DISTRIBUCIONES DE PROBABILIDAD

ÍNDICE

DISTRIBUCIONES DE PROBABILIDAD

CÁLCULO DE PROBABILIDADES

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 Kolmogorov 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 (vale decir, que desagrega el 1 a lo largo de los valores posibles de la variable). 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 con subintervalos de valores. Una forma usual de describir la distribución de probabilidad de una variable aleatoria es mediante la denominada función de densidad, en tanto que lo que se conoce como función de distribución representa las probabilidades acumuladas[2-7].

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 3.1 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.

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, o la probabilidad de que la variable tome valores inferiores o iguales a x (cola izquierda) y la probabilidad de valores superiores a x (cola derecha). En el caso continuo, la probabilidad de que la variable sea igual a cualquier punto es igual a cero; por tanto, no influye en las colas el hecho de incluir o excluir el punto x. Hay un tercer resultado que el programa presenta sólo para las distribuciones continuas simétricas (normal, logística y t de Student): 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.

Asimismo, los resultados de Epidat 3.1 incluyen la media y la varianza de la correspondiente distribución, así como la mediana y/o la moda en el caso de las distribuciones continuas.

Epidat 3.1 también ofrece la posibilidad de representar, gráficamente, las funciones de distribución y densidad.

DISTRIBUCIONES DISCRETAS

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

- Uniforme discreta
- Binomial
- Hipergeométrica
- Geométrica
- Binomial Negativa
- Poisson

Distribución Uniforme discreta (a,b)

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 3.1, 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:

x: *a*, *a*+1, *a*+2, ..., *b,* números enteros

Parámetros:

a: mínimo, *a* entero

b: máximo, *b* entero con *a* < *b*

Ejercicio

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 apruebe el examen?

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 aprueba el examen si le toca un tema 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 3.1 basta con proporcionarle los parámetros de la distribución, y seleccionar calcular probabilidades para el punto 35.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Uniforme discreta (a,b)
a : Mínimo                          1
b : Máximo                         50
Punto K                            35

Probabilidad   Pr[X=k]          0,0200
Cola Izquierda Pr[X<=k]         0,7000
Cola Derecha   Pr[X>k]           0,3000

Media                           25,5000
Varianza                       208,2500
```

La persona tiene una probabilidad de aprobar igual a 0,7.

Distribución Binomial (n,p)

La distribución binomial es una distribución discreta muy importante que surge en muchas aplicaciones bioestadísticas.

Esta distribución aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificada como "éxito" o "fracaso". Por ejemplo, esa respuesta puede 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. Este modelo se aplica a poblaciones finitas de las que se toma 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 ingresados en una unidad hospitalaria que desarrollan una infección nosocomial.

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

Valores:
x: 0, 1, 2, ..., *n*

Parámetros:
n: número de pruebas, $n > 0$ entero
p: probabilidad de éxito, $0 < p < 1$

Ejercicio

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 qué probabilidad hay de que tenga al menos 14 aciertos.

Hay que proporcionarle a Epidat 3.1 los parámetros de la distribución 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 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Binomial (n,p)
n: Número de pruebas                        20
p: Probabilidad de éxito          0,7500
Punto K                              14

Probabilidad    Pr[X=k]           0,1686
Cola Izquierda  Pr[X<=k]          0,3828
Cola Derecha    Pr[X>k]           0,6172

Media                            15,0000
Varianza                          3,7500
```

La probabilidad de que el alumno tenga más de 14 aciertos se sitúa en 0,61.

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.

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.

Valores:

x: *max{0,n-(N-R)}, ..., 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, $N>0$ entero

R: número de éxitos en la población, $R\geq 0$ entero

n: número de pruebas, $n>0$ entero

Ejercicio

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 defectuosos.

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 $P\{\text{número de defectuosos} \leq 2\}$.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Hipergeométrica (N,R,n)
N : Tamaño de la población         100
R : Número éxitos en la pob.         7
n : Número de pruebas               10
Punto K                              2

Probabilidad   Pr[X=k]          0,1235
Cola Izquierda Pr[X<=k]         0,9792
Cola Derecha   Pr[X>k]          0,0208

Media                           0,7000
Varianza                        0,5918
```

La probabilidad de que a lo sumo haya dos útiles defectuosos en el lote es aproximadamente 0,98.

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 hasta obtener un éxito por primera vez. 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 denominada "propiedad de Harkov" o de falta de memoria, que implica que la probabilidad de tener que esperar un tiempo *t* no depende del tiempo que ya haya transcurrido.

Valores:

x: 0, 1, 2, ...

Parámetros:

p: probabilidad de éxito, 0<*p*<1

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Geométrica (p)
p : Probabilidad de éxito       0,8000
Punto K                              4

Probabilidad   Pr[X=k]           0,0013
Cola Izquierda Pr[X<=k]          0,9997
Cola Derecha   Pr[X>k]           0,0003

Media                            0,2500
Varianza                         0,3125
```

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

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 para que ocurran r éxitos, por lo que también se denomina "distribución binomial de tiempo de espera".

La distribución binomial negativa fue propuesta, originalmente, 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 S. Gosset) 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 distribuida en la gota, es decir, la tasa de intensidad no es homogénea.

Por ejemplo, la distribución binomial negativa es más adecuada que la de Poisson para modelar 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*")[8,9]. 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.

Valores:

x: 0, 1, 2, ...

Parámetros:

p: probabilidad de éxito, $0<p<1$

r: número de éxitos, $r\geq 0$

Ejercicio

Se sabe que, en promedio, de cada 100 placas de rayos X que se realizan, una 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.

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Binomial negativa (r,p)
r : Número de éxitos                9
p : Probabilidad de éxito      0,0100
Punto K                             1

Media                         891,0000
Varianza                    89100,0000
```

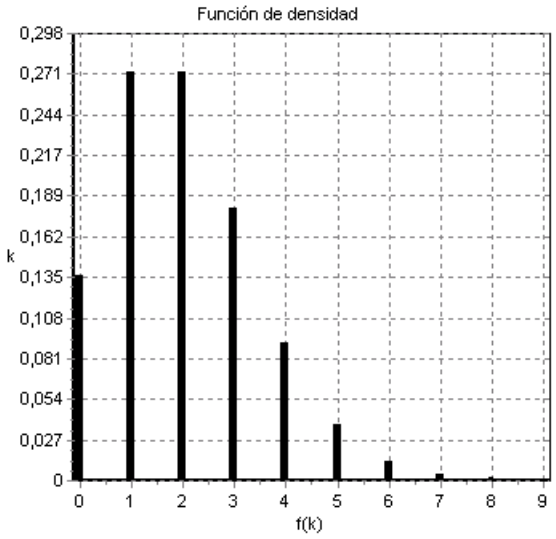
Entre 10 placas defectuosas se producen, en promedio, unas 891 placas útiles.

Distribución Poisson (lambda)

La distribución de Poisson, que debe su nombre al matemático francés Simeón Denis Poisson (1781-1840), ya había sido introducida en 1718 por Abraham De Moivre como una forma límite de la distribución binomial que surge cuando se observa un evento raro después de un número grande de repeticiones[10]. En general, la distribución de Poisson 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; una regla es 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 densidad obtenido con Epidat 3.1:



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, *lambda*, representa el número promedio de eventos esperados por unidad de tiempo o de espacio, por lo que también se suele hablar de *lambda* 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, habitualmente 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 *lambda* (el riesgo, en el caso de la prevalencia, y la densidad de incidencia, en el de incidencia).

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.

Valores:

x: 0, 1, 2, ...

Parámetros:

lambda: media de la distribución, *lambda* > 0

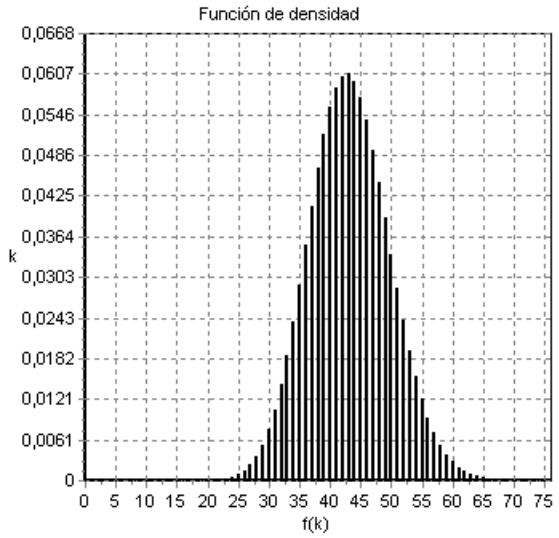
Ejercicio

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 densidad de probabilidad.

Para calcular la probabilidad pedida y, además, representar la función de densidad 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 3.1

```
Cálculo de probabilidades. Distribuciones discretas

Poisson (lambda)
lambda : Media                  43,2000
Punto K                              50

Probabilidad   Pr[X=k]           0,0339
Cola Izquierda Pr[X<=k]          0,8657
Cola Derecha   Pr[X>k]           0,1343

Media                           43,2000
Varianza                        43,2000
```



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

DISTRIBUCIONES CONTINUAS

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

- Uniforme
- Normal
- Lognormal
- Logística
- Beta
- Gamma
- Exponencial
- Ji-cuadrado
- t de Student
- F de Snedecor

Distribución Uniforme (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. Esta distribución presenta una peculiaridad importante: la probabilidad de un suceso dependerá exclusivamente de la amplitud del intervalo considerado y no de su posición en el campo de variación de la variable.

Cualquiera 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.

Campo de variación:

$a \leq x \leq b$

Parámetros:

a: mínimo del recorrido
b: máximo del recorrido

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Uniforme (a,b)
a : Mínimo                  380,0000
b : Máximo                   1200,0000
Punto X                      1000,0000

Cola Izquierda Pr[X<=k]         0,7561
Cola Derecha   Pr[X>=k]         0,2439

Media                         790,0000
Varianza                    56033,3333
Mediana                       790,0000
```

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.

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Uniforme (a,b)
a : Mínimo                      55,0000
b : Máximo                      75,0000
Punto X                         60,0000

Cola Izquierda   Pr[X<=k]        0,2500
Cola Derecha     Pr[X>=k]       0,7500


Media                           65,0000
Varianza                        33,3333
Mediana                         65,0000
```

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.

Distribución Normal (Mu, Sigma)

La distribución normal es, sin duda, la distribución de probabilidad más importante del Cálculo de probabilidades y de la Estadística. Fue descubierta por De Moivre (1773), como aproximación de la distribución binomial. De todas formas, 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 y F) 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 (*Mu*) y la desviación estándar (*Sigma*).

Campo de variación:

$-\infty < x < \infty$

Parámetros:

Mu: media de la distribución, $-\infty < Mu < \infty$

Sigma: desviación estándar de la distribución, $Sigma > 0$

Ejercicio

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.

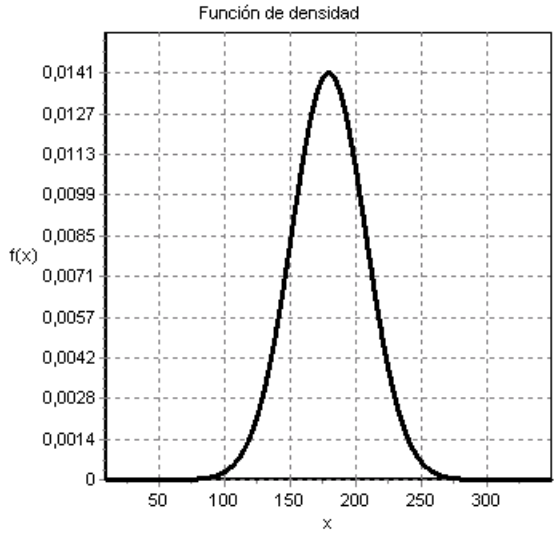
1. Calcule el porcentaje de enfermos con un nivel de colesterol inferior a 169 mg/dL.

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. Represente la función de densidad.

En este caso, se tendrá que ejecutar Epidat 3.1 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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Normal (Mu,Sigma)
Mu    : Media                  179,1000
Sigma : Desviación estándar     28,2000
Punto X                        169,0000

Cola Izquierda Pr[X<=k]          0,3601
Cola Derecha   Pr[X>=k]          0,6399
Dos Colas      1-Pr[|X|<=k]      0,7202
```



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

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas
Normal (Mu,Sigma)
Mu    : Media                  179,1000
Sigma : Desviación estándar     28,2000
Cola Izquierda Pr[X<=k]          0,9000
Cola Derecha   Pr[X>=k]          0,1000
Dos Colas      1-Pr[|X|<=k]      0,2000

Punto X                        215,2398
```

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

Distribución Lognormal (Mu, Sigma)

La variable resultante al aplicar la función exponencial a una variable que se distribuye normal con media *Mu* y desviación estándar *Sigma*, sigue una distribución lognormal con parámetros *Mu* (escala) y *Sigma* (forma). Dicho de otro modo, si una variable X se distribuye normalmente, la variable *ln*X, sigue una distribución lognormal.

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.

Campo de variación:

$0 < x < \infty$

Parámetros:

Mu: parámetro de escala, $-\infty < Mu < \infty$

Sigma: parámetro de forma, $Sigma > 0$

Ejercicio

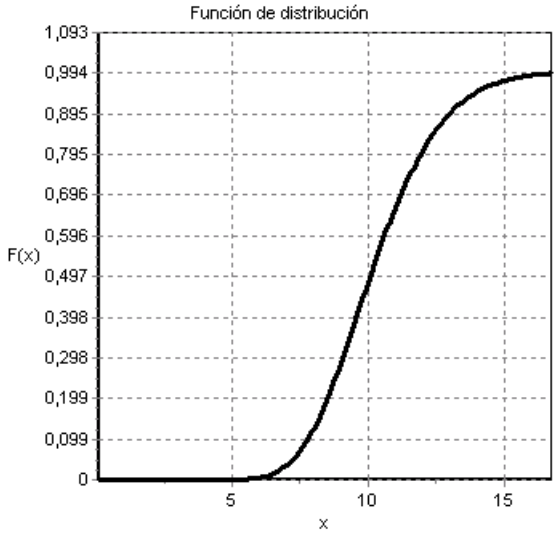
Supóngase que la supervivencia, en años, luego de una intervención quirúrgica (tiempo que pasa hasta que ocurre la muerte del enfermo) en una cierta población sigue una distribución lognormal de parámetro de escala 2,32 y de forma 0,20. Calcúlese la probabilidad de supervivencia a los 12 años, la mediana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

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Lognormal (Mu,Sigma)
Mu    : Escala                  2,3200
Sigma : Forma                   0,2000
Punto X                        12,0000

Cola Izquierda Pr[X<=k]         0,7952
Cola Derecha   Pr[X>=k]         0,2048

Media                          10,3812
Varianza                        4,3982
Mediana                        10,1757
Moda                            9,7767
```

La probabilidad de supervivencia a los 12 años se sitúa próximo a 0,20.

La función de distribución de la supervivencia a la intervención quirúrgica se presenta a continuación:



Distribución Logística (a, b)

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.

Campo de variación:

$-\infty < x < \infty$

Parámetros:

a: parámetro de posición, $-\infty < a < \infty$

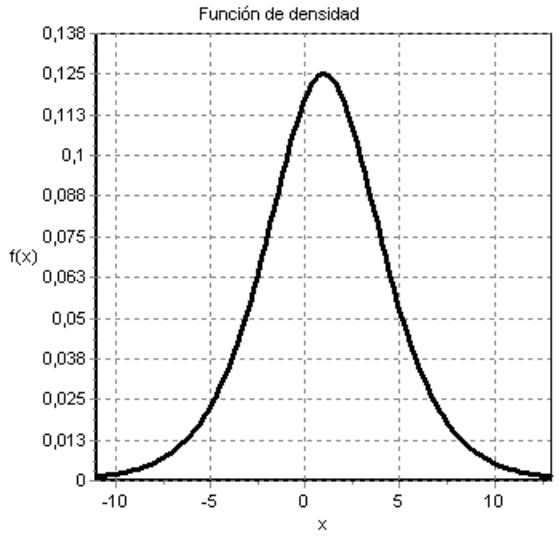
b: parámetro de escala, $b > 0$

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas
Logística (a,b)
a : Posición                    1,0000
b : Escala                      2,0000
Punto X                         5,0000

Cola Izquierda Pr[X<=k]         0,8808
Cola Derecha   Pr[X>=k]         0,1192
Dos Colas      1-Pr[|X|<=k]     0,2384
Media                           1,0000
Varianza                       13,1595
Mediana                         1,0000
Moda                            1,0000
```



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

Distribución Beta (p,q)

La distribución beta es posible para una variable aleatoria continua que toma 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.

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, denotada Beta(1,1).

Campo de variación:

$0 \leq x \leq 1$

Parámetros:

p: parámetro de forma, $p > 0$

q: parámetro de forma, $q > 0$

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Beta (p,q)
p : Forma                       2,0000
q : Forma                       2,0000
Punto X                         0,2500

Cola Izquierda Pr[X<=k]         0,1563
Cola Derecha   Pr[X>=k]         0,8438

Media                           0,5000
Varianza                        0,0500
Moda                            0,5000
```

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%.

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 *lambda*, la variable que mide el tiempo transcurrido hasta obtener n ocurrencias del evento sigue una distribución gamma con parámetros $a = n \times lambda$ (escala) y $p=n$ (forma). Se denota Gamma(a,p).

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

Esta distribución presenta como propiedad interesante la "falta de memoria". Por esta razón, es muy utilizada en las teorías de la fiabilidad, mantenimiento y fenómenos de espera (por ejemplo en una consulta médica "tiempo que transcurre hasta la llegada del segundo paciente").

Campo de variación:

$0 < x < \infty$

Parámetros:

a: parámetro de escala, $a > 0$

p: parámetro de forma, $p > 0$

Ejercicio 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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Gamma (a,p)
a : Escala                      6,0000
p : Forma                       2,0000
Punto X                         1,0000

Cola Izquierda Pr[X<=k]         0,9826
Cola Derecha   Pr[X>=k]         0,0174

Media                           0,3333
Varianza                        0,0556
Moda                            0,1667
```

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

Ejercicio 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, calcúlese:

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Gamma (a,p)
a : Escala                      0,8100
p : Forma                       7,8100
Cola Izquierda Pr[X<=k]         0,9000
Cola Derecha   Pr[X>=k]         0,1000

Punto X                         14,2429

Media                           9,6420
Varianza                        11,9037
Moda                            8,4074
```

El tiempo medio de supervivencia es de, aproximadamente, 10 años.

Distribución Exponencial (lambda)

La distribución exponencial es 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.

La distribución exponencial 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, *lambda*, 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.

También se puede ver como un caso particular de la distribución gamma(a,p), con a=*lambda* y p=1.

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

Campo de variación:

$0 < x < \infty$

Parámetros:

lambda: tasa, *lambda* > 0

Ejercicio

Se ha comprobado que el tiempo de vida de cierto tipo de marcapasos sigue una distribución exponencial con media de 16 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*=1/16=0,0625

Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Exponencial (lambda)
lambda : Tasa                  0,0625
Punto X                       20,0000

Cola Izquierda Pr[X<=k]       0,7135
Cola Derecha   Pr[X>=k]        0,2865
```

La probabilidad de que se le tenga que implantar otro marcapasos antes de los 20 años se sitúa en un entorno a 0,71.

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) = 0,71.

Distribución Ji-cuadrado (n)

Un caso especial, muy importante, de la distribución Gamma se obtiene cuando a=1/2 y p=n/2. La distribución resultante se conoce con el nombre de Ji-cuadrado con n grados de libertad. Es la distribución que sigue la suma de los cuadrados de n variables independientes N(0,1).

La Ji-cuadrado es una distribución fundamental en inferencia estadística y en los tests estadísticos de bondad de ajuste. Se emplea, entre muchas otras aplicaciones, para determinar los límites de confianza de la varianza de una población normal, para contrastar la hipótesis de homogeneidad o de independencia en una tabla de contingencia y para pruebas de bondad de ajuste.

La distribución Ji-cuadrado queda totalmente definida mediante sus grados de libertad n.

Campo de variación:

$0 \leq x < \infty$

Parámetros:

n: grados de libertad, $n>0$

Ejercicio

Considere la distribución Ji-cuadrado con 2 grados de libertad.

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Ji-cuadrado (n)
n : Grados de libertad              2
Punto X                         9,2100

Cola Izquierda Pr[X<=k]         0,9900
Cola Derecha   Pr[X>=k]         0,0100
```

El 1% del área bajo la curva se ubica a la derecha de 9,21.

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Ji-cuadrado (n)
n : Grados de libertad              2
Cola Izquierda Pr[X<=k]         0,9000
Cola Derecha   Pr[X>=k]         0,1000

Punto X                         4,6052
```

El valor 4,6052 divide a la distribución en dos partes: el 90% de ésta queda a la izquierda de dicho punto y el 10% a la derecha.

Distribución t de Student (n)

La distribución t de Student se construye como un cociente entre una normal y la raíz de una Ji-cuadrado independientes. Esta distribución desempeña un papel importante en la inferencia estadística asociada a la teoría de muestras pequeñas. Se usa habitualmente en el contraste de hipótesis para la media de una población, o para comparar las medias de dos poblaciones, y viene definida por sus grados de libertad *n*.

A medida que aumentan los grados de libertad, la distribución t de Student se aproxima a una normal de media 0 y varianza 1 (normal estándar).

Campo de variación:

$-\infty < x < \infty$

Parámetros:

n: grados de libertad, $n>0$

Ejercicio

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= 500.

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

Normal (Mu,Sigma)
Mu    : Media                   0,0000
Sigma : Desviación estándar     1,0000
Cola Izquierda Pr[X<=k]         0,9500
Cola Derecha   Pr[X>=k]         0,0500
Dos Colas      1-Pr[|X|<=k]     0,1000

Punto X                         1,6449

Media                           0,0000
Varianza                        1,0000
```

En el segundo apartado se ejecutará dos veces Epidat 3.1: la primera vez con una distribución t de Student con 10 grados de libertad y la segunda vez con 500 grados de libertad.

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

t de Student (n)
n : Grados de libertad             10
Punto X                          1,6449

Cola Izquierda Pr[X<=k]          0,9345
Cola Derecha   Pr[X>=k]          0,0655
Dos Colas      1-Pr[|X|<=k]      0,1310
```

```
Cálculo de probabilidades. Distribuciones continuas

t de Student (n)
n : Grados de libertad            500
Punto X                          1,6449

Cola Izquierda Pr[X<=k]          0,9497
Cola Derecha   Pr[X>=k]          0,0503
Dos Colas      1-Pr[|X|<=k]      0,1006
```

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.

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 con distribución Ji-cuadrado divididas por sus respectivos grados de libertad, n y m. En este caso la variable aleatoria sigue una distribución F de Snedecor de parámetros n y m. Hay muchas aplicaciones de la F en estadística y, en particular, tiene un papel importante en las técnicas del análisis de la varianza y del diseño de experimentos.

Campo de variación:

$0 \leq x < \infty$

Parámetros:

n: grados de libertad del numerador, $n>0$

m: grados de libertad del denominador, $m>0$

Ejercicio

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 3.1

```
Cálculo de probabilidades. Distribuciones continuas

F de Snedecor (n,m)
n : Grados libertad del num.   10,0000
m : Grados libertad del denom.  12,0000
Cola Izquierda Pr[X<=k]         0,9500
Cola Derecha   Pr[X>=k]         0,0500

Punto X                        2,7534
```

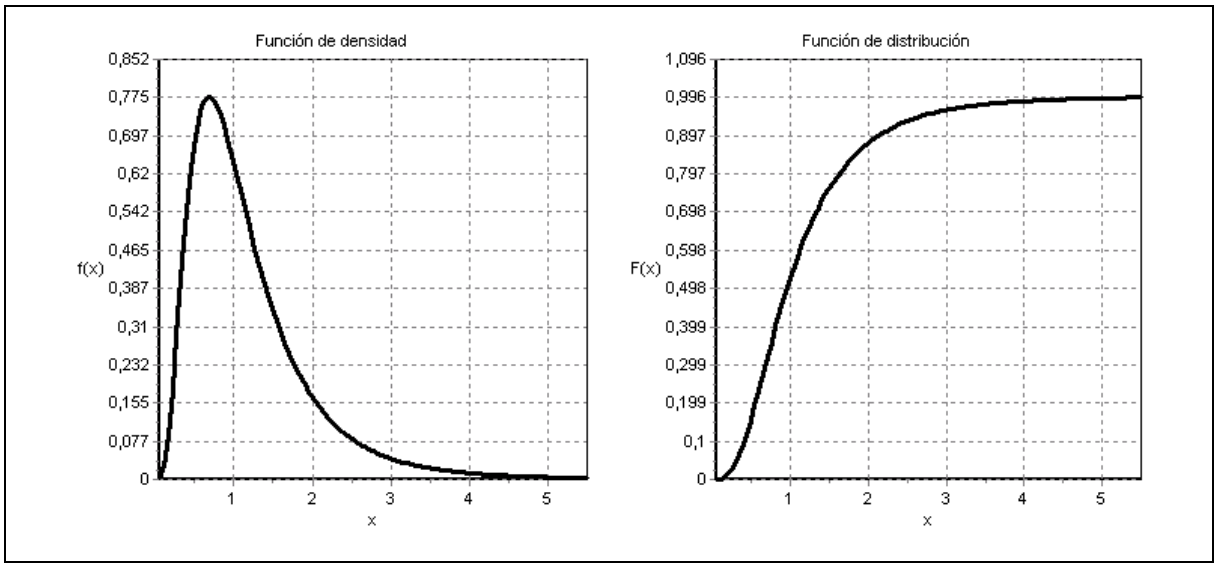
El valor que deja a la derecha una probabilidad de 0,05 es 2,75.

2. Resultados con Epidat 3.1

```
Cálculo de probabilidades. Distribuciones continuas

F de Snedecor (n,m)
n : Grados libertad del num.   10,0000
m : Grados libertad del denom.  12,0000
Punto X                         4,3000

Cola Izquierda Pr[X<=k]         0,9900
Cola Derecha   Pr[X>=k]        0,0100

Media                           1,2000
Varianza                        0,7200
Moda                            0,6857
```

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:



GENERACIÓN DE DISTRIBUCIONES

Conceptos generales

Epidat 3.1 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[11], 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.

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.

DISTRIBUCIONES DISCRETAS

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

- Uniforme discreta
- Binomial
- Multinomial
- Hipergeométrica
- Geométrica
- Binomial Negativa
- 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.

Distribución Multinomial

Generaliza la distribución binomial al caso en que la población se divida en m>2 grupos mutuamente exclusivos y exhaustivos.

Se supone un proceso estable y sin memoria que genera elementos que pueden clasificarse en m clases distintas. Supóngase que se toma una muestra de n elementos y se definen m variables aleatorias X_i=número de elementos de la clase i (i=1, ..., m), entonces el vector de m-variables es una 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 de la clase 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 8 descendientes, el vector de variables (X_1, X_2, X_3) donde:

X_1= Nº de descendientes rojos

X_2= Nº de descendientes negros

X_3= Nº de descendientes blancos

sigue una distribución multinomial con parámetros n=8; $p_1 = 8/16 = 0{,}5$; $p_2 = 4/16 = 0{,}25$ 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, ... (i = 1, ..., m)

Parámetros:

n: número de pruebas, n>0 entero

m: número de clases, m>0 entero

p_i: probabilidad de la clase i, $0<p_i<1$ (i=1, ..., m), donde $\sum_{i=1}^{m} p_i = 1$.

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

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.

Distribución Normal bivariante

Fue introducida por Gauss a principios del siglo XIX en su estudio de errores de medida en las observaciones astronómicas y de cálculo de órbitas de cuerpos celestes. 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, *Mu*, y su varianza, *Sigma*, 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 $Sigma=(Sigma_1, Sigma_2)$ y el coeficiente de correlación *Ro* 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 $Ro=0$.

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, $Sigma_1 > 0$, $Sigma_2 > 0$

Ro: coeficiente de correlación, $-1 \leq Ro \leq 1$

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 ejercicio en el que se aplica la distribución normal bivariante.

Ejercicio

Suponga que la distribución de la variable peso de una población de jóvenes sigue una distribución normal de media $\mu=65$ kg y desviación estándar $\sigma=15$ kg. Suponga, además, que la variable altura en dicha población sigue una distribución normal de media $\mu=1{,}68$ m y desviación estándar $\sigma=0{,}20$ 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 3.1

```
Generación de distribuciones. Distribuciones continuas

Normal bivar. (Mu,Sigma,Ro)
Mu : Vector de medias
  65,0000    1,6800
Sigma : Vector de desviaciones estándar
  15,0000    0,2000
Ro : Coeficiente de correlación
   0,7500

Vector de medias
  65,0000    1,6800

Matriz de dispersión
 225,0000    2,2500
   2,2500    0,0400

Tamaño de la muestra que se simula       1000

Valores de la distribución
-------------------------
  72,0752     1,7142
  82,8499     1,8594
  59,8245     1,4226
  61,4830     1,8272
…
```

Con los 1.000 valores simulados se obtiene un porcentaje de sujetos con un *IMC* superior a 30 kg/m^2 del 9%.

Notas: 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. Martín Pliego FJ, Ruiz-Maya L. *Estadística I: Probabilidad*. Madrid: Editorial AC; 1997.

3. Meyer PL. *Probabilidad y Aplicaciones Estadísticas*. México: Addison-Wesley Iberoamericana; 1986.

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

5. Katz DL. *Epidemiology, Biostatistics and Preventive Medicine Review*. USA: W.B. Saunders Company; 1997.

6. Doménech JM. *Métodos Estadísticos en Ciencias de la Salud*. Barcelona: Signo; 1997.

7. Hospital Ramón y Cajal. Material docente de la unidad de bioestadística clínica. Disponible en: http://www.hrc.es/bioest/M_docente.html

8. Biggeri A. *Negative Binomial Distribution*. En: Armitage P, Colton T editores. *Encyclopedia of Biostatistics.* Vol. 4. Chichester: John Wiley & Sons; 1998. p. 2962-7.

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

10. Palmgren J. *Poisson Distribution*. En: Armitage P, Colton T editores. *Encyclopedia of Biostatistics.* Vol. 4. Chichester: John Wiley & Sons; 1998. p. 3398-3402.

11. 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. p. 141-6.