Distribuciones de probabilidad: introducción práctica

00R Team

Marzo 2016

Índice

Introducción

Es este documento discutiremos de forma práctica los aspectos y conceptos básicos de la distribuciones de probabilidad.

Veremos cómo se pueden simular datos con determinadas distribuciones y cómo calcular los valores teóricos de estas.

Como veremos más adelante el fundamento de las pruebas estadísticas está en comprender que los estadísticos pueden entenderse como valores de distribuciones específicas; entender las ideas de distribución de variables aleatorias es pues un paso previo al estudio de los test estadísticos.

Un dado es un dado...

Si disponemos de un dado no cargado de seis caras asumimos que la probabilidad de aparecer de cada una de ellas tras el lanzamiento es de un sexto. Podemos simular el lanzamiento de este dado miles de veces con la ayuda de R:

```
ncaras <- 6
n      <- 10000
x <- sample( ncaras, n, replace = TRUE )
table( x )
```

```
## x
##    1    2    3    4    5    6
## 1682 1702 1617 1642 1643 1714
```

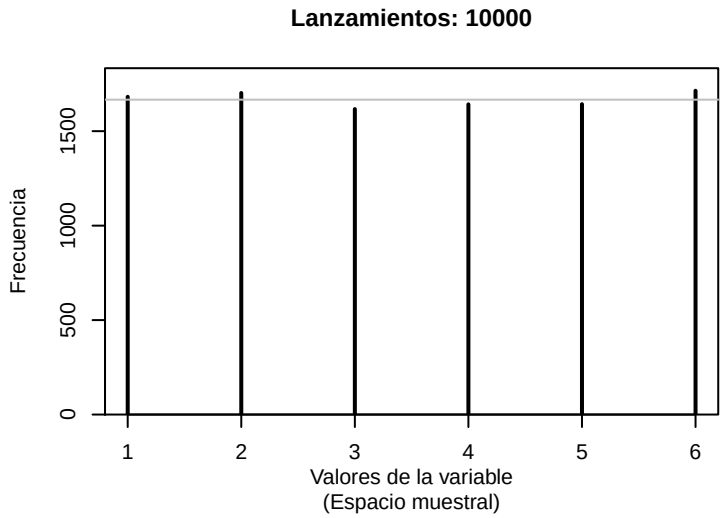
- experimento: *valor de la cara superior tras el lanzamiento de un dado no cargado*

- espacio muestral: 1, 2, 3, 4, 5 y 6
- tamaño muestral: 10000 observaciones.
- variable: x de tipo cualitativo.

Para este tipo de variables sólo cabe tabular los datos y determinar la frecuencia para cada elemento del espacio muestral.

Gráficamente podemos ver cómo la distribución de estos valores se aproxima al valor teórico esperado: 1666.667, es decir, una proporción de 1/6.

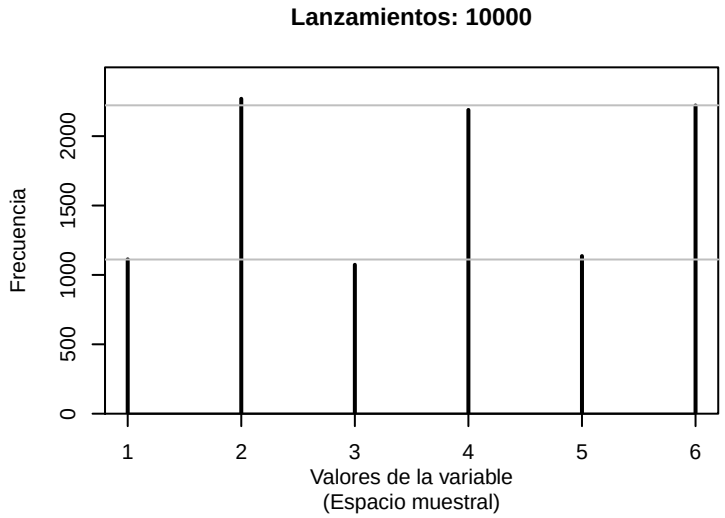
```
par( cex = 0.65 )
plot( table( x ), lwd = 2, ylim = c( 0, n / ncaras * 1.1),
      ylab = "Frecuencia", yaxs='i', xaxs='r',
      xlab = "Valores de la variable\n(Espacio muestral)",
      main = paste( "Lanzamientos:", n ) )
abline( h = n / ncaras, col = "grey75" )
```



Un dado cargado

Sin embargo el dado podría estar construido para proporcionar distintos pesos (y por tanto probabilidad) a las distintas caras. En este ejemplo vemos como los valores pares tiene un peso doble que el de los impares.

```
par( cex = 0.65 )
ncaras <- 6
n      <- 10000
pesos  <-  rep( 1:2, 3 )
x      <- sample( ncaras, n, replace = TRUE, prob = pesos )
tabla  <- table( x )
plot( tabla, lwd = 2, ylim = c( 0, max( tabla ) * 1.1),
      ylab = "Frecuencia", yaxs = 'i', xaxs = 'r',
      xlab = "Valores de la variable\n(Espacio muestral)",
      main = paste( "Lanzamientos:", n ) )
abline( h = n * unique( pesos / sum( pesos ) ), col = "grey75" )
```



La variable, `x`, es una variable aleatoria cualitativa, y se podrían generar simulaciones con distintos pesos, buscando que su comportamiento sea semejante al esperado en un proceso natural. Este comportamiento es crucial por dos aspectos:

1. Los procesos aleatorios esconden procesos naturales de los que sólo vemos el patrón que reflejan las tablas, gráficos y otros estadísticos.
2. El objetivo de la investigación debe contemplar el conocimiento de las causas del fenómeno estudiado.

Un ejemplo clásico: estudios de Mendel

Mendel descifró el misterio de los mecanismos de la herencia gracias a plantear adecuadamente la distribución que debemos esperar de los genotipos/fenotipos del cruzamiento de padres homocigotos o heterocigotos.

Supuso que la herencia estaba ligada a alelos, que definían características en los individuos. Estos tenían variantes, y una de ellas, el alelo dominante, se expresaba por encima de otras, alelos recesivos. Esto daba lugar a un imposibilidad de apreciar a simple vista el genotipo, o composición de alelos del organismo, y solo podía apreciarse el fenotipo, la expresión de aquel.

Mendel propuso (I) la segregación independiente de los caracteres (alelos) y (II) la herencia independiente de estos.

Simulemos el caso de la situación que muestra la independencia de la descendencia de heterocigotos.

Los parentales con genotipo para un alelo dado: Aa A y a, dominante y recesivo respectivamente, generan gametos con igual probabilidad. Los posibles resultados de la fecundación son: AA, Aa, aA y aa, todos con la misma probabilidad, por lo tanto:

```
n        <- 100000
x <- sample( c( "AA", "Aa", "aA", "aa" ), n, replace = TRUE )
table( x )
```

```
## x
##    aa    aA    Aa    AA
## 24927 25025 25001 25047
```

```
table( x ) / n
```

```
## x
##      aa      aA      Aa      AA
## 0.24927 0.25025 0.25001 0.25047
```

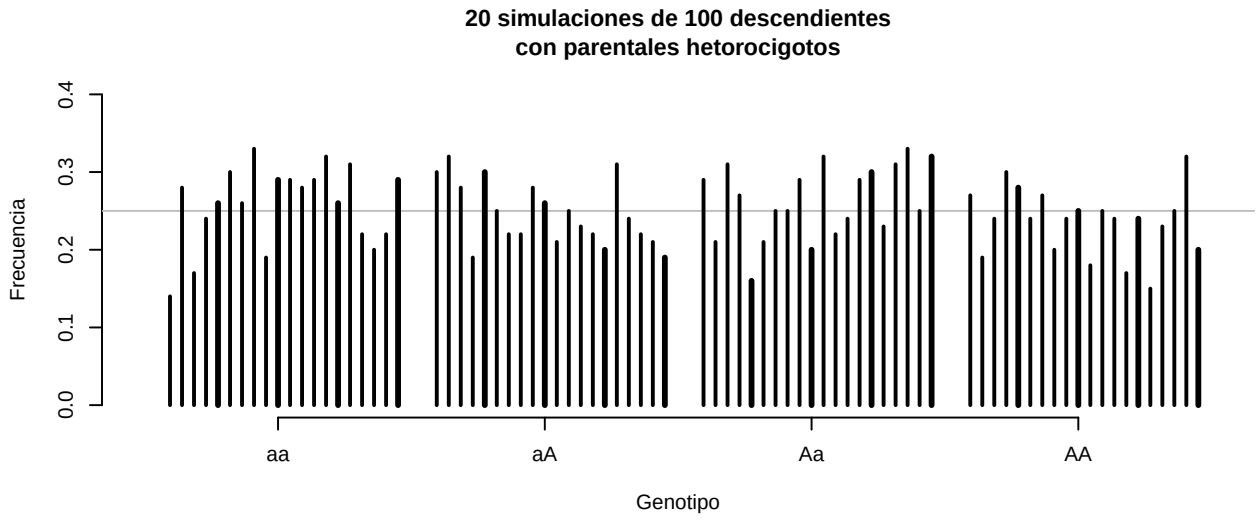
Vemos que los homocigotos son aproximadamente la mitad de los simulados y los heterocigotos son la otra mitad, tal como cabía esperar. El problema experimental surge cuando el tamaño muestral (n) no puede ser grande y produce resultados que se alejan del esperado por puro azar, veamos:

```
par( cex = 0.65 )
n        <- 100
simulaciones <- 20
genotipos    <- c( "AA", "Aa", "aA", "aa" )
ngenotipos   <- length( genotipos )

x <- sample( genotipos, n, replace = TRUE )

plot( table( x ) / n, xlim = c( 0.5, ngenotipos + 0.5 ),
      ylim = c( 0, 0.4 ) , type = "n",
      xlab = "Genotipo",
      ylab = "Frecuencia",
      main = paste( simulaciones, "simulaciones de", n,
                    "descendientes\ncon parentales hetorocigotos" ) )
abline( h = 1/ngenotipos, col = "grey75" )

incremento <- 0.045
desplaza <- - incremento * simulaciones / 2
for(i in 1:simulaciones ) {
    x <- table( sample( genotipos, n, replace = TRUE ) ) / n
    x <- as.numeric( x )
    lines( 1:ngenotipos + incremento * i + desplaza,  x, type = "h",
           lwd = 2 + ( i %% 5 == 0 ) )
    }
```



Dada un situación experimental real, que podría parecerse a una de nuestras 20 realizadas nos preguntamos *¿cómo medir que el alejamiento de valor esperado puede achacarse a un aleatorio o no*? Afortunadamente, podemos encontrar regularidad en el comportamiento de las diferencias entre lo observado y lo esperado. Esa regularidad tiene una distribución teórica como veremos en el curso.

A modo de reflexión

Dado el experimento estudiado describe, desde el punto de vista estadístico, y realiza las precisiones que consideres oportunas:

- El experimento
- El espacio muestral
- El tamaño muestral
- Tipo de variable

Distribuciones para variables discretas

Cuando el espacio muestral es numérico y solo se pueden tomar valores enteros como, por ejemplo, en conteos: número de hijos, número de intentos, etc., hablamos de distribuciones aleatorias discretas.

Una de las más importantes es la distribución binomial o binomial positiva. La variable puede tomar los valores 0, 1, ..., n. Cuidado, este valor n representa el valor máximo que puede tomar la variable y no debe confundirse con el tamaño muestral.

Veamos un ejemplo para ilustrar esta distribución. La probabilidad de tener un hijo varón en una familia con un sólo descendiente es 0.5, asumiendo que la probabilidad de tener un hijo varón sea la misma que la de tener una hija; para familias con 2 descendientes este valor es el mismo, la composición de la descendencia puede ser hijo-hijo, hijo-hija, hija-hijo e hija-hija, es decir, en la mitad de los casos se produce el resultado de un hijo varón. Podemos ver estos resultados como el número de hijos: hijo-hijo = 2, hijo-hija = 1, hija-hijo = 1 e hija-hija = 0; En resumen, la variable *número de hijos varones* puede tomar los valores: 0, 1 y 2, con peso 1, 2, 1.

Hablamos de una distribución binomial de parámetros:

- n: número de éxitos, en nuestro caso consideramos éxito al tener un descendiente varón en el parto.
- p: probabilidad del éxito, probabilidad de que el descendiente en el parto sea varón.

Como hemos visto es posible calcular la probabilidad de tener: 0 éxitos (sólo hijas), 1 éxito (el primero o el segundo de los descendientes es varón) o 2 éxitos (los dos son varones).

En general hablamos de n posibles descendientes y de p como la probabilidad de que sea varón. En la literatura estadística encontramos la siguiente notación para hablar de una variable aleatoria discreta que sigue una binomial: $X \sim \mathrm{B}(n, p)$ para hablar de una distribución binomial con los parámetros indicados, en nuestro ejemplo: $\mathrm{B}(2, 0{,}5)$

La función de probabilidad es

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

donde la variable puede tomar los valores: $x = \{0, 1, 2, \ldots, n\}$. Entonces puede calcularse el valor de probabilidad para 0, 1, 2, ... éxitos. La notación habitual es $P(X = x)$, es decir, probabilidad de que la variable tome un valor particular dado.

Afortunadamente, con R es muy fácil calcular los valores asociados a una distribución recurriendo a las funciones asociadas a ella. Para la binomial tenemos:

- `dbinom`: calcula $P(X = x)$, por ejemplo probabilidad de tener un hijo varón en el caso de una familia de dos descendientes:

```
par( cex = 0.65 )
dbinom( 1, 2, 0.5 )
```

```
## [1] 0.5
```

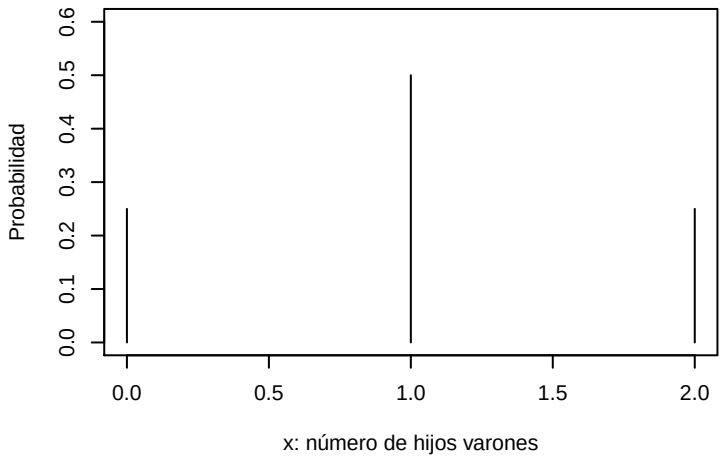
```
dbinom( 0:2, 2, 0.5 ) # todas las situaciones posibles
```

```
## [1] 0.25 0.50 0.25
```

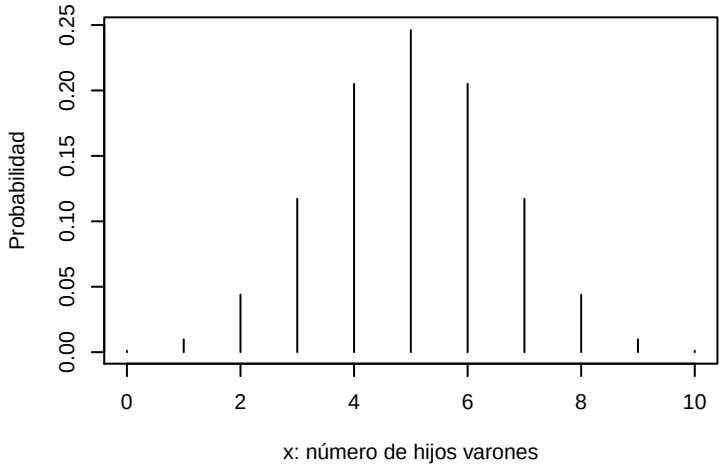
```
dbinom( 0:5, 5, 0.5 ) # para familias de 5 hijos
```

```
## [1] 0.03125 0.15625 0.31250 0.31250 0.15625 0.03125
```

```
plot( 0:2, dbinom( 0:2, 2, 0.5 ), type = "h", ylim = c(0, 0.6 ),
      ylab = "Probabilidad", xlab = "x: número de hijos varones" )
```




```
m <- 10
plot( 0:m, dbinom( 0:m, m, 0.5 ), type = "h",
      ylab = "Probabilidad", xlab = "x: número de hijos varones" )
```



- `rbinom` proporciona valores aleatorios de una distribución binomial:

```
table( rbinom( 1000, 2, 0.5 ) ) # mil valores aleatorios
```

```
##
##   0   1   2
## 278 473 249
```

A modo de reflexión

1. Supongamos que tener una hija es más probable que tener un hijo ¿qué ocurre en los ejemplos anteriores cuando la probabilidad no es 0.5?
2. ¿Qué aspecto tiene la gráfica de la probabilidad asociada a cada valor con valores grandes del parámetro n?

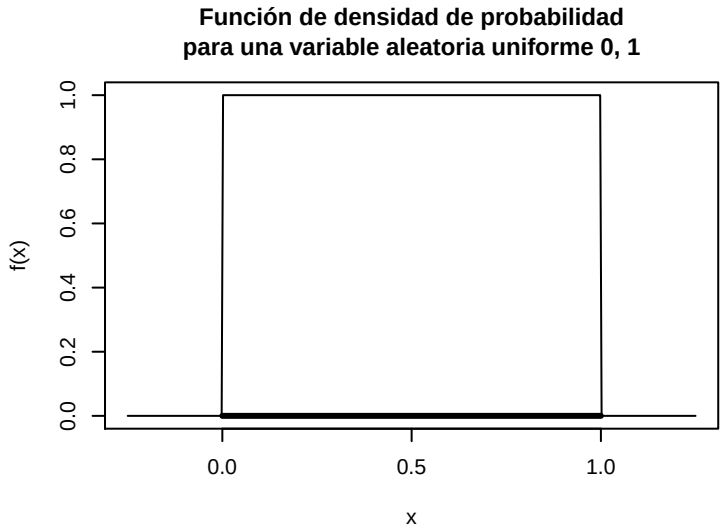
Distribución uniforme

Cuando una variable aleatoria es continua asumimos que puede tomar cualquier valor para un intervalo dado, por ejemplo, la longitud de dos individuos de una población puede ser cualquiera entre el más pequeño y el más grande y cabe pensar que no hay dos individuos exactamente iguales. Un problema derivado de la observación es utilizar una unidad de medida y redondear (por ejemplo a centímetros o milímetros) en función de los intereses de estudio.

La distribución uniforme, si bien no aparece de forma natural, nos resulta muy útil por ser el equivalente a la idea de un dado no cargado. Así, dos intervalos distintos de posibles valores de la variable con la misma longitud tienen la misma probabilidad.

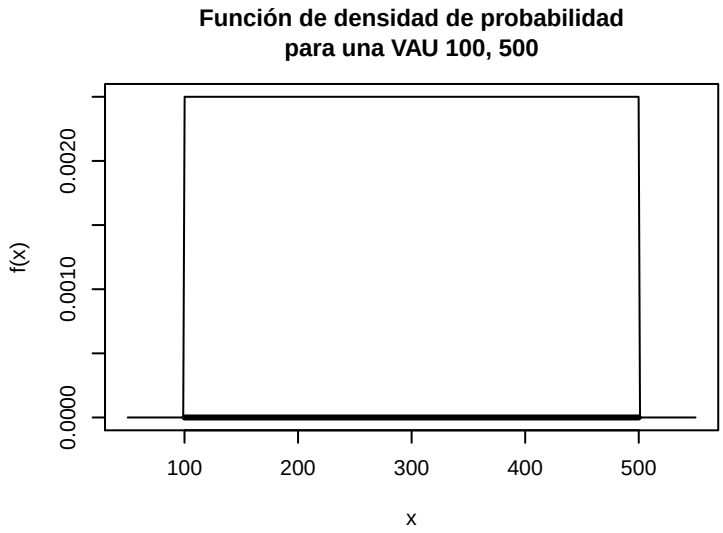
La probabilidad en variables aleatorias continuas se asocia al área que deja por debajo la curva de la función de distribución. Para el caso de una variable uniforme en el rango a, b hablamos de $x \sim \mathrm{U}(a, b)$; así esta distribución queda descrita por estos dos parámetros. Gráficamente para $\mathrm{U}(0, 1)$:

```
par( cex = 0.65 )
curve( dunif, -0.25, 1.25, ylab= "f(x)", n = 400,
       main = "Función de densidad de probabilidad\npara una variable aleatoria uniforme 0, 1" )
lines( c( 0, 1 ), c( 0, 0 ), lwd = 3 )
```



La curva de este tipo de distribución es semejante en otros casos y la región por debajo de la curva tiene superficie de área unidad, cosa que puede comprobarse fácilmente multiplicando la base (rango) por la altura:

```
par( cex = 0.65 )
curve( dunif( x, 100, 500 ), 50, 550, ylab= "f(x)", n = 400,
       main = "Función de densidad de probabilidad\npara una VAU 100, 500" )
lines( c( 100, 500 ), c( 0, 0 ), lwd = 3 )
```



Para variables continuas no podemos hablar de la probabilidad de que la variable tome un valor dado y debemos recurrir a la probabilidad de que la variables sea mayor o menor que un valor ($P(x > a)$, $P(x < a)$) o que este valor se encuentre en un determinado intervalo $P(a \leq x \leq b)$.

En el caso de la distribución uniforme es muy fácil calcular la probabilidad, ya que, esta es proporcional al intervalo considerado. También puede recurrirse a la función `punif` en R. Veamos para el caso $\mathrm{U}(0, 100)$:

- La probabilidad de obtener un valor menor que 50, puesto que hablamos del valor intermedio: 0.5

```
punif( 50, 0, 100 )
```

```
## [1] 0.5
```

- La probabilidad de obtener un valor menor que 10, en este caso es la décima parte del rango: 0.1

```
punif( 10, 0, 100 )
```

```
## [1] 0.1
```

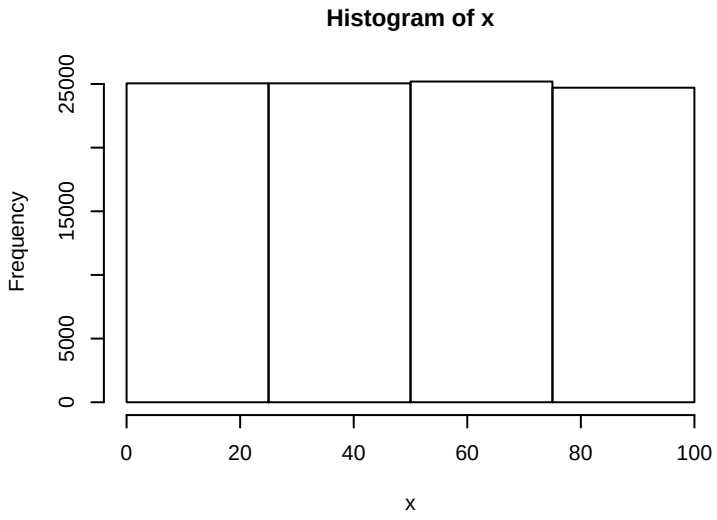
- Probabilidad de obtener un valor entre 35 y 40. tenemos una probabilidad asociada a un 0.05 del rango: 0.05

```
punif( 40, 0, 100 ) - punif( 35, 0, 100 )
```

```
## [1] 0.05
```

- Podemos simular valores de una distribución uniforme gracias a la función `runif`:

```r
par( cex = 0.65 )
x <- runif( 100000, 0, 100 )
hist( x, seq( 0, 100, 25 ) )
```




```r
summary( x )
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   0.00247  24.95000  49.88000  49.93000  74.68000 100.00000
```

```r
sd( x )
```

```
## [1] 28.82194
```

```r
sqrt( (100 - 0)^2 /12 )
```

```
## [1] 28.86751
```

A modo de reflexión

1. ¿Qué media tiene una distribución uniforme entre a y b?
2. En el ejemplo anterior ¿coincide la media obtenida por simulación con la esperada?
3. ¿Qué ocurre si el histograma resumen de la simulación tiene un mayor número de intervalos?
4. ¿Cuáles son las probabilidades teóricas asociadas a los cuartiles?
5. La desviación típica de una variable con distribución uniforme se calcula como $s_x = \sqrt{\frac{(b-a)^2}{12}}$. En el ejemplo anterior, ¿coincide la desviación típica obtenida por simulación con la esperada?

La distribución normal

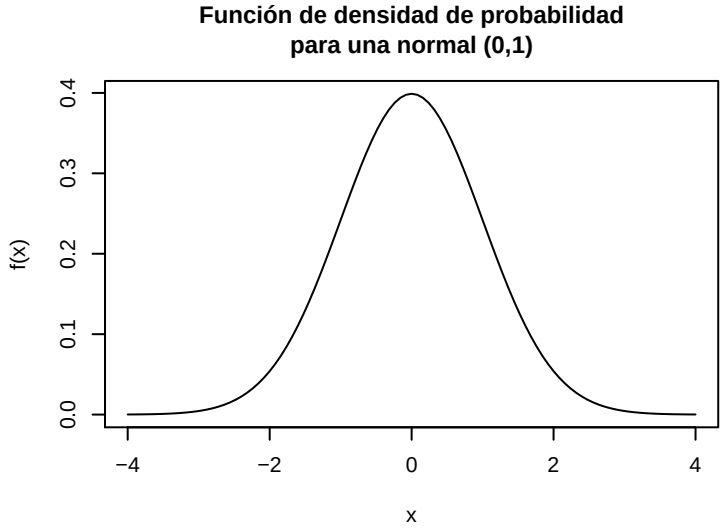
Se trata de la más conocida de la distribuciones aleatorias de variables continuas. Posee dos parámetros: la media, μ, y la desviación típica, σ. Hablamos de $x \sim \mathrm{N}(\mu, \sigma)$, si bien es muy habitual hablar de la variable estandarizada $z = x - \mu/\sigma$, entonces $z \sim \mathrm{N}(0, 1)$.

La curva de esta distribución es:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Gráficamente:

```
par( cex = 0.65 )
curve( dnorm, -4, 4, ylab= "f(x)",
       main = "Función de densidad de probabilidad\npara una normal (0,1)"
       )
```



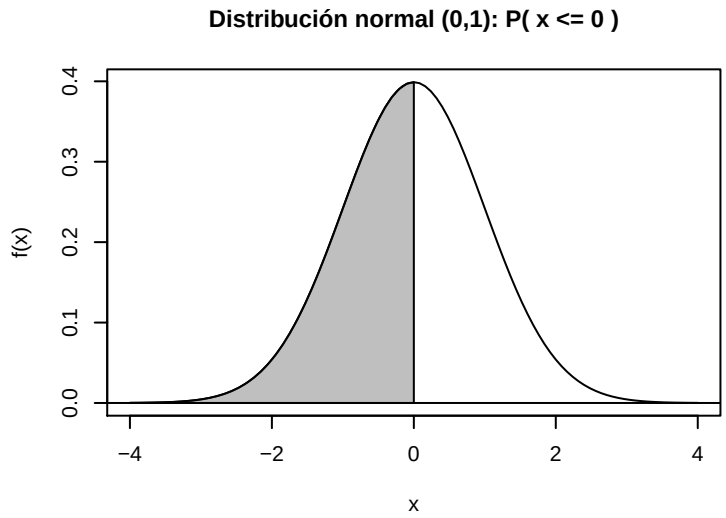
Trabajaremos con la normal de media 0 y desviación típica 1, recurriendo a las funciones:

- `pnorm`, para la probabilidad de que el valor sea menor que a: $P(z < a)$, para $z = 0$, estamos en mitad de la curva y la probabilidad es 0.5.

  ```
  pnorm( 0 )
  ```

  ```
  ## [1] 0.5
  ```

  Visualmente, el área sombreada se corresponde con la probabilidad indicada:



- Probabilidad de obtener un valor entre -1.0 y 1.0, esto es, la proporción de valores de una distribución normal entre la media menos una desviación típica y la media más una desviación típica:

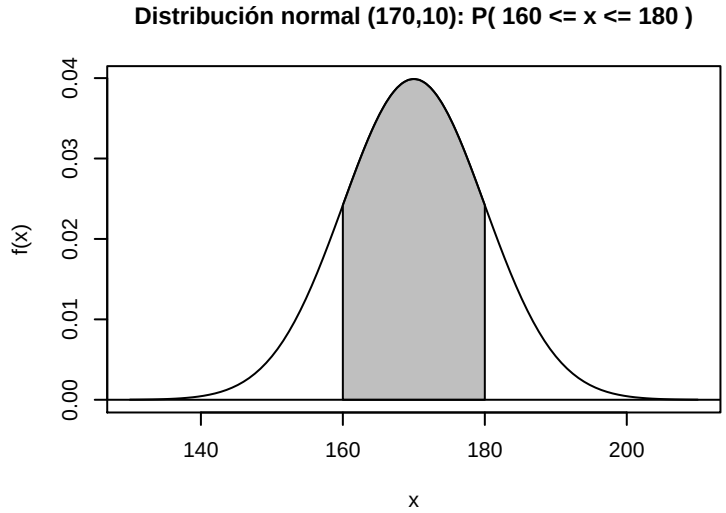
```
pnorm( 1 ) - pnorm( -1.0 )
```

```
## [1] 0.6826895
```

Podemos comprobar que para una variable normal con media 170 y desviación típica de 10 se produce el mismo resultado:

```
pnorm( 180, 170, 10 ) - pnorm( 160, 170, 10 )
```

```
## [1] 0.6826895
```



- Para el cálculo inverso, dada una probabilidad o percentil, calculamos el valor de la variable asociado con la función `qnorm`, el valor que deja el 99 por cien a su izquierda es:

```r
qnorm( 0.99 )
```

```
## [1] 2.326348
```

```r
pnorm( qnorm( 0.99 ) ) # comprobación inversa
```

```
## [1] 0.99
```

Como la distribución es simétiva podemos comprobar que se obtiene un resultado similar considerando sólo el uno por ciento de los valores:

```r
qnorm( 0.01 )
```

```
## [1] -2.326348
```

Pero para determinar el intervalo que recoge el 99 por ciento de los datos entorno a la media:

```r
qnorm( 0.005 )
```

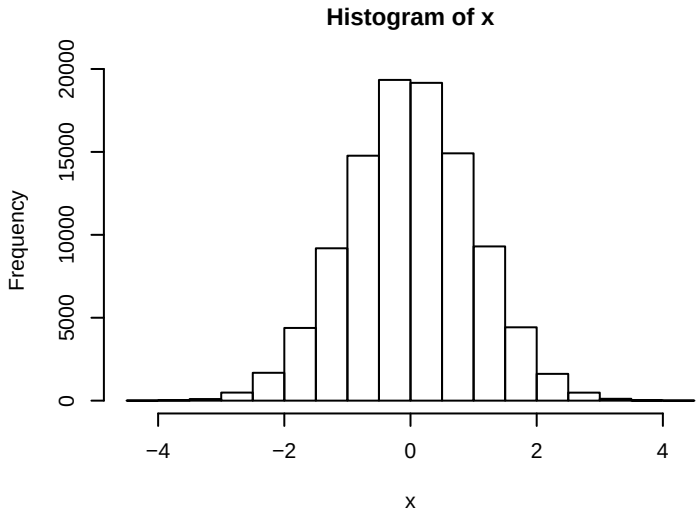
```
## [1] -2.575829
```

```r
qnorm( 0.995 )
```

```
## [1] 2.575829
```

- Simulamos valores de una distribución normal gracias a la función `rnorm`:

```r
par( cex = 0.65 )
x <- rnorm( 100000 )
hist( x )
```



```r
summary( x )
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -4.181000 -0.672800  0.000859  0.001200  0.676500  4.257000
```

```r
sd( x )
```

```
## [1] 0.9990276
```

A modo de reflexión

1. Calcular los valores de z que dejan el 0.25, el 0.5 y el 0.75 de los datos a su izquierda. ¿Coinciden con los cuartiles y mediana obtenidos en la simulación? Justifica la respuesta.
2. Intenta representar el histograma de los valores simulados junto a la curva de la distribución normal en un gráfico (pista: utiliza frecuencias relativas).
3. ¿Qué intervalo deja al 95 por ciento de los datos entorno a la media en una normal?
4. ¿Coinciden los estadísticos calculados para los datos simulados con los esperados? Justifica la respuesta.