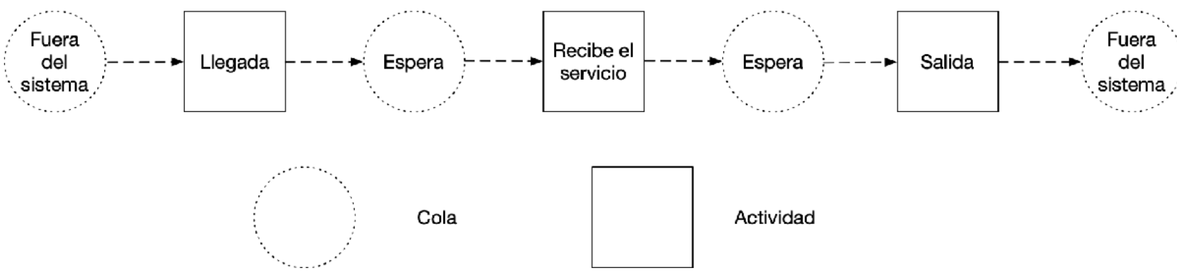
Distribuciones de probabilidad para DEVS

Distribuciones de probabilidad

Generalidades

Supongamos una cola cualquiera que se forma para recibir un servicio. Podemos caracterizar este proceso con un esquema como el de la figura:



Si nos fijamos en las actividades, «Llegada», «Recibe el servicio», «Salida», nos daremos cuenta de que es muy difícil establecer una regla por la que regir dichas actividades. Es decir, no es fácil predecir ni la llegada de clientes a la cola, ni el tiempo de servicio ni la salida de los clientes del servicio. Observamos, pues, cierto azar o aleatoriedad en estas actividades.

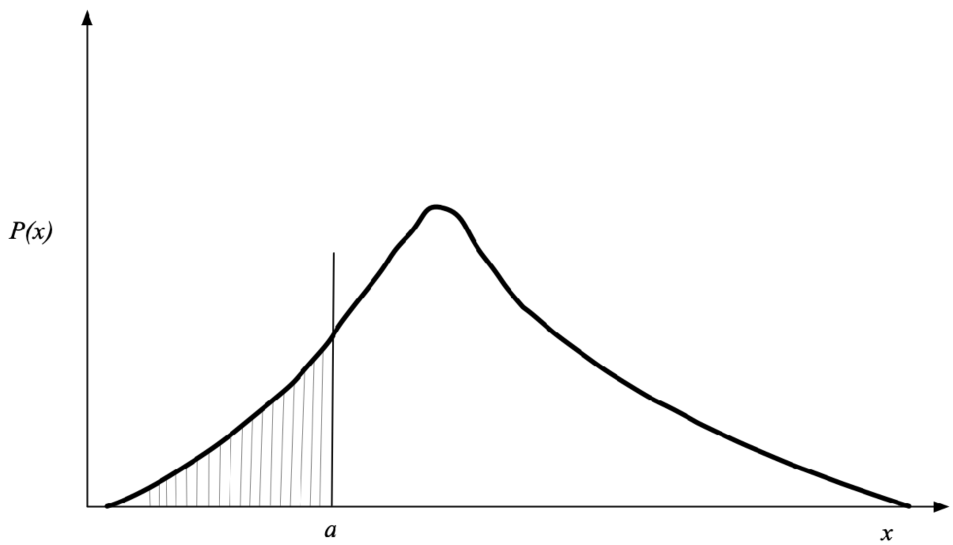
En la simulación discreta tenemos una ventaja. Muy rara vez querremos saber el tiempo exacto que una determinada entidad (persona, producto, repuesto, pieza, ...) pasa en una actividad. La información que buscamos en los procesos de eventos discretos es el *patrón medio de comportamiento* para algunos de los componentes del modelo. Esto significa que los patrones individuales de comportamiento no son importantes, especialmente si el número de entidades a considerar es muy grande. También significa que somos libres en elegir la variabilidad del muestreo que vamos a elegir para nuestro modelo. En otras palabras, somo libres de utilizar una distribución de probabilidad que mejor se ajuste a nuestro problema atendiendo a sus características especiales.

En este capítulo vamos a hacer un repaso de las Distribuciones de Probabilidad más comunmente usadas para modelar las actividades de nuestro modelo. Podríamos decir que la elección de una u otra, con sus parámetros asociados, no es una labor fácil.

Depende de muchos factores, aunque quizá el más significativo sea la experiencia. Para esto, la simulación es nuestra mejor herramienta, ya que nos permitirá elegir una u otra distribución y observar los resultados, compararlos, sacar conclusiones, presentarlos al cliente, confrontarlos con la realidad y por último, elegir.

Distribuciones para DEVS

La siguiente figura muestra una *Función de Densidad de Probabilidad* (PDF, en inglés). Ella indica que la probabilidad de que un determinado valor x de la variable ocurra, viene dado por $P(x)$ en cualquier punto. La probabilidad de que un valor x ocurra entre dos valores cualquiera a y b, está dada por el área bajo la curva $P(x)$ entre esos dos puntos. La probabilidad de que x ocurra entre $-\infty$ y $+\infty$ viene dada por el área total bajo la curva $P(x)$, y ya que ello trae consigo la totalidad de los valores de x, la probabilidad será de 1.0. La PDF no es muy útil en términos de las aplicaciones en simulación de sistemas discretos.



Lo que para nosotros es más útil es la *Función de Densidad Acumulada* (CDF, en inglés). Para un valor a, dentro de todos los posibles valores de x, la CDF representa la probabilidad total de que x se encuentre en algún lugar dentro del rango $(-\infty, a]$. En términos matemáticos:

$$C(x)\,|_{x=a} = \int_{-\infty}^{a} P(x)\ \mathrm{dx}$$

Distribución uniforme

Esta distribución puede aparecer, por ejemplo, en el estudio de los errores de redondeo cuando las medidas son tomadas con cierto grado de precisión. Si tuviéramos que tomar datos de pesos redondeados al gramo más próximo, sería razonable asumir que el error se encontraría entre +0,5 gramos y -0,5 gramos distribuido de una manera uniforme a lo largo de este intervalo.

Para una distribución uniforme, la probabilidad de que un valor se encuentre en medio del intervalo de valores $[a, b]$, está dada por la PDF:

$$P(x) = \tfrac{1}{b-a}, \text{para } a \leq x \leq b$$
$$P(x) = 0 \text{ en otro caso}$$

y la CDF:

$$C(x) = \int_a^x \frac{\mathrm{du}}{b-a} = \frac{1}{b-a} |u|_a^x = \frac{x-a}{b-a}, \quad \text{con } (0 \leq C(x) \leq 1)$$

Una aplicación mucho más importante de la función de distribución uniforme es la generación de números aleatorios. En simulación dicha generación es de una importancia capital, ya que son los que verdaderamente van a introducir el azar en el modelo. Si la probabilidad de cada número generado dentro del intervalo es verdaderamente aleatoria, la PDF tendrá la forma de la figura (a), normalmente en el intervalo [0,1]. Si quisieramos que la generación de números aleatorios fuera en otro rango distinto, digamos $[a, b]$,bastaría con hacer

$$x_1 = a + (b - a) R$$

siendo R el número aleatorio generado en el intervalo $[0, 1]$.

La interpretación gráfica de la CDF se muestra en la figura (b).

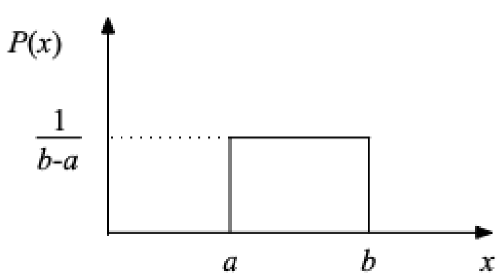


Figura (a)

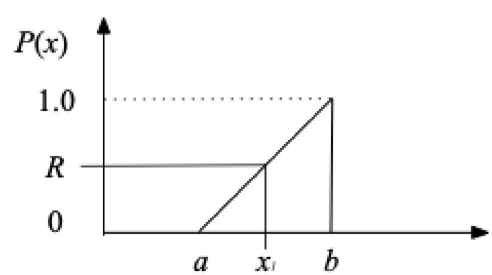


Figura (b)

Toda distribución está caracterizada por su media (μ)y su desviación típica (σ)o varianza (σ^2):

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

Distribución binomial

Esta distribución es muy útil en los procesos de control de calidad, en los que los elementos muestreados tienen dos valores de atributos (si/no, con defecto/sin defecto, etc.) La distribución binomial es una distribución discreta que nos da un indicador de que la probabilidad de un evento o un «éxito» ocurra x veces después de n ensayos, donde la probabilidad general de un éxito es p, de lo que se sigue que la probabilidad general de fracaso o «no éxito» es $q = 1 - p$.

La PDF de esta distribución está dada por:

$$P\left(x\right) = \left[\begin{array}{c} n \\ x \end{array}\right] p^x q^{n-x} \ \text{ donde } \ (x = 0, 1, 2, 3, \ldots, n)$$

Su media y varianza son:

$$\begin{array}{l} \mu = n\,p \\ \sigma^2 = n\,p\,q \end{array}$$

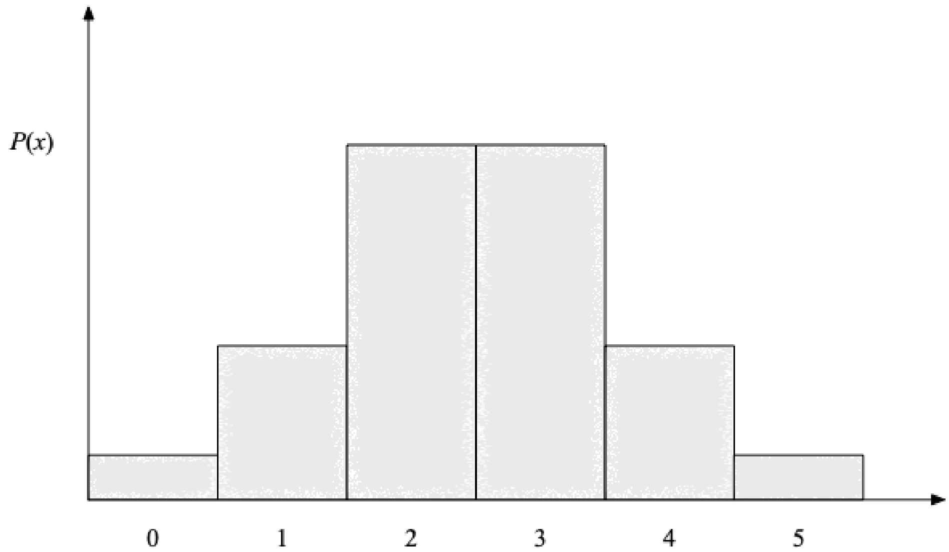
Pongamos un ejemplo cuando tenemos una situación de «defecto/sin defecto» con una probabilidad $p = \frac{1}{2}$ y una probabilidad $q = 1 - p = \frac{1}{2}$. Entonces, usando la ecuación general de la distribución binomial, tenemos una probabilidad de obtener $0, 1, 2, \ldots n$ items sin defecto después de un muestreo de n items de

$$q^n, \left[\begin{array}{c} n \\ n-1 \end{array}\right] q^{n-1} p, \left[\begin{array}{c} n \\ n-2 \end{array}\right] q^{n-2} p^2, \ldots, p^n$$

Ya que es una distribución discreta, es más apropiado representar esta distribución como un histograma. Para el caso de una muestra de 5 items, la probabilidad de obtener $0, 1, 2, 3, 4, 5$ sin defecto viene dada por las probabilidades:

$$\frac{1}{32}, \frac{5}{32}, \frac{10}{32}, \frac{10}{32}, \frac{5}{32}, \frac{1}{32}$$

cuyo histograma será:



Distribución de Poisson

Esta distribución es una de las más importantes en la simulación discreta, no tanto en términos de aplicación directa, sino más en virtud de la aplicación de su denominada distribución «inversa», la distribución exponencial negativa, como veremos más tarde. Mientras tanto, podemos pasar a considerar qué tipos de fenómenos se ajustan a una distribución de Poisson, qué es esta distribución y cómo podemos derivarla de una manera simple.

En física nuclear se realiza un experimento para determinar la frecuencia de detección de partículas en la radiación de fondo del universo. Analizada esa frecuencia nos damos cuenta de que sigue una distribución de Poisson. Es interesante analizar porqué debería ser así. Una de las características de esta radiación es el hecho de que la llegada de una partícula individual es algo que tiene probabilidad de ocurrencia extremadamente pequeña. Además, a lo largo del tiempo del experimento, se habrían registrado un gran número de detecciones. La conclusión que podemos sacar de este fenómeno es que, en este experimento, registraríamos un gran número de llegadas de partículas, cada una de las cuales tiene una probabilidad muy pequeña de ocurrir dentro del marco de tiempo del experimento.

Parece haber algún tipo de conexión con la distribución binomial y veamos el camino para obtener la distribución de Poisson.

Consideremos el flujo de tráfico de una carretera observado desde una cámara de tráfico situada en un punto determinado. Supongamos que ese flujo tiene un valor promedio de sólo un vehículo por minuto, con vehículos que llegan al azar y sin ningún tipo de relación entre ellos. En breves espacios de tiempo, como algunos minutos sucesivos, no esperaríamos que llegaran exactamente el mismo número de vehículos, y mucho menos $1/60$ vehículos en cada segundo sucesivo. Sin embargo es posible que uno o más vehículos lleguen en un segundo en particular.

Si mantenemos que la probabilidad de que llegue un vehículo por segundo es $1/60$, entonces en el periodo de un minuto hay 60 intentos de un evento con una probabilidad de $1/60$ o 120 intentos de un evento con una probabilidad de $1/120$. Lo que tenemos aquí, por tanto, es una situación en la que el número de intentos del evento multiplicado por la probabilidad de una llegada durante cada periodo de prueba es constante. Podríamos utilizar la distribución binomial estudiada en la sección anterior para calcular las probabilidades de $0, 1, 2, \ldots$vehículos por minuto mediante las fórmulas de las ecuaciones de dicha sección:

$$q^n, \begin{bmatrix} n \\ n-1 \end{bmatrix} q^{n-1}p, \begin{bmatrix} n \\ n-2 \end{bmatrix} q^{n-2}p^2, \ldots, p^n$$

Ahora podemos escribir:

$$\begin{aligned} \ln(q^n) &= n \ \ln \ (q) \\ &= n \ \ln(1-p) \\ &= -n \ p \ \ (\text{si } p \text{ es muy pequeño}) \end{aligned}$$

Esto significa que, ya que $n\ p$ es constante, cuando $n \to \infty, p \to 0$. Por tanto

$$q^n \to e^{-n\,p} = e^{-\mu}$$

donde μ es el valor medio o esperado. La distribución binomial tiende en el límite a la distribución de Poisson. Esto reflejaría los tipos de condiciones presentes en la situación de flujo de tráfico en la que un número promedio de incidentes (llegadas al punto de observación) es el resultado de una gran cantidad de oportunidades para que ocurra la llegada. Este gran número resulta de la subdivisión de un periodo de observación en intervalos de tiempo muy pequeños. Esto, a su vez, significa que la probabilidad de que tenga lugar una llegada dentro de uno de esos periodos de tiempo de minutos es muy pequeña. Esta situación, como se recordará, es exactamente la que caracterizó el experimento de conteo de partículas de radiación.

Dado que cuando $p \to 0, q \to 1$,

$$\sigma^2 = n\ p\ q = n\ p = \mu$$

Por tanto, la distribución de Poisson nos dice que la probabilidad de $0, 1, 2, \ldots, n$ exitos es:

$$e^{-\mu}, \mu e^{-\mu}, \frac{\mu^2}{2!}e^{-\mu}, \frac{\mu^3}{3!}e^{-\mu}, \ldots, \frac{\mu^n}{n!}e^{-\mu} \qquad (1)$$

donde μ es igual a la media e igual a la varianza.

Esta característica de que la media y la varianza sean iguales es la característica principal y más importante de la distribución de Poisson.

Para poder utilizar este resultado en una simulación discreta, necesitamos expresar la distribución de Poisson de una manera diferente, en términos de un parámetro que llamaremos *tasa de llegada* (que podemos asimilar a la frecuencia). Si llamamos a esta tasa de llegada λ, entonces podemos expresar el intervalo entre llegadas correpondiente como $\frac{1}{\lambda}$ (que se puede asimilar al periodo).

Si contamos una sola llegada como uno de los éxitos mencionados anteriormente para la distribución de Poisson, entonces el número medio de llegadas en un intervalo de tiempo $t = \lambda\ t = \mu$ (que es la media de la distribución de Poisson). Por lo tanto, podemos reescribir el último término en la ecuación (1) para darnos la probabilidad de que ocurran x llegadas en un intervalo de tiempo t como la ecuación:

$$P\left(x,t\right) = \left[\frac{\left(\lambda t\right)^{x}}{x!}\right] e^{-\lambda t} \qquad (2)$$

Esta relación se mantiene si podemos satisfacer las siguientes condiciones:

- Los eventos ocurren de manera completamente independiente entre sí. Esto significa que, dentro de un intervalo de tiempo, el número de llegadas que tienen lugar no está influenciado de ninguna manera por llegadas anteriores en intervalos de tiempo anteriores.
- La tasa media de llegada λ permanece constante. Este es el concepto de una distribución estacionaria.

El muestreo de una distribución de Poisson dentro de una simulación no es una cosa fácil de hacer. Sin embargo, hay una forma de solucionar el problema si presuponemos que lo siguiente es cierto: Si de alguna manera somos capaces de muestrear una distribución para darnos el intervalo entre llegadas como una variable, entonces varios de esos intervalos dentro de un período de tiempo t como se usa en la ecuación (2) constituirán una muestra x de una distribución de Poisson cuya media o el valor esperado es λt. Al comienzo de la sección, se mencionó que la distribución exponencial negativa es la «inversa» de la distribución de Poisson. Si las dos condiciones anteriores para la distribución de Poisson son verdaderas, entonces se puede demostrar que la PDF del intervalo de tiempo t entre la ocurrencia de llegadas sucesivas está dada por:

$$P\left(t\right) = \lambda e^{-\lambda t} \qquad (3)$$

mientras que en la distribución de Poisson correspondiente dada por la ecuación (2), tenemos una expresión para la PDF de la probabilidad de que ocurran x llegadas dentro del intervalo de tiempo t. De esta manera, tenemos una función como la inversa de la otra. Como veremos en la siguiente sección, la distribución exponencial negativa es, con mucho, la más fácil de implementar dentro de una simulación por ordenador y, por lo tanto, las variables de Poisson deben generarse como se describió anteriormente a partir de una serie de variables exponenciales negativas.

Distribución exponencial negativa

Si la tasa de ocurrencia de llegadas tiene una distribución de Poisson, entonces se puede demostrar que la PDF del intervalo de tiempo t entre llegadas sucesivas está dada por (3). Esta ecuación, sin más, es la conocida como distribución exponencial negativa (o simplemente, distribución exponencial).

La CDF correspondiente es la probabilidad total acumulada de que todos los intervalos de tiempo generados se encuentren entre los tiempos 0 y t, y está dada por:

$$C\left(t\right) = \int_0^t \lambda e^{-\lambda t}\,\mathrm{dt} = \left[-e^{-\lambda t}\right]_0^t = 1 - e^{-\lambda t} \qquad (4)$$

Es interesante intentar visualizar lo que esto significa en la práctica. Si miramos la figura (a), podemos ver que la PDF para la distribución exponencial negativa es una exponencial decreciente que intercepta el eje Y en un valor λ, la tasa media de llegada. Lo que representa la curva es la probabilidad de que un intervalo de un tamaño determinado t encaje en todos los intervalos que se generan mediante el muestreo.

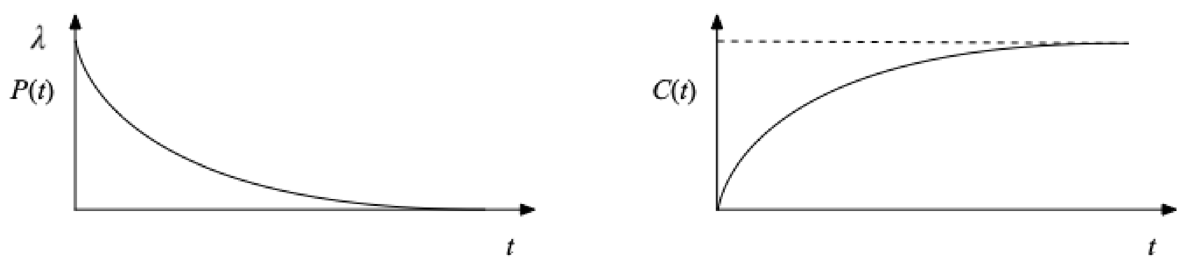


Figura (a) Figura (b)

Un intervalo de duración cero se ajustará a todos los intervalos generados; por lo tanto, la PDF para esto es la misma que la tasa de llegada media, λ. Cuando aumentamos el tamaño de nuestro intervalo de «prueba» más allá de cero a cualquier valor t, ya no se ajusta a todos los intervalos de tiempo generados, y cuanto más larga sea la longitud del intervalo, menor es la posibilidad de encajar en una de las muestras. Un intervalo de tiempo cero tendrá la mayor probabilidad (λ) de encajar entre llegadas sucesivas. Esta probabilidad disminuye exponencialmente a medida que aumenta el tamaño del intervalo de tiempo. La figura (b) muestra la CDF para la distribución exponencial negativa. Continuando con nuestra analogía de un intervalo de «prueba», $C\left(t\right)$ representa la probabilidad total acumulada de que todos los intervalos generados encajen en nuestro intervalo de prueba de tamaño t. A medida que aumenta el tamaño del intervalo, $C\left(t\right)$ se acerca asintóticamente a un valor de 1.

Podemos usar el mismo método para extraer una muestra de una distribución exponencial negativa que usamos para las variables uniformes. Si tomamos muestras de una distribución uniforme con un rango entre 0 y 1 (es decir, usamos un generador de números aleatorios para hacer esto por nosotros), entonces tenemos un valor que es igualmente probable que ocurra dentro del rango de 0 a 1 y lo podemos representar por un valor R en el eje $C\left(t\right)$ de la figura (c).

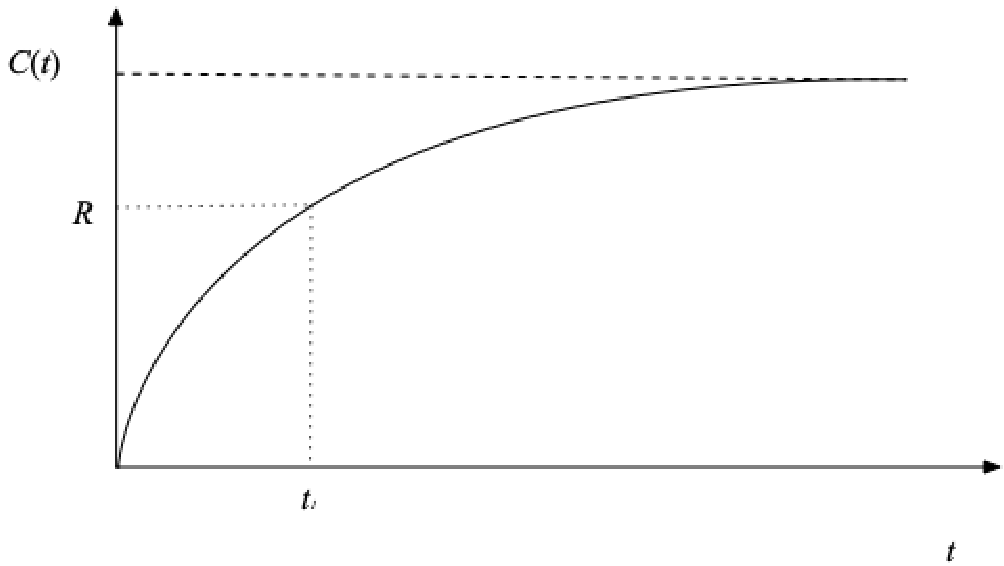


Figura (c)

El valor correspondiente de t producido por la construcción que se muestra en la figura es un intervalo de tiempo t_1 que es una muestra de la distribución exponencial negativa. Dado que R puede ocurrir en cualquier lugar del eje $C\left(t\right)$ con la misma probabilidad, la forma de la CDF asegura que es más probable que se generen intervalos más pequeños que los más grandes. El cincuenta por ciento de las muestras se producirá en el extremo inferior del rango, mientras que el cincuenta por ciento restante abarcará un amplio rango de posibles intervalos de tiempo. En términos computacionales, es muy fácil extraer muestras de una distribución exponencial negativa.

Si generamos un número aleatorio R donde $0 \leq R \leq 1$ entonces

$$\begin{aligned}
&C\left(t\right) = R = 1 - e^{-\lambda t} \\
&e^{-\lambda t} = 1 - R \\
&\lambda t = -\ln\left(1 - R\right)
\end{aligned}$$

Ya que R está uniformemente distribuido en el intervalo $0 \leq R \leq 1$, $(1 - R)$ también es uniforme en dicho intervalo. Por tanto, tendremos el mismo efecto que usar directamente la expresión:

$$\lambda t = -\ln\left(R\right)$$

de la cual podemos extraer la relación de trabajo:

$$t = -\frac{1}{\lambda} \ln (R)$$

donde t es un intervalo de tiempo muestreado de una distribución exponencial negativa.

Si llamamos μ a la *tasa de servicio* del servidor, el tiempo teórico de estancia en la cola de una entidad es:

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

y el *tiempo en el servicio* de una entidad será:

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

Distribución normal o gaussiana

Se trata de una distribución simétrica continua muy utilizada en el modelado y simulación por ordenador. Podemos usarlo para describir la variación aleatoria en la mayoría de los fenómenos de medición, por ejemplo:

- Errores de medición angulares o lineales (no es lo mismo que los errores de redondeo que están distribuidos uniformemente)
- Variaciones en las calificaciones de los exámenes grupales
- Variabilidad en atributos humanos como la altura o el peso corporal para grupos dados.

El PDF de la distribución normal viene dado por la fórmula de aspecto bastante intimidante:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty < x < +\infty)$$

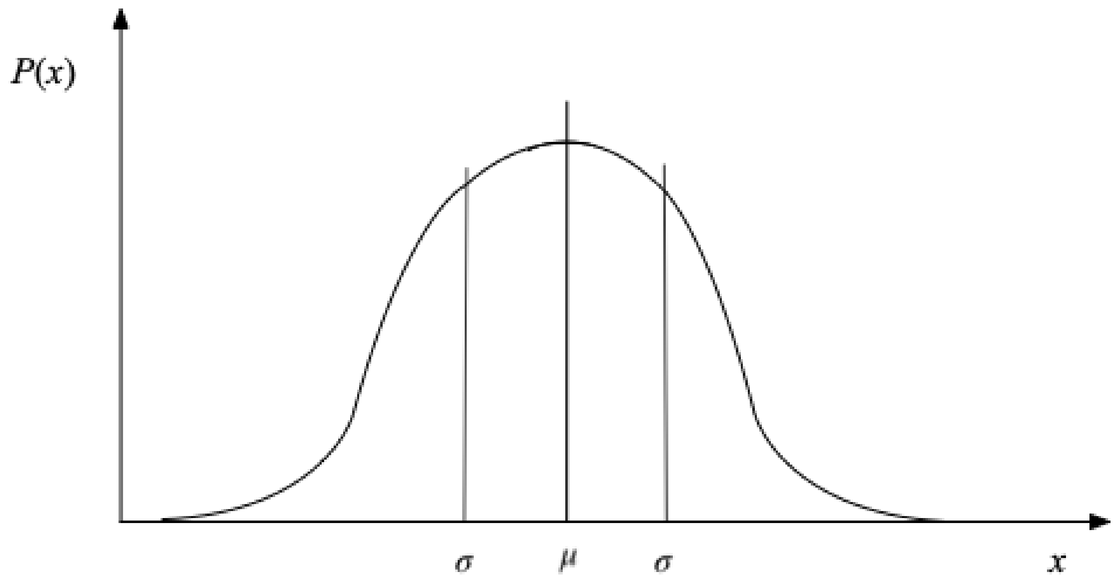
donde

μ = media o valor esperado.

σ = desviación típica.

σ^2 = varianza.

La figura siguiente muestra la forma general de la distribución normal, siendo el valor medio también el eje de simetría de la curva. Las líneas de desviación estándar se muestran para indicar el efecto que tendrá el valor de σ en la forma de la curva. Los valores altos de σ harán que la curva se extienda un poco, mientras que los valores pequeños producirán una forma de campana estrecha.



Podemos simplificar la fórmula PDF hasta cierto punto dando valores especiales a la media y la desviación estándar. Para $\mu = 0$ y $\sigma = 1$, podemos escribir una fórmula conocida como *distribución normal estándar* dada por

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (-\infty < z < +\infty)$$

Podemos convertir cualquier distribución normal en una distribución estándar sin más que hacer el cambio o sustitución:

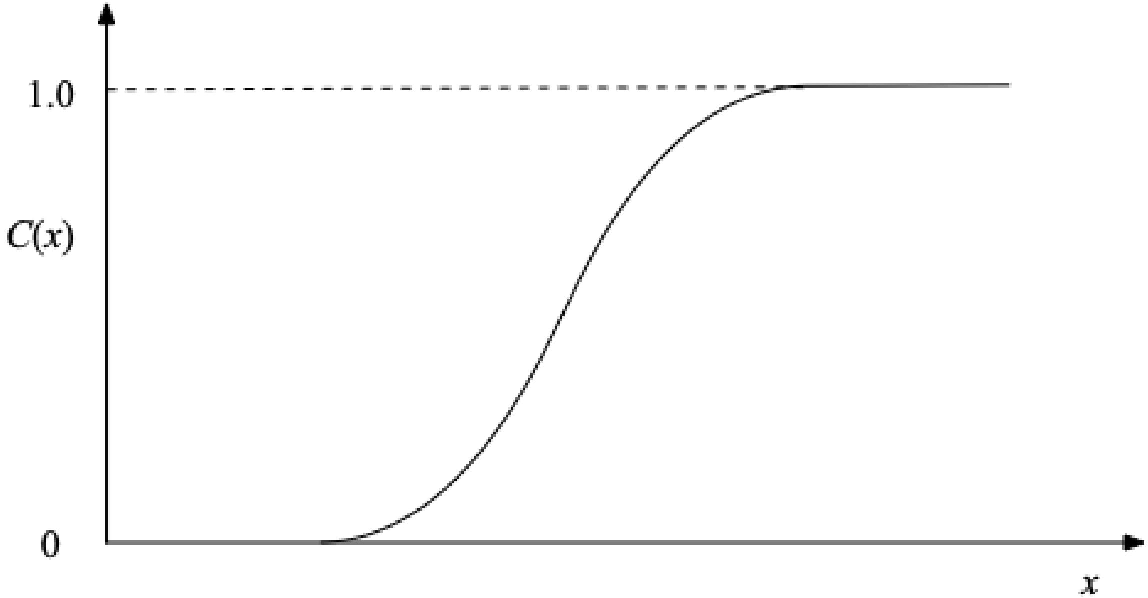
$$z = \frac{x - \mu}{\sigma}$$

La CDF de una distribución normal viene dada por

$$C(x) = \int_{-\infty}^{x} P(x)\,\mathrm{dx} =$$

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{dx}$$

Que tiene la forma que se muestra a continuación:



En términos computacionales, no es fácil tratar de generar la variable normal x (o z) por transformación inversa usando la curva CDF. En lugar de este complicado procedimiento, podemos usar un método más simple basado en reunir un grupo de variables uniformes (muestras de una distribución uniforme). Se puede demostrar que si se genera tal serie de variables uniformes R_k, se puede derivar una variable normal estándar z en un grado cercano de aproximación, usando la fórmula

$$z = \frac{\sum_{k=1}^{m} R_k - \frac{m}{2}}{\sqrt{\frac{m}{12}}}$$

si $m = 12$, entonces

$$z = \sum_{k=1}^{12} R_k - 6$$

Si necesitáramos convertir este muestreo en una distribución normal no estándar, tendremos que hacer la sustitución:

$$x = \mu + \sigma z$$