno de los resultados fundamentales del la teoría de colas. La explicación para este comportamiento se puede deducir analizando el comportamiento en situaciones de alta y baja carga:

- Cuando la carga sea elevada, ambos sistemas tienen todos sus recursos ocupados, por lo que la salida de usuarios se produce a la máxima capacidad μ.
- Sin embargo, en situaciones de baja carga, en un sistema M/M/1 *siempre* se aprovechará toda la capacidad μ a partir del primer usuario (Fig. 7.7), mientras que en un sistema M/M/m esto sólo ocurre si hay m o más usuarios en el mismo, por lo que se desaprovecharán recursos.

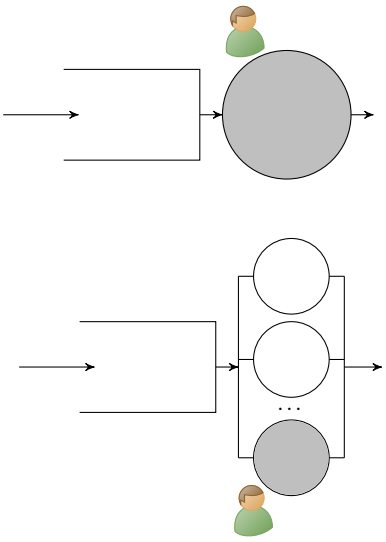


Figura 7.7: Comparación M/M/1 con M/M/m cuando $\rho \to 0$. Un usuario ocupa sólo un recurso.

No obstante, para elegir *el mejor* sistema a igualdad de capacidad, en un caso general habrá que tener en consideración otros parámetros, como p.ej. el coste o la necesidad de tener cierta fiabilidad frente a fallos (en un sistema M/M/1, el fallo del servidor resulta catastrófico).

El sistema M/M/∞ (*)

Se trata de nuevo de un sistema con múltiples servidores, con llegadas de Poisson a tasa λ y tiempos de servicio exponenciales de media $1/\mu$. En este caso, el número de servidores se puede considerar infinito, por lo que los usuarios una vez que llegan al sistema nunca tienen que esperar para ser atendidos. Se trata de un sistema que podría modelar, por ejemplo, el comportamiento de servicios "en la nube" que escalan automáticamente con el tráfico, donde no hay tiempo de espera (p.ej., servicios de video bajo demanda).

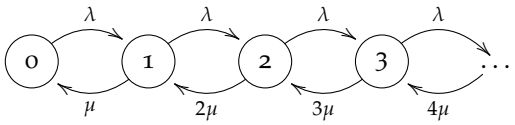
Dado que siempre hay un servidor disponible para cualquier petición, el tiempo medio de estancia en el sistema es el tiempo medio de servicio

$$T = W + t_s = t_s = \frac{1}{\mu}$$

Por lo que el número medio de usuarios en el sistema viene dado por

$$N = \lambda T = \frac{\lambda}{\mu}$$

Si bien ha sido posible calcular N y T sin necesidad de deducir la probabilidad de n usuarios en el sistema (p_n), dichas probabilidades son interesantes para obtener otras métricas de interés: por ejemplo, para un servicio de video bajo demanda, con los valores de p_n se puede calcular el consumo energético esperable del centro de cómputo. La cadena de Markov que modela el comportamiento del sistema se corresponde con la del sistema M/M/m para los estados $n \leq m$, esto es:



Por lo que para todos los estados se tiene que:

$$p_1 = \frac{\lambda}{\mu} p_0, \quad p_2 = \frac{\lambda}{2\mu} p_1, \quad p_3 = \frac{\lambda}{3\mu} p_2, \quad \ldots$$

Es decir:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \cdot p_0$$

Añadida la condición de normalización

$$\sum_{n=0}^{\infty} p_n = 1$$

Se tiene que, por inspección, el número de usuarios en el sistema sigue una distribución de Poisson de media λ/μ

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

Ejemplo 7.9. Sea un pequeño servicio de video bajo demanda en la nube, donde cada servidor consume 200 W si está encendido y tiene la capacidad de servir hasta tres usuarios a la vez. Los usuarios llegan según un proceso de Poisson a tasa 6 usuarios/día y la duración de una sesión sigue una v.a. exponencial de media 2 horas. Suponiendo que no hay límite en el número de servidores que se pueden encender, se puede modelar como un sistema M/M/∞, con un número medio de usuarios igual a

$$N = \frac{\lambda}{\mu} = \frac{6}{12} = \frac{1}{2}$$

La probabilidad de tener n usuarios se puede obtener con la v.a. discreta de Poisson, con los resultados de la Tabla 7.11 (con los 6 casos considerados se tiene en cuenta el 99,999 % de las posibilidades). Sabiendo que un servidor sirve hasta tres usuarios, a partir de dicha tabla se puede calcular la probabilidad de tener k servidores encendidos.

n	0	1	2	3	4	5
Pr(n)	0.60653	0.30327	0.07582	0.01264	0.00158	0.00016
k	0	1			2	
Pr(k)	0.60653	0.39173			0.00174	

Tabla 7.11: Probabilidad de n usuarios en el sistema M/M/∞ del Ejemplo 7.9 y probabilidad de k servidores encendidos.

Con los valores obtenidos, se puede obtener que el consumo medio de dicho servicio de video bajo demanda será

$$P = 200 \cdot \Pr(k = 1) + 400 \cdot \Pr(k = 2) = 85{,}306\ W$$

Sistemas con capacidad finita: sistemas con rechazo

Los sistemas con rechazo son aquellos en los que el tamaño de la cola tiene un valor finito, por lo que una llegada que se encuentre el sistema a su capacidad máxima no será admitida (será rechazada). Únicamente se tendrán en cuenta los sistemas en los que la petición rechazada es la que intenta acceder al sistema y no otros sistemas como, por ejemplo, cuando hay un mecanismo de prioridad en el que la llegada de un nuevo usuario provoca la expulsión de otro usuario de menor prioridad.

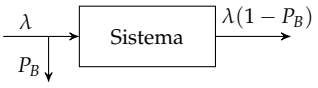


Figura 7.8: (Repetida) Tasa cursada en un sistema con bloqueo.

En estos sistemas la tasa efectiva $\overline{\lambda}$ que cursa el sistema no es igual que la tasa ofrecida λ al mismo, sino que están relacionadas por la probabilidad de bloqueo P_B (Figura 7.8):

$$\overline{\lambda} = \lambda \cdot (1 - P_B)$$

En el modelado de los sistemas M/M/1 y M/M/m, dado que el número de estados de la cadena de Markov es infinito, para poder obtener la distribución de probabilidades de estado estacionaria era preciso que

$$\frac{\lambda}{m \cdot \mu} < 1, \text{ es decir, } \lambda < m \cdot \mu,$$

para que la tasa media de entrada fuese inferior a la tasa máxima de salida. En los sistemas con una capacidad finita K esta condición no es necesaria, dado que el número de estados es finito. Por lo tanto, independientemente del valor de λ y μ será posible obtener las prestaciones del sistema, incluyendo los casos en que $\lambda > n\mu$. En el caso límite $p_K \to 1$ la probabilidad de rechazo será muy alta y el número medio de usuarios tenderá a K, pero siempre será posible analizar el sistema.

Recordatorio: la propiedad PASTA

En un sistema con capacidad finita K, la probabilidad P_B de rechazar un usuario es una probabilidad *condicionada*: se define como la probabilidad de que llegue un usuario al sistema y *vea* que el sistema está completo, por lo que se produce un bloqueo y no es admitido.

Por otra parte, se puede definir la probabilidad de que haya K usuarios en el sistema, esto es, que esté lleno: P_K. A diferencia de la anterior, se trata de una definición temporal: es la probabilidad de que en un instante de tiempo al azar el sistema esté completo, independientemente de la llegada o no de usuarios.

Las probabilidades P_B y P_K son probabilidades diferentes, por lo que sus valores no tendrían por qué coincidir. Sin embargo, por la propiedad PASTA de los procesos de Poisson (página 56), en un sistema donde el tiempo entre llegadas siga una variable aleatoria exponencial, se tiene que ambas coinciden. Por lo tanto, para calcular la probabilidad de rechazo bastará con calcular la probabilidad de que haya K usuarios en el sistema:

$$P_B = P_K \ .$$

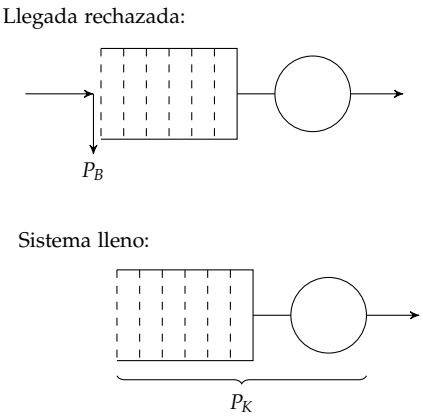


Figura 7.9: Rechazo de una llegada (arriba) vs. sistema lleno (abajo).

Utilización de recursos

En un sistema M/M/m (incluyendo m=1) el término ρ es el cociente entre la tasa de llegadas y la tasa máxima de salidas

$$\rho = \frac{\lambda}{m\mu},$$

y coincide, como se vio en Ejemplo: Aplicación de Little para un sistema con un recurso (página 72), con la utilización media de un recurso. Sin embargo, en un sistema con rechazo es preciso tener en cuenta la tasa efectiva de llegadas:

$$\bar{\lambda} = \lambda(1 - P_B) \ .$$

Sea un sistema con rechazo y m recursos en paralelo, del que se pretende calcular ρ, esto es, la utilización media de los recursos. Por definición de ρ, la tasa media de salidas es

$$\rho \cdot m \cdot \mu.$$

Dado que el sistema no crea peticiones, debe cumplirse que la tasa media de entrada sea igual a la tasa de salida, por lo tanto:[12]

[12] Nótese que no se realiza ninguna suposición sobre el proceso de llegadas o los tiempos de servicio.

$$\bar{\lambda} = \rho \cdot m \cdot \mu.$$

A partir de esta expresión, ρ se puede calcular como

$$\rho = \frac{\lambda(1 - P_B)}{m\mu} = \frac{\bar{\lambda}}{m\mu}.$$

Teorema de Little y tiempos de estancia

Al igual que para el caso de la utilización de los recursos, el teorema de Little relaciona los retardos medios en el sistema con el número medio de usuarios que lo atraviesan. Por lo tanto, la expresión a emplear dada una probabilidad de rechazo P_B será

$$N = \bar{\lambda} \cdot T = \lambda(1 - P_B) \cdot T.$$

De hecho, cuando se habla de tiempos medios de estancia, ya sea en la cola, en el recurso o en el sistema, se deberá entender que son los tiempos medios de estancia de aquellas peticiones que son atendidas por el sistema, dado que no tiene sentido calcular el tiempo medio para una petición que es rechazada.

Ejemplo 7.10. Sea una gasolinera similar a la del Ejemplo 6.4, con un surtidor y una capacidad máxima para dos coches (incluyendo al que está repostando). La gasolinera siempre está llena y el tiempo medio de estancia es de 5 minutos. En estas condiciones, la tasa de coches que *entran* a la gasolinera se puede calcular como:

$$\bar{\lambda} = \frac{N}{T} = \frac{2}{5} = 24 \text{ coches/hora}.$$

Si la tasa de coches que *quieren* entrar a la gasolinera es de $\lambda = 30$ coches/hora, la probabilidad de rechazo es de

$$P_B = 1 - \frac{\bar{\lambda}}{\lambda} = 20\,\%$$

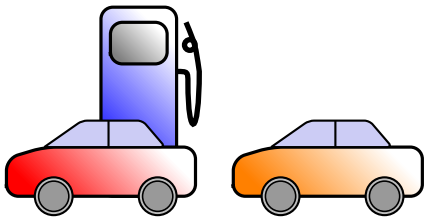
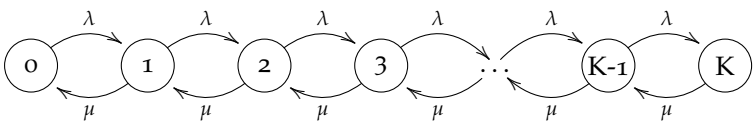


Figura 7.10: Gasolinera con un surtidor y sitio para dos coches.

El sistema M/M/1/K

El sistema M/M/1/K es muy similar al M/M/1, pero el número máximo de usuarios en el sistema está limitado por K. Por lo tanto, la longitud máxima de la cola es $K - 1$, dado que el recurso puede albergar a un usuario. Este sistema resulta apropiado para modelar situaciones de capacidad finita, como por ejemplo una línea de transmission con un *buffer* de un tamaño limitado, o una sala de espera que no puede albergar a más de un número dado de usuarios.

La cadena de Markov para modelar el sistema tiene una estructura muy parecida a la empleada al modelar el M/M/1, si bien acaba en el estado K (por lo que hay $K + 1$ estados):



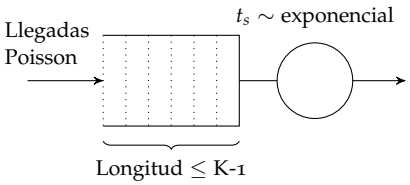


Figura 7.11: Sistema $M/M/1/K$.

Las relaciones entre estados son las mismas que para el caso del $M/M/1$, es decir

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0, \quad \forall 1 \leq n \leq K,$$

Empleando el parámetro $I = \lambda/\mu$ (que *no* coincide con el parámetro ρ, según se verá más adelante), las relaciones se pueden expresar como

$$p_n = I^n p_0, \quad \forall 1 \leq n \leq K,$$

A diferencia del M/M/1, calcular el valor de p_0 hay que hacer una suma finita

$$\sum_{n=0}^{K} p_n = \sum_{n=0}^{K} I^n p_0 = 1$$

quedando la expresión para p_0 como

$$p_0 = \left(\sum_{n=0}^{K} I^n\right)^{-1} \tag{7.11}$$

cuya solución dependerá, como se verá a continuación, del valor de $I = \lambda/\mu$

Caso $I \neq 1$ (es decir, $\lambda \neq \mu$)

En este caso, el cálculo del valor de p_0 según la expresión (7.11) resulta[13]

$$p_0 = \frac{1 - I}{1 - I^{K+1}}$$

Por lo que la distribución de probabilidades en estado estacionario es

$$p_n = \frac{1 - I}{1 - I^{K+1}} I^n, \quad N = 0, 1, \ldots, K$$

La tasa efectiva de entrada al sistema viene dada por

$$\bar{\lambda} = \lambda(1 - p_K) = \lambda \frac{1 - I^K}{1 - I^{K+1}}$$

mientras que la ocupación media del recurso viene dada por la expresión habitual

$$\rho = \frac{\bar{\lambda}}{\mu}$$

A partir de la expresión para p_n, se puede obtener el número medio de usuario en el sistema, calculando

$$N = \sum_{n=0}^{K} n \cdot p_n,$$

que queda

$$N = \frac{I}{1 - I} - \frac{(K+1)I^{K+1}}{1 - I^{K+1}},$$

A partir de N se puede obtener T aplicando el teorema de Little

$$T = \frac{N}{\lambda(1 - p_K)},$$

[13] La suma finita de la serie geométrica, para cualquier $r \neq 1$, es:

$$\sum_{n=0}^{N} r^n = \frac{1 - r^{N+1}}{1 - r}$$

mientras que el valor de Q se puede obtener, bien restando del número medio de usuarios en el sistema la ocupación media del recurso

$$Q = N - (1 - p_0),$$

o bien a partir del tiempo medio de espera en cola

$$W = T - t_s,$$

y aplicando Little (con la tasa efectiva de entrada)

$$Q = \lambda(1 - p_K) \cdot W$$

Ejemplo 7.11. Sea una empresa donde se comparte una impresora en red con una capacidad máxima de 4 trabajos de impresión, donde el tiempo para imprimir se puede modelar con una variable aleatoria exponencial de media 24 segundos. La tasa de generación de trabajos es de 5 peticiones/minuto, y se pretende conocer la probabilidad de que un trabajo se pierda y el tiempo medio para imprimirlo

Se puede modelar como un sistema M/M/1/4, donde el tiempo de servicio tiene por media

$$1/\mu = 24 \text{ segundos} \equiv 0.4 \text{ minutos},$$

por lo que $I = \lambda/\mu = 2$. La probabilidad de que el sistema esté lleno es

$$p_k = \frac{\left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^k}{1 - \left(\frac{\lambda}{\mu}\right)^{k+1}} = \frac{(1-2)2^4}{1-2^5} = \frac{16}{31} \ ,$$

que coincide con la probabilidad de un trabajo se pierda (por la propiedad PASTA).

El número medio de trabajos en la impresora viene dado por

$$N = \frac{I}{1-I} - \frac{(K+1)I^{K+1}}{1-I^{K+1}} = \frac{2}{1-2} - \frac{(4+1)2^{4+1}}{1-2^{4+1}} = \frac{98}{31}$$

por lo que el tiempo medio para imprimir es

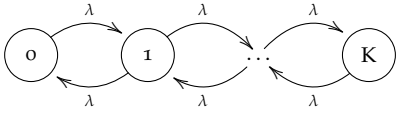
$$T = \frac{N}{\lambda(1-p_K)} = \frac{98/31}{5(1-16/31)} = \frac{98}{75} \text{ minutos } (78.4 \text{ s})$$

Caso $I = 1$ (es decir, $\lambda = \mu$)

En este caso, el incremento de usuarios en la cadena de Markov tiene la misma tasa que su decremento. Se puede ver[14] que, dado que $\lambda = \mu$, la cadena se encuentra "equilibrada" sin existir una tendencia hacia el estado 0 o el estado K. Las relaciones entre las probabilidades de estado pasan a ser

$$p_n = p_0, \ \ \forall n$$

[14] La cadena de Markov se puede representar como:



por lo que la ecuación $\sum p_n = 1$ se convierte en

$$(K+1)p_0 = 1$$

de donde se obtiene que:

$$p_n = \frac{1}{K+1}, \quad \forall n = 0, 1, \ldots, K$$

Se puede deducir fácilmente que

$$N = \frac{K}{2}$$

mientras que el resto de parámetros del sistema se calculan de una forma similar al caso anterior.

Ejemplo 7.12. Sea el caso de la impresora compartida como en el Ejemplo 7.11, pero con la mitad de tasa la tasa, esto es,

$$\lambda = 2{,}5 \text{ peticiones/minuto.}$$

Sigue siendo un sistema M/M/1/4 con $\mu = 0{,}4$ minutos, pero ahora $I = 1$, por lo el número medio de usuarios en una impresora es

$$N = K/2 = 2 \text{ trabajos de impresión.}$$

La probabilidad de rechazo viene dada por

$$p_k = \frac{1}{k+1} = \frac{1}{5} \ ,$$

mientras que el tiempo medio para imprimir pasa a ser

$$T = \frac{N}{\lambda(1-p_k)} = \frac{2}{2{,}5(1-(1/5))} = 1 \text{ minuto.}$$

El sistema M/M/m/m

Se trata de un sistema donde el número de recursos en paralelo m es igual al número máximo de usuarios en el sistema m. Por lo tanto, no hay cola en la que los usuarios aguarden a esperar: o son atendidos al entrar al sistema, o son directamente rechazados (Figura 7.12). Se trata de un sistema que serviría para modelar, por ejemplo, una central de conmutación de circuitos con m líneas, o una trama TDMA de m ranuras.

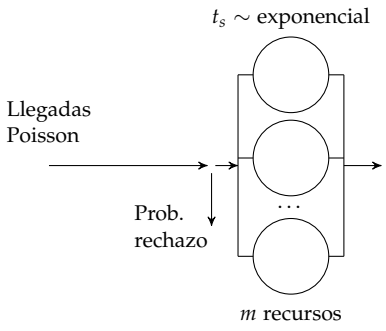


Figura 7.12: Sistema $M/M/m/m$.

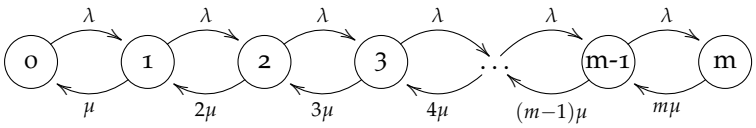
Dado que no hay cola, el número medio de usuarios en cola así como el tiempo medio de espera es cero:

$$W = 0$$
$$Q = 0$$

Por lo tanto, el tiempo medio de estancia en el sistema (para los usuarios que no son rechazados) es el tiempo medio de servicio:

$$T = 1/\mu = t_s.$$

Para obtener N a partir de T mediante el teorema de Little es preciso calcular la probabilidad de bloqueo P_B, para lo que es preciso modelar el sistema con una cadena de Markov. Esta cadena tiene un número de estados finito, como en el caso anterior, si bien las tasas de transiciones del estado n al $n-1$ se producen a tasa $n\mu$, dado que hay n usuarios siendo servidos:



Las relaciones entre las probabilidades de estados son similares a las del sistema M/M/m cuando $n < m$, esto es

$$p_1 = \frac{\lambda}{\mu}p_0, \quad p_2 = \frac{\lambda}{2\mu}p_1, \quad p_3 = \frac{\lambda}{3\mu}p_2, \quad p_4 = \frac{\lambda}{4\mu}p_3, \ldots$$

A partir de estas relaciones, de $\sum p_n = 1$ se obtiene que

$$p_0 = \left(\sum_{n=0}^{m} \frac{I^n}{n!}\right)^{-1}$$

mientras que para el resto de estados

$$p_n = \frac{I^n}{n!}p_0, \quad n = 0, 1, \ldots, m$$

En general, en un sistema M/M/m/m la variable de mayor interés es la probabilidad de bloqueo P_B, que coincide con la probabilidad de que haya m usuarios en el sistema (p_m).Particularizando la expresión de p_n para el caso de m usuarios se obtiene otra de las expresiones más importantes de la teoría de colas, la ecuación conocida como *Erlang-B*:

$$B(m, I) \triangleq p_m = \frac{\frac{I^m}{m!}}{\sum_{n=0}^{m} \frac{I^n}{n!}} \tag{7.12}$$

y que, por lo tanto, permite calcular la probabilidad de que un usuario no sea admitido al sistema, dados un número de recursos m y el cociente I.

Ejemplo 7.13. Para una población que genera llamadas a una tasa de 4 llamadas/minuto (de Poisson) se instala una centralita con dos circuitos. La duración de una llamada se puede modelar con una variable aleatoria exponencial de media 30 segundos. En estas condiciones, se tiene que

$$\lambda = 4 \text{ llamadas/minuto}, \mu = 2 \text{ llamadas/minuto}, I = \frac{4}{2} = 2\,.$$

La probabilidad de que un intento de llamada sea rechazado se calcula como

$$p_m = \frac{\frac{I^m}{m!}}{\sum_{n=0}^{m} \frac{I^n}{n!}} = \frac{\frac{2^2}{2!}}{1 + 2 + \frac{2^2}{2!}} = \frac{2}{5}\,,$$

esto es, del 40 %, por lo que se cursan $\bar{\lambda} = 2.4$ llamadas/minuto.

Si se instalase un circuito adicional para aumentar la facturación ($m = 3$), la probabilidad de rechazo sería

$$p_m = \frac{\frac{2^3}{3!}}{1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!}} = \frac{4}{19},$$

esto es, aproximadamente del 20 %, por lo que en estas circunstancias se cursarían unas 3.2 llamadas/minuto.

Con otro circuito adicional ($m = 4$), la probabilidad pasa a ser $2/21$, esto es, aproximadamente un 10 % de rechazo de las llamadas, mientras que con $m = 5$ dicha probabilidad pasaría a ser $4/109$ (esto es, menos del 4 %). En la siguiente tabla se representa la reducción de la probabilidad de bloqueo en función del número de circuitos empleados.

m	(Δm)	p_m	(∇p_m)	$p_m - p_{m-1}$
2		0.400		
3	(50 %)	0.211	(52 %)	-0.189
4	(33 %)	0.095	(45 %)	-0.116
5	(25 %)	0.037	(39 %)	-0.061

Resulta interesante comprobar que, en términos relativos, el incremento en el número de circuitos conlleva una reducción similar de la probabilidad de bloqueo; sin embargo, en términos absolutos la reducción de p_m resulta cada vez menor, lo cual es lógico dado que su valor tiende a o conforme m aumente: por lo tanto, en valores de m relativamente elevados cada vez resultará más costoso disminuir la probabilidad de bloqueo.

Resumen del tema

- Agregar recursos en un único servidor mejora las prestaciones del sistema.
- Para una tasa de llegadas λ, un tiempo medio de servicio $1/\mu$ y m servidores en paralelo (incluyendo $m = 1$) se define

$$I = \frac{\lambda}{\mu}, \quad \rho = \frac{\bar{\lambda}}{\mu}$$

- En un sistema M/M/1:

$$p_n = \rho^n(1-\rho), \quad T = \frac{1/\mu}{1-\rho} = \frac{1}{\mu - \lambda}$$

$$F_T(t) = 1 - e^{-(\mu-\lambda)t}.$$

- En un sistema M/M/m:

$$p_0 = \left(\left(\sum_{n=0}^{m-1} \frac{I^n}{n!} \right) + \frac{I^m}{m!} \cdot \frac{1}{1-\rho} \right)^{-1}, \quad Q = \frac{I^m \cdot \rho}{m!(1-\rho)^2} p_0.$$

$$p_n = \begin{cases} \frac{I^n}{n!} \cdot p_0 & \text{Si } n \leq m \\ \rho^n \frac{m^m}{m!} \cdot p_0 & \text{Si } n \geq m \end{cases}$$

$$P_Q = E_c(m, I) = \frac{\frac{I^m}{m!}}{(1-\rho)\left(\sum_{n=0}^{m-1} \frac{I^n}{n!}\right) + \frac{I^m}{m!}}$$

- En un sistema M/M/1/K:

Si $\lambda \neq \mu$ $\quad p_n = \frac{\left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n}{1-\left(\frac{\lambda}{\mu}\right)^{K+1}}$ $\quad N = \frac{I}{1-I} - \frac{(K+1)I^{K+1}}{1-I^{K+1}}$

Si $\lambda = \mu$ $\quad p_n = \frac{1}{K+1}$ $\quad N = \frac{K}{2}$

- En un sistema M/M/m/m:

$$p_0 = \left(\sum_{n=0}^{m} \frac{I^n}{n!}\right)^{-1} \qquad p_n = \frac{I^n}{n!} p_0$$

Teoría de colas: sistemas avanzados

En todos los sistemas vistos en el tema anterior, independientemente de su capacidad o número de recursos en paralelo, se cumplía que: (*i*) las llegadas siguen un proceso de Poisson, (*ii*) el tiempo de servicio se puede modelar con una única variable aleatoria exponencial, (*iii*) todas las peticiones y usuarios son del mismo tipo, sin prioridades, y (*iv*) únicamente se analiza un sistema. En este tema se aborda el estudio de sistemas donde no se cumpla alguna de las tres últimas características, esto es, sistemas donde las llegadas sigan siendo de Poisson, pero ocurra que:

- Los tiempos de servicio no siguen una variable aleatoria exponencial, sino que pueden ser fijos, estar distribuidos uniformemente, ser mezclas de diferentes variables aleatorias, etc. Relajar esta suposición puede servir, p. ej., para modelar flujos donde se mezclan tramas de datos y de asentimientos, o para el caso de agregados de tramas de voz de longitud constante. Estas situaciones se analizarán con el *sistema M/G/1*.
- Hay usuarios con mayor prioridad que otros, por lo que cuando distintos tipos de usuarios coinciden en la cola, el de mayor prioridad tendrá preferencia para ser atendido. Esto puede servir para modelar un router que proporcione "diferenciación de servicio" entre tipos de trama, p.ej., si las tramas de un usuario *premium* tienen prioridad sobre las de un usuario *normal*. Este caso se analizará con el *sistema M/G/1 con prioridades*.
- Dos o más sistemas se encuentran conectados, de modo tal que tras ser servidos por un sistema, los usuarios pueden ser servidos por otro diferente (o más de uno). Esto servirá, bajo ciertas suposiciones, para modelar el caso de una red de comunicaciones, tratándose de *redes de colas*.

El sistema M/G/1

Como su notación indica, en este sistema las llegadas son de Poisson, únicamente hay un recurso, y el número máximo de usuarios en el sistema no está acotado. La diferencia respecto al M/M/1 es que el tiempo de servicio no se distribuye necesariamente como una variable aleatoria exponencial, por lo que no se puede partir de

la propiedad "sin memoria" que facilitaba el análisis (p.ej., en un M/M/1 ocupado, la probabilidad de que un usuario sea servido antes de t unidades de tiempo es siempre la misma).[1]

Algunos escenarios que se pueden modelar mediante un M/G/1 son: un router de salida que recibe el agregado de varios flujos con tramas de longitud no exponencial (p. ej., constante), o un punto de acceso de una red inalámbrica, donde los usuarios descargan archivos de longitud uniformemente distribuida.

[1] Cuando el tiempo de servicio sea una variable aleatoria exponencial será un caso particular del sistema M/G/1, por lo que los resultados de ese análisis también serán válidos para un M/M/1.

Planteamiento básico

Dado que los tiempos de servicio no siguen una variable aleatoria exponencial, el sistema no puede modelarse con una cadena de Markov, como los sistemas básicos. Dado que las llegadas son de Poisson, para analizar el sistema se aplicará la propiedad PASTA (página 56): la esperanza de lo que "ve" un usuario al llegar al sistema coincide con la media temporal de la variable considerada (*Poisson Arrivals See Time Averages*).

Sea una llegada cualquiera al sistema. Lo dicha llegada "ve" que tiene que esperar antes de llegar al recurso (ver Figura 8.1) se puede dividir en dos componentes:

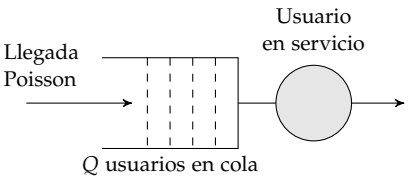


Figura 8.1: Estado visto por una llegada al sistema

- El tiempo para dar servicio a las Q peticiones que ya se encontrasen en la cola cuando llegó, cada una reclamando un tiempo medio de servicio t_s.

- El tiempo que falta para que el recurso atienda al siguiente usuario (el primero que está en la cola). Si el recurso estuviese ocupado, este tiempo *residual* de servicio R es el tiempo en terminar de servir la petición (que se analizará en el siguiente apartado).

Por lo tanto, la esperanza del tiempo de espera en cola de la petición que llega al sistema es la suma de estas dos componentes:

$$W = Q \cdot t_s + R$$

Dado que, por el teorema de Little, se cumple que $Q = \lambda \cdot W$, de lo anterior se tiene que

$$W = \lambda \cdot W \cdot t_s + R \ .$$

Además, si se llama a la ocupación $\rho = \lambda \cdot t_s$, se deduce la siguiente expresión para el tiempo medio de espera en cola:

$$W = \frac{R}{1-\rho}. \tag{8.1}$$

Por lo tanto, una vez calculado R, con la expresión (8.1) se puede obtener el tiempo medio de espera en el sistema, y a partir de éste el tiempo total de estancia en el sistema T (sumando el tiempo medio de servicio) y resto de variables.

Cálculo del tiempo residual

La variable R es la esperanza de lo que quede por servir a un usuario que *pudiese* estar en el recurso cuando llega un nuevo usuario al sistema (es una esperanza condicionada). Dado que la condición es que llegue un usuario, hay que tener en cuenta que cuanto mayor sea el tiempo que falta por servir al usuario en el recurso, más probable será que se produzca una llegada (y viceversa).

Ejemplo 8.1. La Barbería Zeus (Leganés) únicamente cuenta con Fran para atender a los clientes, que pueden pedir arreglarse la barba o cortarse el pelo. Ambos tipos de clientes acuden en igual proporción, y el tiempo para atender a un cliente es de 10′ si hay que arreglar la barba y de 50′ si hay que cortar el pelo. En estas condiciones, el tiempo medio de servicio por cliente es:

$$t = \Pr(\text{barba})t_{\text{barba}} + \Pr(\text{pelo})t_{\text{pelo}} = \frac{1}{2}10 + \frac{1}{2}50 = 30 \text{ minutos.}$$

Suponiendo que los clientes llegan según un proceso de Poisson, dado que la peluquería siempre está llena, el tiempo residual es el tiempo que pasa desde que un nuevo cliente llega hasta que el peluquero termina de atender al cliente que está siendo atendido. Un posible razonamiento (incorrecto) sería suponer que el cliente llegará, en media, cuando quede la mitad del tiempo de servicio. Por lo tanto, la media de este tiempo residual sería la mitad del tiempo medio de servicio por cliente, es decir, $t/2 = 15$ minutos. Este razonamiento es incorrecto dado que es más probable que un cliente llegue cuando Fran está cortando el pelo que cuando está arreglando una barba:

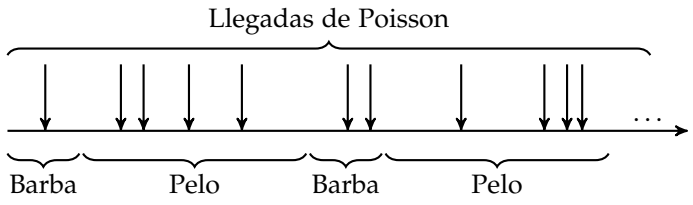


Figura 8.2: Llegadas de Poisson a una peluquería con dos tiempos de servicio. Es cinco veces más probable que un allegada de Poisson suceda cuando está cortando el pelo que cuando está arreglando una barba.

Dado que una llegada de Poisson puede suceder en cualquier momento, es cinco veces más probable que un cliente llegue cuando Fran está cortando el pelo que cuando está arreglando una barba.[2] En el primer caso, el tiempo medio hasta que termine será 50′/2 = 25 minutos, mientras que en el segundo caso dicho tiempo será 10′/2 = 5 minutos. Por lo tanto, el tiempo residual de servicio será en este caso

$$R = \Pr(\text{llegada cortando pelo})\frac{t_{\text{pelo}}}{2} + \Pr(\text{llegada arreglando barba})\frac{t_{\text{barba}}}{2}$$
$$= \frac{5}{6}25 + \frac{1}{6}5 = \frac{65}{3} \text{ minutos} \equiv 21' \, 40 \text{ s},$$

que resulta bastante más que la mitad del tiempo medio de servicio por cliente (que era 15 minutos).

[2] Distribución condicionada de una llegada, página 54.

Ejemplo 8.2. Suponga que un explorador tiene planeado realizar un viaje muy largo y no tiene ni idea sobre cuándo volverá a casa. Una vez que llegue, el *tiempo hasta mañana* es una variable aleatoria que será cercana a cero si llega justo antes de medianoche, cercana a 24 horas si llega justo después de medianoche. Suponiendo que probabilidad de llegar en un instante dado es siempre la misma, la esperanza de este tiempo "residual" hasta mañana es 12 h.

EN UN CASO GENERAL, donde el sistema no siempre esté ocupado, un usuario que llegue a un sistema puede encontrarlo de dos formas:

- Libre, por lo que el tiempo residual de servicio es cero.
- Ocupado, siendo el tiempo residual de servicio el tiempo que aún le queda al usuario en el recurso.

A continuación se presenta un cálculo gráfico del tiempo residual medio de servicio R. Para ello, se define $R(\tau)$ como el tiempo residual instantáneo de servicio que "ve" un usuario que llegue al sistema en un instante de tiempo τ. Se supone que, dado un valor de $R(\tau) > 0$, el usuario en el recurso recibe servicio a una tasa constante, por lo que $R(\tau)$ decrece de forma lineal con el tiempo, con pendiente -1.

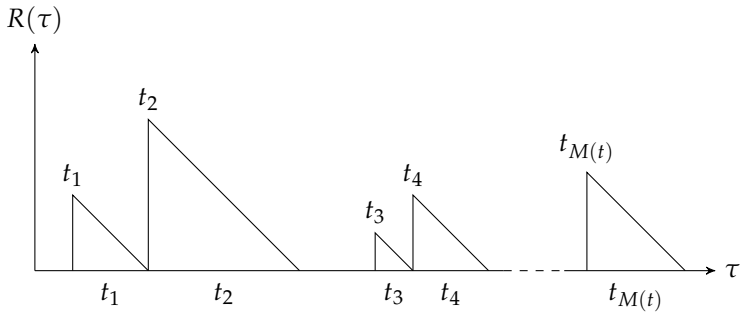


Figura 8.3: Cálculo del tiempo residual de servicio.

En la Figura 8.3 se representa $R(\tau)$ para un caso general con $M(t)$ llegadas de usuarios, donde el usuario i requiere un tiempo de servicio t_i.[3] Nótese que hay periodos donde el sistema está completamente vacío (por ejemplo, entre la segunda y la tercera llegada), y que el tiempo de inicio de servicio para un usuario no tienen por qué coincidir con su instante de llegada (p.ej., el segundo usuario pudo haber llegado justo después del primero).

[3] De esta forma, la primera llegada require un tiempo de servicio t_1, mientras que la última llegada (la llegada $M(t)$) requiere un tiempo de servicio $t_{M(t)}$.

La media del tiempo residual de servicio se puede calcular como

$$R = \lim_{t\to\infty} \frac{1}{t} \int_0^t R(\tau) d\tau,$$

y supone calcular la integral de $R(\tau)$ a lo largo del tiempo. Dicha integral se puede expresar como la suma de las áreas de los triángulos correspondientes a cada usuario

$$\int_0^t R(\tau) d\tau \approx \sum_{i=1}^{M(t)} \frac{t_i^2}{2},$$

por lo que el tiempo residual de servicio queda como

$$R = \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{M(t)} \frac{t_i^2}{2} .$$

Multiplicando y dividiendo esta expresión por $M(t)$ y re-ordenando términos y límites, se tiene que

$$R = \frac{1}{2} \left(\lim_{t\to\infty} \frac{M(t)}{t} \right) \left(\lim_{t\to\infty} \frac{1}{M(t)} \sum_{i=1}^{M(t)} t_i^2 \right) ,$$

expresión que resulta similar a la del número medio de usuarios N en el Teorema de Little (página 70).[4]

[4] En el teorema de Little, dicha expresión era

$$N = \lim_{t\to\infty} \frac{\alpha(t)}{t} \cdot \lim_{t\to\infty} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

Analizando la expresión para el cálculo de R, aparecen por tanto dos límites:

- El primer límite es el cociente del número total de usuarios servidos sobre el tiempo total. Por lo tanto, dicho cociente es, de nuevo, la tasa media de llegadas al sistema:

$$\lim_{t\to\infty} \frac{M(t)}{t} \triangleq \lambda$$

- El segundo término es la suma de todos los tiempos de estancia en el sistema al cuadrado t_i^2 dividida por el número total de llegadas al sistema. A partir de un número de llegadas elevado, este término se podrá aproximar por el momento de segundo orden de los tiempos de servicio

$$\lim_{t\to\infty} \frac{1}{M(t)} \sum_{i=1}^{M(t)} t_i^2 \approx \mathbb{E}[t_i^2].$$

A partir del cálculo de estos dos límites, se tiene que el tiempo medio residual de servicio se puede expresar como

$$R = \frac{1}{2} \cdot \bar{\lambda} \cdot \mathbb{E}[t_s^2] \tag{8.2}$$

Fórmula de Pollaczek-Khintchine (P-K)

Una vez obtenida la expresión para el tiempo medio residual R (8.2), su sustitución en la expresión para el tiempo medio de espera en cola W (8.1) resulta en la fórmula de Pollaczek-Khintchine[5]

$$W = \frac{R}{1-\rho} = \frac{\lambda \mathbb{E}[t_s^2]}{2(1-\rho)}, \tag{8.3}$$

[5] Fórmula publicada por primera vez por el matemático e ingeniero austriaco Felix Pollaczek en 1930 y redefinida en términos probabilísticos por el matemático soviético Aleksandr Khintchine dos años más tarde.

que permite calcular el tiempo medio de espera en cola en un sistema M/G/1 a partir de la tasa media de llegadas y los momentos de primer y segundo orden del tiempo de servicio.

Al igual que en el caso de los anteriores sistemas, a partir de W se puede calcular los otros parámetros de rendimiento del sistema, como el tiempo medio de estancia ($T = W + t_s$) o el número medio de usuarios en el sistema ($N = \lambda T$).

Ejemplo 8.3. A un enlace de 100 Mbps llega un flujo de 80 Mbps, donde las tramas tienen un tamaño fijo de 1500 B y el proceso de llegada se puede considerar de Poisson.

Para calcular el retardo total T al atravesar el enlace, es preciso primero obtener los términos de la expresión (8.3). Por una parte

$$\rho = \frac{80 \text{ Mbps}}{100 \text{ Mbps}} = 0.8,$$

mientras que la tasa de llegadas y el tiempo (constante) de servicio se obtienen como

$$\lambda = \frac{80 \text{ Mbps}}{1500 \cdot 8 \text{ bits/trama}} \approx 6.66 \text{ tramas/ms}, \qquad t_s = \frac{1500 \cdot 8}{100 \text{ Mbps}} = 0.12 \text{ ms}.$$

A partir de estos valores, se obtiene[6]

$$W = \frac{6{,}66 \cdot (0,12)^2}{2(1 - 0{,}8)} = 0.24 \text{ ms}.$$

[6] Dado que el tiempo de servicio es constante, $\mathbb{E}[t_s^2]$ coincide con t_s^2.

Por lo tanto, el retardo total al atravesar el sistema es

$$T = W + t_s = \ 0.36 \text{ ms}.$$

AL APLICAR LA FÓRMULA DE P-K es preciso obtener el momento de segundo orden del tiempo de servicio. Para ello, resulta útil recordar la expresión que lo relaciona con el momento de primer orden (esto es, la media) y la desviación típica

$$\mathbb{E}[t_s^2] = t_s^2 + \sigma_{t_s}^2$$

De hecho, la fórmula de P-K se puede escribir como

$$W = \frac{\lambda \mathbb{E}[t_s^2]}{2(1-\rho)} = \frac{\lambda(t_s^2 + \sigma_{t_s}^2)}{2(1-\rho)},$$

que, reordenando términos, se puede expresar como

$$W = \frac{\rho \cdot t_s}{2(1-\rho)} \left(1 + \left(\frac{\sigma_{t_s}}{t_s}\right)^2\right) \tag{8.4}$$

donde σ_{t_s}/t_s es el *coeficiente de variación* de la variable aleatoria t_s (para la variable aleatoria exponencial, por ejemplo, su valor es 1).[7] Esta expresión pone de manifiesto que, dada una carga en el sistema ρ y un tiempo de servicio medio t_s, el tiempo medio de espera aumenta conforme aumenta la varianza de t_s. De esta forma, cuando el tiempo de servicio sea un valor constante (esto es, en un sistema M/D/1) el tiempo medio será mínimo.

[7] Estrictamente, el coeficiente de variación se define como el cociente entre la desviación típica y el valor absoluto de la media.

Ejemplo 8.4. En un sistema M/M/1 el tiempo de servicio es una variable aleatoria exponencial, que cumple que $\sigma_{t_s} = t_s$. Por lo tanto, la expresión (8.4) se convierte en

$$W = \frac{\rho \cdot t_s}{(1-\rho)},$$

que es la expresión ya conocida para dicho sistema, según lo visto en el capítulo anterior.

En el caso de un sistema M/D/1 se tiene que $\sigma_{t_s} = 0$, por lo que el tiempo medio de espera en cola es

$$W = \frac{1}{2} \cdot \frac{\rho \cdot t_s}{(1-\rho)},$$

esto es, la mitad que para el caso del M/M/1. Estos resultados confirman que cuanto mayor es la variabilidad en el tiempo de servicio, peores serán las prestaciones que recibirán los usuarios.

Usuarios con diferentes tipos de tiempos de servicio

El sistema M/G/1 también sirve para modelar escenarios donde diferentes peticiones sigan diferentes variables aleatorias, como p.ej. el agregado de varios flujos de comunicaciones de diferentes tipos. Para ello, hay que suponer que dicho agregado se ha realizado de forma tal que, desde el punto de vista del recurso, la probabilidad de que una determinada petición sea de un tipo es constante e independiente a cada petición.

Sea un escenario con N flujos diferentes, donde el tipo i aparece con probabilidad α_i y tiene un tiempo medio de servicio t_i, con $i = \{1, \ldots, N\}$. En estas condiciones, el tiempo medio de servicio puede obtenerse como

$$t_s = \sum_{i=1}^{N} \alpha_i \cdot t_i \ ,$$

Dada una tasa total de llegadas λ, la tasa de llegadas del tipo de flujo i viene dada por

$$\lambda_i = \alpha_i \lambda \ ,$$

por lo que, dadas las tasas de llegadas de todos los tipos de flujo $\{\lambda_k\}$, la probabilidad del tipo i se puede calcular como la proporción de dicho flujo sobre el total

$$\alpha_i = \frac{\lambda_i}{\sum_1^N \lambda_i}$$

A partir de estas variables, para aplicar la fórmula de P-K es preciso obtener el momento de segundo orden del agregado, lo que se consigue mediante la ley de la probabilidad total

$$\mathbb{E}[t_s^2] = \sum_{i=1}^{N} \alpha_i \cdot \mathbb{E}[t_i^2] \ ,$$

lo que permite calcular el tiempo medio de espera, que es el mismo para todos los usuarios

$$W_i = \frac{\lambda \mathbb{E}[t_s^2]}{2(1-\rho)} \quad \forall i$$

A partir del tiempo medio de espera, se puede calcular el tiempo medio de estancia total para los usuarios de tipo i

$$T_i = W + t_i \ ,$$

que, en general, no coincidirá con el tiempo medio de estancia para una petición genérica

$$T = W + t_s$$

Ejemplo 8.5. (Ej. 13/14) Sea un control de seguridad de un aeropuerto, con una única cola de espera (Figura 8.4). Los viajeros llegan según un proceso de Poisson de tasa 24 viajeros / hora, y el tiempo que tardan en pasar el control se puede modelar con una variable aleatoria exponencial. Hay dos tipos de viajeros: frecuentes (el 40 %), que tardan en media 30 segundos en pasar el control, y no frecuentes (el 60 %), que tardan en media 3 minutos en pasar el control.

Se trata de un sistema $M/G/1$, dado que el tiempo de servicio es a veces exponencial de media 30 s, con probabilidad $\alpha_f = 0{,}4$, y otras veces exponencial pero de media 3 min, con probabilidad $\alpha_n = 0{,}6$, por lo que en ningún caso es exponencial. El tiempo medio de servicio es

$$E[t_s] = E[t_s^{freq}]\alpha_f + E[t_s^{nofreq}]\alpha_n = 2 \text{ min.},$$

mientras que el momento de segundo orden de dicho tiempo es (teniendo en cuenta que para la variable aleatoria exponencial la media y la desviación típica coinciden)

$$E[t_s^2] = E[(t_s^{freq})^2]\alpha_f + E[(t_s^{nofreq})^2]\alpha_n = 11 \text{ min.}^2,$$

Con esto, la fórmula del tiempo medio de espera en cola (para todo tipo de viajero) resulta

$$W = \frac{\lambda E[t_s^2]}{2(1-\rho)} = 11 \text{ min.},$$

por lo que el tiempo medio para pasar el control un viajero frecuente es

$$T^{freq} = W + t_s^{freq} = 11{,}5 \text{ min.},$$

mientras que para un viajero no frecuente es

$$T^{nofreq} = W + t_s^{nofreq} = 14 \text{ min.}$$

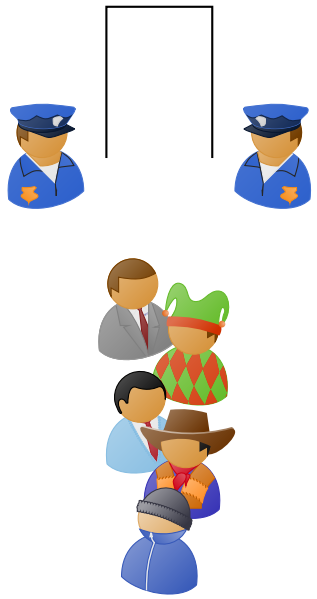


Figura 8.4: Control de seguridad con dos tipos de viajero en una única cola.

El sistema M/G/1 con prioridades

Sea ahora un sistema con distintas clases de tráfico y que, en vez de una única cola, tiene N colas diferentes, una para cada clase de tráfico (Figura 8.5). Cada clase tiene una diferente prioridad y existe un orden estricto: si hay un usuario de la primera clase, será el siguiente en acceder al recurso una vez que éste quede libre; si el recurso queda libre y no hay ningún usuario de la primera clase, será atendido el primer usuario de la segunda clase, etc.

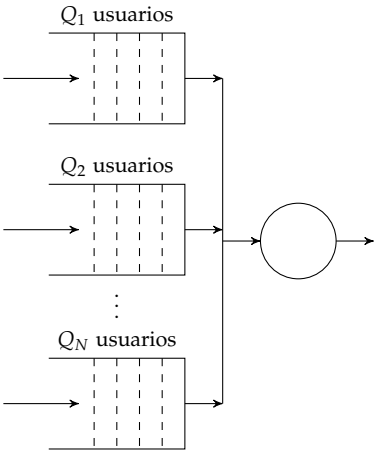
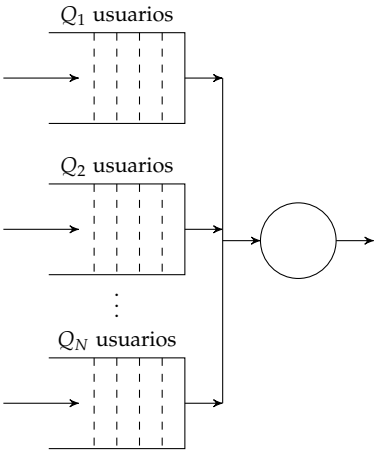


Figura 8.5: Sistema con N diferentes clases.

Se trata un sistema *sin apropiación*: una petición siempre espera a que termine de ser servida la petición que estuviese en el recurso, aunque fuese de menor prioridad (en un sistema con apropiación, una petición de baja prioridad es expulsada del recurso cuando llega una petición de mayor prioridad).



Siguiendo el mismo razonamiento que en el análisis del M/G/1, el tiempo medio de espera en cola para los usuarios de la primera clase (W_1), que llegan a una tasa λ_1, viene dado por:

- El tiempo medio residual de servicio del usuario que pudiera estar en el recurso R (que puede ser de cualquier clase y se calculará más adelante)
- El tiempo medio de servicio que precisan los Q_1 usuarios por delante en la cola, cada uno t_1.

Por lo tanto, el tiempo medio de espera en cola se puede expresar como

$$W_1 = R + Q_1 \cdot t_1 \ . \tag{8.5}$$

A partir de esta expresión, aplicando el teorema de Little sobre la cola de máxima prioridad se tiene

$$Q_1 = \lambda_1 W_1 \ ,$$

por lo que definiendo $\rho_1 = \lambda_1 t_1$, que se puede interpretar como la "carga relativa" de la primera clase, se puede expresar W_1 como

$$W_1 = \frac{R}{1-\rho_1} \ . \tag{8.6}$$

Sea ahora el caso de la segunda clase. De forma similar al cálculo de W_1, dos de los componentes del tiempo de espera en cola de W_2 serán:

- El tiempo residual del usuario en el recurso R
- El tiempo demandado por los usuarios que estén esperando, que son los Q_1 usuarios de la primera clase (que tienen más prioridad) y los Q_2 usuarios que se encuentren por delante de este usuario.

Además de estos factores, para toda clase que no sea la de máxima prioridad aparece otro componente: los usuarios mayor prioridad que llegan mientras se está esperando (y que pasarán al recurso en cuanto quede libre). Para el caso de W_2, este tiempo es:

- El tiempo demandado por los usuarios de la primera clase que lleguen (a tasa λ_1) mientras dicho usuario está esperando (W_2), y que cada uno requerirá un tiempo t_1.

Por lo tanto, la expresión para W_2 queda como:

$$W_2 = R + Q_1 \cdot t_1 + Q_2 \cdot t_2 + \lambda_1 \cdot W_2 \cdot t_1 \ ,$$

que puede expresarse, gracias a la ecuación (8.5), como

$$W_2 = W_1 + Q_2 \cdot t_2 + \lambda_1 \cdot W_2 \cdot t_1 \ .$$

Sustituyendo $Q_2 = \lambda_2 W_2$ (teorema de Little) y re-ordenando términos se llega a

$$W_2 = \frac{W_1}{(1-\rho_1-\rho_2)} \ ,$$

y sustituyendo W_1 por el valor dado por (8.6) queda

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)} \ .$$

Realizando un análisis similar al anterior, se puede comprobar que el retardo en cola para los usuarios de la tercera clase es:

$$W_3 = \frac{R}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} \ ,$$

siendo la expresión del tiempo medio en cola para una clase k cualquiera:

$$W_k = \frac{R}{(1-\rho_1-\rho_2-\ldots-\rho_{k-1})(1-\rho_1-\rho_2-\ldots-\rho_k)} \ .$$

De forma similar al M/G/1 con distintos tipos de usuarios (Usuarios con diferentes tipos de tiempos de servicio, página 153), el retardo total de la clase k se puede obtener como

$$T_k = W_k + t_k \ .$$

Por último, falta obtener el tiempo medio residual de servicio R, que resulta ser la misma que en el sistema M/G/1, ya que en el cálculo presentado no se hizo ninguna suposición sobre la disciplina de la cola:

$$R = \frac{1}{2}\lambda \mathbb{E}[t_s^2] \ , \tag{8.7}$$

Para calcular $\mathbb{E}[t_s^2]$ suele ser preciso aplicar la ley de la probabilidad total, para lo que se tiene en cuenta la proporción relativa del flujo i sobre el total, esto es, $\alpha_i = \lambda_i/\lambda$, por lo que (8.7) se convierte en

$$R = \frac{1}{2}\lambda \sum_{i=1}^{N} \alpha_i \mathbb{E}[t_i^2] = \frac{1}{2}\sum_{i=1}^{N} \lambda_i \mathbb{E}[t_i^2] \ ,$$

donde $\mathbb{E}[t_i^2]$ el momento de segundo orden del tiempo de servicio de la clase i. Por lo tanto, el tiempo medio de espera en cola de un usuario de la prioridad k puede expresarse como

$$W_k = \frac{\frac{1}{2}\sum_{i=1}^{N} \lambda_i \mathbb{E}[t_i^2]}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)} \ .$$

Ejemplo 8.6. Sea un control de seguridad similar al del Ejemplo 8.5, si bien en este caso llegan viajeros frecuentes a una tasa de 8 viajeros/hora y viajeros no frecuentes a una tasa de 16 viajeros/hora, y se forma una cola para cada tipo de viajero (Figura 8.6). Siempre que hay un viajero frecuente esperando, se le da prioridad sobre el viajero no frecuente. Como en el caso anterior, los viajeros frecuentes tardan en media 30 segundos en pasar el control, mientras que los no frecuentes tardan en media 3 minutos.

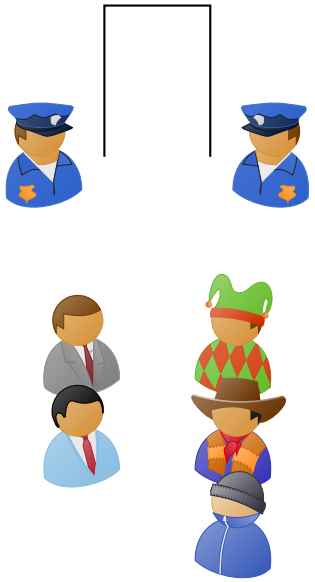


Figura 8.6: Control de seguridad con dos tipos de viajeros y prioridad.

Con estos datos, la aplicación de () queda como

$$R = \frac{1}{2}\left(\frac{8}{60}\cdot 1/2 + \frac{16}{60}\cdot 18\right) = \frac{73}{30}\text{ min.}$$

por lo que si $\rho_1 = 2/30$, el tiempo de espera para los viajeros frecuentes (es decir, de la primera cola) será

$$W_1 = \frac{R}{1-\rho_1} = \frac{73}{28} \approx 2{,}6\text{ minutos,}$$

por lo que su tiempo total en atravesar el control será de $T_1 \approx 3{,}1$ minutos. Para los viajeros no frecuentes, dado que $\rho_2 = 24/30$, su tiempo de espera será

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{73\cdot 30}{4\cdot 28} \approx 19{,}55\text{ minutos,}$$

por lo que el tiempo total para atravesar el control será de $T_2 \approx 22{,}55$ minutos.

Redes de colas

Una red de colas es un sistema donde los usuarios atraviesan una o más colas, cada una de ellas una o más veces. Se distinguen dos tipos de redes de colas:

- Redes cerradas (Figura 8.7), donde los usuarios siempre permanecen en el sistema, por lo que su número permanece constante a lo largo del tiempo. Este tipo de redes no se consideran en este texto.
- Redes abiertas (Figura 8.8), donde un usuario entra al sistema en un momento dado y con total certeza lo abandonará pasado un tiempo. A su vez, se puede diferenciar entre
  - Redes *acíclicas*, si es imposible que una petición pase más de una vez por un mismo recurso (esto es, no hay "bucles" alrededor de uno o más recursos)
  - Redes *cíclicas*, si es posible que una petición sea atendida varias veces por un recurso (no necesariamente de forma consecutiva).

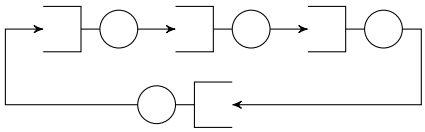


Figura 8.7: Red de colas cerrada.

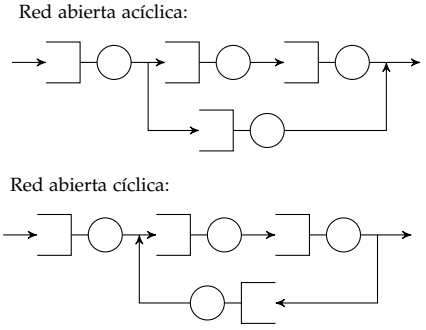


Figura 8.8: Redes abiertas.

En general, se supondrá que los tiempos de servicio que demanda un usuario en diferentes sistemas son variables independientes, lo que podrá ser más o menos acertado según el tipo de red considerada. A continuación, se inicia el análisis de las redes de colas con el caso acíclico más sencillo: dos sistemas en tándem.

Dos sistemas en tándem

Sea el caso de dos sistemas en *tándem*, como el ilustrado en la Figura 8.9, donde el proceso de llegada es de Poisson a tasa λ y el tiempo de servicio se distribuye según una variable aleatoria exponencial, de media $1/\mu_1$ en el primer sistema y $1/\mu_2$ en el segundo sistema.

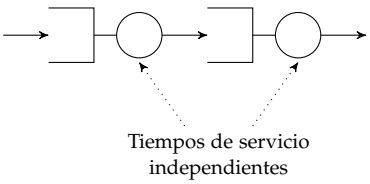


Figura 8.9: Sistemas M/M/1 en tándem.

Considerando el primer sistema, se trata de un M/M/1, por lo que la probabilidad de que haya n usuarios en el mismo (según lo visto en el tema anterior) viene dada por

$$p_1(n) = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) = \rho_1^n(1-\rho_1)$$

Además, dado que la salida del primer sistema también es un proceso de Poisson (Proceso de salida de un M/M/1 (*), página 128), el segundo sistema es otro M/M/1, por lo que la probabilidad de que haya m usuarios en el segundo sistema tiene una expresión similar

$$p_2(m) = \left(\frac{\lambda}{\mu_2}\right)^n \left(1 - \frac{\lambda}{\mu_2}\right) = \rho_2^n(1-\rho_2)$$

Para que la probabilidad de tener n usuarios en el primer sistema y m en el segundo se pueda calcular como el producto de las mismas, es necesario que $p_1(n)$ y $p_2(m)$ sean independientes, lo cual no es del todo intuitivo: se podría pensar que una alta ocupación del primer sistema conlleva una alta ocupación del segundo. Sin embargo, se puede demostrar que son probabilidades independientes gracias al *teorema de Burke*,[8] que amplía y generaliza la caracterización del proceso de salida de un M/M/1.

[8] Publicado por primera vez por Paul J. Burke en 1959 cuando trabajaba en los laboratorios Bell.

Teorema de Burke En un sistema M/M/1 o M/M/m en condiciones estacionarias y con un proceso de llegadas de Poisson a tasa λ se cumple que:

- El proceso de salida es de Poisson a la misma tasa λ.
- En cualquier instante de tiempo t, el número de usuarios en el sistema es independiente de la secuencia de instantes de salida de usuarios hasta t.

Nótese que el segundo resultado no resulta demasiado intuitivo, dado que una ráfaga de salidas en $t_1, t_2, \ldots, t_5$ con poco tiempo entre ellas podría indicar que la cola se encontrase muy ocupada en t_5, pero –según el teorema– esta ocupación es independiente del proceso de salidas.[9]

[9] Lo que el teorema no garantiza es la independencia del estado de la cola en t_1 (es decir, al principio) respecto al proceso de salidas: en general, varias salidas de forma consecutiva vendrá causado por una alta ocupación *anterior*.

Dado que el teorema de Burke garantiza que el estado del primer sistema no depende de la secuencia de las salidas desde dicho sistema, se tiene que $p_1(n)$ es independiente de $p_2(m)$, por que la probabilidad de tener n usuarios en el primer sistema y m en el segundo se puede calcular como:

$$\Pr(n,m) = \rho_1^n(1-\rho_1) \cdot \rho_2^n(1-\rho_2)$$

El número medio de usuarios en el sistema se puede calcular como

$$N = \sum_{n,m}(n+m)\Pr(n,m) = \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2}$$

y, aplicando el teorema de Little, el tiempo medio de estancia en el sistema se puede obtener como

$$T = \frac{N}{\lambda} = \frac{1/\mu_1}{1-\rho_1} + \frac{1/\mu_2}{1-\rho_2}\ ,$$

lo que equivaldría a obtener las prestaciones de cada sistema por separado, esto es,

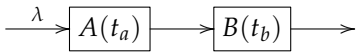
$$T_1 = \frac{1/\mu_1}{1-\rho_1}, N_1 = \frac{\rho_1}{1-\rho_1}, T_2 = \frac{1/\mu_2}{1-\rho_2}, N_2 = \frac{\rho_2}{1-\rho_2}\ \ ,$$

y a partir de estos valores, obtener las prestaciones globales como

$$N = N_1 + N_2\ \ ,$$
$$T = T_1 + T_2\ \ ,$$

expresiones que *no siempre* se podrán aplicar para analizar una red de colas (en concreto, el cálculo del retardo total en el sistema T).

Ejemplo 8.7. Sea la red indicada a continuación, a la que se producen llegadas según un proceso de Poisson a tasa $\lambda = 2$ llegadas/segundo, y los tiempos de servicio en cada sistema se pueden modelar con variables aleatorias exponenciales independientes, de media $t_a = 0.1$ segundos y $t_b = 0.2$ segundos.



En estas condiciones, la carga de cada sistema puede calcularse como

$$\rho_a = \frac{\lambda}{1/t_a} = \frac{1}{5},\ \ \rho_b = \frac{\lambda}{1/t_b} = \frac{2}{5},$$

por lo que el número medio de usuarios en cada sistema es

$$N_a = \frac{\rho_a}{1-\rho_a} = \frac{1}{4},\ \ N_b = \frac{\rho_b}{1-\rho_b} = \frac{2}{3},$$

lo que lleva a que el retardo total para atravesar la red es de

$$T = \frac{N_a + N_b}{\lambda} = \frac{11}{24} \approx 0.458 \text{ segundos.}$$

Redes abiertas acíclicas

El anterior resultado de dos sistemas M/M/1 en tándem se puede generalizar para el caso en que las colas tengan más de un recurso en paralelo (es decir, sistemas M/M/m) y que haya más de dos colas conectadas en cascada sin bucles. Como se ha mencionado anteriormente, este tipo de red se denomina *red acíclica*, ilustrándose un ejemplo en la Figura 8.10.

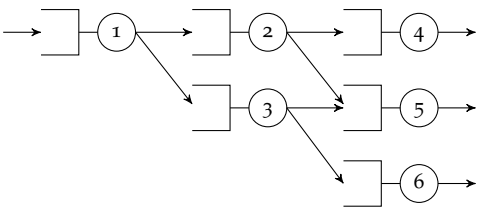


Figura 8.10: Ejemplo de red acíclica

La ausencia de bucles en una red se define a través del concepto de *camino de bajada*: se dice que un nodo k está en un camino de bajada desde un nodo i si la probabilidad de que una petición que salga del nodo i llegue al nodo k no es cero (pudiendo pasar por cualquier otro nodo).[10]

[10] En la Figura 8.11, los nodos 4, 5 y 6 están en el camino de bajada del nodo 1, pero el nodo 4 no está en el camino de bajada del nodo 3.

Definición: Una red abierta acíclica es aquella red en la que se cumplen las siguientes propiedades:

- La salida del nodo i va hacia un nodo j directamente conectado con una probabilidad $p_{ij} > 0$, que es independiente a cada petición.
- Si k está en el camino de bajada de i, entonces i no está en el camino de bajada de k (esto es, no hay bucles).

Si una red es acíclica y el proceso de entrada es de Poisson, entonces el proceso de entrada a todos los sistemas también es de Poisson: por el teorema de Burke, la salida λ_i del nodo i es también un proceso de Poisson. Dado que, desde dicho nodo, la probabilidad de ir al nodo j es p_{ij} y es independiente a cada petición, por la Descomposición de procesos de Poisson (página 48) el proceso de entrada al nodo j será también de Poisson a tasa $\lambda_i \cdot p_{ij}$ (y si varios nodos llevasen al nodo j, sería el agregado de procesos de Poisson).

Ejemplo 8.8. Sea la siguiente red acíclica de cuatro nodos, con una proceso de entrada de Poisson a tasa λ, tiempos de servicio exponenciales, y donde p_{ij} indica la probabilidad de ir al nodo j desde el nodo i:

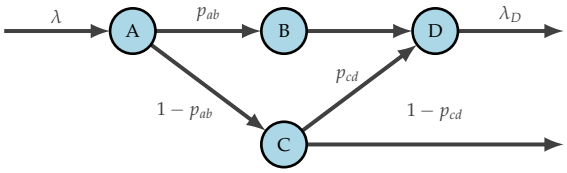


Figura 8.11: Red acíclica para el Ejemplo 8.8.

Según lo visto anteriormente, el proceso de salida desde A hacia B será de Poisson, a tasa λp_{ab}, al igual que el proceso de salida desde B hacia D. El proceso desde A hacia C también será de Poisson, a tasa $\lambda(1 - p_{ab})$, mientras que el de C hacia D será de poisson a tasa $\lambda(1 - p_{ab})p_{cd}$. Por lo tanto, el proceso de salida del nodo D será un proceso de Poisson a tasa:

$$\lambda_D = \lambda p_{ab} + \lambda(1 - p_{ab})p_{cd}$$

Además de garantizar que el proceso de entrada a cada sistema sea de Poisson, el teorema de Burke también garantiza la independencia entre el estado de un sistema y las salidas que de él se hayan producido. Por lo tanto, de forma análoga al sistema en tándem, en una red de colas acíclica de sistemas de tipo M/M/1 o M/M/1, la

probabilidad conjunta de tener n_1 usuarios en el primer sistema, n_2 usuarios en el segundo sistema, etc., se puede calcular como el producto de dichas probabilidades considerando cada sistema de forma independiente.

Ejemplo 8.9. En el caso del anterior Ejemplo 8.8, si los tiempos medios de servicio en los sistemas A, B, C y D fuesen, respectivamente, t_a, t_b, t_c y t_d (todos ellos exponenciales), se pueden definir las siguientes ocupaciones

$$\rho_a = \lambda_a t_a, \quad \rho_b = \lambda_b t_b, \quad \rho_c = \lambda_c t_c, \quad \rho_d = \lambda_d t_d$$

donde

$$\lambda_a = \lambda, \quad \lambda_b = \lambda_a p_{ab}, \quad \lambda_c = \lambda_a (1 - p_{ab}), \quad \lambda_d = \lambda p_{ab} + \lambda (1 - p_{ab}) p_{cd}$$

La probabilidad de tener n_1 usuarios en el sistema A, n_2 usuarios en el sistema B, etc., se puede calcular como

$$\Pr(n_A = n_1, n_B = n_2, n_C = n_3, n_D = n_4) = \rho_a^{n_1}(1-\rho_a)\rho_b^{n_2}(1-\rho_b)\rho_c^{n_3}(1-\rho_c)\rho_d^{n_4}(1-\rho_d)$$

mientras que la probabilidad de que todos los sistemas se encontrasen ocupados es

$$\Pr(n_a > 0, n_b > 0, n_c > 0, n_d > 0) = \rho_a \rho_b \rho_c \rho_d$$

En un caso general de una red abierta acíclica con N sistemas M/M/1, donde λ_i sea la tasa de entrada al sistema i y μ_i su tasa máxima de salida, la probabilidad de n_1 usuarios en el primer sistema, n_2 en el segundo, etc., sería (definiendo $\rho_i = \lambda_i / \mu_i$):

$$\Pr(n_1, n_2, \ldots, n_N) = \prod_{i=1}^{N} \rho^{n_i}(1 - \rho_i)$$

Si se trata de una red donde no todos los sistemas son M/M/1, será preciso acudir a las fórmulas correspondientes para el caso M/M/m.

Ejemplo 8.10. Sea la red de la de la figura, donde $p_{ab} = 3/4$ y $p_{de} = 1/2$. Todos los tiempos de servicios son exponenciales de media 20 ms, y todos los sistemas disponen de un único recurso salvo D, que dispone de dos recursos idénticos en paralelo. Las llegadas son de Poisson a tasa $\lambda = 40$ usuarios/segundo.

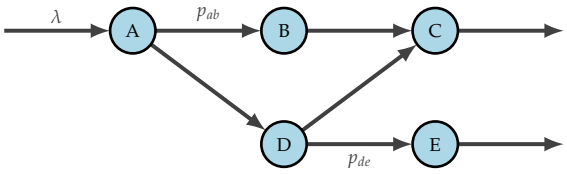


Figura 8.12: Red acíclica para el Ejemplo 8.10.

Las tasas de entrada a cada sistema son

$$\lambda_a = \lambda = 40\ u/s,\quad \lambda_b = 30\ u/s, \lambda_d = 10\ u/s, \lambda_c = 35\ u/s, \lambda_e = 5\ u/s,$$

mientras que la tasa máxima de salida para todos los sistemas es

$$\mu = 1/t_s = 50\ u/s.$$

Para los sistemas A, B, C y E el número medio de usuarios se puede obtener mediante la expresión del M/M/1, lo que lleva a[11]

$$N_a = 4, N_b = 3/2, N_c = 7/3, N_e = 1/9\ .$$

El sistema D se trata de un M/M/m, con una probabilidad de estar vacío y número medio de usuarios en cola igual a[12]

$$p_0 = \frac{40}{49},\quad Q = \frac{1}{392},$$

de lo que se obtiene

$$W = \frac{Q}{\lambda} = \frac{1}{3920},\quad T = W + t_s = \frac{794}{39200}\ ,$$

lo que lleva a que el número medio de usuarios en el sistema D es

$$N_d = \lambda \cdot T = \frac{197}{1960} \approx \frac{1}{10}$$

Por lo tanto, el número medio total de usuarios en la red es

$$N = \sum N_i = \frac{1448}{180} = \frac{362}{45} \approx 8 \text{ usuarios},$$

y el tiempo medio para atravesarla es, aplicando el teorema de Little

$$T = \frac{N}{\lambda} = 0{,}2 \text{ segundos}$$

Redes abiertas cíclicas

A diferencia del caso anterior, en una red cíclica sí que aparecen bucles en la topología de la red, lo que posibilita que un usuario pase más de una vez por un mismo sistema. Esto complica el análisis, ya que una consecuencia es que el proceso de entrada a la cola deja de ser necesariamente de Poisson. Para ilustrar esto, sea el caso de la Figura 8.13 a continuación, donde se producen llegadas nuevas al sistema una tasa λ según un proceso de Poisson, el tiempo de servicio es exponencial, de media t_s, y existe la probabilidad p de que un usuario vuelva al sistema tras ser servido (por lo que, con probabilidad $1 - p$, lo abandona).

Para analizar lo que sucede a la entrada de la cola, supóngase que la probabilidad de volver a entrar en el sistema es relativamente alta

$$p \gg 0\ ,$$

[11] Número medio de usuarios en un M/M/1:

$$N = \frac{\rho}{1-\rho},\ \text{con}\ \rho = \lambda/\mu\ .$$

[12] En un sistema M/M/m, el número medio de usuarios en cola es

$$Q = \frac{I^m \cdot \rho}{m!(1-\rho)^2} p_0\ ,$$

con

$$p_0 = \left(\left(\sum_{n=0}^{m-1} \frac{I^n}{n!} \right) + \frac{I^m}{m!} \cdot \frac{1}{1-\rho} \right)^{-1}$$

donde $I = \lambda/\mu$ y $\rho = I/m$.

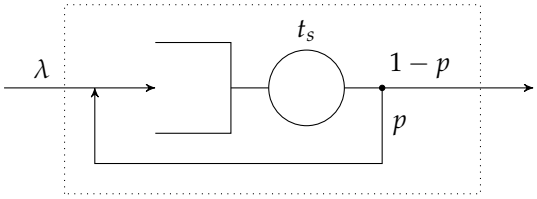


Figura 8.13: Sistema compuesto por una cola con realimentación.

y que la tasa de entrada es mucho menor que la tasa máxima de salida,

$$\lambda \ll \mu = 1/t_s \ .$$

por lo que el tiempo medio entre llegadas nuevas al sistema es mucho mayor que el tiempo medio de servicio

$$\lambda^{-1} \gg t_s \ .$$

En estas condiciones, la probabilidad de que el sistema esté vacío es muy alta (por ser μ mucho mayor que λ), por lo que una nueva llegada atravesará el sistema sin esperar en cola y, con alta probabilidad, lo volverá a atravesar varias veces (por ser p alta), según lo ilustrado en la Figura 8.14:

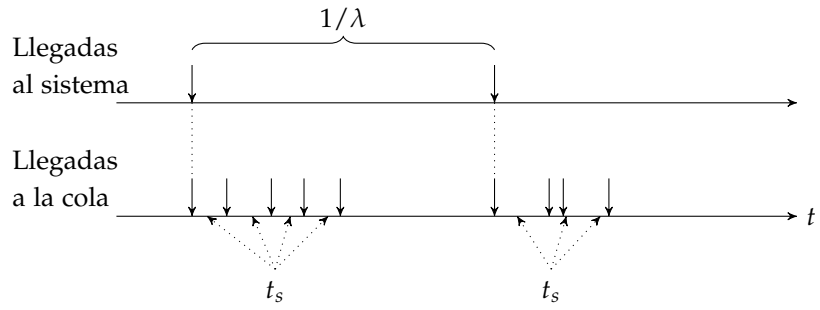


Figura 8.14: Procesos de entrada al sistema y a la cola para un sistema con re-alimentación como el de la Fig. 8.13, con $p \gg 0$ y $t_s \ll \lambda^{-1}$.

Por cada por cada petición que llegue al sistema se genera una *ráfaga* de llegadas a la cola, separadas entre sí el tiempo de servicio de la llegada original. Este patrón a ráfagas incumple claramente las propiedades de incrementos estacionarios e independientes, por lo que el proceso de llegadas a la cola no será de Poisson, así que el sistema no puede analizarse partiendo del M/M/1.

A diferencia del caso de las redes acíclicas, en una red cíclica, en general, no se podrá suponer que los procesos de llegada sean de Poisson. Sólo cuando se puedan hacer algunas suposiciones, orientadas a que la naturaleza "a ráfagas" de las llegadas se debilite, se podrá analizar el comportamiento de una red de colas partiendo del análisis de cada cola por separado –consideraciones que se verán a continuación.

Definición de red abierta. Teorema de Jackson

Sea una red de N colas, en la que una petición, tras ser servida por una cola i, pasa a la cola j con una probabilidad constante

e independiente p_{ij}. El tráfico de entrada a un nodo j se puede expresar como la suma del tráfico desde los otros nodos y el tráfico que llegue desde fuera de la red al nodo j (denotado como r_j):

$$\lambda_i = \sum_{j=1, j\neq i}^{N} \lambda_j p_{ji} + r_j \tag{8.8}$$

Ejemplo 8.11. Sea el caso de la red de la Figura 8.15.

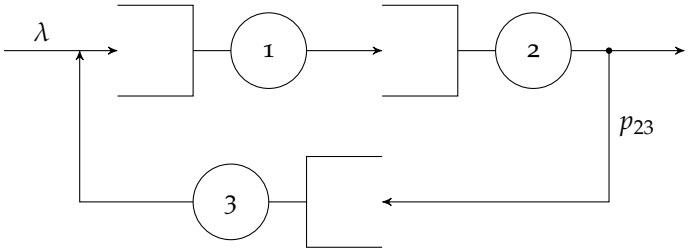


Figura 8.15: Ejemplo de red abierta cíclica compuesta por tres colas.

El tráfico de entrada al nodo 1 viene dado por el tráfico que pasa por el nodo 3 más el tráfico que llega desde fuera de la red, por lo que (8.8) quedaría en este caso como:

$$\lambda_1 = \lambda_3 + \lambda$$

donde el tráfico que pasa por el nodo 3 viene dado por

$$\lambda_3 = \lambda_2 p_{23} \, .$$

mientras que el tráfico que pasa por el nodo 2 es el mismo que el que pasa por el nodo 1 (en tasa)

$$\lambda_2 = \lambda_1 \, .$$

Dadas las anteriores expresiones, λ_1 se puede expresar como

$$\lambda_1 = \lambda_1 p_{23} + \lambda \rightarrow \lambda_1 = \frac{\lambda}{1 - p_{23}}$$

Una red es *abierta* si las peticiones "pasan" a través de dicha red, lo que sucede si (*i*) *entran* a través de (al menos) un nodo, y (*ii*) *salen* de la red a través de (al menos) otro nodo. Esto se puede expresar como:

- Existe alguna cola i donde se cumple que $r_i > 0$
- Existe alguna cola j donde no todas las peticiones vuelven al sistema, esto es, la probabilidad de salir desde dicho nodo, denotada como p_j^s, no es cero:

$$p_j^s = 1 - \sum_{k=1, k\neq j}^{N} p_{jk} > 0 \, ,$$

- Desde cualquier cola k existe la posibilidad de alcanzar una cola j donde p_j^s no es cero.[13]

[13] De manera algo más formal: existe al menos un camino posible desde k hasta j, esto es, una secuencia de nodos

$$k, \; k_1, \; k_2, \; \ldots, \; k_n, j$$

tal que

$$p_{k,k_1} > 0, \; p_{k_1,k_2} > 0, \; \ldots, \; p_{k_n,j} > 0.$$

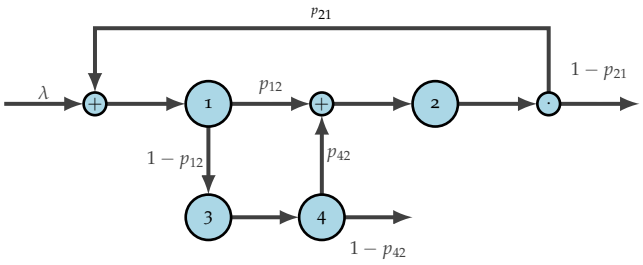


Figura 8.16: Red abierta de colas con cuatro nodos.

Ejemplo 8.12. Sea la red ilustrada en la Figura 8.16. Resulta sencillo comprobar que se trata de una red abierta, dado que hay una fuente externa de tráfico en el nodo 1 (a tasa λ) y, suponiendo $p_{21} < 1$ y $p_{42} < 1$, desde cualquier nodo existe la posibilidad de alcanzar los nodos 2 o 4, que llevan a la salida de dicha red.

La tasa de entrada al nodo 1 es la tasa de entrada al sistema λ más la tasa de salida del nodo 2 que vuelva al sistema, con probabilidad p_{21}:

$$\lambda_1 = \lambda + \lambda_2 p_{21}$$

La tasa de entrada al nodo B viene dada por dos componentes: la tasa de entrada al nodo A que pasa a B con probabilidad p_{12}, y la tasa de entrada al nodo 4 que pasa al nodo 2 con probabilidad p_{42}

$$\lambda_2 = \lambda_1 p_{12} + \lambda_4 p_{42} \ ,$$

donde la tasa de entrada al nodo 4 es la proporción de tráfico que sale desde 1 hacia 3

$$\lambda_4 = \lambda_3 = \lambda_1 (1 - p_{12})$$

UNA VEZ OBTENIDAS las tasas de entrada a cada nodo, si el proceso de llegadas al sistema es de Poisson y los tiempos de servicio son variables aleatorias exponenciales, el teorema de Jackson permite aplicar la misma metodología que en el caso de las redes abiertas acíclicas para analizar el sistema, a pesar de que los procesos de entrada puedan dejar de ser de Poisson.

Teorema de Jackson Sea una red abierta cíclica de N servidores, cada uno con una cola infinita, por la que pasa un tráfico de Poisson. Si los tiempos de servicio son exponenciales y la disciplina de la cola es FCFS, la probabilidad conjunta de que haya n_1 usuarios en la cola 1, n_2 usuarios en la cola 2, etc., se puede obtener como el producto de las probabilidades de que haya n_i usuarios en el sistema i, analizado cada uno de forma independiente:

$$\Pr(n_1, n_2, \ldots, n_N) = P_1(n_1) \cdot P_2(n_2) \cdots P_N(n_N) \ ,$$

donde $P_i(n)$ representa la probabilidad de que haya n usuarios en el sistema i. Además, para calcular esta probabilidad se puede emplear la expresión correspondiente a un sistema M/M/1

$$P_i(n) = \left(\frac{\lambda_i}{\mu_i}\right)^n \left(1 - \frac{\lambda_i}{\mu_i}\right)$$

donde λ_i es la tasa de entrada al servidor i y $1/\mu_i$ su tiempo medio de servicio.

Ejemplo 8.13. Sea el caso de la red de la Figura 8.17, a la que llegan usuarios a una tasa $\lambda = 8$ tramas/ms según un proceso de Poisson. El tiempo de servicio en cada sistema se puede modelar con una variable aleatoria exponencial, de media $t_A = (1/20)$ ms y $t_B = (1/15)$ ms, respectivamente, mientras que $p = 0,2$.

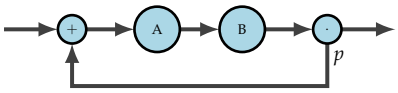


Figura 8.17: Red abierta cíclica.

Analizando la topología de la red, la tasa de entrada al sistema A es la suma de la tasa global de entrada más la parte que sale de B y vuelve a entrar:

$$\lambda_A = \lambda + p \cdot \lambda_B.$$

Además, si ningún sistema está congestionado (lo que se puede comprobar *a posteriori*), la tasa de entrada del sistema A es igual a su tasa de salida, que es la tasa de entrada y salida del sistema B, por lo que

$$\lambda_A = \lambda + p \cdot \lambda_A,$$

de lo que se obtiene que

$$\lambda_A = \frac{\lambda}{1-p} = \frac{8}{1-0,2} = 10 \text{ tramas/ms}.$$

A partir de este resultado se puede obtener la ocupación media de cada recurso

$$\rho_A = \frac{10}{20} = 1/2, \quad \rho_B = \frac{10}{15} = 2/3\ ,$$

el número medio de usuarios en cada sistema

$$N_A = \frac{\rho_A}{1-\rho_A} = 1, \quad N_B = \frac{\rho_B}{1-\rho_B} = 2$$

y el número medio total de usuarios en toda la red:

$$N = N_A + N_B = 3.$$

Una vez obtenido N, por el teorema de Little se puede obtener el tiempo medio en atravesar todo el sistema

$$T = N/\lambda = \frac{3}{8} \text{ ms},$$

que *no coincide* con la suma de los tiempos medios necesarios para atravesar cada sistema T_A y T_B,

$$T_A + T_B = \frac{t_A}{1-\rho_A} + \frac{t_B}{1-\rho_B} = \frac{1/20 \text{ ms}}{1-1/2} + \frac{1/15 \text{ ms}}{1-2/3} = \frac{3}{10} \text{ ms},$$

dado que, por efecto de la re-alimentación, en media un usuario visita los sistemas más de una vez. De hecho, para el caso de la Figura 8.17, el número de veces que un usuario pasa por los dos sistemas es una variable aleatoria geométrica, de media $1/(1-p)$, por lo que para este caso se puede ver que

$$T = \frac{1}{1-p}(T_A + T_B)$$

Modelado de redes de comunicaciones

En el caso de las redes de comunicaciones, donde en general los "usuarios" son tramas de datos y los "recursos" son canales de transmisión (por ejemplo, una red de routers), hay algunas hipótesis de los análisis anteriores que hay que reconsiderar:

1. Las llegadas de peticiones siguen un proceso de Poisson.

Esta hipótesis resulta relativamente razonable, sobre todo si se tiene el agregado de muchos flujos de datos independientes, dado que, en este caso, el agregado *tenderá* a ser un proceso de Poisson por el teorema de Palm-Khintchine

2. Cada petición pasa de un nodo i a un nodo j con una probabilidad independiente (esto es, hay un enrutamiento *probabilístico*).

Esta hipótesis puede resultar algo más discutible, ya que las tramas pertenecientes a un mismo flujo siguen el mismo camino desde el origen al destino.[14] Si se trata de un agregado de varios flujos, cada uno con un destino diferente, cuanto mayor sea la variedad de destinos y más mezcladas estén las tramas, más razonable será realizar esta suposición.

[14] Salvo que aparezca algún tipo de mecanismo de balanceo de carga o las rutas cambien durante un flujo.

3. El tiempo demandado por un usuario en un recurso es una variable aleatoria exponencial.

Modelar un tiempo de transmisión con una variable aleatoria continua, teniendo en cuenta que la longitud de una trama es una variable discreta, será más o menos acertado en función de la distribución de tamaños de las tramas (a mayor tamaño, menor error relativo de redondeo). Además, hay algunos flujos de tráfico donde el tamaño de las tramas dista mucho de seguir dicha variable (p.ej., una descarga TCP, donde básicamente sólo hay segmentos de 1500 B y asentimientos de 40 B). Por último, para poder suponer que dicha variable continua es una variable aleatoria exponencial, habrá que considerar el impacto mecanismos como el protocolo de control de acceso al medio, retardos de procesamiento, etc.

4. El tiempo demandado por cada petición *en cada recurso* es una variable aleatoria independiente.

En una red de comunicaciones, una trama larga siempre necesita más tiempo para ser transmitida que una trama corta, independientemente de que el enlace vaya a 1 Gbps o 10 Mbps. Sin embargo, esta hipótesis (la más discutible de todas) supone que los tiempos de servicio no van asociados nunca al usuario (esto es, a las tramas), sino únicamente al recurso, y son variables aleatorias independientes a cada petición. De hecho, en una red cíclica un mismo usuario puede atravesar el mismo servidor dos veces y necesitará dos tiempos de servicio diferentes. A continuación se presentan las dificultades que supone no hacer esta hipótesis.

Tiempos de servicio asociados a usuarios

En todos los sistemas considerados hasta ahora siempre se ha supuesto que la variable *tiempo de servicio*, aleatoria o no, siempre iba asociada a un recurso, y no a un usuario. Para ilustrar las dificultades de no realizar esta suposición, que resulta algo discutible en redes de comunicaciones, sea el ejemplo de dos sistemas en tándem ilustrado en la Figura 8.18 a continuación:

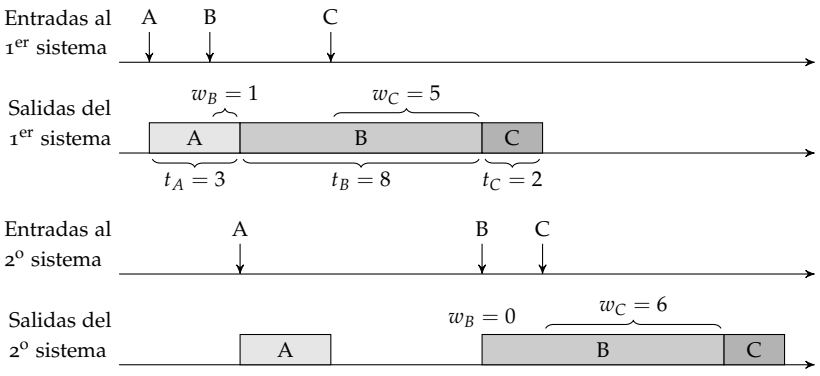


Figura 8.18: Sistema en tándem con tiempos de servicio no independientes.

Según se ilustra en la figura, al primer sistema llega una trama *corta* (tres unidades de longitud, como se indica) seguida de una trama *larga* (ocho unidades) y posteriormente otra trama *corta* (dos unidades), siendo los tiempos entre tramas de dos y cuatro unidades de tiempo, respectivamente. Una vez que pasan dicho sistema, las tramas atraviesan otro sistema idéntico.

Considerando el primer sistema, la segunda y tercera trama tienen que esperar en cola antes de ser transmitidas –sobre todo la tercera trama, dado el tiempo de transmisión de la segunda. Sin embargo, en el segundo sistema la situación cambia para la segunda trama (la trama larga), que ya no tiene que esperar en cola antes de ser transmitida, dado que en el primer sistema su tiempo de transmisión dejó "margen" para que la primera trama fuese enviada. Por otro lado, la tercera trama está "condenada" a tener que esperar en cola a cada sistema que pase, ya que llega mientras la segunda trama está siendo transmitida.[15]

Como ilustra este ejemplo, cuando los tiempos de servicio están asociados a los usuarios (y no a los recursos) aparecen relaciones entre dichos tiempos y los tiempos de estancia en el sistema, lo que dificulta el modelado de redes de comunicaciones –en contraposición con, por ejemplo, dos sistemas M/M/1 en tándem, que se pueden analizar de forma independiente. Para poder analizar cada sistema por separado, es preciso realizar una aproximación de independencia.

[15] La situación guarda cierta analogía con una carretera de un carril, donde un vehículo largo (típicamente, lento) no tiene vehículos por delante, pero sí una cola que le sigue.

Aproximación de independencia de Kleinrock

Según visto en el ejemplo anterior, en una red de comunicaciones aparece correlación entre tiempos de servicio y los tiempos de

espera en cola: en general, los paquetes largos tienden a no esperar en cola, mientras que los paquetes cortos sí que esperan. Una forma en la que esta correlación puede disminuir es que se produzcan *agregados* de flujos de diferentes fuentes, de modo tal que las tramas de dichos flujos se intercalen salto a salto y así se "rompa" esta dependencia. Esta disminución en la correlación será mayor cuanto mayor sea el volumen de tráfico "nuevo" frente al tráfico que ha pasado por un sistema.

Ejemplo 8.14. Sea el caso de la Figura 8.19 con dos enlaces. Al primer enlace llega un flujo a tasa 50 tramas/ms, y a continuación el 40 % de dicho tráfico se agrega con otro flujo a tasa 80 tramas/ms antes de atravesar el segundo enlace.

En este caso, el 40 % de salida del primer enlace supone un flujo a 20 tramas/ms, que se agrega a un flujo "nuevo" a 80 tramas/ms. Dada esta gran proporción del flujo nuevo sobre el anterior, será razonable suponer que no existirá mucha dependencia entre los tiempos de servicio y de espera.

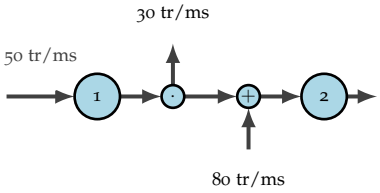


Figura 8.19: Agregado y diezmado de flujos en una red de comunicaciones.

Fue Leonard Kleinrock[16] quien sugirió que el agregado de flujos en una red de conmutación de paquetes puede debilitar la correlación entre longitudes de tramas y tiempos de espera, esto es, "restaurar" la suposición de independencia en los tiempos de servicio de cada usuario. De ahí el nombre de dicha aproximación (que no es un teorema que establezca de forma rigurosa las condiciones cuando se puede aplicar), que se puede definir de la siguiente forma

[16] Ingeniero neoyorquino (n. 1934), pionero en el análisis matemático de las prestaciones de redes de conmutación de paquetes: se considera que su tesis doctoral (MIT, mayo 1961) estableció los fundamentos matemáticos de dicho campo.

Aproximación de Kleinrock En una red de comunicaciones con varias líneas de transmisión conectadas, en la que los tiempos de servicio no son independientes en cada enlace, el agregado de una cantidad *suficiente* de flujos de datos de diferentes fuentes tiene un efecto similar a restaurar la independencia entre los tiempos de llegada y la longitud de tramas.

Por lo tanto, si cada enlace se considera como un recurso, las llegadas son de Poisson y los tiempos de servicio se pueden aproximar con una variable aleatoria exponencial, esta aproximación permite analizar las prestaciones de la red como si cada enlace se pudiese modelar como un sistema M/M/1 independiente. Cuanto más densamente conectada esté la red y, por tanto, más flujos diferentes sean agregados en cada enlace, mejor será la aproximación (y, por motivos similares, la aproximación será mejor para cargas de tráfico medias-altas).

Ejemplo 8.15. Sea la red de comunicaciones de la Figura 8.20 en la que $\lambda_1 = 3$ tramas/ms, $\lambda_2 = 4$ tramas/ms, $p_1 = p_2 = 1/4$ y la longitud de todas las tramas se puede modelar con una variable aleatoria exponencial de media 1000 bytes.

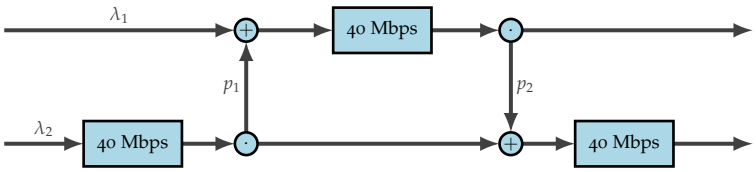


Figura 8.20: Red abierta.

La tasa máxima de salida de cada enlaces es

$$\mu = \frac{40 \text{ Mbps}}{1500 \cdot 8} = 5 \text{ tramas/ms}$$

El primer enlace inferior se puede analizar sin mayor consideración, dado que el proceso de llegadas λ_2 es externo a la red.

Al enlace superior llega el agregado de un flujo externo (λ_1), y el 25 % de un flujo que ya ha pasado por el enlace inferior izquierda (λ_2), y que supone un 25 % del flujo total (por lo que su peso relativo es bajo). Además, el flujo total por el enlace es de

$$\lambda_1 + \frac{1}{4} \cdot \lambda_2 = 4 \text{ tramas/ms},$$

lo que supone una ocupación relativamente alta, $\rho = \lambda/\mu = 4/5$, por lo que, por la aproximación de Kleinrock, también podría analizarse como un sistema tradicional.

Por último, el segundo enlace inferior recibe el 75 % del flujo λ_2 tras pasar por un enlace anterior, y el 25 % del flujo del anterior enlace. Aún suponiendo que el routing fuese aleatorio, en valores absolutos supone el agregado de dos flujos a 3 tramas/ms y 1 trama/ms, respectivamente, no siendo ninguno de ellos externo a la red: sería preciso ser cuidadosos a la hora de aplicar la aproximación de Kleinrock en este enlace.

Resumen del tema

- En un sistema M/G/1 el tiempo medio de espera en cola viene dado por
$$W = \frac{\lambda \mathbb{E}[t_s^2]}{2(1-\rho)},$$
- En un sistema con prioridades, el tiempo de espera en cola de la clase k viene dado por
$$W_k = \frac{\frac{1}{2}\sum_{i=1}^{K} \lambda_i \mathbb{E}[t_i^2]}{(1-\rho_1-\ldots-\rho_{k-1})(1-\rho_1-\ldots-\rho_k)}$$
- En una red de colas abierta acíclica, la probabilidad conjunta de $(n_1, n_2, \ldots)$ usuarios en el sistema $1, 2, \ldots$ viene dada por el producto de las probabilidades de $n_1, n_2, \ldots$ usuarios en cada uno de los sistemas, analizados estos de forma independiente
$$\Pr(n_1, n_2, \ldots, n_N) = P_1(n_1)P_2(n_2)\cdots P_N(n_N)$$

- En una red de colas abierta cíclica también se puede calcular la probabilidad conjunta mediante el análisis de cada sistema por separado, empleando las expresiones vistas en el tema anterior.
- En una red de comunicaciones hay que ser cautos a la hora de modelar el sistema. Si el proceso de llegadas es de Poisson y la carga y conectividad de la red son relativamente altas, la aproximación de Kleinrock permite emplear los resultados anteriores.