Tema 5

Modelos de distribuciones continuas

En este capítulo seguiremos estudiando los modelos probabilísticos más importantes que aparecen en computación. Como es de esperar, el método se va a basar en el aprendizaje por indagación por parte del alumno —con la ayuda de un profesor inasequible al desaliento—.

5.1 Distribución uniforme continua

Problema 5.1.1 Supongamos que tenemos una variable aleatoria que solo toma valores en el intervalo (a, b). Sabemos que la variable aleatoria asigna probabilidades a los intervalos en (a, b) de manera proporcional a la longitud del intervalo. Obtened su función de densidad. ¿En qué contextos generales puede surgir esta situación?

Ejercicio 5.1.2 Para la función de densidad del problema anterior:

(a) Hallad su función de distribución.

(b) Hallad $E(X), \sigma^2, CV$ y la mediana.

Problema 5.1.3 El tiempo de acceso a un registro de una base de datos, en segundos, sigue una distribución $U(0, b)$. Sabiendo que el 10% de los accesos tarda 2 segundos o menos, calculad:

(a) El valor de b;

(b) La probabilidad de que dicho tiempo esté entre 5 y 20 segundos;

(c) El tiempo total esperado de acceso a 100 registros, situados al azar, y la varianza de dicho tiempo;

(d) Probabilidad de que el tiempo total de acceso a 100 registros sea mayor que 1100 segundos.

5.2 Distribución exponencial

La distribución exponencial, como era de esperar por su inclusión aquí, modeliza muchas situaciones interesantes que surgen en computación. Una enumeración exhaustiva sería prolija,

pero déjenos el lector ofrecer un ejemplo rápido. Dentro de la Ingeniería del *Software*, en concreto en el campo de la fiabilidad del *software*, esta distribución se emplea con harta frecuencia. Por ejemplo, la NASA la usó (y usa) para estimar las tasas de error de los sistemas informáticos de sus transbordadores espaciales. En un experimento tomaron datos de 200 horas de vuelo de misiones espaciales y los ajustaron al modelo exponencial. El ajuste fue particularmente bueno, como se puede ver en la figura 5.1[1].

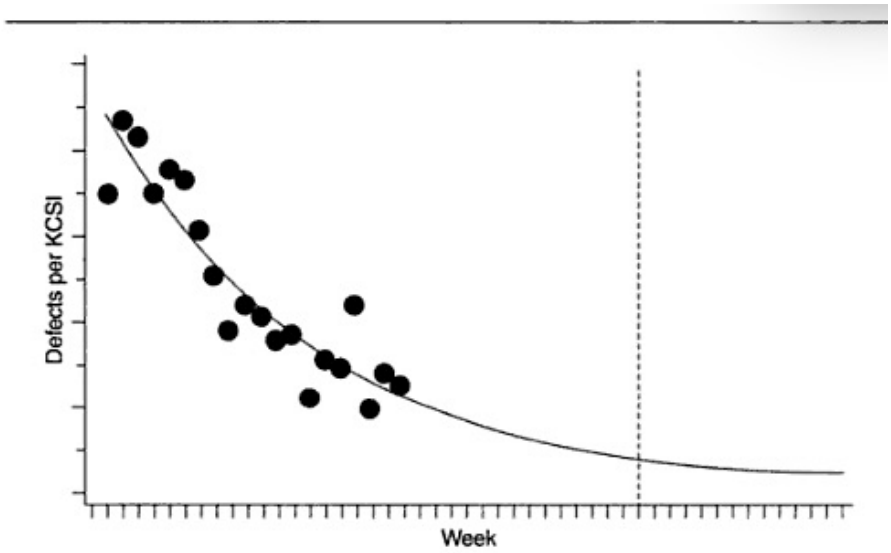

Figura 5.1: Ajuste de datos en fiabilidad de software con la distribución exponencial

Otros ejemplos de fenómenos modelizados por la distribución exponencial son los siguientes:

(a) Tiempo entre fallos consecutivos de un sistema *software* o *hardware*;

(b) Tiempo entre peticiones a un servicio informático (un servidor, una página web, etc.);

(c) Tiempo de servicio en una cola;

(d) Tiempo entre dos mutaciones en una hebra de ADN;

(e) Tiempo hasta que un núcleo radioactivo deja de ser perjudicial para el ser humano;

(f) Tiempo de fallo de dispositivos (bombillas, discos duros, preservativos, etc.);

(g) Duración de una llamada telefónica.

Procedamos, pues, a la deducción de la variable aleatoria exponencial. Como el lector habrá advertido, los ejemplos anteriores se refieren al tiempo entre dos fenómenos o a tiempo hasta que ocurre un cierto suceso. En efecto, la variable exponencial es continua —el tiempo es una cantidad continua por antonomasia—. La definición de la variable exponencial parte de un experimento de Poisson. Supongamos que X es la variable aleatoria número de sucesos por unidad de tiempo de cierto fenómeno y que además $X \sim P(\lambda)$. Definamos la variable T tiempo

[1]Este ejemplo está tomado del libro de *informática* (que no de uno de aburridas, innecesarias y tediosas matemáticas) *Metrics and Models in Software Quality Engineering* de Stephen H. Kan

entre dos sucesos consecutivos en el experimento de Poisson. Esta variable solo toma valores en el intervalo $(0, \infty)$. Empezamos por observar un suceso y contamos el tiempo hasta que ocurra el siguiente suceso. Consideremos la probabilidad de que en la primera unidad de tiempo no se produzca otro suceso; dicha probabilidad es $P(T \geq 1)$. Su relación con la variable X es

$$P(T \geq 1) = P(X = 0) = e^{-\lambda}$$

Según la definición de experimento de Poisson (definición 4.4.1) el número medio de sucesos por unidad de tiempo es constante y estos ocurren de manera independiente. Esto significa que la probabilidad del suceso $\{T \geq 2\}$ es la probabilidad de que se observen cero sucesos en una variable de Poisson $X_2 \sim P(2\lambda)$, número de sucesos observados en dos unidades de tiempo. Entonces, tenemos que:

$$P(T \geq 2) = P(X_2 = 0) = e^{-2\lambda}$$

Esta igualdad es cierta en virtud de la reproductividad de la distribución de Poisson (teorema 4.4.7).

Siguiendo este razonamiento, concluimos que para t unidades de tiempo, podemos escribir

$$P(T \geq t) = P(X_t = 0) = e^{-t\lambda}$$

donde X_t es una distribución de Poisson $P(\lambda t)$ que es el número de sucesos observados en t unidades de tiempo. A partir de esta última expresión se puede obtener la función de distribución de la variable T:

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

donde $t > 0$. Si ahora derivamos esta función, obtenemos la función de densidad:

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Si una variable X sigue una distribución exponencial se escribe $X \sim \exp(\lambda)$. En la figura 5.2 tenemos ejemplos de función de densidad de la exponencial para varios valores del parámetro λ. La función $P(T \geq t)$ se llama **función de supervivencia** o **función de fiabilidad**, dependiendo del contexto y la aplicación concretos.

Ejercicio 5.2.1 Probad que, en efecto, la función $f(t)$ anterior es una función de densidad. Un repaso de integrales impropias puede ser oportuno aquí.

Por razones de claridad y notación, no se usa el parámetro λ, sino β (más adelante se entenderá por qué) y se suele escribir $X \sim \exp(\beta)$, convención que seguiremos nosotros también.

Ejercicio 5.2.2 Sea $X \sim \exp(\beta)$. Hallad $E(X), \sigma^2$ y CV. Analizad el significado de estos momentos.

Problema 5.2.3 El tiempo que tarda un cliente en ser atendido en una tienda sigue una distribución exponencial de media 30 segundos. Sabiendo que dicho tiempo es independiente del número de clientes que están esperando, calculad la probabilidad de que 5 clientes, entre los 9 que hay en la tienda, sean atendidos en menos de 20 segundos cada uno.

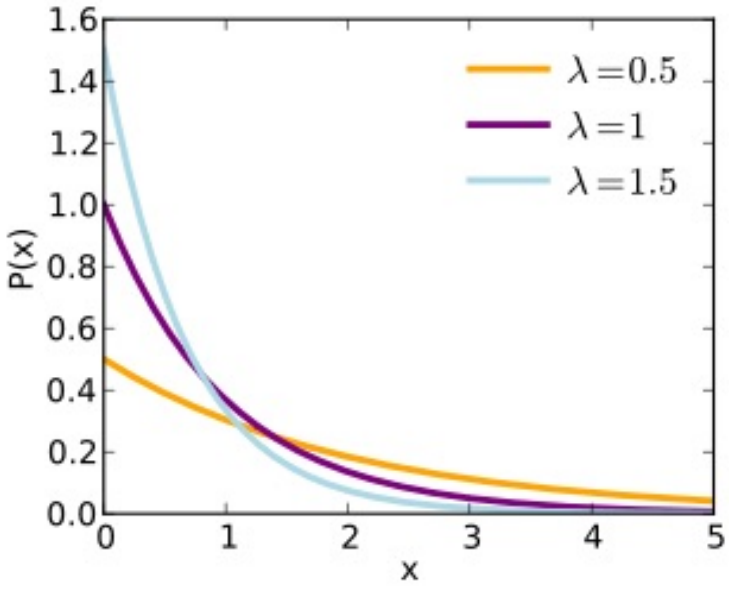

Figura 5.2: Ejemplos de la función de densidad de la distribución exponencial

Teorema 5.2.4 *La distribución exponencial no tiene memoria:*

$$P(X \leq t + h \,|\, X > h) = P(X \leq t)$$

Problema 5.2.5 Explicad la diferencia entre independencia y falta de memoria en variables aleatorias.

Problema 5.2.6 El tiempo que tardan en llegar los mensajes enviados a un ordenador sigue una distribución exponencial de media 20 segundos. Calculad:

(a) La probabilidad de que el tiempo que tarda en llegar un mensaje sea menor que 25 segundos;

(b) El tiempo, k, tal que el 80% de los mensajes enviados tardan en llegar un tiempo menor que k;

(c) La probabilidad de que un mensaje tarde en llegar más de 25 segundos sabiendo que no ha llegado en los primeros 15 segundos. Comparadla con la probabilidad de que un mensaje tarde en llegar al ordenador más de 10 segundos;

(d) La probabilidad de que el tiempo total que tardan en llegar tres mensajes enviados desde tres terminales sea mayor que 50 segundos.

Problema 5.2.7 Un sistema esta formado por 2 componentes independientes montados en paralelo. La duración de cada componente sigue una distribución exponencial, de media 24 meses para el primero y 36 meses para el segundo.

(a) Hallad la probabilidad de que el sistema funcione más de un año.

(b) Sabiendo que el sistema ha funcionado más de un año, calculad la probabilidad de que el primer componente no se haya estropeado durante ese tiempo.

(c) Un mecanismo está formado por 100 sistemas independientes del tipo anterior. El mecanismo se para cuando dejan de funcionar 15 o mas sistemas. Obtened la probabilidad de que el mecanismo se pare antes de 1 año.

5.3 Distribución gamma

En esta sección vamos a presentar la distribución gamma. Hay varios enfoques para hacer esto. Por ejemplo, se puede definir como una distribución muy general que aparece en multitud de contextos y argumentar su bondad a partir de su ubicuidad, pero nosotros preferimos un enfoque más inductivo y la presentaremos como la suma de variables independientes exponenciales. Si consideramos un proceso de Poisson y una variable $X \sim P(\lambda)$, sabemos, por la sección anterior, que la variable T tiempo entre dos sucesos consecutivos es una variable exponencial $T \sim \exp(\lambda)$. Ahora generalizamos y queremos que T mida el tiempo entre α sucesos consecutivos (α es entero, por supuesto). ¿Cómo se hace esto?

Calcularemos la función de distribución de T. Si $t > 0$, entonces para que ocurra el suceso $\{T \leq t\}$ ha de haber menos de α sucesos en el intervalo de tiempo $(0, t]$. Sea X la variable número de sucesos en el intervalo de tiempo $(0, \lambda t]$; sabemos que X es una Poisson $P(\lambda t)$ y entonces:

$$F(t) = P(T \leq t) = 1 - P(T > t) = 1 - P(X < \alpha) = 1 - \bigcup_{k=0}^{\alpha-1} P(X = k)$$

La probabilidad de $P(X = k)$ es la función de masa de la Poisson $P(\lambda t)$:

$$\begin{aligned} F(t) =& P(T \leq t) = 1 - \bigcup_{k=0}^{\alpha-1} P(X = k) = 1 - \sum_{i=0}^{\alpha-1} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \\ =& 1 - e^{-\lambda t} - \sum_{i=1}^{\alpha-1} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \end{aligned}$$

Para hallar la función de masa de T, derivamos su función de distribución:

$$\begin{aligned} f(t) =& F'(t) = 0 + \lambda e^{-\lambda t} - \sum_{i=1}^{\alpha-1} \left(\frac{(\lambda t)^i e^{-\lambda t}}{i!} \right)' \overset{(1)}{=} \lambda e^{-\lambda t} - \sum_{i=1}^{\alpha-1} \frac{\lambda \cdot i \cdot (\lambda t)^{i-1} e^{-\lambda t} - \lambda e^{-\lambda t} (\lambda t)^i}{i!} \\ \overset{(2)}{=}& \lambda e^{-\lambda t} - \lambda e^{-\lambda t} \sum_{i=1}^{\alpha-1} \frac{i \cdot (\lambda t)^{i-1} - (\lambda t)^i}{i!} \\ \overset{(3)}{=}& \lambda e^{-\lambda t} - \lambda e^{-\lambda t} \left(1 - \lambda t + \lambda t - \frac{(\lambda t)^2}{2} + \frac{(\lambda t)^2}{2!} - \frac{(\lambda t)^3}{3!} + \ldots + \frac{(\lambda t)^{\alpha-2}}{(\alpha-2)!} - \frac{(\lambda t)^{\alpha-1}}{(\alpha-1)!} \right) \\ \overset{(4)}{=}& \lambda e^{-\lambda t} - \lambda e^{-\lambda t} \left(1 - \frac{(\lambda t)^{\alpha-1}}{(\alpha-1)!} \right) = \frac{\lambda^\alpha t^{\alpha-1}}{(\alpha-1)!} e^{-\lambda t} \end{aligned}$$

donde: en (1) se ha efectuado la derivada del producto respecto a t; en (2) se ha sacado $\lambda e^{-\lambda t}$ fuera del sumatorio; en (3) se ha desarrollado el sumatorio, que ha resultado ser una serie telescópica en que se cancelan todos los términos menos el primero y el último; en (4) se han efectuado operaciones para dejar la fórmula en la expresión más cerrada posible.

Esta nueva distribución recibe el nombre de **distribución gamma**. Una vez más y como pasaba con la distribución exponencial, es costumbre escribir $X \sim \mathrm{G}(\alpha, \beta)$ para una variable distribuida según la gamma. El parámetro α recibe el nombre de **parámetro de forma** y β el **parámetro de escala**. En la figura 5.3 tenemos las gráficas de esta función de densidad para varios valores de los parámetros de forma y escala.

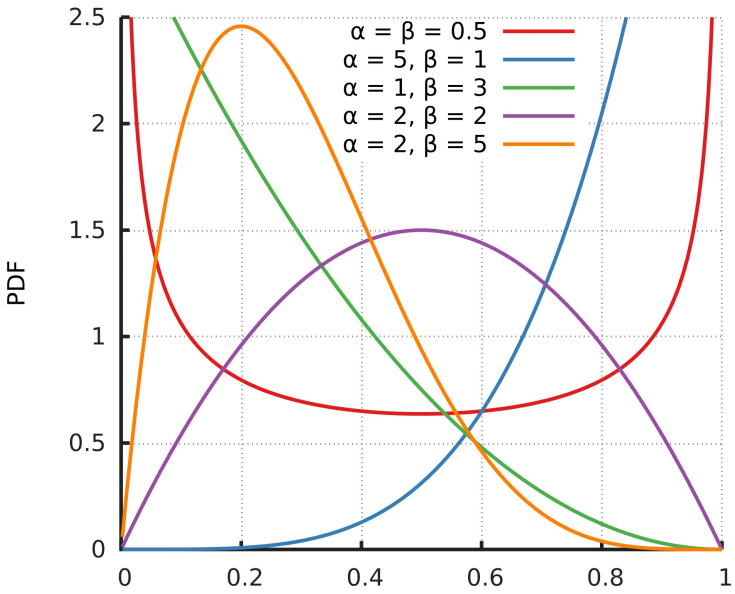

Figura 5.3: Ejemplos de la función de densidad de la distribución gamma

Problema 5.3.1 Probad que la función de densidad obtenida antes

$$\frac{\beta^\alpha t^{\alpha-1}}{(\alpha-1)!}e^{-\beta t}$$

lo es en efecto, es decir, que la integral de esta función entre 0 e infinito es 1. Para ello, os hará falta los números gamma (se dieron en AM: ¡Dios santo!, sí, cualquiera que sea el número de años en que la cursasteis o, incluso peor, el número de años que os faltan por cursarla). Los números gamma se escriben $\Gamma(\alpha)$ y se definen por

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\,dx$$

Como recordatorio nauseabundo, he aquí unas cuantas propiedades que os serán de utilidad

(a) Para todo $\alpha > 0$, $\Gamma(\alpha) > 0$.

(b) Para todo $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

(c) Para todo $n \in \mathbb{N}$ se verifica que $\Gamma(n+1) = n!$

(d) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Los principales momentos de la distribución $\mathrm{G}(\alpha, \beta)$ son: $E(X) = \frac{\alpha}{\beta}, \sigma^2 = \frac{\alpha}{\beta^2}, CV = \frac{1}{\sqrt{\alpha}}$ y $CA_F = \frac{2}{\sqrt{\alpha}}$.

Ejercicio 5.3.2 Analizad el coeficiente de variación y de asimetría de la distribución gamma. ¿Tienen alguna característica especial?

Ejercicio 5.3.3 Dada las siguientes variables, modelizadlas con la gamma adecuada.

(a) Se sabe que en una ciudad ocurre una gran inundación en media una vez cada seis años. ¿Qué distribución seguirá la variable tiempo que pasará antes de las siguientes cuatro grandes inundaciones?

(b) En una cola hay 12 personas y el tiempo medio de servicio es de 5 minutos. ¿Qué distribución seguirá la variable tiempo de servicio para esas 12 personas?

La distribución gamma $\mathrm{G}(\alpha, \beta)$ se llama **Erlang** cuando α es un número natural positivo. ¿Por qué tendría α que ser un número no natural si se dedujo la gamma como suma de α exponenciales, con α un número natural? Cierto, sin embargo, dado que la función $\frac{\lambda^\alpha t^{\alpha-1}}{(\alpha-1)!}e^{-\lambda t}$ es función de densidad independientemente de los parámetros α y β, la gamma $\mathrm{G}(\alpha, \beta)$ tiene sentido para $\alpha, \beta > 0$, con $\alpha, \beta \in \mathbb{R}$. La distinción terminológica es pertinente entonces.

5.4 Distribución de Pareto

Esta distribución recibe su nombre por su descubridor, Vilfredo Pareto (1848–1932), un hombre multifacético, que fue ingeniero civil, economista, sociólogo y filósofo. Investigó con profundidad las distribuciones de probabilidad de la forma x^α, las llamadas leyes potenciales. Encontró que este tipo de distribuciones aparecían en la descripción de fenómenos de las ciencias puras, ciencias sociales, geofísica, entre otros.

El primer trabajo de Pareto con su distribución fue la del estudio del reparto de la riqueza. El modelo de Pareto predice que la mayor parte de la riqueza de una sociedad estará en manos de un pequeño porcentaje de individuos (y así, lamentablemente, es). En general, esta distribución modeliza situaciones en las que hay un equilibrio entre la distribución de cantidades pequeñas frente a cantidades grandes. La distribución de Pareto también se conoce por distribución de Bradford. Los siguientes ejemplos son variables donde la distribución de Pareto explica razonablemente el fenómeno:

(a) El tamaño de los asentamientos humanos;

(b) El tamaño de los ficheros que circulan por internet con el protocolo TCP (muchos ficheros pequeños y pocos muy grandes).

(c) Las tasas de error en los discos duros;

(d) Las cantidades de petróleo en las reservas (de nuevo, pocas reservas grandes frente a muchas reservas pequeñas);

(e) El tamaño de tareas asignadas a superordenadores;

(f) El tamaño de los meteoritos;

(g) El tamaño de granos de arena en una playa;

(h) El tamaño de las áreas quemadas en un bosque tras un incendio.

La distribución de Pareto tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \alpha \dfrac{k^\alpha}{x^{\alpha+1}} & \text{si } x > k \\ \\ 0 & \text{en otro caso} \end{cases}$$

con $\alpha, k > 0$. La figura 5.4 muestra la gráfica de esta función de densidad para varios valores de α. Una variable aleatoria de Pareto se escribe $X \sim \text{Par}(\alpha, k)$, donde α recibe el nombre de **parámetro de forma** y k **parámetro de escala**.

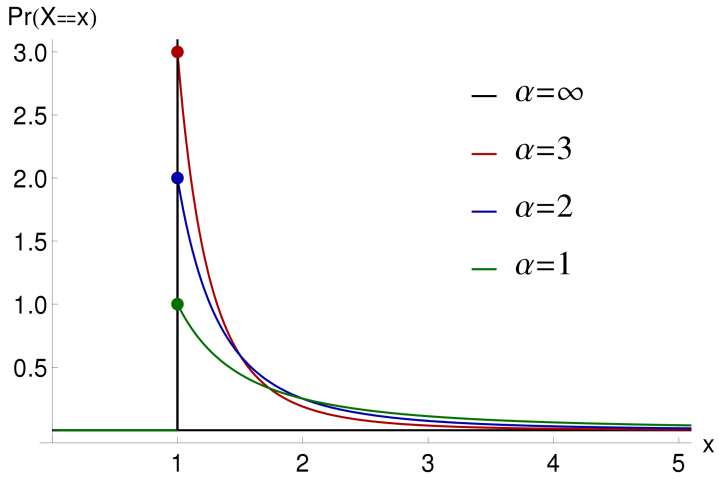

Figura 5.4: Ejemplos de la función de densidad de la distribución de Pareto

Ejercicio 5.4.1 Probad que la función anterior es de densidad para todos los valores posibles de α y k. Hallad su función de distribución.

Los momentos de la distribución de Pareto son los siguientes:

(a) Si $\alpha > 1$ la esperanza existe y es igual a $E(X) = k\dfrac{\alpha}{\alpha - 1}$

(b) Si $\alpha > 2$, la varianza es finita e igual a $V(X) = k^2 \dfrac{\alpha}{(\alpha-1)^2(\alpha-2)}$

(c) Para $\alpha > 3$, el coeficiente de asimetría existe y es $CA_F = \dfrac{2(\alpha+1)}{\alpha-3}\sqrt{\dfrac{\alpha-2}{\alpha}}$

Problema 5.4.2 Supongamos que los salarios mensuales de una empresa siguen una distribución de Pareto Par$(1000, 20)$

(a) Calculad la probabilidad de que una persona gane por encima de la media.

(b) Calculad la probabilidad de que una persona gane por debajo de 1500 euros.

(c) Calculad la probabilidad de que una persona gane entre 3000 y 6000 euros.

(d) Calculad la mediana del salario.

Teorema 5.4.3 *Sea X una variable aleatoria con distribución de Pareto Par(α, k). Si $m, x \in \mathbb{R}$, con $m > 1$ y $x/m > k$, entonces se cumple la siguiente propiedad:*

$$P(X > mx \,|\, X > x) = P(X > x \,|\, X > x/m)$$

Problema 5.4.4 Desentrañad el significado del teorema anterior.

Ejercicio 5.4.5 Se sabe que el tamaño de los mensajes, medidos en kilobytes, que pasan por un cierto nodo de internet siguen una distribución de Pareto Par$(2{,}5, 2)$. Calculad la probabilidad de que un mensaje tenga más 1000 kilobytes si sabemos que ya es mayor de 10 kilobytes. Ahora calculad la probabilidad de que el mensaje 20000 kilobytes si nos informan de que tiene más de 200 kilobytes.

La distribución de Pareto pertenece a una familia de distribuciones de las llamadas de **cola pesada**. Esto significa que la probabilidad de los **valores anómalos** —entendidos estos por aquellos que están lejos de la media— es relativamente grande. La definición formal de distribución de cola pesada es la siguiente.

Definición 5.4.6 Distribución de cola pesada. Una distribución se dice que tiene cola pesada si, para todo $\lambda > 0$, se cumple

$$\lim_{x \to \infty} e^{\lambda x} P(X > x) = +\infty$$

La definición anterior establece una comparación entre $P(X > x)$ y la función $e^{-\lambda x}$ cuando tienden a cero (siendo x que tiende a $+\infty$). Dice que $P(X > x)$ va más lento a cero que $e^{-\lambda x}$.

Ejercicio 5.4.7 Probad que la distribución de Pareto es de cola pesada.

Problema 5.4.8 Sea $X \sim \text{Par}(3, 2)$ e $Y \sim \exp(1/9)$. Calculad para ambas variables la probabilidad de que haya valores por encima de $\mu + 3\sigma$. Comparad ambas probabilidades y sacad conclusiones.