DISTRIBUCIONES DE PROBABILIDAD

ÍNDICE

13.1. Cálculo de probabilidades ... 3
13.1.0. Conceptos generales ... 3
13.1.1. Distribuciones discretas ... 5
13.1.1.1. Distribución uniforme discreta (a,b) ... 5
13.1.1.2. Distribución binomial (n,p) ... 6
13.1.1.3. Distribución hipergeométrica (N,R,n) ... 8
13.1.1.4. Distribución geométrica (p) ... 9
13.1.1.5. Distribución binomial negativa (r,p) ... 10
13.1.1.6. Distribución Pascal (r,p) ... 12
13.1.1.7. Distribución Poisson (λ) ... 13
13.1.2. Distribuciones continuas ... 16
13.1.2.1. Distribución uniforme o rectangular (a,b) ... 16
13.1.2.2. Distribución normal (μ, σ) ... 18
13.1.2.3. Distribución lognormal (μ, σ) ... 20
13.1.2.4. Distribución logística (a, b) ... 21
13.1.2.5. Distribución beta (p,q) ... 22
13.1.2.6. Distribución gamma (a,p) ... 23
13.1.2.7. Distribución exponencial (λ) ... 25
13.1.2.8. Distribución ji-cuadrado (n) ... 26
13.1.2.9. Distribución t de Student (n) ... 29
13.1.2.10. Distribución F de Snedecor (n,m) ... 31
13.1.2.11. Distribución Cauchy (μ, θ) ... 33
13.1.2.12. Distribución Weibull (a, b) ... 34
13.1.2.13. Distribución Laplace (a, b) ... 36
13.1.2.14. Distribución Pareto (α, x_0) ... 37
13.1.2.15. Distribución triangular (a, c, b) ... 39
13.2. Generación de distribuciones ... 40
13.2.0. Conceptos generales ... 40
13.2.1. Distribuciones discretas ... 41
13.2.1.1. Distribución multinomial ... 41
13.2.2. Distribuciones continuas ... 44
13.2.2.1. Distribución normal bivariante ... 44
Bibliografía ... 46
Anexo 1: Novedades del módulo de distribuciones de probabilidad ... 48
Anexo 2: Fórmulas del módulo de distribuciones de probabilidad ... 49
Anexo 3: Resumen de las distribuciones discretas ... 71
Anexo 4: Resumen de las distribuciones continuas ... 72

13.1. Cálculo de probabilidades

13.1.0. Conceptos generales

Uno de los objetivos de la estadística es el conocimiento cuantitativo de una determinada parcela de la realidad. Para ello, es necesario construir un modelo de esta realidad particular objeto de estudio, partiendo de la premisa de que lo real es siempre más complejo y multiforme que cualquier modelo que se pueda construir. De todas formas, la formulación de modelos aceptados por las instituciones responsables y por los usuarios, permite obviar la existencia del error o distancia entre la realidad y el modelo.

Los modelos teóricos a los que se hace referencia se reducen en muchos casos a (o incluyen en su formulación) funciones de probabilidad. La teoría de la probabilidad tiene su origen en el estudio de los juegos de azar, que impulsaron los primeros estudios sobre cálculo de probabilidades en el siglo XVI, aunque no es hasta el siglo XVIII cuando se aborda la probabilidad desde una perspectiva matemática con la demostración de la "ley débil de los grandes números" según la cual, al aumentar el número de pruebas, la frecuencia de un suceso tiende a aproximarse a un número fijo denominado probabilidad. Este enfoque, denominado *enfoque frecuentista*, se modela matemáticamente en el siglo XX cuando el matemático ruso Andrei Nikolaevich Kolmogorov (1903-1987) formula la *teoría axiomática* de la probabilidad [1]. Dicha teoría define la probabilidad como una función que asigna a cada posible resultado de un experimento aleatorio un valor no negativo, de forma que se cumpla la propiedad aditiva. La definición axiomática establece las reglas que deben cumplir las probabilidades, aunque no asigna valores concretos.

Uno de los conceptos más importantes de la teoría de probabilidades es el de variable aleatoria que, intuitivamente, puede definirse como cualquier característica medible que toma diferentes valores con probabilidades determinadas. Toda variable aleatoria posee una distribución de probabilidad que describe su comportamiento. Si la variable es discreta, es decir, si toma valores aislados dentro de un intervalo, su distribución de probabilidad especifica todos los valores posibles de la variable junto con la probabilidad de que cada uno ocurra. En el caso continuo, es decir, cuando la variable puede tomar cualquier valor de un intervalo, la distribución de probabilidad permite determinar las probabilidades correspondientes a subintervalos de valores. Una forma usual de describir la distribución de probabilidad de una variable aleatoria es mediante la denominada función de densidad en el caso de variables continuas y función de masa de probabilidad en el caso de variables discretas, en tanto que lo que se conoce como función de distribución representa las probabilidades acumuladas [2][3][4][5][6][7].

Una de las preocupaciones de los científicos ha sido construir modelos de distribuciones de probabilidad que pudieran representar el comportamiento teórico de diferentes fenómenos aleatorios que aparecían en el mundo real. La pretensión de modelar lo observable ha constituido siempre una necesidad básica para el científico empírico, dado que a través de esas construcciones teóricas, los modelos, podía experimentar sobre aquello que la realidad no le permitía. Por otra parte, un modelo resulta extremadamente útil, siempre que se corresponda con la realidad que pretende representar o predecir, de manera que ponga de relieve las propiedades más importantes del mundo que nos rodea, aunque sea a costa de la simplificación que implica todo modelo.

En la práctica hay unas cuantas leyes de probabilidad teóricas, como son, por ejemplo, la ley binomial o la de Poisson para variables discretas o la ley normal para variables continuas, que sirven de modelo para representar las distribuciones empíricas más frecuentes.

Así, por ejemplo, la variable "talla de un recién nacido" puede tener valores entre 47 cm y 53 cm, pero no todos los valores tienen la misma probabilidad, porque las más frecuentes son las tallas próximas a los 50 cm. En este caso la ley normal se adapta satisfactoriamente a la distribución de probabilidad empírica, que se obtendría con una muestra grande de casos.

Epidat 4 ofrece, en este módulo, procedimientos usuales para calcular probabilidades y sus inversas, para un conjunto bastante amplio de funciones de distribución, discretas y continuas, que son habituales en el proceso de modelación. Por ejemplo, el conjunto de distribuciones pertenecientes a la familia exponencial es de uso frecuente en metodologías como el análisis de supervivencia o el Modelo Lineal Generalizado. Otras distribuciones son comunes y habituales en el campo de actuación de disciplinas tales como la economía, la biología, etc. La lista de distribuciones disponibles en Epidat 4 ha sido ampliada con respecto a la versión anterior del programa.

Cuando la opción elegida es el cálculo de una probabilidad dado un punto x de la distribución, se presentan en todos los casos dos resultados: la probabilidad acumulada hasta ese punto, dicho de otra manera, la probabilidad de que la variable tome valores inferiores o iguales a x (cola izquierda); y la probabilidad de que la variable tome valores superiores a x (cola derecha), es decir, el complementario de la cola izquierda. En el caso discreto, a mayores se presenta la probabilidad de que la variable sea igual al punto x; este resultado no tiene sentido cuando estamos ante una distribución continua ya que la probabilidad de que la variable sea igual a un punto es igual a cero, lo que hace que la inclusión o exclusión del punto x no influya en el cálculo de las colas. Para ciertas distribuciones continuas simétricas (normal, logística y t de Student) el programa también presenta la probabilidad de dos colas, es decir, la probabilidad que queda a ambos lados del intervalo (-x, x) ó (x, -x) según el punto sea positivo o negativo, respectivamente. Epidat 4 permite calcular probabilidades para varios puntos a la vez.

La otra opción permitida en Epidat consiste en calcular un punto a partir de una probabilidad, bien sea la probabilidad de la cola izquierda, de la cola derecha o de las dos colas, siempre que sea posible.

Asimismo, los resultados de Epidat 4 incluyen la media, la varianza, la asimetría y la curtosis de la correspondiente distribución, así como la mediana y la moda en el caso de las distribuciones continuas.

Epidat 4 también ofrece la posibilidad de representar gráficamente la función de distribución y la función de densidad, o de masa de probabilidad, de cada una de las distribuciones. Estas gráficas pueden ser personalizadas por medio de un editor de gráficos que se inicia cada vez que se genera una gráfica.

Aunque cada distribución fue estudiada de forma independiente, en general el programa representa las funciones en el intervalo $(\mu - 3\sigma\,,\,\mu + 3\sigma)$, que puede ser ampliado por el usuario hasta $(\mu - 10\sigma\,,\,\mu + 10\sigma)$ desde el editor. La justificación para elegir estos intervalos se basa en la desigualdad de Chebyshev, que establece que [8]:

$$\Pr\{|X - \mu| \geq r\sigma\} \leq \frac{1}{r^2}$$

donde X es una variable aleatoria de media μ y varianza σ^2, y r es un número positivo.

Teniendo en cuenta esta desigualdad, se obtiene que en el intervalo $(\mu - 3\sigma\,,\,\mu + 3\sigma)$ queda sin representar una probabilidad de 0,11, que se reduce a 0,01 para el intervalo $(\mu - 10\sigma\,,\,\mu + 10\sigma)$.

13.1.1. Distribuciones discretas

Las distribuciones discretas incluidas en el módulo de "Cálculo de probabilidades" son:

- Uniforme discreta
- Binomial
- Hipergeométrica
- Geométrica
- Binomial negativa
- Pascal
- Poisson

En el Anexo 3 se incluye una tabla que resume las características de estas distribuciones.

13.1.1.1. Distribución uniforme discreta (a,b)

La distribución uniforme discreta describe el comportamiento de una variable discreta que puede tomar n valores distintos con la misma probabilidad cada uno de ellos. Un caso particular de esta distribución, que es la que se incluye en este módulo de Epidat 4, ocurre cuando los valores son enteros consecutivos. Esta distribución asigna igual probabilidad a todos los valores enteros entre el límite inferior y el límite superior que definen el recorrido de la variable. Si la variable puede tomar valores entre a y b, debe ocurrir que b sea mayor que a, y la variable toma los valores enteros empezando por a, a+1, a+2, etc. hasta el valor máximo b. Por ejemplo, cuando se observa el número obtenido tras el lanzamiento de un dado perfecto, los valores posibles siguen una distribución uniforme discreta en {1, 2, 3, 4, 5, 6}, y la probabilidad de cada cara es 1/6.

Valores:

k: a, a+1, a+2, ..., b, números enteros

Parámetros:

a: mínimo, a entero
b: máximo, b entero con $a < b$

Ejemplo

El temario de un examen para un proceso selectivo contiene 50 temas, de los cuales se elegirá uno por sorteo. Si una persona no ha estudiado los 15 últimos temas ¿cuál es la probabilidad de que salga un tema que haya estudiado?

La variable que representa el número del tema seleccionado para el examen sigue una distribución uniforme con parámetros a = 1 y b = 50. La persona ha estudiado los temas del 1 al 35; por tanto, la probabilidad que se pide es la cola a la izquierda de 35. Para obtener los resultados en Epidat 4 basta con proporcionarle los parámetros de la distribución, y seleccionar la opción de calcular probabilidades para el punto 35.

Resultados con Epidat 4:

Datos:

Distribución uniforme discreta (a, b)

Parámetros:

a: Mínimo 1

b: Máximo 50

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
35	0,02	0,7	0,3

Media	Varianza	Asimetría	Curtosis
25,5	208,25	0	-1,201

La persona tiene una probabilidad del 70% de que el tema elegido sea uno de los que haya estudiado.

13.1.1.2. Distribución binomial (n,p)

La distribución binomial es una distribución discreta muy importante que surge en muchas aplicaciones bioestadísticas. Fue obtenida por Jakob Bernoulli (1654-1705) y publicada en su obra póstuma *Ars Conjectandi* en 1713.

Esta distribución aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificada como "éxito" o "fracaso"; este experimento recibe el nombre de experimento de Bernoulli. Ejemplos de respuesta binaria pueden ser el hábito de fumar (sí/no), si un paciente hospitalizado desarrolla o no una infección, o si un artículo de un lote es o no defectuoso. La variable discreta que cuenta el número de éxitos en n pruebas independientes de ese experimento, cada una de ellas con la misma probabilidad de "éxito" igual a p, sigue una distribución binomial de parámetros n y p, que se denota por (Bi(n,p)). Este modelo se aplica a poblaciones finitas de las que se toman elementos al azar con reemplazo, y también a poblaciones conceptualmente infinitas, como por ejemplo las piezas que produce una máquina, siempre que el proceso de producción sea estable (la proporción de piezas defectuosas se mantiene constante a largo plazo) y sin memoria (el resultado de cada pieza no depende de las anteriores).

Un ejemplo de variable binomial puede ser el número de pacientes con cáncer de pulmón ingresados en una unidad hospitalaria.

Un caso particular se tiene cuando n=1, que da lugar a la distribución de Bernoulli.

En Epidat 4 el número de pruebas de la distribución binomial está limitado a 1.000; para valores superiores no es posible realizar el cálculo. Esta restricción no debe ser considerada un inconveniente dado que, cuando se tiene un número de pruebas "grande", la distribución binomial se aproxima a una distribución normal de media np y varianza $np(1-p)$ [8].

Valores:

k: 0, 1, 2, ..., n

Parámetros:

n: número de pruebas, $n \geq 1$ entero

p: probabilidad de éxito, $0 < p < 1$

Ejemplo

En un examen formado por 20 preguntas, cada una de las cuales se responde declarando "verdadero" o "falso", el alumno sabe que, históricamente, en el 75% de los casos la respuesta correcta es "verdadero" y decide responder al examen tirando dos monedas: pone "falso" si ambas monedas muestran una cara y "verdadero" si al menos hay una cruz. Se desea saber cual es la probabilidad de que tenga más de 14 aciertos.

Hay que proporcionarle a Epidat 4 los parámetros de la distribución binomial y el punto k a partir del cual se calculará la probabilidad. En este caso n = 20, p = 0,75 y el punto k = 14.

Resultados con Epidat 4:

Datos:

Distribución binomial (n, p)

Parámetros:

n: Número de pruebas 20

p: Probabilidad de éxito 0,75

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
14	0,1686	0,3828	0,6172

Media	Varianza	Asimetría	Curtosis
15	3,75	-0,2582	-0,0333

La probabilidad de que el alumno tenga más de 14 aciertos es del 62%.

El programa, además de calcular probabilidades, proporciona los valores característicos de la distribución (media, varianza, asimetría y curtosis) como información complementaria. Esta información depende solo de los parámetros de la distribución, no se ve influida por la opción elegida a la hora de realizar el cálculo (probabilidades o puntos) ni por el punto o probabilidad sobre el que se realiza dicho cálculo.

En este ejemplo, la media indica que 15 es el número medio de aciertos mediante la técnica de tirar dos monedas.

13.1.1.3. Distribución hipergeométrica (N,R,n)

La distribución hipergeométrica suele aparecer en procesos muestrales sin reemplazo, en los que se investiga la presencia o ausencia de cierta característica. Piénsese, por ejemplo, en un procedimiento de control de calidad en una empresa farmacéutica, durante el cual se extraen muestras de las cápsulas fabricadas y se someten a análisis para determinar su composición. Durante las pruebas, las cápsulas son destruidas y no pueden ser devueltas al lote del que provienen. En esta situación, la variable que cuenta el número de cápsulas que no cumplen los criterios de calidad establecidos sigue una distribución hipergeométrica. Por tanto, esta distribución es la equivalente a la binomial, pero cuando el muestreo se hace sin reemplazo, de forma que la probabilidad de éxito no permanece constante a lo largo de las n pruebas, a diferencia de la distribución binomial.

Esta distribución se puede ilustrar del modo siguiente: se tiene una población finita con N elementos, de los cuales R tienen una determinada característica que se llama "éxito" (diabetes, obesidad, hábito de fumar, etc.). El número de "éxitos" en una muestra aleatoria de tamaño n, extraída sin reemplazo de la población, es una variable aleatoria con distribución hipergeométrica de parámetros N, R y n.

Cuando el tamaño de la población es grande, los muestreos con y sin reemplazo son equivalentes, por lo que la distribución hipergeométrica se aproxima en tal caso a la binomial.

En el caso de esta distribución, Epidat 4 limita el cálculo a valores del tamaño de población (N) menores o iguales que 1.000.

Valores:

k: $max\{0,n-(N-R)\}, \ldots, min\{R,n\}$, donde $max\{0,n-(N-R)\}$ indica el valor máximo entre 0 y $n-(N-R)$ y $min\{R,n\}$ indica el valor mínimo entre R y n.

Parámetros:

N: tamaño de la población, $N \geq 1$ entero
R: número de éxitos en la población; $1 \leq R \leq N$, N entero
n: número de pruebas; $1 \leq n \leq N$, n entero

Ejemplo

Se sabe que el 7% de los útiles quirúrgicos en un lote de 100 no cumplen ciertas especificaciones de calidad. Tomada una muestra al azar de 10 unidades sin reemplazo, interesa conocer la probabilidad de que no más de dos sean defectuosas.

El número de útiles defectuosos en el lote es $R = 0{,}07 \times 100 = 7$. Para un tamaño muestral de $n = 10$, la probabilidad buscada es P{número de defectuosos ≤ 2}.

Resultados con Epidat 4:

Datos:

Distribución hipergeométrica (N, R, n)
Parámetros:
N: Tamaño de la población 100
R: Número de éxitos en la población 7
n: Número de pruebas 10

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
2	0,1235	0,9792	0,0208

Media	Varianza	Asimetría	Curtosis
0,7	0,5918	0,9126	0,4529

La probabilidad de que, a lo sumo, haya dos útiles defectuosos en el lote es aproximadamente 0,98. Además, puede decirse que la media y la varianza de la distribución hipergeométrica (100, 7, 10) son 0,7 y 0,59, respectivamente; en este caso, la media de útiles quirúrgicos defectuosos en 10 pruebas es de 0,7 y la varianza de 0,59.

13.1.1.4. Distribución geométrica (p)

Supóngase que se efectúa repetidamente un experimento o prueba, que las repeticiones son independientes y que se está interesado en la ocurrencia o no de un suceso al que se refiere como "éxito", siendo la probabilidad de este suceso p. La distribución geométrica permite calcular la probabilidad de que tenga que realizarse un número k de repeticiones antes de obtener un éxito por primera vez; esta probabilidad decrece a medida que aumenta k con lo que la función de masa de probabilidad es siempre decreciente. Así pues, se diferencia de la distribución binomial en que el número de repeticiones no está predeterminado, sino que es la variable aleatoria que se mide y, por otra parte, el conjunto de valores posibles de la variable es ilimitado.

Para ilustrar el empleo de esta distribución, se supone que cierto medicamento opera exitosamente ante la enfermedad para la cual fue concebido en el 80% de los casos a los que se aplica; la variable aleatoria "intentos fallidos en la aplicación del medicamento antes del primer éxito" sigue una distribución geométrica de parámetro p = 0,8. Otro ejemplo de variable geométrica es el número de hijos hasta el nacimiento de la primera niña.

La distribución geométrica se utiliza en la distribución de tiempos de espera, de manera que si los ensayos se realizan a intervalos regulares de tiempo, esta variable aleatoria proporciona el tiempo transcurrido hasta el primer éxito.

Esta distribución presenta la propiedad denominada "falta de memoria", que implica que la probabilidad de tener que esperar un tiempo t no depende del tiempo que ya haya transcurrido.

Valores:

k: 0, 1, 2, ...

Parámetros:

p: probabilidad de éxito, 0 < *p* < 1

Ejemplo

La probabilidad de que cierto examen médico dé lugar a una reacción "positiva" es igual a 0,8, ¿cuál es la probabilidad de que ocurran menos de 5 reacciones "negativas" antes de la primera positiva?

La variable aleatoria "número de reacciones negativas antes de la primera positiva" sigue una distribución geométrica con parámetro *p* = 0,8.

Resultados con Epidat 4:

Datos:

Distribución geométrica (p)

Parámetros:

p: Probabilidad de éxito 0,8

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
4	0,0013	0,9997	0,0003

Media	Varianza	Asimetría	Curtosis
0,25	0,3125	2,6833	9,2

La probabilidad de que ocurran menos de 5 reacciones "negativas" antes de la primera positiva es casi 1 (0,9997).

13.1.1.5. Distribución binomial negativa (r,p)

Una generalización obvia de la distribución geométrica aparece si se supone que un experimento se continúa hasta que un determinado suceso, de probabilidad *p*, ocurre por *r*-ésima vez. La variable aleatoria que proporciona la probabilidad de que se produzcan *k* fracasos antes de obtener el *r*-ésimo éxito sigue una distribución binomial negativa de parámetros *r* y *p*, BN(*r*,*p*). La distribución geométrica corresponde al caso particular en que *r*= 1. Un ejemplo es el número de lanzamientos fallidos de un dado antes de obtener un 6 en tres ocasiones, que sigue una BN(3,1/6).

En el caso de que los sucesos ocurran a intervalos regulares de tiempo, esta variable proporciona el tiempo total hasta que ocurren *r* éxitos, por lo que también se denomina "distribución binomial de tiempo de espera".

La distribución binomial negativa aparece en un estudio de Pierre Rémond de Montmort (1678-1719) sobre los juegos de azar en 1714, pero años antes ya había sido descrita por Blaise Pascal (1623-1662). Más adelante, esta distribución fue propuesta como una alternativa a la distribución de Poisson para modelar el número de ocurrencias de un suceso cuando los datos presentan lo que se conoce como variación extra-Poisson o sobredispersión. En estas

situaciones, la varianza es mayor que la media, por lo que se incumple la propiedad que caracteriza a una distribución de Poisson, según la cual la media es igual a la varianza. La primera aplicación en bioestadística la realizó Student (William Sealy Gosset (1876-1937)) a principios de siglo cuando propuso esta distribución para modelar el número de glóbulos rojos en una gota de sangre. En este caso, la variabilidad extra se debe al hecho de que esas células no están uniformemente distribuidas en la gota, es decir, la tasa de intensidad no es homogénea.

La distribución binomial negativa es más adecuada que la de Poisson para modelar, por ejemplo, el número de accidentes laborales ocurridos en un determinado lapso. La distribución de Poisson asume que todos los individuos tienen la misma probabilidad de sufrir un accidente y que ésta permanece constante durante el período de estudio; sin embargo, es más plausible la hipótesis de que los individuos tienen probabilidades constantes en el tiempo, pero que varían de unos sujetos a otros; esto es lo que se conoce en la literatura como la propensión a los accidentes ("*accident proneness*") [9][10]. Esta hipótesis se traduce en una distribución de Poisson mixta, o de efectos aleatorios, en la que se supone que las probabilidades varían entre individuos de acuerdo a una distribución gamma y esto resulta en una distribución binomial negativa para el número de accidentes.

El número máximo de éxitos permitidos en Epidat 4, para realizar cálculos de la distribución binomial negativa, es 1.000.

Valores:

k: 0, 1, 2, ...

Parámetros:

r: número de éxitos, $r \geq 1$ entero

p: probabilidad de éxito, $0 < p < 1$

Ejemplo

Se sabe que, en promedio, una de cada 100 placas de rayos X que se realizan es defectuosa. ¿Cuál es el número medio de placas útiles que se producen entre 10 defectuosas?

Si se considera el primer fallo como punto de inicio, hay que considerar la variable "número de placas útiles antes de 9 defectuosas", que sigue una distribución binomial negativa de parámetros $r = 9$ y $p = 0{,}01$.

Es necesario hacer notar que, cuando se está interesado en obtener alguno de los valores característicos de la distribución objeto de estudio (en este ejemplo, número medio de placas útiles), es indiferente calcular probabilidades o puntos, ya que el programa presenta los valores característicos de la distribución en ambos casos. En este ejemplo, se seleccionó la opción de calcular la probabilidad del punto 1, aunque se trata de un dato irrelevante en el cálculo del número medio de placas útiles. Se puede comprobar fácilmente que la modificación del punto o de la opción de cálculo no influyen en los valores característicos.

Resultados con Epidat 4:

Datos:

Distribución binomial negativa (r, p)
Parámetros:
r: Número de éxitos 9
p: Probabilidad de éxito 0,01

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
1	0	0	1

Media	Varianza	Asimetría	Curtosis
891	89.100	0,6667	0,6667

Entre 10 placas defectuosas se producen, en promedio, unas 891 placas útiles.

13.1.1.6. Distribución Pascal (r,p)

La distribución de Pascal debe su nombre al matemático francés Blaise Pascal (1623-1662), uno de los matemáticos que creó las bases de la teoría de la probabilidad.

El número de pruebas necesarias para obtener r éxitos, siendo p la probabilidad de éxito, es una variable aleatoria que sigue una distribución Pascal de parámetros r y p. Por tanto, esta distribución está relacionada con la binomial negativa de idénticos parámetros del modo siguiente[11]:

$$\text{Pascal}(r,p) = \text{BN}(r,p)+r$$

Teniendo en cuenta esta relación, podríamos decir que el número de lanzamientos de un dado realizados antes de obtener un 6 en tres ocasiones sigue una Pascal(3,1/6).

De la misma manera que ocurre en la distribución binomial negativa, Epidat 4 sólo permite realizar el cálculo cuando el número de éxitos considerados es igual o inferior a 1.000.

Valores:

k: r, r+1, r+2, ...

Parámetros:

r: número de éxitos, $r \geq 1$ entero
p: probabilidad de éxito, $0 < p < 1$

Ejemplo

Siguiendo con el ejemplo de la distribución binomial negativa, si en promedio una de cada 100 placas de rayos X que se realizan es defectuosa, ¿cuál es el número medio de placas realizadas entre 10 defectuosas?

Si se considera el primer fallo como punto de inicio, hay que considerar la variable "número de placas realizadas antes de 9 defectuosas", que sigue una distribución de Pascal de parámetros $r = 9$ y $p = 0{,}01$.

Resultados con Epidat 4:

Datos:

Distribución Pascal (r, p)
Parámetros:
r: Número de éxitos 9
p: Probabilidad de éxito 0,01

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
12	0	0	1

Media	Varianza	Asimetría	Curtosis
900	89.100	0,6667	0,6667

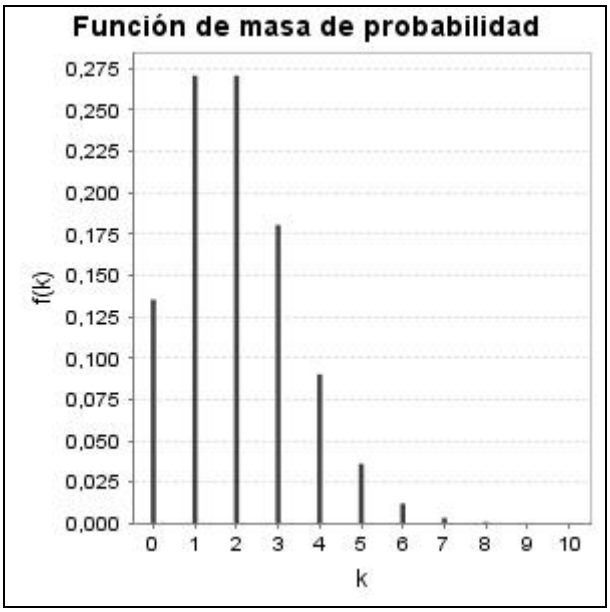
El número medio de placas realizadas entre 10 defectuosas es de 900.

13.1.1.7. Distribución Poisson (λ)

La distribución de Poisson debe su nombre al matemático francés Simeón Denis Poisson (1781-1840), aunque ya había sido introducida en 1718 por Abraham De Moivre (1667-1754) como una forma límite de la distribución binomial que surge cuando se observa un evento raro después de un número grande de repeticiones [12]. En general, la distribución de Poisson de parámetro λ se puede utilizar como una aproximación de la binomial, Bin(n, p), si el número de pruebas n es grande, pero la probabilidad de éxito p es pequeña, siendo $\lambda = np$; podemos considerar que la aproximación Poisson-binomial es "buena" si $n \geq 20$ y $p \leq 0{,}05$ y "muy buena" si $n \geq 100$ y $p \leq 0{,}01$.

La distribución de Poisson también surge cuando un evento o suceso "raro" ocurre aleatoriamente en el espacio o el tiempo. La variable asociada es el número de ocurrencias del evento en un intervalo o espacio continuo, por tanto, es una variable aleatoria discreta que toma valores enteros de 0 en adelante (0, 1, 2,...). Así, el número de pacientes que llegan a un consultorio en un lapso dado, el número de llamadas que recibe un servicio de atención a urgencias durante 1 hora, el número de células anormales en una superficie histológica o el número de glóbulos blancos en un milímetro cúbico de sangre son ejemplos de variables que siguen una distribución de Poisson. En general, es una distribución muy utilizada en diversas áreas de la investigación médica y, en particular, en epidemiología.

El concepto de evento "raro" o poco frecuente debe ser entendido en el sentido de que la probabilidad de observar k eventos decrece rápidamente a medida que k aumenta. Supóngase, por ejemplo, que el número de reacciones adversas tras la administración de un fármaco sigue una distribución de Poisson de media $\lambda = 2$. Si se administra este fármaco a 1.000 individuos, la probabilidad de que se produzca una reacción adversa (k = 1) es 0,27; los valores de dicha probabilidad para k = 2, 3, 4, 5, 6 reacciones, respectivamente, son: 0,27; 0,18; 0,09; 0,03 y 0,01. Para k = 10 o mayor, la probabilidad es virtualmente 0. El rápido descenso de la probabilidad de que se produzcan k reacciones adversas a medida que k aumenta puede observarse claramente en el gráfico de la función de masa de probabilidad obtenido con Epidat 4:



Para que una variable recuento siga una distribución de Poisson deben cumplirse varias condiciones:

1. En un intervalo muy pequeño (p. e. de un milisegundo) la probabilidad de que ocurra un evento es proporcional al tamaño del intervalo.
2. La probabilidad de que ocurran dos o más eventos en un intervalo muy pequeño es tan reducida que, a efectos prácticos, se puede considerar nula.
3. El número de ocurrencias en un intervalo pequeño no depende de lo que ocurra en cualquier otro intervalo pequeño que no se solape con aquél.

Estas propiedades pueden resumirse en que el proceso que genera una distribución de Poisson es estable (produce, a largo plazo, un número medio de sucesos constante por unidad de observación) y no tiene memoria (conocer el número de sucesos en un intervalo no ayuda a predecir el número de sucesos en el siguiente).

El parámetro de la distribución, λ, representa el número promedio de eventos esperados por unidad de tiempo o de espacio, por lo que también se suele hablar de λ como "la tasa de ocurrencia" del fenómeno que se observa.

A veces se usan variables de Poisson con "intervalos" que no son espaciales ni temporales, sino de otro tipo. Por ejemplo, para medir la frecuencia de una enfermedad se puede contar, en un período dado, el número de enfermos en cierta población dividida en "intervalos" de, por ejemplo, 10.000 habitantes. Al número de personas enfermas en una población de tamaño prefijado, en un instante dado, se le denomina prevalencia de la enfermedad en ese instante y es una variable que sigue una distribución de Poisson. Otra medida para la frecuencia de una enfermedad es la incidencia, que es el número de personas que enferman en una población en un periodo determinado. En este caso, el intervalo es de personas-tiempo, generalmente personas-año, y es también una variable con distribución de Poisson. Habitualmente, ambas medidas se expresan para intervalos de tamaño unidad o, dicho de otro modo, en lugar de la variable número de enfermos, se usa el parámetro λ.

La distribución de Poisson tiene iguales la media y la varianza. Si la variación de los casos observados en una población excede a la variación esperada por la Poisson, se está ante la presencia de un problema conocido como sobredispersión y, en tal caso, la distribución binomial negativa es más adecuada.

Para valores de λ mayores de 20 la distribución de Poisson se aproxima a una distribución normal de media y varianza iguales a λ. Por este motivo no se debe considerar una limitación la restricción que tiene Epidat 4 de no realizar el cálculo para valores de λ superiores a 50.

Valores:

k: 0, 1, 2, ...

Parámetros:

λ: tasa de ocurrencia, $\lambda > 0$

Ejemplo

El número de enfermos que solicitan atención de urgencia en un hospital durante un periodo de 24 horas tiene una media de 43,2 pacientes. Se sabe que el servicio se colapsará si el número de enfermos excede de 50. ¿Cuál es la probabilidad de que se colapse el servicio de urgencias del hospital? Representar la función de masa de probabilidad.

Para calcular la probabilidad pedida y, además, representar la función de masa de probabilidad hay que marcar el cuadro situado en la parte inferior derecha de la pantalla: *Obtener las funciones de distribución y densidad.*

Resultados con Epidat 4:

Datos:

Distribución Poisson (λ)

Parámetros:

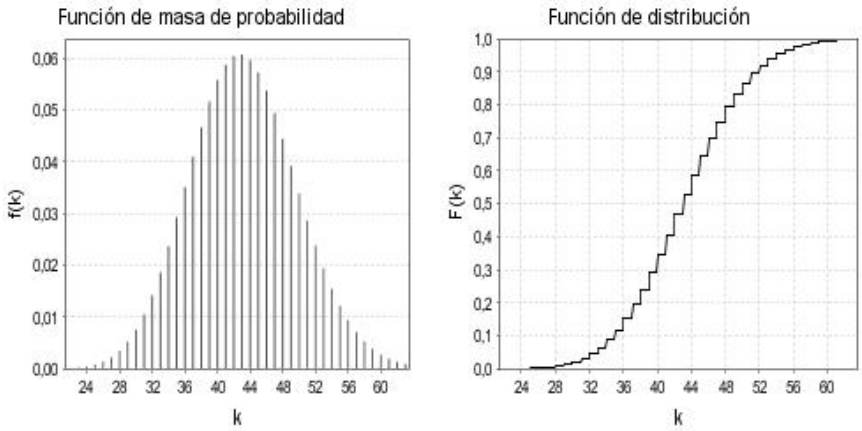
λ: Tasa de ocurrencia 43,2

Resultados:

Punto k	Probabilidad punto k	Cola izquierda Pr[X<=k]	Cola derecha Pr[X>k]
50	0,0339	0,8657	0,1343

Media	Varianza	Asimetría	Curtosis
43,2	43,2	0,1521	0,0231

Gráficos:



La probabilidad de que el servicio colapse está cerca de 0,13.

13.1.2. Distribuciones continuas

Las distribuciones continuas incluidas en el módulo de "Cálculo de probabilidades" son:

- Uniforme o rectangular
- Normal
- Lognormal
- Logística
- Beta
- Gamma
- Exponencial
- Ji-cuadrado
- t de Student
- F de Snedecor
- Cauchy
- Weibull
- Laplace
- Pareto
- Triangular

En el Anexo 4 se incluye una tabla que resume las características de estas distribuciones.

13.1.2.1. Distribución uniforme o rectangular (a,b)

La distribución uniforme es útil para describir una variable aleatoria con probabilidad constante sobre el intervalo (a,b) en el que está definida y se denota por U(a,b). También es conocida con el nombre de distribución rectangular por el aspecto de su función de densidad.

Una peculiaridad importante de esta distribución es que la probabilidad de un suceso depende exclusivamente de la amplitud del intervalo considerado y no de su posición en el campo de variación de la variable.

Cualquiera que sea la distribución F de cierta variable X, la variable transformada Y = F(X) sigue una distribución uniforme en el intervalo (0,1). Esta propiedad es fundamental por ser la base para la generación de números aleatorios de cualquier distribución en las técnicas de simulación, y recibe el nombre de método de inversión.

Campo de variación:

$a < x < b$

Parámetros:

a: mínimo, $-\infty < a < \infty$

b: máximo, $-\infty < b < \infty$ con $a < b$

Ejemplo 1

Supóngase una variable que se distribuye uniformemente entre 380 y 1.200. Determínese:

1. La probabilidad de que el valor de la variable sea superior a mil.
2. La media y la desviación estándar de dicha variable.

A Epidat se le proporcionará el límite superior e inferior del campo de variación de la variable [380, 1.200] y, además, el punto a partir del cual se quiere calcular la probabilidad.

Resultados con Epidat 4:

Datos:

Distribución uniforme o rectangular (a, b)

Parámetros:

a: Mínimo 380

b: Máximo 1.200

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
1.000	0,7561	0,2439

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
790	790	(380 , 1.200)	56.033,3333	0	-1,2

La probabilidad de que la variable sea superior a mil se sitúa en un entorno de 0,24, la media es 790 y la desviación estándar, raíz cuadrada de la varianza, es aproximadamente 237.

Ejemplo 2

Un contratista A está preparando una oferta sobre un nuevo proyecto de construcción. La oferta sigue una distribución uniforme entre 55 y 75 miles de euros. Determínese:

1. La probabilidad de que la oferta sea superior a 60 mil euros.
2. La media y la desviación estándar de la oferta.

A Epidat se le proporcionará el límite superior e inferior del campo de variación de la variable [55, 75] y, además, el punto a partir del cual se quiere calcular la probabilidad.

Resultados con Epidat 4:

Datos:

Distribución uniforme o rectangular (a, b)

Parámetros:

a: Mínimo 55

b: Máximo 75

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
60	0,25	0,75

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
65	65	(55 , 75)	33,3333	0	-1,2

La probabilidad de que la oferta sea superior a 60 mil euros se sitúa en un entorno de 0,75, y la media es 65.

13.1.2.2. Distribución normal (μ, σ)

La distribución normal es, sin duda, la distribución de probabilidad más importante del Cálculo de probabilidades y de la Estadística. Fue descubierta, como aproximación de la distribución binomial, por Abraham De Moivre (1667-1754) y publicada en 1733 en su libro *The Doctrine of Chances*; estos resultados fueron ampliados por Pierre-Simon Laplace (1749-1827), quién también realizó aportaciones importantes. En 1809, Carl Friedrich Gauss (1777-1855) publicó un libro sobre el movimiento de los cuerpos celestes donde asumía errores normales, por este motivo esta distribución también es conocida como distribución Gaussiana.

La importancia de la distribución normal queda totalmente consolidada por ser la distribución límite de numerosas variables aleatorias, discretas y continuas, como se demuestra a través de los teoremas centrales del límite. Las consecuencias de estos teoremas implican la casi universal presencia de la distribución normal en todos los campos de las ciencias empíricas: biología, medicina, psicología, física, economía, etc. En particular, muchas medidas de datos continuos en medicina y en biología (talla, presión arterial, etc.) se aproximan a la distribución normal.

Junto a lo anterior, no es menos importante el interés que supone la simplicidad de sus características y de que de ella derivan, entre otras, tres distribuciones (ji-cuadrado, t de Student y F de Snedecor) que se mencionarán más adelante, de importancia clave en el campo de la contrastación de hipótesis estadísticas.

La distribución normal queda totalmente definida mediante dos parámetros: la media (μ) y la desviación estándar o desviación típica (σ). Su función de densidad es simétrica respecto a la media y la desviación estándar nos indica el mayor o menor grado de apertura de la curva que, por su aspecto, se suele llamar campana de Gauss. Esta distribución se denota por $N(\mu,\sigma)$.

Cuando la distribución normal tiene como parámetros $\mu = 0$ y $\sigma = 1$ recibe el nombre de distribución normal estándar. Cualquier variable X que siga una distribución normal de parámetros μ y σ se puede transformar en otra variable $Y= (X-\mu)/\sigma$ que sigue una distribución normal estándar; este proceso se denomina estandarización, tipificación o normalización.

Campo de variación:

$-\infty < x < \infty$

Parámetros:

μ: media, $-\infty < \mu < \infty$

σ: desviación estándar, $\sigma > 0$

Ejemplo

Se supone que el nivel de colesterol de los enfermos de un hospital sigue una distribución normal con una media de 179,1 mg/dL y una desviación estándar de 28,2 mg/dL.

1. ¿Cuál es el porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dL?
2. ¿Cuál será el valor del nivel de colesterol a partir del cual se encuentra el 10% de los enfermos del hospital con los niveles más altos?
3. Representar la función de densidad.

Para responder a estas preguntas habrá que ejecutar Epidat 4 dos veces: en el primer caso para calcular una probabilidad, en el segundo caso el dato de entrada es una probabilidad,

concretamente la cola de la derecha, lo que permitirá obtener el punto. En ambas ejecuciones se ofrece, de manera opcional, la función de densidad del nivel de colesterol.

1. Resultados con Epidat 4:

Datos:

Distribución normal (μ, σ)
Parámetros:
μ: Media 179,1
σ: Desviación estándar 28,2

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]	Dos colas 1-Pr[|X|<=x]
169	0,3601	0,6399	0,7202

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
179,1	179,1	179,1	795,24	0	0

El porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dL es 36%.

2. Resultados con Epidat 4:

Datos:

Distribución normal (μ, σ)
Parámetros:
μ: Media 179,1
σ: Desviación estándar 28,2
Cola izquierda Pr[X<=x]: 0,9
Cola derecha Pr[X>x]: 0,1
Dos colas 1-Pr[|X|<=x]: 0,2

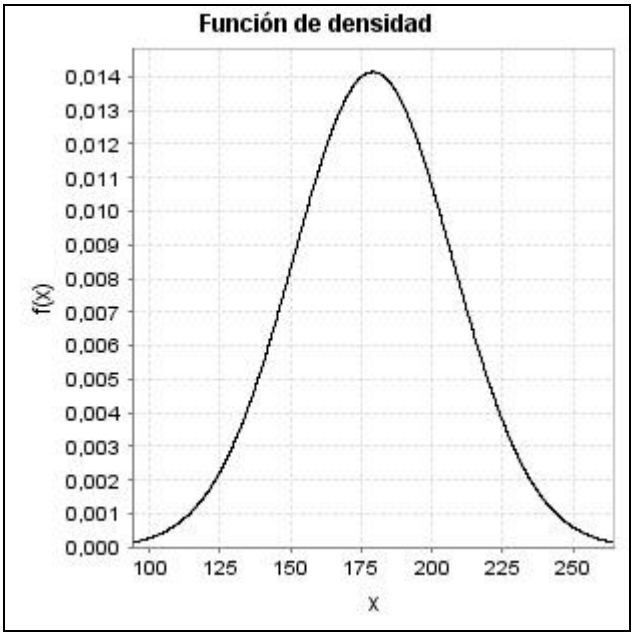
Resultados:

Punto x
215,2398

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
179,1	179,1	179,1	795,24	0	0

A partir de 215,24 mg/dL se encuentran los valores de colesterol del 10% de los enfermos que tienen los valores más altos.

3. Resultados con Epidat 4:



13.1.2.3. Distribución lognormal (μ, σ)

La variable resultante de aplicar la función exponencial a una variable que se distribuye normal con media μ y desviación estándar σ, sigue una distribución lognormal con parámetros μ (escala) y σ (forma). Dicho de otro modo, si una variable X sigue una distribución lognormal entonces la variable *ln*X se distribuye normalmente. Esta variable aleatoria fue propuesta por Francis Galton (1822-1911) en 1879, como consecuencia del estudio de la media geométrica de n variables aleatorias independientes.

La distribución lognormal es útil para modelar datos de numerosos estudios médicos tales como el período de incubación de una enfermedad, los títulos de anticuerpo a un virus, el tiempo de supervivencia en pacientes con cáncer o SIDA, el tiempo hasta la seroconversión de VIH+, etc.

Epidat 4 limita los cálculos para esta distribución a valores del parámetro μ entre -5 y 5, ambos inclusive, y a valores del parámetro σ menores o iguales que 5.

Campo de variación:

$0 < x < \infty$

Parámetros:

μ: escala, $-\infty < \mu < \infty$

σ: forma, $\sigma > 0$

Ejemplo

Supóngase que la supervivencia, en años, luego de una intervención quirúrgica (tiempo que pasa hasta que ocurre la muerte del enfermo) en una cierta población sigue una distribución lognormal de parámetro de escala 2,32 y de forma 0,20. Calcular la probabilidad de supervivencia a los 12 años y la mediana de supervivencia.

Resultados con Epidat 4:

Datos:

Distribución lognormal (μ, σ)
Parámetros:
μ: Escala 2,32
σ: Forma 0,2

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
12	0,7952	0,2048

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
10,3812	10,1757	9,7767	4,3982	0,6143	0,6784

La probabilidad de supervivencia a los 12 años es próxima a 0,20.

A la vista de los resultados también se puede decir que el número medio de años de supervivencia de un paciente tras una intervención quirúrgica es de, aproximadamente, 10 años y medio.

13.1.2.4. Distribución logística (a, b)

Pierre François Verhulst (1804-1849) describió por primera vez la curva logística en un trabajo, publicado en 1845, que versaba sobre las investigaciones matemáticas en las leyes que gobiernan el crecimiento de la población.

La distribución logística se utiliza en el estudio del crecimiento temporal de variables, en particular, demográficas. En biología se ha aplicado, por ejemplo, para modelar el crecimiento de células de levadura, y para representar curvas de dosis-respuesta en bioensayos.

La más conocida y generalizada aplicación de la distribución logística en Ciencias de la Salud se fundamenta en la siguiente propiedad: si U es una variable uniformemente distribuida en el intervalo (0,1), entonces la variable $X = \ln\left(\frac{U}{1-U}\right)$ sigue una distribución logística. Esta transformación, denominada *logit*, se utiliza para modelar datos de respuesta binaria, especialmente en el contexto de la regresión logística.

Los parámetros asociados a esta distribución son situación (a) y escala (b). Su función de densidad es simétrica respecto al parámetro a y presenta un perfil más apuntado que el de la distribución normal con la misma media y desviación estándar (distribución leptocúrtica).

Campo de variación:

$-\infty < x < \infty$

Parámetros:

a: situación, $-\infty < a < \infty$
b: escala, $b > 0$

Ejemplo

El crecimiento relativo anual (%) de la población de un determinado país sigue una distribución logística de parámetro de posición 1 y de escala 2. Calcular la probabilidad de que el crecimiento en un año determinado sea superior al 5% y representar la función de densidad.

Resultados con Epidat 4:

Datos:

Distribución logística (a, b)

Parámetros:

a: Situación 1

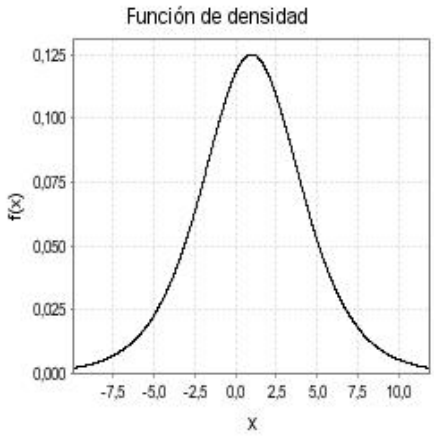
b: Escala 2

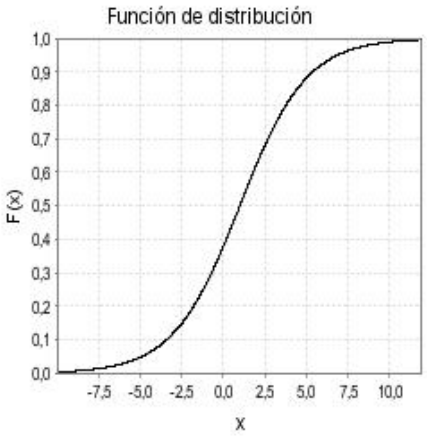
Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]	Dos colas 1-Pr[|X|<=x]
5	0,8808	0,1192	0,2384

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
1	1	1	13,1595	0	1,2

Gráficos:





La probabilidad de que la población tenga un crecimiento superior al 5% es del orden de 0,12.

13.1.2.5. Distribución beta (p,q)

La distribución beta es adecuada para variables aleatorias continuas que toman valores en el intervalo (0,1), lo que la hace muy apropiada para modelar proporciones. En la inferencia bayesiana, por ejemplo, es muy utilizada como distribución a priori cuando las observaciones tienen una distribución binomial.

Uno de los principales recursos de esta distribución es el ajuste a una gran variedad de distribuciones empíricas, pues adopta formas muy diversas dependiendo de cuáles sean los valores de los parámetros de forma p y q, mediante los que viene definida la distribución, denotada por Beta(p,q).

Un caso particular de la distribución beta es la distribución uniforme en (0,1), que se corresponde con una beta de parámetros $p = 1$ y $q = 1$.

La limitación que impone Epidat 4 a los valores que pueden tomar sus parámetros es que no deben ser mayores que 100 para poder realizar los cálculos.

Campo de variación:
$0 < x < 1$

Parámetros:
p: forma, $p > 0$
q: forma, $q > 0$

Ejemplo

En el presupuesto familiar, la porción que se dedica a salud sigue una distribución beta(2,2).

1. ¿Cuál es la probabilidad de que se gaste más del 25% del presupuesto familiar en salud?
2. ¿Cuál será el porcentaje medio que las familias dedican a la compra de productos y servicios de salud?

Resultados con Epidat 4:

Datos:

Distribución beta (p, q)
Parámetros:
p: Forma 2
q: Forma 2

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
0,25	0,1562	0,8438

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
0,5	0,5	0,5	0,05	0	-0,8571

Teniendo en cuenta la distribución beta, la probabilidad de que se gaste más de la cuarta parte del presupuesto en salud será 0,84 y el porcentaje medio que las familias dedican a la compra de productos y servicios de salud será el 50%.

13.1.2.6. Distribución gamma (a,p)

La distribución gamma se puede caracterizar del modo siguiente: si se está interesado en la ocurrencia de un evento generado por un proceso de Poisson de media λ, la variable que mide el tiempo transcurrido hasta obtener n ocurrencias del evento sigue una distribución gamma con parámetros $a = n\lambda$ (escala) y $p = n$ (forma). Se denota por Gamma(a,p).

Por ejemplo, la distribución gamma aparece cuando se realiza el estudio de la duración de elementos físicos (tiempo de vida).

Cuando p es un número entero positivo se tiene un caso particular de la distribución gamma que se denomina distribución de Erlang. Otros casos particulares de la distribución gamma, que se comentarán con detalle más adelante, son la distribución exponencial (Gamma(λ,1)) y la distribución ji-cuadrado (Gamma(1/2,n/2)).

Según los valores que tome el parámetro de forma, p, la función de densidad presenta perfiles muy diversos. Con valores de p menores o iguales que 1, la función de densidad muestra un perfil decreciente; en cambio, si p es mayor que la unidad, la función de densidad crece hasta el valor x= (p-1)/a y decrece a partir de este valor.

Epidat 4 limita los cálculos a valores de los parámetros menores o iguales que 25.

Campo de variación:

$0 < x < \infty$

Parámetros:

a: escala, $a > 0$

p: forma, $p > 0$

Ejemplo 1

El número de pacientes que llegan a la consulta de un médico sigue una distribución de Poisson de media 3 pacientes por hora. Calcular la probabilidad de que transcurra menos de una hora hasta la llegada del segundo paciente.

Debe tenerse en cuenta que la variable aleatoria "tiempo que transcurre hasta la llegada del segundo paciente" sigue una distribución Gamma (6, 2).

Resultados con Epidat 4:

Datos:

Distribución gamma (a, p)

Parámetros:

a: Escala 6

p: Forma 2

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
1	0,9826	0,0174

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
0,3333	0,2797	0,1667	0,0556	1,4142	3

La probabilidad de que transcurra menos de una hora hasta que llegue el segundo paciente es 0,98.

Ejemplo 2

Suponiendo que el tiempo de supervivencia, en años, de pacientes que son sometidos a una cierta intervención quirúrgica en un hospital sigue una distribución gamma con parámetros a = 0,81 y p = 7,81, interesa saber:

1. El tiempo medio de supervivencia.
2. Los años a partir de los cuales la probabilidad de supervivencia es menor que 0,1.

Resultados con Epidat 4:

Datos:

Distribución gamma (a, p)
Parámetros:
a: Escala 0,81
p: Forma 7,81
Cola izquierda Pr[X<=x]: 0,9
Cola derecha Pr[X>x]: 0,1

Resultados:

Punto x
14,2429

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
9,642	9,2337	8,4074	11,9037	0,7157	0,7682

El tiempo medio de supervivencia es de, aproximadamente, 10 años, y a partir de 14,2 años, la probabilidad de supervivencia es menor de 0,1.

13.1.2.7. Distribución exponencial (λ)

La distribución exponencial es un caso particular de la distribución gamma y el equivalente continuo de la distribución geométrica discreta. Esta ley de distribución describe procesos en los que interesa saber el tiempo hasta que ocurre determinado evento; en particular, se utiliza para modelar tiempos de supervivencia. Un ejemplo es el tiempo que tarda una partícula radiactiva en desintegrarse. El conocimiento de la ley que sigue este evento se utiliza, por ejemplo, para la datación de fósiles o cualquier materia orgánica mediante la técnica del carbono 14.

Una característica importante de esta distribución es la propiedad conocida como "falta de memoria". Esto significa, por ejemplo, que la probabilidad de que un individuo de edad t sobreviva x años más, hasta la edad $x+t$, es la misma que tiene un recién nacido de sobrevivir hasta la edad x. Dicho de manera más general, el tiempo transcurrido desde cualquier instante dado t_0 hasta que ocurre el evento, no depende de lo que haya ocurrido antes del instante t_0.

Se cumple que variable aleatoria que tome valores positivos y que verifique la propiedad de "falta de memoria" sigue una distribución exponencial [8].

Esta distribución se puede caracterizar como la distribución del tiempo entre sucesos consecutivos generados por un proceso de Poisson; por ejemplo, el tiempo que transcurre entre dos heridas graves sufridas por una persona. La media de la distribución de Poisson, λ, que representa la tasa de ocurrencia del evento por unidad de tiempo, es el parámetro de la distribución exponencial, y su inversa es el valor medio de la distribución.

El uso de la distribución exponencial ha sido limitado en bioestadística, debido a que la propiedad de falta de memoria la hace demasiado restrictiva para la mayoría de los problemas.

Epidat 4 permite realizar cálculos de esta distribución siempre y cuando el parámetro λ sea menor o igual que 100.

Campo de variación:

$0 < x < \infty$

Parámetros:

λ: tasa, $\lambda > 0$

Ejemplo

Se ha comprobado que el tiempo de vida de cierto tipo de marcapasos sigue una distribución exponencial con media de 14 años. ¿Cuál es la probabilidad de que a una persona a la que se le ha implantado este marcapasos se le deba reimplantar otro antes de 20 años? Si el marcapasos lleva funcionando correctamente 5 años en un paciente, ¿cuál es la probabilidad de que haya que cambiarlo antes de 25 años?

La variable aleatoria "tiempo de vida del marcapasos" sigue una distribución exponencial de parámetro $\lambda = 1/14 \approx 0{,}07$

Resultados con Epidat 4:

Datos:

Distribución exponencial (λ)

Parámetros:

λ: Tasa 0,07

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
20	0,7534	0,2466

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
14,2857	9,9021	No definida	204,0816	2	6

La probabilidad de que se le tenga que implantar otro marcapasos antes de los 20 años se sitúa en un entorno de 0,75.

Teniendo en cuenta la propiedad de "falta de memoria" de la exponencial, la probabilidad de tener que cambiar antes de 25 años un marcapasos que lleva funcionando 5 es igual a la probabilidad de cambio a los 20 años, es decir, $P(X < 25/X > 5) = P(X < 20) = 0{,}75$.

13.1.2.8. Distribución ji-cuadrado (n)

Un caso especial y muy importante de la distribución gamma se obtiene cuando $a = 1/2$ y $p=n/2$, y es conocida por el nombre de distribución ji-cuadrado de Pearson con n grados de libertad (se denota por χ^2_n). Es la distribución que sigue la suma de los cuadrados de n variables independientes e idénticamente distribuidas según una distribución normal estándar, N(0,1).

Esta distribución, que debe su nombre al matemático inglés Karl Pearson (1857-1936), es fundamental en inferencia estadística y en los tests estadísticos de bondad de ajuste. Se emplea, entre otras muchas aplicaciones, para realizar la prueba de hipótesis de homogeneidad, de independencia o la prueba de bondad de ajuste (todas ellas denominadas pruebas ji-cuadrado) y para determinar los límites de confianza de la varianza muestral de una población normal.

Si X sigue una distribución ji-cuadrado con *n* grados de libertad, para valores de *n* grandes (*n* > 100), entonces la variable $Y = \sqrt{2X}$ sigue aproximadamente una distribución normal de media $\sqrt{2n-1}$ y desviación estándar 1.

Epidat 4 realiza los cálculos de esta distribución para valores de *n* menores o iguales que 150.

Campo de variación:

$0 < x < \infty$

Parámetros:

n: grados de libertad, $n \geq 1$ entero

Ejemplo

Para estudiar la relación entre la edad de las mujeres y su aceptación de una ley sobre interrupción del embarazo se ha llevado a cabo una encuesta sobre 400 mujeres cuyos resultados se recogen en la siguiente tabla:

	Aceptación		
Edad	**Baja**	**Media**	**Alta**
0-18	21	34	25
18-35	24	31	25
36-50	30	30	20
51-65	37	30	13
> 65	40	30	10

Como resultado de aplicar la prueba ji-cuadrado de Pearson se obtuvo como valor del estadístico χ^2=19,2828. Este valor por si solo no permite extraer ninguna conclusión; debe compararse con el valor de la distribución ji-cuadrado de (5-1)*(3-1)=8 grados de libertad que deja un 5% de probabilidad a su derecha, fijado un nivel de significación del 5% o, equivalentemente, un nivel de confianza del 95%. Este valor, llamado punto crítico, delimita la zona de rechazo de la hipótesis nula de no asociación entre las variables.

1. Calcular el valor de la ji-cuadrado con 8 grados de libertad que deja a su derecha un área bajo la curva igual a 0,05.
2. Representar la función de densidad y marcar en ella el valor del estadístico y el punto crítico, ¿qué puede concluirse acerca de la relación entre las dos variables?
3. Calcular el valor p del estadístico, es decir, la probabilidad a la derecha del valor del estadístico χ^2=19,2828.

1. Resultados con Epidat 4:

Datos:

Distribución ji-cuadrado (n)
Parámetros:

n: Grados de libertad	8
Cola izquierda Pr[X<=x]:	0,95
Cola derecha Pr[X>x]:	0,05

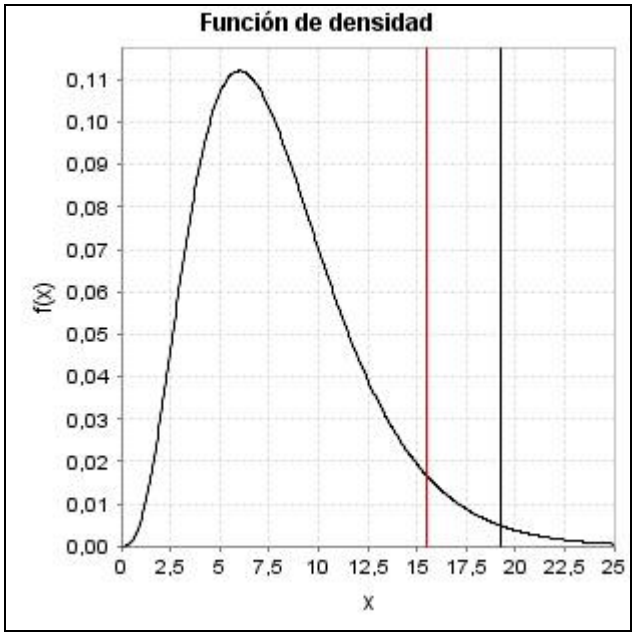
Resultados:

Punto x
15,5073

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
8	7,3441	6	16	1	1,5

El valor 15,5073 es el punto crítico del test para un nivel de significación del 5%, ya que deja a su derecha una cola de probabilidad 0,05.

2. Resultados con Epidat 4:



A la vista de este gráfico se puede observar como el valor del estadístico (línea negra) es superior al punto crítico del test para un nivel de significación del 5% y, por lo tanto, está en la zona de rechazo. Esto significa que hay evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres.

3. Resultados con Epidat 4:

Datos:

Distribución ji-cuadrado (n)
Parámetros:
n: Grados de libertad 8

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
19,2828	0,9866	0,0134

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
8	7,3441	6	16	1	1,5

El valor p del estadístico es 0,0134 y se corresponde con el área bajo la curva a la derecha del valor del estadístico. Si este valor es menor que 0,05, como así ocurre, se rechaza la hipótesis nula de no asociación entre las dos variables.

13.1.2.9. Distribución t de Student (n)

Esta distribución fue propuesta y tabulada por William Sealy Gosset (1876-1937), más conocido por el seudónimo de Student, como resultado de un estudio sobre la estimación de la media cuando el tamaño de muestra es pequeño, estos resultados fueron publicados en 1908 en el artículo *The Probable Error of a Mean* [13].

La distribución t de Student queda completamente definida por medio de sus grados de libertad, *n*, y se denota por t_n. Surge cuando se plantea estudiar el cociente entre una variable aleatoria con distribución normal estándar y la raíz cuadrada del cociente entre una variable aleatoria con distribución ji-cuadrado y sus grados de libertad (*n*), siendo las dos variables independientes. Esta distribución desempeña un papel muy importante en la inferencia estadística asociada a la teoría de muestras pequeñas y es usada habitualmente en el contraste de hipótesis para la media de una población o para comparar medias de dos poblaciones.

En cuanto a la forma que presenta su función de densidad cabe destacar las similitudes que mantiene con la función de densidad de la distribución normal estándar: forma acampanada, simétrica y centrada en el origen; la única diferencia existente entre ambas distribuciones es que la función de densidad de la t de Student presenta unas colas más pesadas (mayor dispersión) que la normal.

Cabe destacar que el programa sólo permite realizar el cálculo para una distribución t de Student con 150 grados de libertad o menos. Esto no supone una limitación ya que, a medida que aumentan los grados de libertad, esta distribución se va aproximando a la normal estándar, de forma que a partir de ese valor de *n* pueden considerarse prácticamente idénticas.

La distribución t de Student con 1 grado de libertad coincide con la distribución de Cauchy estándar.

Campo de variación:

$-\infty < x < \infty$

Parámetros:

n: grados de libertad, $n \geq 1$ entero

Ejemplo

La distribución t de Student se aproxima a la normal a medida que aumentan los grados de libertad.

1. Calcular, para una distribución N(0,1), el punto que deja a la derecha una cola de probabilidad 0,05.
2. Calcular, para una distribución t de Student, la probabilidad de que la variable tome un valor a la derecha de ese punto. Tomar como grados de libertad sucesivamente n = 10 y n = 150.

Para el primer apartado hay que seleccionar en la lista de distribuciones la normal de parámetros $\mu = 0$ y $\sigma = 1$.

1. Resultados con Epidat 4:

Datos:

Distribución normal (μ, σ)

Parámetros:

μ: Media	0
σ: Desviación estándar	1
Cola izquierda Pr[X<=x]:	0,95
Cola derecha Pr[X>x]:	0,05
Dos colas 1-Pr[|X|<=x]:	0,1

Resultados:

Punto x
1,6449

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
0	0	0	1	0	0

En el segundo apartado se ejecutará dos veces Epidat 4: la primera vez para una distribución t de Student con 10 grados de libertad y la segunda vez con 150 grados de libertad.

2. Resultados con Epidat 4:

Datos:

Distribución t de Student (n)
Parámetros:
n: Grados de libertad 10

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]	Dos colas 1-Pr[|X|<=x]
1,6449	0,9345	0,0655	0,131

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
0	0	0	1,25	0	1

Datos:

Distribución t de Student (n)
Parámetros:
n: Grados de libertad 150

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]	Dos colas 1-Pr[|X|<=x]
1,6449	0,949	0,051	0,1021

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
0	0	0	1,0135	0	0,0411

Se aprecia claramente que, al aumentar los grados de libertad de la t de Student, la probabilidad se acerca a la calculada con la distribución normal.

13.1.2.10. Distribución F de Snedecor (n,m)

Otra de las distribuciones importantes asociadas a la normal es la que se define como el cociente de dos variables aleatorias independientes con distribución ji-cuadrado divididas entre sus respectivos grados de libertad, *n* y *m;* la variable aleatoria resultante sigue una distribución F de Snedecor de parámetros *n* y *m* (denotada por $F_{n,m}$). Hay muchas aplicaciones de la F en estadística y, en particular, tiene un papel importante en las técnicas del análisis de la varianza (ANOVA) y del diseño de experimentos. Debe su nombre al matemático y estadístico americano George Waddel Snedecor (1881-1974).

Al igual que en la distribución ji-cuadrado y t de Student, el programa limita los grados de libertad, tanto del numerador como del denominador, no pudiendo exceder el valor 150 para poder realizar los cálculos.

Campo de variación:

$0 < x < \infty$

Parámetros:

n: grados de libertad del numerador, $n \geq 1$ entero
m: grados de libertad del denominador, $m \geq 1$ entero

Ejemplo

En un laboratorio se efectuaron ciertas mediciones y se comprobó que seguían una distribución F con 10 grados de libertad en el numerador y 12 grados de libertad en el denominador.

1. Calcule el valor que deja a la derecha el 5% del área bajo la curva de densidad.
2. ¿Cuál es la probabilidad de que la medición sea superior a 4,30?
3. Represente la función de distribución y de densidad de las medidas.

1. Resultados con Epidat 4:

Datos:

Distribución F de Snedecor (n,m)

n: Grados de libertad del numerador 10

m: Grados de libertad del denominador 12

Cola izquierda Pr[X<=x]: 0,95

Cola derecha Pr[X>x]: 0,05

Resultados:

Punto x
2,7534

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
1,2	0,9886	0,6857	0,72	2,8284	21

El valor que deja a la derecha una probabilidad de 0,05 es 2,75.

2. Resultados con Epidat 4:

Datos:

Distribución F de Snedecor (n,m)

Parámetros:

n: Grados de libertad del numerador 10

m: Grados de libertad del denominador 12

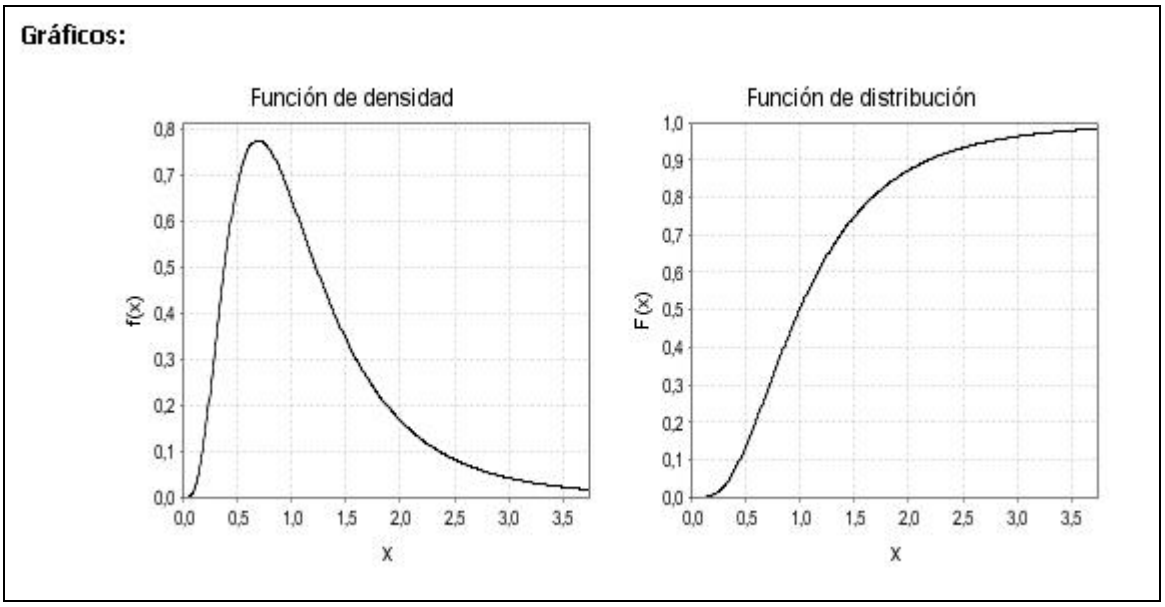
Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
4,3	0,99	0,01

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
1,2	0,9886	0,6857	0,72	2,8284	21

La probabilidad que deja a la derecha 4,30 es 0,01.

3. Las funciones de densidad y distribución de las medidas efectuadas se presentan a continuación:



13.1.2.11. Distribución Cauchy (μ, θ)

Esta distribución fue introducida por Simeón Denis Poisson (1781-1840) en 1824, aunque debe su nombre al matemático francés Augustin Louis Cauchy (1789-1857) quien la reintrodujo en 1853 [14]. En el ámbito de la física también es conocida con el nombre de distribución de Lorentz o distribución de Breit-Wigner.

La distribución de Cauchy depende de dos parámetros: escala (μ) y situación (θ); en el caso particular de que $\mu = 1$ y $\theta = 0$, recibe el nombre de distribución de Cauchy estándar.

Una característica destacable de esta distribución es que carece de momentos, por lo que no existen la media, varianza, asimetría y curtosis de esta distribución. Su función de densidad es simétrica respecto al parámetro de situación θ.

Epidat 4 limita los cálculos de esta distribución a valores del parámetro de escala menores o iguales que 30.

Campo de variación:

$-\infty < x < \infty$

Parámetros:

μ: escala, $\mu > 0$

θ: situación, $-\infty < \theta < \infty$

Ejemplo

Considere la distribución Cauchy de parámetros $\mu = 0{,}75$ y $\theta = 5$.

1. ¿Qué proporción del área bajo la curva se ubica a la derecha de 9,21?
2. ¿Qué valor de la variable aísla el 10% superior de la distribución?

1. Resultados con Epidat 4:

Datos:

Distribución Cauchy (μ, θ)
Parámetros:
μ: Escala 0,75
θ: Situación 5

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
9,21	0,9439	0,0561

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
No definida	5	5	No definida	No definida	No definida

El 5,6% del área bajo la curva se ubica a la derecha de 9,21.

2. Resultados con Epidat 4:

Datos:

Distribución Cauchy (μ, θ)
Parámetros:
μ: Escala 0,75
θ: Situación 5
Cola izquierda Pr[X<=x]: 0,9
Cola derecha Pr[X>x]: 0,1

Resultados:

Punto x
7,3083

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
No definida	5	5	No definida	No definida	No definida

El valor 7,3083 divide a la distribución en dos partes: el 90% de ésta queda a la izquierda de dicho punto y el 10% a la derecha.

13.1.2.12. Distribución Weibull (a, b)

Esta distribución debe su nombre al físico sueco Waloddi Weibull (1887-1979) quien la usó en un artículo publicado en 1939 sobre resistencia de los materiales (*A Statistical Theory of the Strength of Materials*), aunque ya era conocida de años antes.

Esta distribución se utiliza para modelar situaciones del tipo tiempo-fallo, modelar tiempos de vida o en el análisis de supervivencia, a parte de otros usos como, por ejemplo, caracterizar el comportamiento climático de la lluvia en un año determinado.

La distribución Weibull queda totalmente definida mediante dos parámetros, forma (a) y escala (b). En el caso particular de que a=1, se tiene la distribución exponencial, y si a = 2 y b = $\sqrt{2}\sigma$ recibe el nombre de distribución de Rayleigh.

El perfil de la función de densidad presenta formas muy variadas dependiendo del valor que tome su parámetro de forma, a. Si a es menor o igual que 1, la función de densidad es siempre decreciente; en caso de tomar valores mayores que la unidad su función de densidad muestra una forma más acampanada, pero no simétrica, de forma que crece hasta alcanzar el máximo y luego decrece.

En este caso, Epidat 4 limita ambos parámetros inferiormente por el valor 0,2 y superiormente por 200.

Campo de variación:

$0 < x < \infty$

Parámetros:

a: forma, $a > 0$

b: escala, $b > 0$

Ejemplo

La vida útil, en años, de cierto tipo de instrumental médico quirúrgico sigue una distribución de Weibull con parámetros a= 2 y b= 1,75.

1. ¿Cuál es la probabilidad de que el instrumental dure menos de 3 años?
2. Representar la función de densidad y de distribución de su vida útil.

1. Resultados con Epidat 4:

Datos:

Distribución Weibull (a, b)

Parámetros:

a: Forma	2
b: Escala	1,75

Resultados:

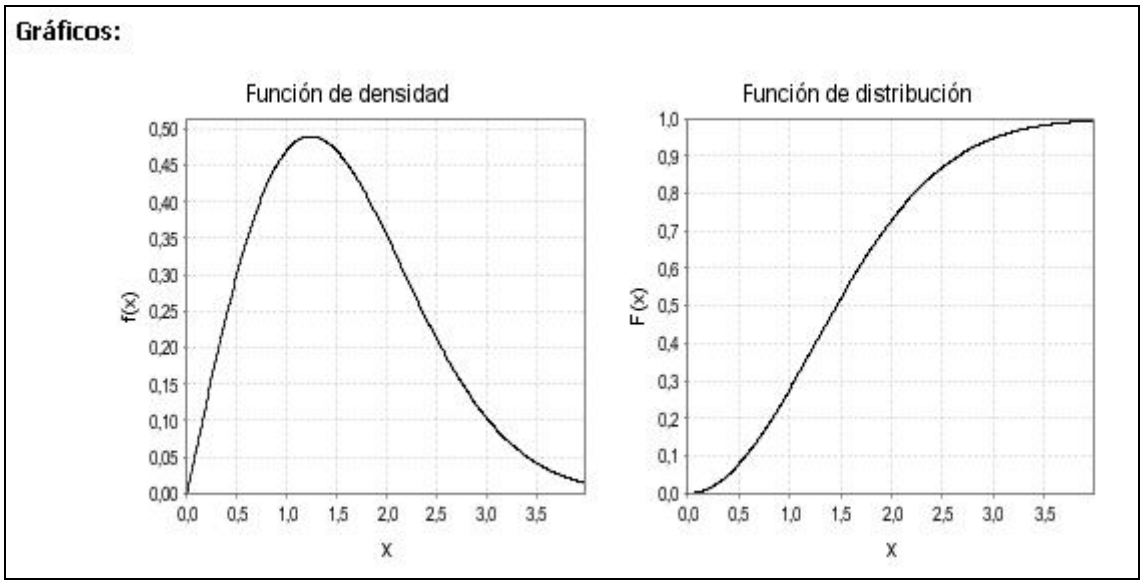
Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
2	0,7291	0,2709

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
1,5509	1,457	1,2374	0,6572	0,6311	0,2451

La probabilidad de que el instrumental dure menos de 3 años, es decir que dure 2 años o menos, es 0,73.

2. Resultados con Epidat 4:

Las funciones de densidad y distribución de la vida útil del instrumental médico son:



13.1.2.13. Distribución Laplace (a, b)

Fue descubierta en 1774 por Pierre-Simon Laplace (1749-1827), a quien debe su nombre, aunque también es conocida por el nombre de distribución doble exponencial.

Esta distribución viene determinada por dos parámetros, uno de situación (a) y otro de escala (b).

Su función de densidad es simétrica y el parámetro de situación determina su eje de simetría, además de ser el punto donde la función alcanza su valor máximo en forma de pico afilado. Independientemente de los valores que tomen sus parámetros, es una distribución leptocúrtica, lo que quiere decir que su función de densidad es más apuntada que la función de densidad de la normal con la misma media y desviación estándar.

Campo de variación:

$-\infty < x < \infty$

Parámetros:

a: situación, $-\infty < a < \infty$

b: escala, $b > 0$

Ejemplo

Una distribución es leptocúrtica si la función de densidad presenta un grado de apuntamiento mayor que el de la distribución normal con la misma media y varianza, lo que se traduce en un coeficiente de curtosis positivo. Comprobar gráficamente el carácter leptocúrtico de la distribución de Laplace para el caso particular en que a= 2 y b= 3.

En primer lugar hay que calcular la media y varianza de esta distribución para luego representar la función de densidad de la distribución normal correspondiente.

Resultados con Epidat 4:

Datos:

Distribución Laplace (a, b)

Parámetros:

a: Situación	2
b: Escala	3
Cola izquierda Pr[X<=x]:	0,5
Cola derecha Pr[X>x]:	0,5

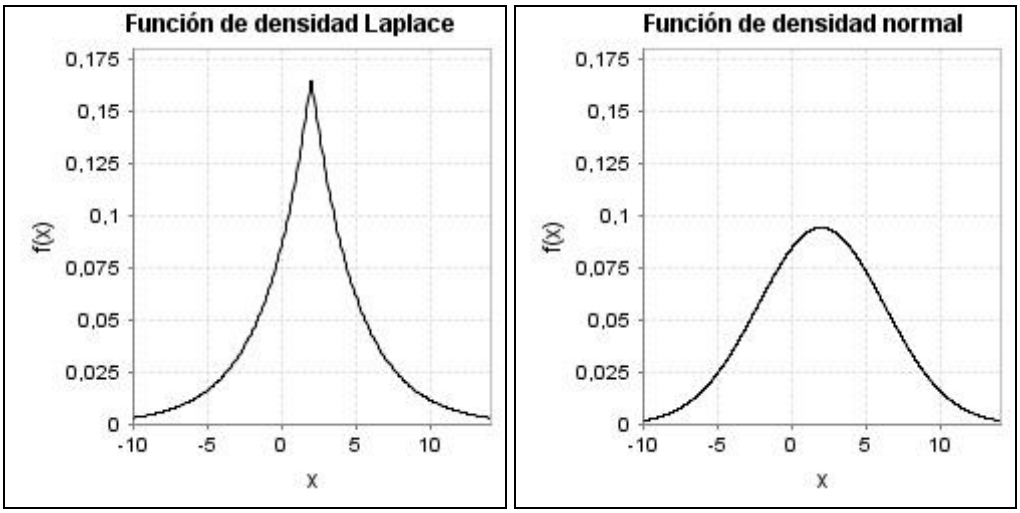
Resultados:

Punto x
2

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
2	2	2	18	0	3

La distribución de Laplace(2,3) se debe comparar gráficamente con la distribución normal de media 2 y varianza 18 (desviación típica 4,24).

Veamos a continuación la representación de ambas funciones de densidad:



A la vista de las gráficas se aprecia claramente que la función de densidad de la distribución Laplace(2,3) es más apuntada que la función de densidad de la normal con su misma media y desviación típica, tal como indicaba el valor del coeficiente de curtosis (3).

13.1.2.14. Distribución Pareto (α, x_0)

La distribución de Pareto fue introducida por el economista italiano Vilfredo Pareto (1848-1923) como un modelo para explicar la distribución de las rentas de los individuos de una población, siempre y cuando se partiera de dos supuestos, la existencia de un umbral inferior (x_0) de forma que no haya rentas inferiores a dicho umbral y el decrecimiento de manera

potencial del porcentaje de individuos con una renta superior o igual a un cierto valor de renta a medida que dicho valor de renta crece [8]. El uso de esta distribución se ha ido ampliando a diferentes ámbitos de estudio.

Se trata de una distribución biparamétrica, con parámetros de forma (α) y de situación (x_0). El parámetro x_0 es un indicador de posición (valor mínimo) que, en términos económicos, puede interpretarse como el ingreso mínimo de la población. El parámetro α está asociado con la dispersión, donde a mayor valor se obtienen densidades de Pareto más concentradas en las proximidades de x_0, es decir, menos dispersas.

Epidat 4 permite valores del parámetro de forma comprendidos entre 0,5 y 100, y valores del parámetro de situación entre 0,1 y 1.000.

Campo de variación:

$x_0 \leq x < \infty$

Parámetros:

α: forma, $\alpha > 0$

x_0: situación, $x_0 > 0$

Ejemplo

Los salarios mensuales, en euros, de una determinada empresa siguen una distribución de Pareto de parámetros α= 2,75 y x_0= 900 ¿Qué porcentaje de individuos tienen un salario superior a 2.000 euros? ¿Y a 3.000 euros?

Resultados con Epidat 4:

Datos:

Distribución Pareto (α , Xo)

Parámetros:

α: Forma	2,75
Xo: Situación	900

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
2.000	0,8887	0,1113
3.000	0,9635	0,0365

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
1.414,2857	1.157,9984	900	969.795,9184	No definida	No definida

Aproximadamente el 11% de los empleados de la empresa tienen un sueldo superior de 2.000 euros, mientras que un 3,7% de los empleados perciben un ingreso mensual superior a 3.000 euros.

13.1.2.15. Distribución triangular (a, c, b)

El nombre de esta distribución viene dado por la forma de su función de densidad. Este modelo proporciona una primera aproximación cuando hay poca información disponible, de forma que sólo se necesita conocer el mínimo (valor pesimista), el máximo (valor optimista) y la moda (valor más probable). Estos tres valores son los parámetros que caracterizan a la distribución triangular y se denotan por a, b y c, respectivamente.

Un ejemplo del uso de esta distribución se encuentra en el análisis del riesgo, donde la distribución más apropiada es la beta pero dada su complejidad, tanto en la su comprensión como en la estimación de sus parámetros, se utiliza la distribución triangular como proxy para la beta [15].

Campo de variación:

$a \leq x \leq b$

Parámetros:

a: mínimo, $-\infty < a < \infty$

c: moda, $-\infty < c < \infty$ con $a \leq c \leq b$

b: máximo, $-\infty < b < \infty$ con $a < b$

Ejemplo

Uno de los problemas de salud que afectan en mayor medida a la población en los meses de verano son los golpes de calor; por ese motivo, es necesario llevar un control de la temperatura atmosférica que alerta, entre otros indicadores, de la presencia de una ola de calor.

Durante el mes de Agosto del año 2010, en Santiago de Compostela, las temperaturas mínima y máxima absolutas fueron de 12,2 ºC y 35,8ºC, respectivamente, y el valor más probable fue de 19,8ºC. Si se asume que la temperatura sigue una distribución triangular de parámetros a=12,2, c=19,8 y b=35,8, ¿cuál es la probabilidad de que supere los 30ºC?

Resultados con Epidat 4:

Datos:

Distribución triangular (a, c, b)

Parámetros:

a: Mínimo	12,2
c: Moda	19,8
b: Máximo	35,8

Resultados:

Punto x	Cola izquierda Pr[X<=x]	Cola derecha Pr[X>x]
30	0,9109	0,0891

Media	Mediana	Moda	Varianza	Asimetría	Curtosis
22,6	22,0595	19,8	24,1867	0,3231	-0,6

La probabilidad de que la temperatura supere los 30 grados es de 0,089.

13.2. Generación de distribuciones

13.2.0. Conceptos generales

Epidat 4 ofrece procedimientos para generar muestras de variables aleatorias que se ajusten a determinadas distribuciones, tanto continuas como discretas. Además de las distribuciones disponibles en el submódulo de "Cálculo de probabilidades", en el presente se incluyen la multinomial, en las discretas, y la normal bivariante, en las continuas.

Este submódulo puede ser útil para realizar ejercicios de simulación (principalmente en estudios de investigación) y, además, para calcular probabilidades asociadas a variables obtenidas a partir de otras cuyas distribuciones sean conocidas, aun cuando la variable resultante tenga distribución desconocida.

El empleo de la simulación para verificar un resultado teórico es, hoy en día, una práctica regular gracias al desarrollo de los ordenadores que permiten obtener, rápida y fácilmente, números aleatorios de cualquier distribución. Esto ha supuesto una auténtica revolución en el campo de la estadística y, en particular, en los métodos bayesianos.

Más que números aleatorios estrictamente, los algoritmos de simulación generan lo que se ha denominado como números *pseudo-aleatorios* a través de fórmulas recursivas que parten de un valor inicial llamado semilla. Existen diferentes métodos de generación que permiten obtener una secuencia de números aleatorios para una distribución dada, pero la mayoría de estos métodos se basan en la generación de observaciones independientes de una distribución uniforme en [0,1]. El generador congruencial, propuesto por Lehmer [16], es uno de los más utilizados para obtener números aleatorios uniformes. Una recomendación muy extendida en la literatura es la de combinar varios generadores de números aleatorios para obtener un generador con mejores características.

Para generar valores de una distribución discreta, uno de los métodos más conocidos es el método de la transformación cuantil o método de inversión generalizada, que se basa en el siguiente resultado: si X es una variable aleatoria con función de distribución F y función cuantil Q y U es una variable aleatoria con distribución uniforme (0,1), entonces la variable Q(U) tiene la misma distribución que X.

La función cuantil de una distribución continua con función de distribución invertible coincide con la inversa de dicha función. Por eso, en este caso, el método de la transformación cuantil recibe el nombre de método de inversión, que es uno de los métodos más importantes en la generación de distribuciones continuas [17]. Su algoritmo se describe de la siguiente manera:

Paso 1: Generar valores de una variable con distribución uniforme (0,1).

Paso 2: Devolver $X=F^{-1}(U)$, siendo F una función de distribución invertible.

De esta forma se generan valores de la variable X con función de distribución F.

Existe otro método adecuado para los casos en que se desconoce la expresión explícita de la función de distribución pero sí se conoce su función de densidad. Este método se denomina método de aceptación-rechazo.

Los métodos de simulación se denominan, de modo general, técnicas de Monte Carlo. Estos métodos se utilizan en la resolución de diferentes problemas en los que la solución analítica exacta es difícil de obtener o consume mucho tiempo. En esos casos, se busca una solución aproximada mediante la simulación. El término Monte Carlo no hace referencia a un algoritmo concreto de simulación, sino más bien al hecho de que se ha aplicado un método

de ese tipo. Una aplicación de estas técnicas se da, por ejemplo, en el campo de la inferencia. El procedimiento se puede describir, de modo general, como sigue: se ajusta un modelo a los datos empíricos y se utiliza este modelo ajustado para simular muestras aleatorias que, a su vez, se usan para estimar los parámetros de la distribución teórica. Este procedimiento general se denomina *bootstrap* paramétrico.

13.2.1. Distribuciones discretas

Las distribuciones discretas incluidas en el submódulo de "Generación de distribuciones" son:

- Uniforme discreta
- Binomial
- Multinomial
- Hipergeométrica
- Geométrica
- Binomial negativa
- Pascal
- Poisson

Con excepción de la multinomial, todas fueron descritas en el submódulo precedente ("Cálculo de probabilidades"), de modo que ahora sólo se explicará dicha distribución.

13.2.1.1. Distribución multinomial

Como ya se comentó anteriormente, la distribución binomial aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, es decir, dos posibles resultados, clasificados generalmente como "éxito" o "fracaso". La distribución multinomial generaliza esta distribución al caso en que la población se divide en $m > 2$ grupos mutuamente excluyentes y exhaustivos o, equivalentemente, a experimentos con m resultados.

Se supone un proceso estable y sin memoria que genera elementos que pueden clasificarse en m grupos distintos o, dicho de otro modo, un experimento que tiene m posibles resultados. Supóngase que se toma una muestra de n elementos, o que el experimento se repite n veces de forma independiente, y se definen m variables aleatorias X_i = número de elementos del grupo i ($i = 1, ..., m$), entonces el vector de m-variables $(X_1, X_2, \ldots, X_m)$ es una nueva variable aleatoria m-dimensional que sigue una distribución multinomial de parámetros $n, p_1, ..., p_m$, donde p_i ($i = 1, ..., m$) es la probabilidad del grupo i.

Véase un ejemplo: de acuerdo con la teoría de la genética, un cierto cruce de conejillo de indias resultará en una descendencia roja, negra y blanca en la relación 8:4:4. Si se tienen 6 descendientes, el vector de variables (X_1, X_2, X_3) donde:

X_1 = Número de descendientes rojos

X_2 = Número de descendientes negros

X_3 = Número de descendientes blancos

sigue una distribución multinomial con parámetros $n = 6$; $p_1 = 8/16 = 0{,}5$; $p_2 = 4/16 = 0{,}25$ y $p_3 = 4/16 = 0{,}25$.

Una situación muy común en la práctica se da cuando se conoce el tamaño de muestra n y se quieren estimar las probabilidades p_i a partir de los valores observados. Pero también hay situaciones en las que se debe estimar el tamaño de muestra n, además de las probabilidades

p_i. Esto ocurre, por ejemplo, en el método de captura-recaptura, que fue desarrollado por zoólogos para estimar poblaciones animales y que ha sido aplicado a poblaciones humanas en estudios epidemiológicos.

Valores:

$x_i = 0, 1, 2, \ldots \quad (i = 1, \ldots, m)$

Parámetros:

n: número de pruebas, $n \geq 1$ entero

m: número de resultados posibles, $m \geq 3$ entero

p_i: probabilidad del suceso i, $0 < p_i < 1$ ($i = 1, \ldots, m$), donde $\sum_{i=1}^{m} p_i = 1$

Ejemplo

Volviendo al ejemplo de los conejillos de indias, supóngase que se está interesado en simular una muestra de tamaño 10 de una distribución multinomial con parámetros $n = 6$; $p_1 = 0{,}5$; $p_2 = 0{,}25$ y $p_3 = 0{,}25$.

Los resultados de Epidat indican que en la primera simulación los 6 conejitos de indias se organizaron de la siguiente manera: tres de ellos fueron descendientes rojos, un descendiente negro y dos descendientes blancos. En la segunda simulación, 4 de los conejitos fueron rojos, uno negro y otro blanco. Y así sucesivamente hasta llegar a la décima simulación donde tres de los conejitos fueron descendientes rojos, dos negros y uno blanco.

Resultados con Epidat 4:

Datos:

Distribución multinomial

Parámetros:

n: Número de pruebas	6
m: Número de resultados posibles	3
Tamaño de la muestra	10

Probabilidades Pi:

	Pi
1	0,5
2	0,25
3	0,25

Resultados:

Vector de medias		
3	1,5	1,5

Matriz de dispersión		
1,5	-0,75	-0,75
-0,75	1,125	-0,375
-0,75	-0,375	1,125

Muestra:

X_1	X_2	X_3
3	1	2
4	1	1
1	3	2
3	2	1
2	1	3
4	2	0
5	0	1
3	0	3
3	1	2
3	2	1

13.2.2. Distribuciones continuas

Las distribuciones continuas incluidas en el módulo de "Generación de distribuciones" son:

- Uniforme
- Normal
- Normal bivariante
- Lognormal
- Logística
- Beta
- Gamma
- Exponencial
- Ji-cuadrado
- t de Student
- F de Snedecor
- Cauchy
- Weibull
- Laplace
- Pareto
- Triangular

Con excepción de la normal bivariante, todas fueron descritas en el submódulo precedente ("Cálculo de probabilidades"), de modo que ahora sólo se explicará dicha distribución.

13.2.2.1. Distribución normal bivariante

Fue introducida por Carl Friedrich Gauss (1777-1855) a principios del siglo XIX en su estudio de errores de medida en las observaciones astronómicas y de cálculo de órbitas de cuerpos celestes, y se trata de la primera distribución continua multivariante estudiada. Como modelo de distribución teórico continuo, se adapta con gran aproximación a fenómenos reales en diversos campos de las ciencias sociales y la astronomía.

De igual modo que la distribución normal univariante está especificada por su media, μ, y su desviación estándar, σ, la función de densidad de la variable aleatoria normal bivariante $X=(X_1, X_2)$, está determinada por el vector de medias $\mu = (\mu_1, \mu_2)$, el vector de desviaciones estándar $\sigma = (\sigma_1, \sigma_2)$ y el coeficiente de correlación ρ entre las variables X_1 y X_2.

Si las variables aleatorias X_1 y X_2 son independientes, el coeficiente de correlación lineal es nulo y por tanto $\rho = 0$.

Por otro lado, al igual que para la distribución normal se tiene el caso particular de la distribución normal estándar, en el caso de la distribución normal bivariante se obtiene la normal bivariante estándar cuando las variables X_1 y X_2 son independientes e idénticamente distribuidas siguiendo una distribución N(0,1).

Campo de variación:

$-\infty < x_1 < \infty$

$-\infty < x_2 < \infty$

Parámetros:

$\mu = (\mu_1, \mu_2)$: vector de medias, $-\infty < \mu_1 < \infty$, $-\infty < \mu_2 < \infty$

$\sigma = (\sigma_1, \sigma_2)$: vector de desviaciones estándar, $\sigma_1 > 0$, $\sigma_2 > 0$

ρ: coeficiente de correlación, $-1 \leq \rho \leq 1$

Aquí, a diferencia de los restantes módulos, no se pondrán ejemplos pues no tiene mayor sentido, ya que la estructura de las aplicaciones siempre es la misma. No obstante, para

ilustrar la solución de un problema práctico por vía de la simulación, se considera el siguiente ejemplo en el que se aplica la distribución normal bivariante.

Ejemplo

Suponga que la distribución de la variable peso de una población de jóvenes sigue una distribución normal de media μ = 65 kg y desviación estándar σ = 15 kg. Suponga, además, que la variable altura en dicha población sigue una distribución normal de media μ = 1,68 m y desviación estándar σ = 0,20 m. La correlación entre las dos variables es alta, de un 0,75. Con estos datos estimar el porcentaje de obesos en la población teniendo en cuenta que la obesidad está definida por un índice de masa corporal (IMC = peso/talla^2) superior a 30 kg/m^2.

Para calcular el porcentaje hay que simular los valores de la variable *IMC*, pues no se dispone de la distribución teórica. Los pasos a seguir serán los siguientes:

1. Simular 1.000 valores de la distribución normal bivariante con los siguientes parámetros: media y desviación estándar del peso, media y desviación estándar de la talla, y el coeficiente de correlación entre la talla y el peso.
2. Llevar los valores de la variable simulada a una hoja de cálculo (por ejemplo) y efectuar el cociente *IMC* = peso/talla^2.
3. Contar el número de valores de la variable *IMC* que superan el umbral 30 kg/m^2 (condición de obesidad).

Resultados con Epidat 4:

Datos:

Distribución normal bivariante (μ, σ, ρ)
Parámetros:

μ: Vector de medias	(68 1,68)
σ: Vector de desviaciones estándar	(15 0,2)
ρ: Coeficiente de correlación	0,75
Tamaño de la muestra	1.000

Resultados:

Matriz de dispersión	
225	2,25
2,25	0,04

Muestra:

*Resultados guardados en: C:\Archivos de programa\Epidat4\Ejemplos\Distribuciones de probabilidad\MuestraNormalBivariante.xls

Con los 1.000 valores simulados se obtiene un porcentaje de sujetos con un *IMC* superior a 30 kg/m^2 del 5,7%.

Nota: Cada vez que se realiza una nueva simulación se obtienen valores diferentes, aunque se mantenga la misma distribución, el valor de sus parámetros y el tamaño de la muestra.

Bibliografía

1 Kolmogorov AN. Grundbegriffe der wahrscheinlichtkeitsrechnung. Berlin: Springer-Verlag; 1933. (Traducido al inglés: Morrison N. Foundations of the theory of probability. New York: Chelsea; 1956).

2 Peña D. Modelos y métodos. 1. Fundamentos. Madrid: Alianza Universidad Textos; 1993.

3 Meyer PL. Probabilidad y aplicaciones estadísticas. 2ª ed. Bógota: Fondo Educativo Interamericano; 1973.

4 Martín-Pliego J, Ruiz-Maya L. Estadística I: probabilidad. 2ª ed. Madrid: Thomson; 2004.

5 Katz DL. Epidemiology, biostatistics and preventive medicine review. USA: W.B. Saunders Company; 1997.

6 Hospital Ramón y Cajal [página en Internet]. Material docente de la unidad de bioestadística clínica. Disponible en: http://www.hrc.es/bioest/M_docente.html

7 Doménech JM. Métodos estadísticos en ciencias de la salud. Barcelona: Signo; 1997.

8 Fernández-Abascal H, Guijarro MM, Rojo JL, Sanz JA. Cálculo de probabilidades y estadística. Barcelona: Editorial Ariel; 1994.

9 Kemp AW, Kemp CD. Accident proneness. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 1. Chichester: John Wiley & Sons; 1998. pp. 35-7.

10 Biggeri A. Negative binomial distribution. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 4. Chichester: John Wiley & Sons; 1998. pp. 2962-7.

11 Canavos GC. Probabilidad y estadística: aplicaciones y métodos. México: McGraw-Hill; 1988.

12 Palmgren J. Poisson distribution. En: Armitage P, Colton T, editores. Encyclopedia of Biostatistics Vol. 4. Chichester: John Wiley & Sons; 1998. pp. 3398-402.

13 Student. The probable error of a mean. Biometrika. 1908;6:1-25.

14 John Aldrich. University of Southampton [página en Internet]. Figures from the history of probability and statistics. Disponible en: http://www.economics.soton.ac.uk/staff/aldrich/Figures.htm

15 Johnson D. The triangular distribution as a Proxy for the beta distribution in risk analysis. The Statistician. 1997;46(3):387-98.

16 Lehmer DH. Mathematical methods in large-scale computing units. En: Proceedings of the second symposium on large scale digital computing units machinery. Cambridge, Mass.: Harvard University Press; 1951. pp. 141-6.

17 Cao Abad R. Introducción a la simulación y a la teoría de colas. A Coruña: Netbiblo; 2002.

Anexo 1: Novedades del módulo de distribuciones de probabilidad

Novedades de la versión 4.0 con respecto a la versión 3.1:

- Se añadió una distribución discreta: Pascal
- Se añadieron las siguientes distribuciones continuas:
    - Cauchy
    - Weibull
    - Laplace
    - Pareto
    - Triangular
- Es posible calcular probabilidades para más de un punto a la vez.
- Los gráficos generados para las funciones de distribución y densidad pueden personalizarse mediante el editor de gráficos.

Anexo 2: Fórmulas del módulo de distribuciones de probabilidad

Esquema del módulo

1. Cálculo de probabilidades
   1.1. Distribuciones discretas
   1.2. Distribuciones continuas
2. Generación de distribuciones
   2.1. Distribuciones discretas
   2.2. Distribuciones continuas

1.- DISTRIBUCIONES DISCRETAS

1. Uniforme discreta
2. Binomial
3. Multinomial
4. Hipergeométrica
5. Geométrica
6. Binomial negativa
7. Pascal
8. Poisson

1.1.- Uniforme discreta en (a,b) [Fernández-Abascal (1994, p. 388-391); Weissten]

X= "Número entero seleccionado aleatoriamente entre a y b"

Parámetros de la distribución:

- a: Mínimo (entero ≥ 1; en Epidat: a ≥ 1)
- b: Máximo (entero ≥ 2, a < b; en Epidat: b ≥ 2)

Función de masa de probabilidad:

$$f(k) = \frac{1}{N}, \; a \le k \le b, \text{ k entero}$$

Valores característicos:

Media: $\frac{a+b}{2}$

Varianza: $\frac{N^2-1}{12}$

Asimetría: 0

Curtosis: $\frac{-6(N^2+1)}{5(N^2-1)}$

Donde N=b-a+1 es el número de enteros entre a y b.

1.2.- Binomial (n,p) [Canavos (1988, p. 89-99)]

X= "Número de éxitos en n pruebas independientes"

Parámetros de la distribución:

- n: Número de pruebas (entero ≥ 1; en Epidat: 1 ≤ n ≤ 1.000)
- p: Probabilidad de éxito (0 < p < 1; en Epidat: 0 < p < 1)

Función de masa de probabilidad:

$$f(k)=\binom{n}{k}p^k(1-p)^{n-k},\ 0\le k\le n,\ k \text{ entero, siendo } \binom{n}{k}=\frac{n!}{k!(n-k)!}$$

Valores característicos:

Media: np

Varianza: np(1-p)

Asimetría: $\dfrac{1-2p}{\sqrt{np(1-p)}}$

Curtosis: $\dfrac{1-6p(1-p)}{np(1-p)}$

1.3.- Multinomial (n, p_1, p_2, …, p_m) [Martín-Pliego y Ruiz-Maya (2004, p. 379-382)]

X = (X_1, X_2, …, X_m) = "Número de veces que ocurren m sucesos disjuntos en n pruebas independientes"

Parámetros de la distribución:

- n: Número de pruebas (entero ≥ 1; en Epidat: n ≥ 1)
- m: Número de resultados posibles (entero ≥ 3; en Epidat: m ≥ 3)
- p_i: Probabilidad del suceso i, i=1, …, m (0 < p_i < 1 y $\sum_{i=1}^{m} p_i = 1$; en Epidat: 0 < p_i < 1)

Función de masa de probabilidad:

$$f(k_1, k_2, \ldots, k_m) = \frac{n!}{k_1! k_2! \ldots k_m!} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m}, \; k_i \in \{0, 1, \ldots, n\} \text{ y } \sum_{i=1}^{m} k_i = n$$

Valores característicos:

Vector de medias: $(np_1 \;\; np_2 \ldots np_m)$

Matriz de dispersión: $\Sigma = \left[Cov(X_i, X_j) \right]_{i,j=1,\ldots,m}$

$$Cov(X_i, X_j) = -np_i p_j \text{ si } i \neq j$$

$$Cov(X_i, X_i) = Var(X_i) = np_i(1 - p_i)$$

1.4.- Hipergeométrica (N, R, n) [Canavos (1988, p. 108-115)]

X= "Número de éxitos en n elementos extraídos, sin reposición, de una población de tamaño N que contiene R éxitos".

Parámetros de la distribución:

- N: Tamaño de la población (entero ≥ 1; en Epidat: 1 ≤ N ≤ 1.000)
- R: Número de éxitos en la población (entero, 1 ≤ R ≤ N; en Epidat: 1 ≤ R ≤ N)
- n: Número de pruebas (entero, 1 ≤ n ≤ N; en Epidat: 1 ≤ n ≤ N)

Función de masa de probabilidad:

$$f(k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}, \; \max(0, n-N+R) \leq k \leq \min(R, n), \text{ k entero}$$

Valores característicos:

Media: np

Varianza: $np(1-p)\dfrac{N-n}{N-1}$

Asimetría: $$\frac{\frac{N-2n}{N-2}(1-2p)}{\sqrt{\frac{N-n}{N-1}np(1-p)}}$$

Curtosis: $$A+\frac{BC}{np(1-p)}-3$$

Donde:

- $p=\frac{R}{N}$
- $A=\frac{3(N-1)(N+6)}{(N-2)(N-3)}$
- $B=\frac{(N-1)N(N+1)}{(N-n)(N-2)(N-3)}$
- $C=1-\frac{6N\left(p(1-p)+n\frac{N-n}{N^2}\right)}{N+1}$

1.5.- Geométrica (p) [Fernández-Abascal (1994, p. 403-407); Weissten]

X= "Número de fracasos antes del primer éxito"

Parámetros de la distribución:

- p: Probabilidad de éxito (0 < p < 1; en Epidat: 0 < p < 1)

Nota: Geométrica (p) = Binomial negativa (1, p)

Función de masa de probabilidad:

$$f(k)=p(1-p)^k,\ k\geq 0,\ k \text{ entero}$$

Valores característicos:

Media: $$\frac{1-p}{p}$$

Varianza: $\frac{1-p}{p^2}$

Asimetría: $\frac{2-p}{\sqrt{1-p}}$

Curtosis: $\frac{p^2-6p+6}{1-p}$

1.6.- Binomial negativa (r, p) [Canavos (1988, p. 115-121)]

X= "Número de fracasos antes de obtener r éxitos"

Parámetros de la distribución:

- r: Número de éxitos (entero ≥ 1; en Epidat: 1 ≤ r ≤ 1.000)
- p: Probabilidad de éxito (0 < p < 1; en Epidat: 0 < p < 1)

Función de masa de probabilidad:

$$f(k)=\binom{k+r-1}{r-1}p^r(1-p)^k,\ k \geq 0,\ k \text{ entero}$$

Valores característicos:

Media: $\frac{r(1-p)}{p}$

Varianza: $\frac{r(1-p)}{p^2}$

Asimetría: $\frac{2-p}{\sqrt{r(1-p)}}$

Curtosis: $\frac{p^2-6p+6}{r(1-p)}$

1.7.- Pascal (r,p) [Meyer (1994, p. 178-179) ; Canavos (1988, p. 115-121)]

X= "Número de pruebas necesarias para obtener r éxitos"

Parámetros de la distribución:

- r: Número de éxitos (entero ≥ 1; en Epidat: 1 ≤ r ≤ 1.000)
- p: Probabilidad de éxito (0 < p < 1; en Epidat: 0 < p < 1)

Nota: $\text{Pascal}(r,p) = \text{BN}(r,p) + r$

Función de masa de probabilidad:

$$f(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \; k \geq r, \text{ k entero}$$

Valores característicos:

Media: $\dfrac{r}{p}$

Varianza: $\dfrac{r(1-p)}{p^2}$

Asimetría: $\dfrac{2-p}{\sqrt{r(1-p)}}$

Curtosis: $\dfrac{p^2 - 6p + 6}{r(1-p)}$

1.8.- Poisson (λ) [Canavos (1988, p. 100-108)]

X= "Número de ocurrencias de un evento raro en un intervalo continuo de tiempo o espacio"

Parámetros de la distribución:

- λ: Tasa de ocurrencia (λ > 0; en Epidat: 0 < λ ≤ 50)

Función de masa de probabilidad:

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ , } k \geq 0, \text{ k entero}$$

Valores característicos:

Media=Varianza: λ

Asimetría: $\frac{1}{\sqrt{\lambda}}$

Curtosis: $\frac{1}{\lambda}$

2.- DISTRIBUCIONES CONTINUAS

1. Uniforme o rectangular
2. Normal
3. Normal bivariante
4. Lognormal
5. Logística
6. Beta
7. Gamma
8. Exponencial
9. Ji-cuadrado
10. t de Student
11. F de Snedecor
12. Cauchy
13. Weibull
14. Laplace
15. Pareto
16. Triangular

2.1.- Uniforme (a, b) o rectangular (a, b) [Canavos (1988, p. 143-147)]

Parámetros de la distribución:

- a: Mínimo ($-\infty < a < \infty$; en Epidat: $-\infty < a < \infty$)
- b: Máximo ($-\infty < b < \infty$, $a < b$; en Epidat: $-\infty < b < \infty$)

Función de densidad:

$$f(x) = \frac{1}{b-a}, \; a < x < b$$

Valores característicos:

Media = Mediana: $\frac{a+b}{2}$

Moda: intervalo (a, b)

Varianza: $\frac{(b-a)^2}{12}$

Asimetría: 0

Curtosis: -6/5

2.2.- Normal (μ, σ) [Canavos (1988, p. 130-143)]

Parámetros de la distribución:

- μ: Media ($-\infty < \mu < \infty$; en Epidat: $-\infty < \mu < \infty$)
- σ: Desviación estándar ($\sigma > 0$; en Epidat: $\sigma > 0$)

Función de densidad:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \; -\infty < x < \infty$$

Valores característicos:

Media = Mediana = Moda: μ

Varianza: σ^2

Asimetría: 0

Curtosis: 0

Nota: con μ=0 y σ=1 se tiene la distribución normal estándar, N(0,1).

2.3.- Normal bivariante (μ, σ, ρ) [Martín-Pliego y Ruiz-Maya (2004, p. 459-460)]

Parámetros de la distribución:

- $\mu=(\mu_x \;\; \mu_y)$: Vector de medias ($-\infty < \mu_x, \mu_y < \infty$; en Epidat: $-\infty < \mu_x, \mu_y < \infty$)
- $\sigma=(\sigma_x \;\; \sigma_y)$: Vector de desviaciones estándar ($\sigma_x, \sigma_y > 0$; en Epidat: $\sigma_x, \sigma_y > 0$)
- ρ: Coeficiente de correlación ($-1 \le \rho \le 1$; en Epidat: $-1 \le \rho \le 1$)

Función de densidad:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{1-\rho^2}\left[\frac{(x-\mu_x)^2}{2\sigma_x^2}-\frac{\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{2\sigma_y^2}\right]\right\}$$

$-\infty < x, y < \infty$

Valores característicos:

Vector de medias: $\mu=(\mu_x \;\; \mu_y)$

Matriz de dispersión: $\Sigma=\begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix}$

2.4.- Lognormal (μ, σ) [Fernández-Abascal (1994, p. 445-448); Weissten]

Parámetros de la distribución:

- μ: Escala ($-\infty < \mu < \infty$; en Epidat: $-5 \le \mu \le 5$)
- σ: Forma ($\sigma > 0$; en Epidat: $0 < \sigma \le 5$)

Nota: Si $X \approx$ Lognormal (μ, σ) $\Rightarrow$ Ln(X) $\approx$ Normal (μ, σ)

Función de densidad:

$$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right],\ x>0$$

Valores característicos:

Media: $e^{\mu+\frac{\sigma^2}{2}}$

Mediana: e^{μ}

Moda: $e^{\mu-\sigma^2}$

Varianza: $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$

Asimetría: $\sqrt{\left(e^{\sigma^2}-1\right)}\left(e^{\sigma^2}+2\right)$

Curtosis: $e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$

2.5.- Logística (a, b) [Fernández-Abascal (1994, p. 464-466); Weissten]

Parámetros de la distribución:

- a: Situación ($-\infty < a < \infty$; en Epidat: $-\infty < a < \infty$)
- b: Escala ($b > 0$; en Epidat: $b > 0$)

Función de densidad:

$$f(x)=\frac{1}{b}\frac{e^{-(x-a)/b}}{\left[1+e^{-(x-a)/b}\right]^2},\ -\infty < x < \infty$$

Valores característicos:

Media = Mediana = Moda: a

Varianza: $\frac{\pi^2}{3}b^2$

Asimetría: 0

Curtosis: $\frac{6}{5}$

2.6.- Beta (p, q) [Canavos (1988, p. 147-151); Weissten]

Parámetros de la distribución:

- p: Forma ($p > 0$; en Epidat: $0 < p \leq 100$)
- q: Forma ($q > 0$; en Epidat: $0 < q \leq 100$)

Función de densidad:

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}, \; 0 < x < 1$$

donde B es la función beta: $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$.

Valores característicos:

Media: $\frac{p}{p+q}$

Mediana: no tiene expresión explícita

Moda: $\frac{p-1}{p+q-2}$ para $p > 1$ y $q > 1$

Varianza: $\frac{pq}{(p+q)^2(1+p+q)}$

Asimetría: $\frac{2(q-p)\sqrt{p+q+1}}{(p+q+2)\sqrt{pq}}$

Curtosis: $6\frac{p(p+1)(p-2q)+q(q+1)(q-2p)}{pq(p+q+2)(p+q+3)}$

2.7.- Gamma Γ(a, p) [Fernández-Abascal (1994, p. 448-452); Weissten]

Parámetros de la distribución:

- a: Escala (a > 0; en Epidat: 0 < a ≤ 25)
- p: Forma (p > 0; en Epidat: 0 < p ≤ 25)

Función de densidad:

$$f(x) = \frac{a^p}{\Gamma(p)} e^{-ax} x^{p-1}, x > 0$$

donde Γ es la función gamma: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$, y si n es un entero: $\Gamma(n) = (n-1)!$

Valores característicos:

Media: $\frac{p}{a}$

Mediana: no tiene expresión explícita

Moda: $\frac{p-1}{a}$ para $p > 1$

Varianza: $\frac{p}{a^2}$

Asimetría: $\frac{2}{\sqrt{p}}$

Curtosis: $\frac{6}{p}$

2.8.- Exponencial (λ) [Fernández-Abascal (1994, p. 452-455); Weissten]

Parámetros de la distribución:

- λ: Tasa (λ > 0; en Epidat: 0 < λ ≤ 100)

Nota: Exponencial (λ) = Gamma (λ, 1).

Función de densidad:

$$f(x) = \lambda e^{-\lambda x}, x > 0$$

Valores característicos:

Media: $\frac{1}{\lambda}$

Mediana: $\frac{\ln 2}{\lambda}$

Moda: no definida

Varianza: $\frac{1}{\lambda^2}$

Asimetría: 2

Curtosis: 6

2.9.- Ji-cuadrado (n) [Fernández-Abascal (1994, p. 473-478); Weissten]

Parámetros de la distribución:

- n: Grados de libertad (entero ≥ 1; en Epidat: 1 ≤ n ≤ 150)

Nota: Ji-cuadrado (n) = Gamma (1/2, n/2).

Función de densidad:

$$f(x)=\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}, x>0$$

donde Γ es la función gamma: $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$, y si n es un entero: $\Gamma(n)=(n-1)!$

Valores característicos:

Media: n

Mediana: no tiene expresión explícita

Moda: n-2 para n > 2

Varianza: 2n

Asimetría: $\sqrt{\frac{8}{n}}$

Curtosis: $\frac{12}{n}$

2.10.- t-Student (n) [Fernández-Abascal (1994, p. 478-481); Weissten]

Parámetros de la distribución:

- n: Grados de libertad (entero ≥ 1; en Epidat: 1 ≤ n ≤ 150)

Función de densidad:

$$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, -\infty<x<\infty$$

Valores característicos:

Media = Mediana = Moda: 0

Varianza: $\dfrac{n}{n-2}$ para $n > 2$

Asimetría: 0

Curtosis: $\dfrac{6}{n-4}$ para $n > 4$

2.11.- F-Snedecor (n, m) [Fernández-Abascal (1994, p. 482-486);Weissten]

Parámetros de la distribución:

- n: Grados de libertad del numerador (entero ≥ 1; en Epidat: $1 \le n \le 150$)
- m: Grados de libertad del denominador (entero ≥ 1; en Epidat: $1 \le m \le 150$)

Función de densidad:

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{\left(\frac{n}{m}\right)^{n/2} x^{n-2/2}}{\left(1+\frac{nx}{m}\right)^{n+m/2}}, \; x > 0$$

Valores característicos:

Media: $\dfrac{m}{m-2}$ para $m > 2$

Mediana: no tiene expresión explícita

Moda: $\dfrac{m(n-2)}{n(m+2)}$ para $n > 2$

Varianza: $\dfrac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ para $m > 4$

Asimetría: $\dfrac{(2n+m-2)\sqrt{8(m-4)}}{(m-6)\sqrt{n(n+m-2)}}$ para $m > 6$

Curtosis: $\dfrac{12\left[(m-2)^2(m-4)+n(n+m-2)(5m-22)\right]}{n(m-6)(m-8)(n+m-2)}$ para m > 8

2.12.- Cauchy (μ, θ) [Fernández-Abascal (1994, p. 461-463)]

Parámetros de la distribución:

- μ: Escala (μ > 0; en Epidat: 0 < μ ≤ 30)
- θ: Situación (-∞ < θ < ∞; en Epidat: -∞ < θ < ∞)

Función de densidad:

$$f(x)=\frac{\mu}{\pi}\frac{1}{(\mu^2+(x-\theta)^2)},\ -\infty<x<\infty$$

Valores característicos:

Media: no definida

Mediana = Moda: θ

Varianza: no definida

Asimetría: no definida

Curtosis: no definida

Nota: Con μ=1 y θ=0 se tiene la distribución de Cauchy estándar.

2.13.- Weibull (a, b) [Canavos (1988, p. 159-163)]

Parámetros de la distribución:

- a: Forma (a > 0; en Epidat: 0,2 ≤ a ≤ 200)
- b: Escala (b > 0; en Epidat: 0,2 ≤ b ≤ 200)

Función de densidad:

$$f(x) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} \exp\left(-\left(\frac{x}{b}\right)^{a}\right), x > 0$$

Valores característicos:

Media: $b\,\Gamma\left(\frac{1}{a}+1\right)$

Mediana: $b(\ln 2)^{1/a}$

Moda: $b\left(\frac{a-1}{a}\right)^{1/a}$ para a > 1

Varianza: $b^2\left[\Gamma\left(\frac{2}{a}+1\right)-\left[\Gamma\left(\frac{1}{a}+1\right)\right]^2\right]$

Asimetría: $$\frac{2\Gamma^3\left(\frac{1}{a}+1\right)-3\Gamma\left(\frac{1}{a}+1\right)\Gamma\left(\frac{2}{a}+1\right)+\Gamma\left(\frac{3}{a}+1\right)}{\left[\Gamma\left(\frac{2}{a}+1\right)-\Gamma^2\left(\frac{1}{a}+1\right)\right]^{3/2}}$$

Curtosis:

$$\frac{-6\Gamma^4\left(\frac{1}{a}+1\right)+12\Gamma^2\left(\frac{1}{a}+1\right)\Gamma\left(\frac{2}{a}+1\right)-3\Gamma^2\left(\frac{2}{a}+1\right)-4\Gamma\left(\frac{1}{a}+1\right)\Gamma\left(\frac{3}{a}+1\right)+\Gamma\left(\frac{4}{a}+1\right)}{\left[\Gamma\left(\frac{2}{a}+1\right)-\Gamma^2\left(\frac{1}{a}+1\right)\right]^2}$$

2.14.- Laplace (a, b) [Weissten]

Parámetros de la distribución:

- a: Situación (-∞ < a < ∞; en Epidat: -∞ < a < ∞)
- b: Escala (b > 0; en Epidat: b > 0)

Función de densidad:

$$f(x) = \frac{1}{2b}\exp\left(-\frac{|x-a|}{b}\right), -\infty < x < \infty$$

Valores característicos:

Media = Mediana = Moda: a

Varianza: $2b^2$

Asimetría: 0

Curtosis: 3

2.15.- Pareto (α, x_0) [Fernández-Abascal (1994, p. 459-461); Weissten]

Parámetros de la distribución:

- α: Forma ($\alpha > 0$; en Epidat: $0{,}5 \le \alpha \le 100$)
- x_0: Situación ($x_0 > 0$; en Epidat: $0{,}1 \le x_0 \le 1.000$)

Función de densidad:

$$f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, x \ge x_0$$

Valores característicos:

Media: $\frac{\alpha x_0}{\alpha - 1}$ para $\alpha > 1$

Mediana: $x_0 2^{1/\alpha}$

Moda: x_0

Varianza: $\frac{\alpha x_0^2}{(\alpha-2)(\alpha-1)^2}$ para $\alpha > 2$

Asimetría: $\frac{2(1+\alpha)}{\alpha-3}\sqrt{\frac{\alpha-2}{\alpha}}$ para $\alpha > 3$

Curtosis: $\frac{6(\alpha^3+\alpha^2-6\alpha-2)}{\alpha(\alpha-3)(\alpha-4)}$ para $\alpha > 4$

2.16.- Triangular (a, c, b) [Herrerías y Palacios (2007, p. 5-6)]

Parámetros de la distribución:

- a: Mínimo ($-\infty < a < \infty$; en Epidat: $-\infty < a < \infty$)
- c: Moda ($-\infty < c < \infty$, $a \le c \le b$; en Epidat: $-\infty < c < \infty$)
- b: Máximo ($-\infty < b < \infty$, $a < b$; en Epidat: $-\infty < b < \infty$)

Función de densidad:

$$f(x) = \frac{2(x-a)}{(b-a)(c-a)} \quad \text{para} \quad a \le x \le c$$

$$f(x) = \frac{2(b-x)}{(b-a)(b-c)} \quad \text{para} \quad c < x \le b$$

Valores característicos:

Media: $\frac{a+b+c}{3}$

Mediana: $\begin{cases} b-\sqrt{\frac{(b-a)(b-c)}{2}} & \text{si} \quad c \le \frac{a+b}{2} \\ a+\sqrt{\frac{(b-a)(c-a)}{2}} & \text{si} \quad c > \frac{a+b}{2} \end{cases}$

Moda: c

Varianza: $\frac{(b-c)^2+(c-a)^2+(b-c)(c-a)}{18}$

Asimetría: $\dfrac{\sqrt{2}(a+b-2c)(b+c-2a)(2b-c-a)}{5\left[(b-a)^2-(c-a)(b-c)\right]^{3/2}}$

Curtosis: $-\dfrac{3}{5}$

Bibliografía

- Canavos GC. Probabilidad y estadística: aplicaciones y métodos. Madrid: McGraw-Hill; 1988.

- Fernández-Abascal H, Guijarro MM, Rojo JL, Sanz JA. Cálculo de probabilidades y estadística. Barcelona: Editorial Ariel; 1994.

- Herrerías Pleguezuelo R, Palacios González F. Curso de inferencia estadística y del modelo lineal simple. Madrid: Delta, Publicaciones Universitarias; 2007.

- Martín-Pliego J, Ruiz-Maya L. Estadística I: Probabilidad. 2ª ed. Madrid: Thomson; 2004.

- Meyer PL. Probabilidad y aplicaciones estadísticas. 2ª ed. Bogotá: Fondo Educativo Interamericano; 1973.

- Weisstein EW. From MathWorld-A Wolfram Web Resource [página en internet]. Statistical Distribution. Disponible en:
http://mathworld.wolfram.com/topics/StatisticalDistributions.html

Anexo 3: Resumen de las distribuciones discretas

Distribución	Valores	Parámetros	Definición de la variable	Observaciones
Uniforme discreta	a, a+1, a+2, …,b	a: mínimo b: máximo	Variable que puede tomar n valores distintos con la misma probabilidad cada uno de ellos	
Binomial	0, 1, 2, …, n	n: número de pruebas p: probabilidad de éxito	Número de éxitos en n pruebas independientes de un experimento con probabilidad de éxito constante	Esta distribución se aplica a poblaciones finitas cuando los elementos se toman al azar y con reemplazo, y a poblaciones conceptualmente infinitas cuando el proceso es estable y sin
Multinomial	X_i: 0, 1, 2, … (i= 1, …, m)	n: número de pruebas m: nº de resultados posibles p_i: probabilidad del suceso i	Número de veces que ocurren m sucesos disjuntos en n pruebas independientes	Se aplica cuando se tiene un proceso estable y sin memoria
Hipergeométrica	de max{0,n-(N-R)} a min{R,n}	N: tamaño de la población R: número de éxitos n: número de pruebas	Número de éxitos en una muestra de tamaño n, extraída sin reemplazo de una población de tamaño N que contiene R éxitos	Es equivalente a la distribución binomial cuando el muestreo se hace sin reemplazo. Si el tamaño de la población es grande ambas distribuciones se pueden considerar prácticamente iguales
Geométrica	0, 1, 2, …	p: probabilidad de éxito	Número de fracasos antes de obtener un éxito por primera vez	Se utiliza en la distribución de tiempos de espera y tiene la propiedad de "falta de memoria"
Binomial negativa	0, 1, 2, …	r: número de éxitos p: probabilidad de éxito	Número de fracasos antes de obtener el r-ésimo éxito	Cuando r=1 se obtiene la distribución geométrica
Pascal	r, r+1, r+2, …	r: número de éxitos p: probabilidad de éxito	Número de pruebas necesarias para obtener r éxitos	Se relaciona con la binomial negativa de la siguiente manera: Pascal(r,p)=BN(r,p)+r
Poisson	0, 1, 2, …	λ: tasa de ocurrencia	Número de ocurrencias de un evento "raro" o poco frecuente en un intervalo o espacio continuo de tiempo	El proceso que genera una distribución de Poisson es estable y no tiene memoria. La distribución binomial se aproxima por la Poisson si n es grande y p pequeña, siendo λ=np

Anexo 4: Resumen de las distribuciones continuas

Distribución	Campo de variación	Parámetros	Observaciones
Uniforme	(a, b)	a: mínimo	Distribución clave en la generación de distribuciones
		b: máximo	
Normal	(-∞, ∞)	μ: media	Si μ=0 y σ=1 se denomina distribución normal estándar De ella derivan las distribuciones ji-cuadrado, t de Student y F de Snedecor
		σ: desviación estándar	
Normal bivariante	$X_1 \in (-\infty, \infty)$ $X_2 \in (-\infty, \infty)$	$\mu=(\mu_1, \mu_2)$: media	
		$\sigma=(\sigma_1, \sigma_2)$: desviación estándar	
		ρ: coeficiente de correlación	
Lognormal	(0, ∞)	μ: escala	Si X sigue una distribución lognormal entonces su logaritmo neperiano sigue una distribución normal
		σ: forma	
Logística	(-∞, ∞)	a: situación	Si U sigue una distribución uniforme en el intervalo (0, 1) entonces X=ln(U/(1-U)) sigue una distribución logística
		b: escala	
Beta	(0, 1)	p: forma	Es adecuada para modelar proporciones Si p=q=1 se obtiene la distribución uniforme en (0, 1)
		q: forma	
Gamma	(0, ∞)	a: escala	Es adecuada para modelar tiempos de vida Si p es un nº entero se denomina distribución de Erlang
		p: forma	
Exponencial	(0, ∞)	λ: tasa	Equivalente continuo de la distribución geométrica, también posee la propiedad de "falta de memoria"
Ji-cuadrado	(0, ∞)	n: grados de libertad	Distribuciones importantes en la contrastación de hipótesis estadísticas
t de Student	(-∞, ∞)	n: grados de libertad	
F de Snedecor	(0, ∞)	n: grados de libertad	
		m: grados de libertad	
Cauchy	(-∞, ∞)	μ: escala	Si μ=1 y θ=0 se denomina distribución de Cauchy estándar
		θ: situación	
Weibull	(0, ∞)	a: forma	Si a=1 se tiene la distribución exponencial Otro caso particular es la distribución de Rayleigh
		b: escala	
Laplace	(-∞, ∞)	a: situación	
		b: escala	
Pareto	$[x_0, \infty)$	α: forma	
		x_0: situación	
Triangular	[a, b]	a: mínimo	Se emplea cuando hay poca información disponible de la variable
		c: moda	
		b: máximo	