DISTRIBUCIONES DE PROBABILIDAD

2023-11-11

ÍNDICE

DISTRIBUCIONES DISCRETAS

DISTRIBUCIÓN BINOMIAL

Mide el número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí con una probabilidad p de éxito entre los ensayos.

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

Es la probabilidad de tener x éxitos en n ensayos.

EJEMPLO

La probabilidad de que un jugador de baloncesto enceste un triple es del 30 %, es decir, $p = 0.3$. Suponemos que lanza 5 veces, $n = 5$.

a) Si quieremos calcular la probabilidad de que enceste tres tiros ($P(X = 3)$):

```
dbinom(3, size = 5, prob = 0.3)
```

```
## [1] 0.1323
```

b) Si queremos calcular todas las probabilidades de golpe en forma de tabla:

```
RBinom <- data.frame(Pr = dbinom(0:5, size = 5, prob = 0.3))
rownames(RBinom) <- 0:5
RBinom
```

```
##        Pr
## 0 0.16807
## 1 0.36015
## 2 0.30870
## 3 0.13230
## 4 0.02835
## 5 0.00243
```

c) La probabilidad de encestar más de 3 triples ($P(X \geq 3)$):

```
pbinom(3, size = 5, prob = 0.3, lower.tail = FALSE)
```

```
## [1] 0.03078
```

d) Si queremos saber todas las probabilidades acumuladas (más de 0, más de 1, ...), calculamos la cola derecha ($P(X > x)$):

```
pbinom(0:5, size = 5, prob = 0.3, lower.tail = FALSE)
```

```
## [1] 0.83193 0.47178 0.16308 0.03078 0.00243 0.00000
```

e) Si queremos calcular las colas izquierdas ($P(X \leq x)$), esto es, la probabilidad de encestar menos de 1, menos de 2, ...:

```
pbinom(0:5, size = 5, prob = 0.3, lower.tail = TRUE)
```

```
## [1] 0.16807 0.52822 0.83692 0.96922 0.99757 1.00000
```

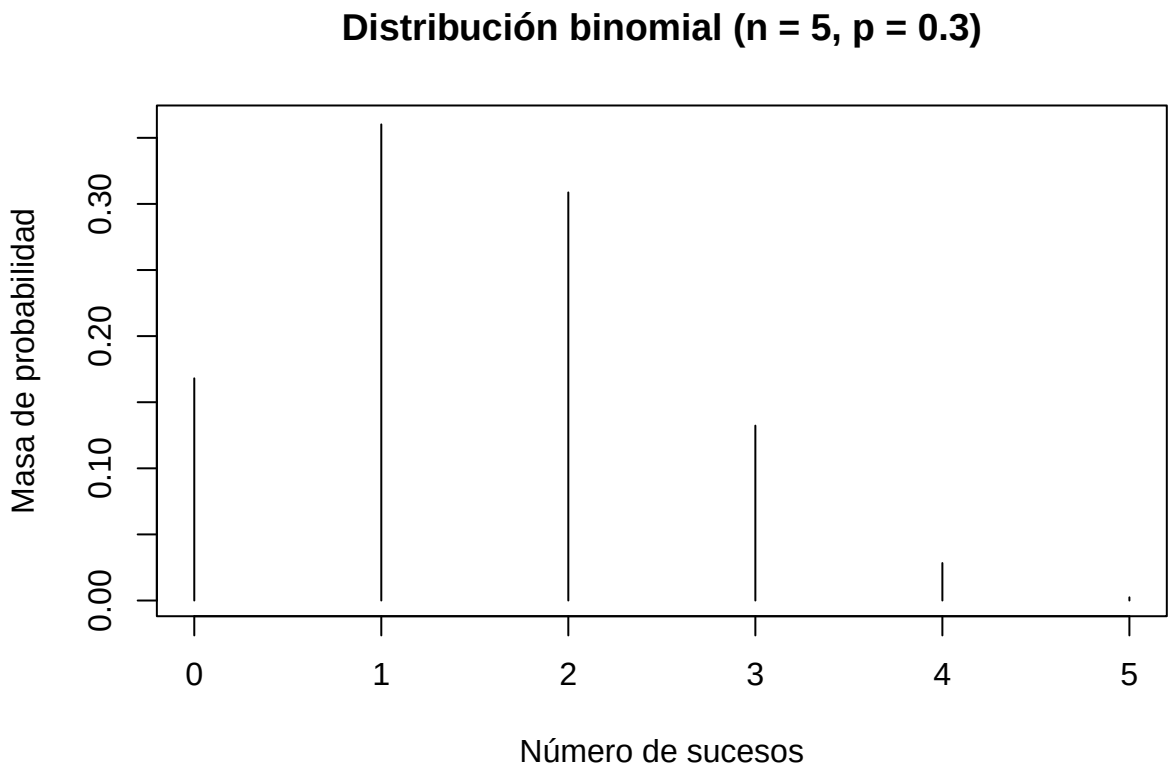
f) La probabilidad de encestar menos de dos tiros ($P(X \leq 1)$):

```
pbinom(1, size = 5, prob = 0.3, lower.tail = TRUE)
```

```
## [1] 0.52822
```

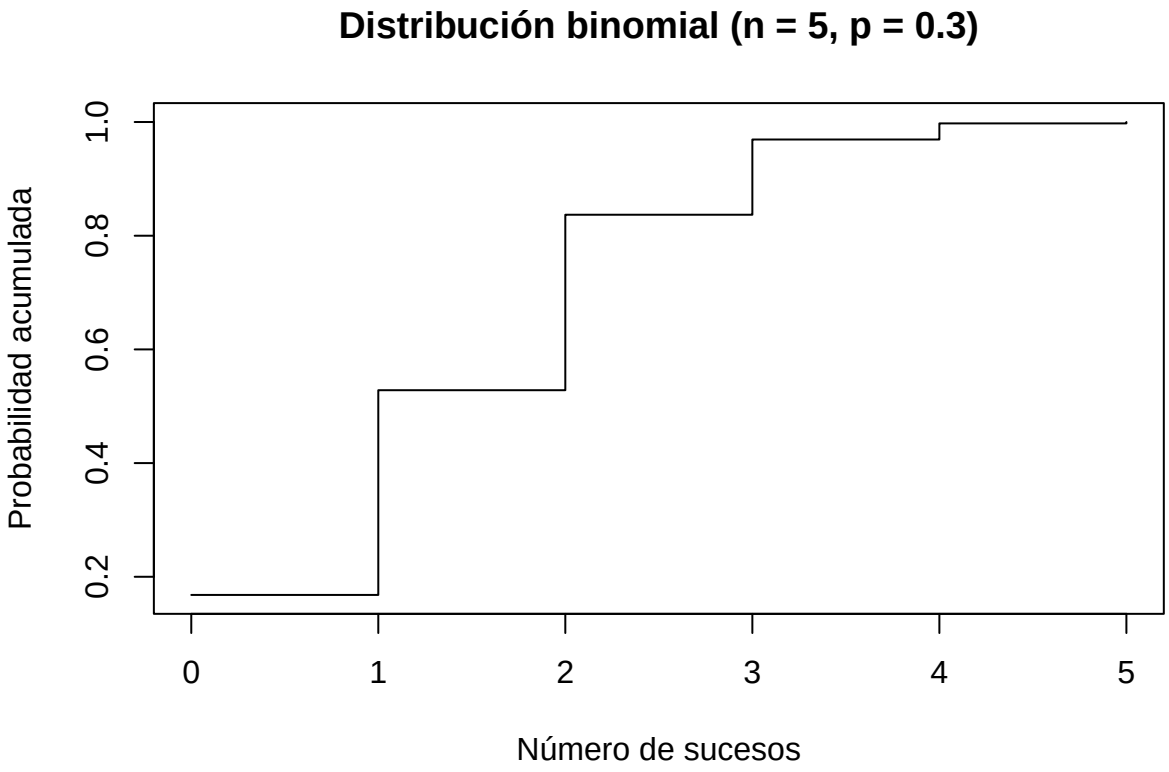
g) Para representar gráficamente la función de probabilidad:

```
x <- 0:5
Binom_plot <- dbinom(x, size = 5, prob = 0.3)
plot(x, Binom_plot, main = "Distribución binomial (n = 5, p = 0.3)",
     xlab = "Número de sucesos", ylab = "Masa de probabilidad",
     type = "h")
```



h) Para representar gráficamente la función de probabilidad acumulada:

```
x <- 0:5
Acum <- pbinom(x, size = 5, prob = 0.3)
plot(x, Acum, main = "Distribución binomial (n = 5, p = 0.3)",
     xlab = "Número de sucesos", ylab = "Probabilidad acumulada",
     type = "s")
```



DISTRIBUCIÓN GEOMÉTRICA

Describe la probabilidad del número de ensayos de Bernoulli necesarios para obtener un éxito.

$$P(X = x) = (1 - p)^{x-1}p$$

$$P(X \leq x) = F(x) = 1 - (1 - p)^{x}$$

En este caso, x es el número de intento en el que el jugador tendrá éxito (encestará el triple) y p es la probabilidad de encestar.

EJEMPLO

Definimos una variable aleatoria X, que será el número del intento en el que el jugador encesta el primer triple, es decir, el número de ensayos necesarios hasta encestar.

a) Si queremos saber la probabilidad de que el primer triple que enceste sea en el sexto intento con una probablidad de encestar de $p = 0.3$ ($P(X \leq 6) = F(6)$):

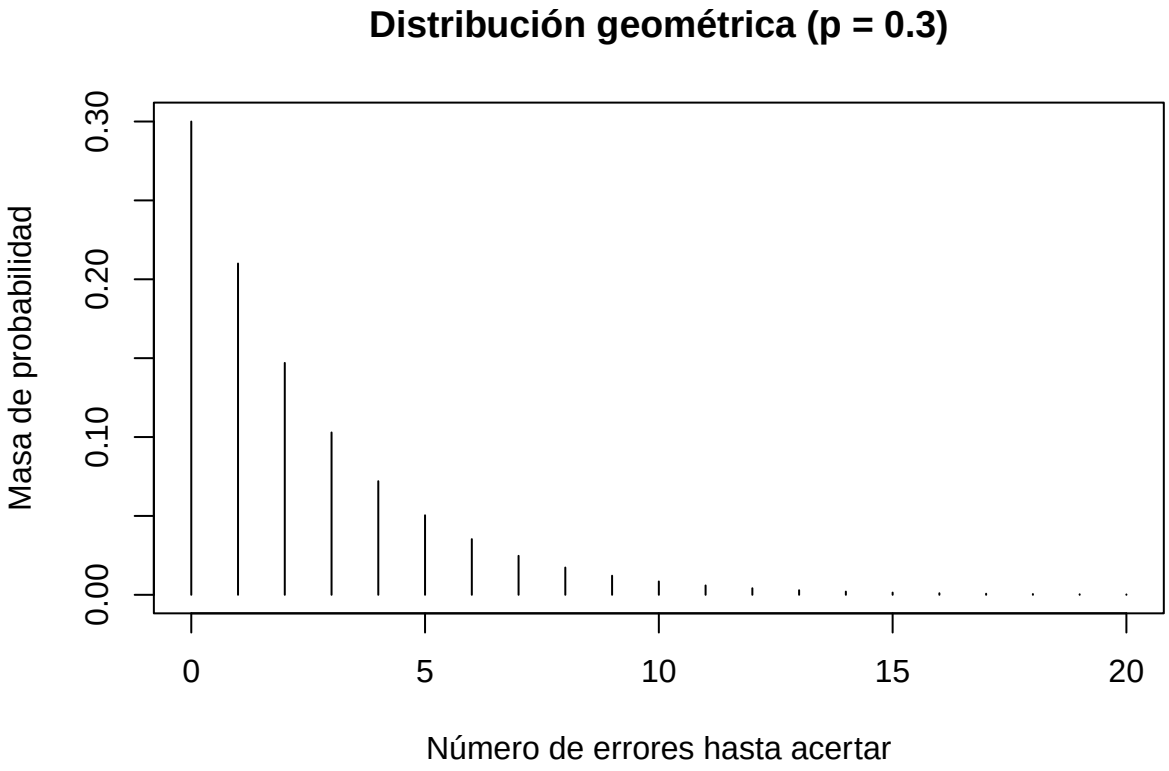
```
pgeom(5, prob = 0.3, lower.tail = TRUE)
```

```
## [1] 0.882351
```

Donde el primer número (5) es el **número de fallos** antes del primer acierto, esto es $x - 1$. Estamos calculando la **cola izquierda** ($P(X \leq x)$).

b) Para hacer la representación gráfica tomando 20 como el número máximo de intentos antes de acertar, definimos primero el vector x que contiene el número de intentos, y el vector y, que contiene la probabilidad de acertar el primer triple en el i-ésimo tiro:

```r
x <- 0:20
y <- dgeom(x, prob = 0.3)
plot(x, y, main = "Distribución geométrica (p = 0.3)",
     xlab = "Número de errores hasta acertar",
     ylab = "Masa de probabilidad", type = "h")
```



DISTRIBUCIÓN HIPERGEOMÉTRICA

Se da cuando las observaciones no son independientes. Supongamos que tenemos una muestra de N bolas, de las cuales N_1 son verdes y N_2 son rojas, de modo que $N_1 + N_2 = N$. Si extraemos n bolas (sin retorno), la variable X será el número de bolas verdes obtenidas.

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$$

EJEMPLO

Un estudiante de oposiciones ha preparado doce de los dieciocho temas de los que consta el examen. La prueba consiste en tres temas elegidos de manera aleatoria. ¿Qué probabilidad tiene el estudiante de conocer los tres temas? Sabemos que $N_1 = 12$, $N_2 = 6$ y $N = 18$. Además, la extracción es $n = 3$.

```r
RHiper <- data.frame(Pr = dhyper(0:3, m = 12, n = 6, k = 3))
rownames(RHiper) <- 0:3
RHiper
```

```
##          Pr
## 0 0.0245098
## 1 0.2205882
## 2 0.4852941
## 3 0.2696078
```

La probabilidad de que los tres temas que ha estudiado entren en el examen es del 26.9 %. La probabilidad más alta es $P(X = 2) = 0.485$, es decir, que haya estudiado dos de los tres temas que saldrán.

a) Si queremos saber la probabilidad de que sepa como mínimo dos temas ($P(X \geq 2)$), necesitamos la probabilidad acumulada de la cola derecha. *R* calcula $P(X > x)$, así que tendremos que calcular la función de distribución con $x = 1$ porque $P(X > 1) = P(X \geq 2)$:

```
phyper(1, m = 12, n = 6, k = 3, lower.tail = FALSE)
```

```
## [1] 0.754902
```

Esto significa que tiene una probabilidad del 75 % de que conozca dos o tres temas en el examen.

DISTRIBUCIÓN DE POISON

Como las anteriores, también se deriva de la distribución binomial.

EJEMPLO

Tomamos una variable X, definida como el tiempo de espera del autobús en minutos, y que tiene como único parámetro λ, que representa tanto la media, como la varianza de la variable aleatoria.

Supongamos que el tiempo de espera es de 3 minutos, o sea, $\lambda = 3$.

a) Para saber la probabilidad de que el autobús tarde 5 minutos:

$$P(X = 5) = \frac{\lambda^k}{k!} e^{-\lambda} = \frac{3^5}{5!} e^{-3} = 0.1008$$

b) Si queremos saber la probabilidad de los once primeros valores (desde $P(X = 0)$ hasta $P(X = 10)$):

```
RPoiss <- data.frame(Pr = dpois(0:10, lambda = 3))
rownames(RPoiss) <- 0:10
RPoiss
```

```
##              Pr
## 0  0.0497870684
## 1  0.1493612051
## 2  0.2240418077
## 3  0.2240418077
## 4  0.1680313557
## 5  0.1008188134
## 6  0.0504094067
## 7  0.0216040315
## 8  0.0081015118
## 9  0.0027005039
## 10 0.0008101512
```

Al igual que en las distribuciones anteriores, se pueden calcular las probabilidades acumuladas, gráficos y simulaciones muestrales.

DISTRIBUCIONES CONTINUAS

DISTRIBUCIÓN UNIFORME

El dominio de esta distribución está definido por los valores máximo y mínimo a y b. Su función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{para} \quad a \leq x \leq b \\ 0 & \text{para} \quad x < a \ \text{ ó } \ x > b \end{cases}$$

EJEMPLO

Una mujer está dando a luz a un bebé, y la hora exacta del alumbramiento (X) sucederá en cualquier momento entre la hora 0 (ahora) y la hora 24 (la misma hora del día siguiente).

a) Para calcular la probabilidad de que la madre de a luz dentro de las primeras cinco horas ($P(X < 5)$) escribiremos:

```
punif(5, min = 0, max = 24, lower.tail = TRUE)
```

```
## [1] 0.2083333
```

En distribuciones continuas, $P(X < x)$ equivale a $P(X \leq x)$. En una distribución continua el cálculo de una probabilidad del tipo $P(X = x)$ será siempre nula, es decir, $P(X = x) = 0$.

b) Si queremos calcular la probabilidad de que la madre para en las últimas dos horas (a partir de la hora 22), estaremos calculando ($P(X > 22)$):

```
punif(22, min = 0, max = 24, lower.tail = FALSE)
```

```
## [1] 0.08333333
```

DISTRIBUCIÓN EXPONENCIAL

Esta distribución está relacionada con la distribución de Poisson.

Con esta distribución se modeliza el intervalo de tiempo que transcurre entre dos sucesos. Tiene un único parámetro λ y está definida para valores no negativos de la variable aleatoria.

Sus funciones de densidad y de distribución son:

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

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$$

EJEMPLO

En un hospital, el tiempo que transcurre entre dos partos (medido en horas) sigue una distribución exponencial con un parámetro $\lambda = 4$. Esto significa que la esperanza viene definida como $E(X) = \frac{1}{\lambda} = \frac{1}{4} = 0.25$, de media transcurre un cuarto de hora ($\frac{1}{4} = 0.25$) entre un parto y otro.

a) La probabiliad de que transcurra una hora o más entre dos partos ($P(X > 1)$):

```
pexp(1, rate = 4, lower.tail = FALSE)
```

```
## [1] 0.01831564
```

Hay aproximadamente un 1.8 % de probabilidad de que transcurra una hora o más entre dos partos.

b) La probabilidad de que transcurra como máximo una hora entre dos partos ($P(X \leq 1) = 1 - P(X > 1)$):

```
pexp(1, rate = 4, lower.tail = TRUE)
```

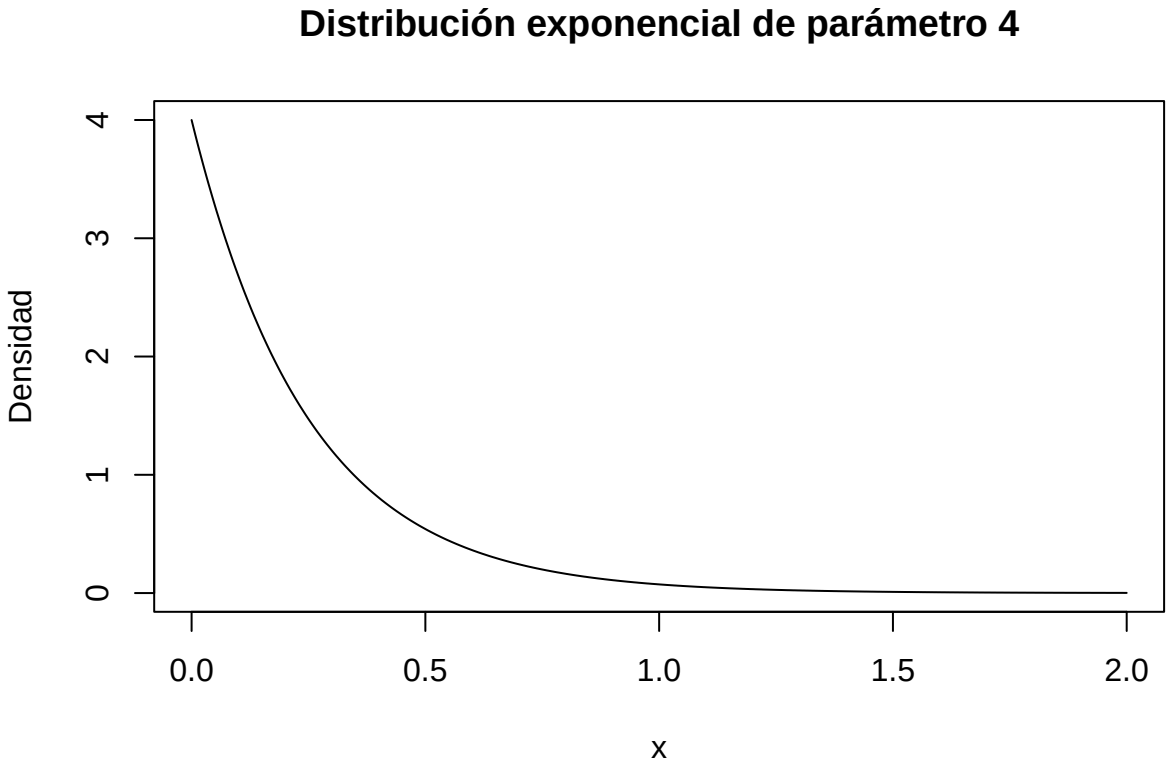
```
## [1] 0.9816844
```

Como vemos, se cumple:

$$P(X \leq x) + P(X > x) = 1$$

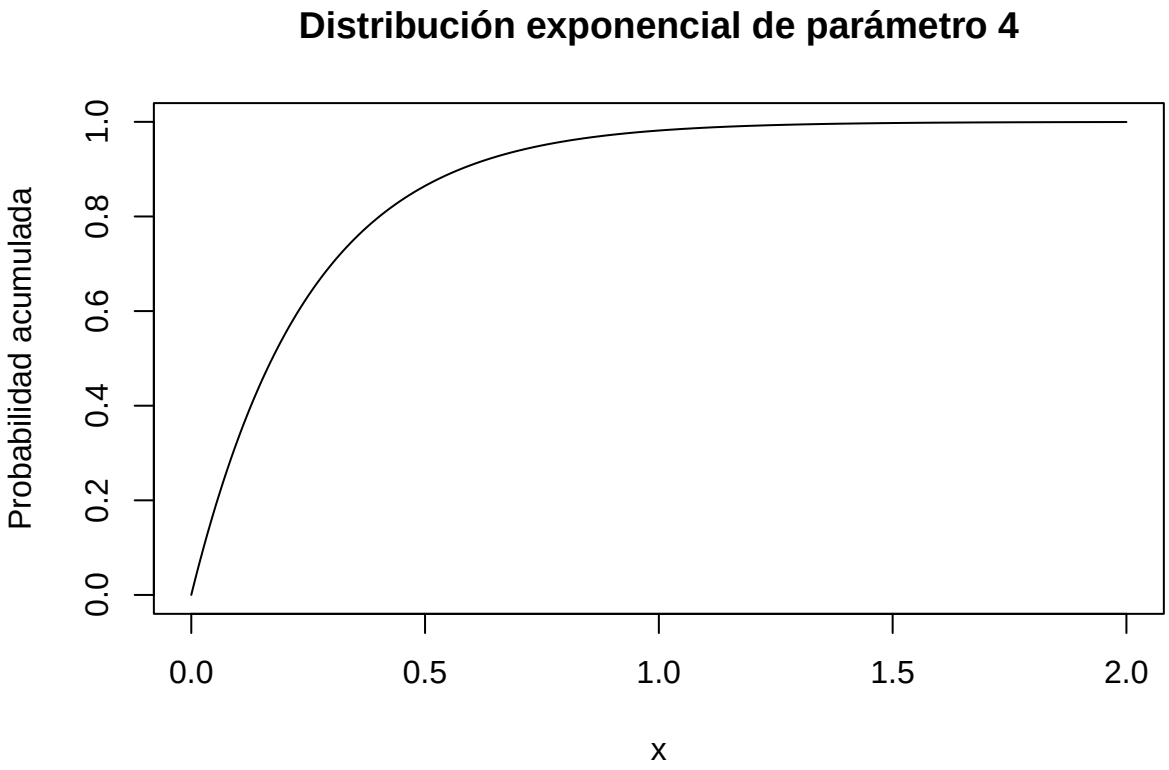
c) Para obtener la gráfica de la función de densidad, definimos un vector x de 1000 valores, equiespaciados entre 0 y 2 y luego calculamos la función de densidad correspondiente a cada valor, que guardamos en otro vector:

```
x <- seq(0, 2, length.out = 1000)
y <- dexp(x, rate = 4)
plot(x, y, main = "Distribución exponencial de parámetro 4",
     xlab = "x", ylab = "Densidad", type = "l")
```



d) Para la función de distribución tendremos que guardar en el vector los valores de la función de distribución correspondientes a cada uno de los valores de x:

```
x <- seq(0, 2, length.out = 1000)
y <- pexp(x, rate = 4, lower.tail = TRUE)
plot(x, y, main = "Distribución exponencial de parámetro 4",
     xlab = "x", ylab = "Probabilidad acumulada", type = "l")
```



DISTRIBUCIÓN NORMAL

La distribución normal o gaussiana tiene dos parámetros: la media μ y la desviación típica o estándar σ. Su función de densidad viene dada por la siguiente expresión:

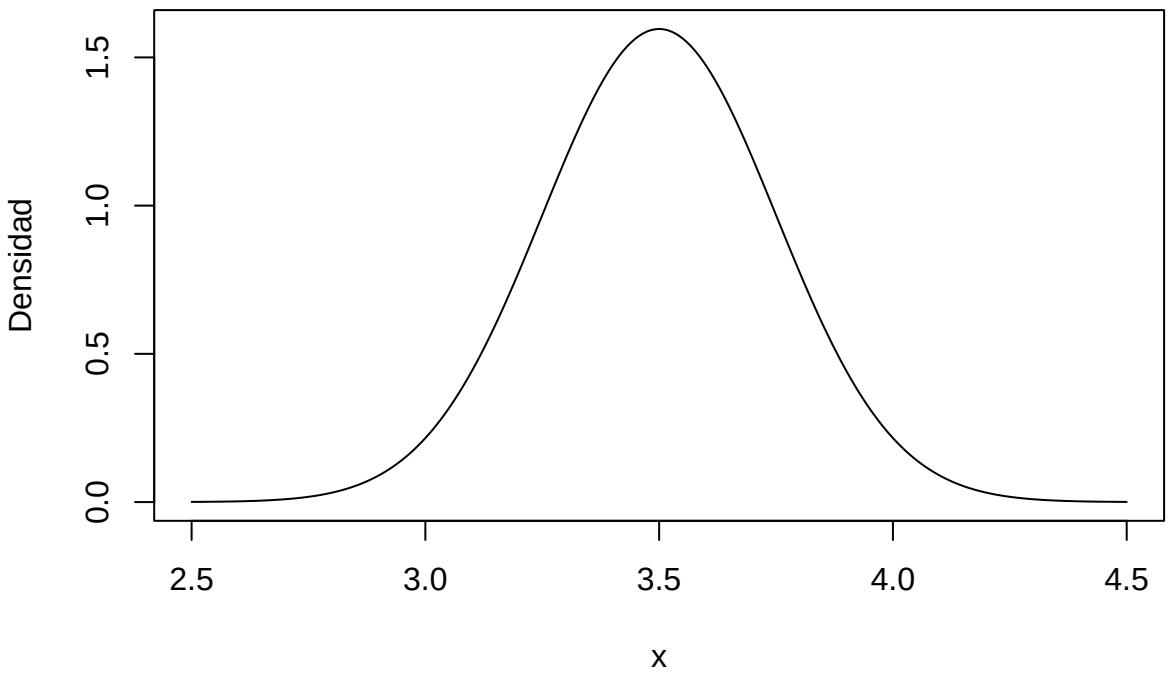
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

EJEMPLO

En un hospital, el peso de los bebés al nacer sigue una distribución normal con media $\mu = 3.5$ y desviación típica $\sigma = 0.25$.

a) Queremos visualizar la forma de la función de densidad. Para eso, tenemos que definir un vector x de forma que su mediana sea 3.5. En este caso, podríamos tomar 1000 número equidistantes entre 2.5 y 4.5 (para asegurarnos de que el 3.5 queda en la mitad del vector). Luego, calculamos el vector y, donde cada componenete será la función de densidad de la coordenada de x correspondiente.

```
x <- seq(2.5, 4.5, length.out = 1000)
y <- dnorm(x, mean = 3.5, sd = 0.25)
plot(x, y, main = "Distribución normal de media 3.5 y desviación típica 0.25",
     xlab = "x", ylab = "Densidad", type = "l")
```



Vemos cómo la mayor parte de los niños pesan entre 3 y 4 kilos al nacer.

b) Calculamos la probabilidad $P(3 \leq X \leq 4))$

Al ser una distribución simétrica, vemos que si trazamos una línea vertical en la media aritmética (3.5), en las dos partes de la distribución queda el 50 % de la masa probabilística. Esto también se puede ver si nos fijamos en que $\mu = 3.5$ se sitúa en medio de los valores 3 y 4; entonces, el área que quedará a la izquierda de 3 y el área que quedará a la derecha de 4 serán iguales debido a la simetría de la distribución. Por tanto, si queremos calcular el área que queda entre 3 y 4, es decir $P(3 \leq X \leq 4)$, una opción es calcular uno menos las dos colas de los extremos, $P(X \leq 3)$ y $P(X \geq 4)$:

$$P(3 \leq X \leq 4) = 1 - P(X \leq 3) - P(X \geq 4)$$

Como sabemos que estos dos extremos han de tener el mismo área o valor, es decir, $P(X \leq 3) = P(X \geq 4)$, la fórmula anterior se puede simplificar de la manera siguiente:

$$P(3 \leq X \leq 4) = 1 - 2P(X \leq 3)$$

Otras formas alternativas de obtener esta probabilidad son:

$$P(3 \leq X \leq 4) = 1 - 2P(x \geq 4)$$

$$P(3 \leq X \leq 4) = P(X \leq 4) - P(X \leq 3)$$

Calculamos la probabilidad $P(X \leq 3)$ con R siguiendo la primera opción:

```
pnorm(3, mean = 3.5, sd = 0.25, lower.tail = TRUE)
```

```
## [1] 0.02275013
```

De esta forma, obtenemos que $P(3 \leq X \leq 4) = 1 - 2 \cdot 0.023 = 0.954$, es decir, más del 95 % de los bebés nacerán dentro del intervalo de pesos $[3, 4]$.

c) Supongamos que el jefe de pediatría quiere separar a los bebés en tres grupos: a) muy poco peso, b) poco peso y c) peso normal. El doctor decide que habrá un 10 % de los bebés en el primer grupo, un 20 % en el segundo y el resto (un 70 %) en el tercero. ¿Cuáles serán los puntos de corte entre los tres grupos?

Para contestar a esta pregunta tenemos que calcular los percentiles 0.1 y 0.2, que dividen el área total en las tres partes que nos interesan (es decir, 10 %, 20 % y 70 %, en este orden):

```
qnorm(c(0.1, 0.2), mean = 3.5, sd = 0.25, lower.tail = TRUE)
```

```
## [1] 3.179612 3.289595
```

Estos percentiles indican que los bebés con muy poco peso serán los que pesan menos de 3.18 kilos; los bebés con poco peso serán los que pesen entre 3.18 y 3.29 kilos y los bebés con peso normal serán los que pesen más de 3.29 kilos.

DISTRIBUCIÓN DE LA MEDIA MUESTRAL

Es un caso relevante en el estudio de la distribución normal.

EJEMPLO

Los doctores del ejemplo anterior quieren hacer estadísticas de control del peso de los bebés que nacen. Interpretando los n bebés que nacen cada día como una muestra aleatoria independiente, se analiza para cada una de estas muestras la media muestral. Ya que cada día (es decir, cada muestra) se obtendrá una media muestral distinta, esta se puede analizar como una variable aleatoria denominada $\overline{X}$, que tiene su propia distribución. A la hora de analizar y calcular probabilidades sobre $\overline{X}$, es imprescindible distinguir si conocemos o no la dispersión de la población (σ) de la que se obtiene $\overline{X}$.

Si esta dispersión es conocida, $\overline{X}$ se distribuye como una normal, tal que:

$$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Si n es el tamaño de la muestra, suponemos que disponemos de una muestra de 150 bebés ($n = 150$) y queremos calcular $P(\overline{X} \geq 3.75)$, sabiendo que $\mu = 3.5$ y que $\sigma = 0.25$.

Primero, calculamos cuál es la desviación típica de la variable media muestral $\overline{X}$:

$$\frac{\sigma}{\sqrt{n}} = \frac{0.25}{\sqrt{150}} = 0.02$$

Si sabemos que $\overline{X} \sim N(3.5, 0.02)$, calculamos esta probabilidad en R de la siguiente manera:

```
pnorm(3.75, mean = 3.5, sd = 0.02, lower.tail = FALSE)
```

```
## [1] 3.732564e-36
```

El resultado es prácticamente cero. Este resultado indica que la probabilidad de que, en un día, la media muestral de los $n = 150$ bebés nacidos sea de 3.75 kilos ($P(\overline{X} \geq 3.75)$) es prácticamente nula.