Capítulo 1

Simulación de distribuciones de probabilidad univariantes conocidas

En esta primera práctica vamos a explicar cómo dar los primeros pasos en la simulación de datos provenientes de variables aleatorias que siguen una distribución de probabilidad conocida, como puede ser el caso de la normal, la exponencial, ... Para ello, en primer lugar revisaremos los conceptos previos necesarios para desarrollar esta práctica, para posteriormente ver cómo hacer cálculos con distribuciones de probabilidad conocidas, y finalmente veremos cómo simulares valores aleatorios que provienen de una distribución conocida.

1.1. Conceptos previos

Comenzamos con una variable aleatoria (unidimensional) X definida sobre el espacio probabilístico $(\Omega, \mathcal{A}, P)$, y denotamos por P_X su probabilidad inducida. Definimos la función de distribución de la variable aleatoria X, denotada por F_X, como:

$$F_X(t) = P_X\big((-\infty, t]\big) = P(X \leq t) = P\big(X \in (-\infty, t]\big) = P\big(\{\omega \in \Omega \mid X(\omega) \leq t]\big) \quad \forall t \in \mathbb{R}.$$

$F_X(t)$ nos indica la probabilidad acumulada hasta el valor t por la variable aleatoria X. Sabemos que toda función de distribución satisface una serie de propiedades:

Creciente: F_X es una función creciente (no necesariamente estricta).

Normalización: $\lim_{t\to\infty} F_X(t) = 1$ y $\lim_{t\to-\infty} F_X(t) = 0$.

Continuidad por la izquierda: $\lim_{\varepsilon\to 0} F_X(t+\varepsilon) = F_X(t)$.

Cuando la variable X es continua, la función de distribución F_X es continua, mientras que para variables discretas, la función de distribución tiene "saltos" en los puntos del soporte, mientras que en el resto de valores se mantiene constante.

Haciendo uso de la función de distribución F_X podemos calcular probabilidades de intervalos de la variable X:

$$\begin{array}{ll} P(X \leq t) = F_X(t), & P(X > t) = 1 - F_X(t). \\ P(X < t) = F_X^-(t), & P(X \geq t) = 1 - F_X^-(t). \\ P(t_1 < X \leq t_2) = F_X(t_2) - F_X(t_1), & P(t_1 \leq X \leq t_2) = F_X(t_2) - F_X^-(t_1). \\ P(t_1 < X < t_2) = F_X^-(t_2) - F_X(t_1), & P(t_1 \leq X < t_2) = F_X^-(t_2) - F_X^-(t_1). \end{array}$$

Una función asociada a la función de distribución y que también tiene relevancia es la función cuantil, denotada por $F_X^{(-1)} : (0,1) \to \mathbb{R}$, y definida por:

$$F_X^{(-1)}(p) = \text{ínf}\{x \in \mathbb{R} \mid F_X(x) \geq p\}.$$

Dicho de otra forma, la función cuantil en el valor p indica el menor valor de manera que por debajo suyo hay una probabilidad acumulada de al menos p.

Cuando la función de distribución es continua y estrictamente creciente, la función cuantil coincide con la inversa de F en los puntos t donde $F(t) \in (0,1)$. En otro caso, $F_X^{(-1)}$ también recibe el nombre de cuasi-inversa de F_X. Además, la función cuantil en algunos valores concretos es conocida, como por ejemplo:

$F_X^{(-1)}(0{,}5) = \text{ínf}\{x \in \mathbb{R} \mid F_X(x) \geq 0{,}5\}$=Mediana$(X)$
$F_X^{(-1)}(0{,}25) = \text{ínf}\{x \in \mathbb{R} \mid F_X(x) \geq 0{,}25\}$=Primer cuartil de X
$F_X^{(-1)}(0{,}75) = \text{ínf}\{x \in \mathbb{R} \mid F_X(x) \geq 0{,}75\}$=Tercer cuartil de X

1.2. Distribuciones univariantes habituales

Hay una serie de distribuciones de probabilidad conocidas y de manejo habitual. En el siguiente cuadro listamos algunas de las distribuciones continuas más relevantes:

Distribución	Notación	Densidad	Valores
Normal	$\mathcal{N}(\mu,\sigma)$	$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu \in \mathbb{R}, \sigma > 0$
Uniforme	$\mathcal{U}(a,b)$	$f(t) = \frac{1}{b-a}, \quad t \in (a,b)$	$a, b \in \mathbb{R}$, $a < b$
Exponencial	$exp(\lambda)$	$f(t) = \lambda e^{-\lambda t}, \quad t > 0$	$\lambda > 0$
Gamma	$\Gamma(k,\beta)$	$f(t) = e^{-t\beta} \frac{t^{k-1}\beta^k}{\Gamma(k)}, \quad t > 0$	$k, \beta > 0$
Beta	$\beta(a,b)$	$f(t) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}, \quad t \in (0,1)$	$a, b > 0$
Chi-cuadrado	χ^2_k		$k \in \mathbb{N}$
F de Snedecor	F_{d_1,d_2}		$d_1, d_2 > 0$
t de student	t_k		$k \in \mathbb{N}$

En el caso de las distribuciones chi-cuadrado, F de Snedecor y t de Student, omitimos la expresión de la función de densidad al ser demasiado compleja. En el siguiente cuadro enumeramos algunas distribuciones discretas:

Distribución	Notación	Soporte	Probabilidad	Valores
Binomial	$\mathcal{B}(n,p)$	$k = 0, 1, \ldots, n$	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$	$n \in \mathbb{N}, p \in (0,1)$
Poisson	$\mathbb{P}(\lambda)$	$k \in \mathbb{N}$	$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$	$\lambda > 0$
Geométrica	$\mathcal{G}(p)$	$k \in \mathbb{N}$	$P(X = k) = (1-p)^k p$	$p \in (0,1)$
Distribución discreta	$\mathcal{U}(x_1, \ldots, x_n)$	$k = x_1, \ldots, x_n$	$P(X = x_i) = p_i$	$x_1 \leq \ldots \leq x_n$ $p_1 + \ldots + p_n = 1$

1.3. Cálculos de probabilidades, cálculo de la función de distribución y de la función cuantil

Para las distribuciones continuas y discretas enumeradas en el apartado anterior, podemos hacer uso de las funciones propias de `R` para hacer cálculos relativos a probabilidades o valores de las funciones de distribución o cuantil. Para ello, debemos utilizar las siguientes instrucciones:

Distribución	Probabilidad/densidad	CDF	Inversa
Normal	`dnorm`	`pnorm`	`qnorm`
Uniforme	`dunif`	`punif`	`qunif`
Exponencial	`dexp`	`pexp`	`qexp`
Gamma	`dgamma`	`pgamma`	`qgamma`
Beta	`dbeta`	`pbeta`	`qbeta`
Chi-cuadrado	`dchisq`	`pchisq`	`qchisq`
F de Snedecor	`df`	`pf`	`qf`
t de Student	`tpdf`	`tcdf`	`tinv`
Binomial	`dbinom`	`pbinom`	`qbinom`
Poisson	`dpois`	`ppois`	`qpois`
Geométrica	`dgeom`	`pgeom`	`qgeom`

Hemos de hacer un pequeño matiz en el comando `pdf`. Para distribuciones continuas, este comando proporciona el valor de la función de densidad en el valor indicado ($f_X(t)$), mientras que para distribuciones discretas este comando proporciona la probabilidad en el punto de interés ($P(X = t)$).

Vamos a detallar cómo utilizar los comandos anteriores:

Probabilidad/densidad En este caso, debemos de utilizar la abreviatura del nombre de la distribución (por ejemplo, `norm` en el caso de la normal), añadiendo la como primera letra `d`. Debemos indicar primero el valor en el que queremos calcular el la densidad, si la distribución es continua, o el valor de la probabilidad, si la distribución es discreta; a continuación indicaremos el valor de los parámetros que determinan la distribución (por ejemplo, en el caso de la normal, primero indicaremos el valor de la media y a continuación el valor de la desviación típica). Ejemplos:

$f_{\mathcal{N}(0,1)}(0{,}3)$: Valor de la densidad de una $\mathcal{N}(0,1)$ en el punto 0.3:

```
dnorm(0.3,0,1)
```

$f_{\mathcal{U}(0,10)}(2)$: Valor de la densidad de una $\mathcal{U}(0,10)$ en el punto 2:

```
dunif(2,0,10)
```

$P\big(\mathcal{B}(5,0{,}2) = 2\big)$: Probabilidad del punto 2 de una $\mathcal{B}(5,0{,}2)$:

```
dbinom(2,5,0.2)
```

Función de distribución (CDF) Para calcular el valor de la función de distribución, utilizaremos la abreviatura del nombre de la distribución y añadiremos como primera letra la `p`. Como primer valor de entrada, indicamos el punto donde queremos calcular la función de distribución y a continuación los parámetros que caracterizan la distribución. Ejemplos:

$F_{\mathcal{N}(0,1)}(0{,}3) = P\big(\mathcal{N}(0,1) \leq 0{,}3\big)$: Valor de la CDF de una $\mathcal{N}(0,1)$ en el punto 0.3:

```
pnorm(0.3,0,1)
```

$F_{\mathcal{U}(0,10)}(2) = P\big(\mathcal{U}(0,1) \leq 2\big)$: Valor de la CDF de una $\mathcal{U}(0,10)$ en el punto 2:

```
punif(2,0,10)
```

$F_{\mathcal{B}(5,0{,}2)}(2) = P\big(\mathcal{B}(5,0{,}2) \leq 2\big)$: Valor de la CDF en el punto 2 de una $\mathcal{B}(5,0{,}2)$:

```
pbinom(2,5,0.2)
```

Función cuantil Al igual que en los apartados anteriores, utilizaremos la abreviatura del nombre de la distribución y añadiremos la primera letra `q`. Indicamos primero el valor donde queremos calcular la inversa y a continuación los parámetros que determinan la distribución. Ejemplos:

Primer cuartil de una $\mathcal{N}(0,1)$:

```
qnorm(0.25,0,1)
```

Mediana de una $\mathcal{U}(0,10)$:

```
qunif(0.5,0,10)
```

Tercer cuantil de una $\mathcal{B}(5,0{,}2)$:

```
qbinom(0.75,5,0.2)
```

A continuación mostramos algunos ejemplos sobre cómo utilizar los comandos anteriores.

Ejemplo 1.1. El número de días mensuales con lluvia en una determinada región sigue una distribución Poisson con parámetro 6. Se pide calcular:

a) ¿Cuál es la probabilidad de que en un mes haya 10 días con lluvia?

b) ¿Cuál es la probabilidad de que en un mes haya a lo sumo 5 días con lluvia?

c) En el 10 % de los días con menos lluvia el consumo de agua aumenta en dicha región y se producen cortes de agua. ¿Cuántos días tiene que llover en un mes para que no haya cortes de agua?

Tenemos una variable X con distribución $\mathcal{P}(6)$. En el primer apartado, nos piden calcular $P(X = 10)$. Para calcular esta probabilidad, usamos el comando `poisspdf`, indicando primero el valor donde queremos calcular la probabilidad y a continuación el parámetro:

```
dpois(10,6)
[1] 0.04130309
```

A continuación, nos piden calcular $P(X \leq 5) = F_X(5)$. Para calcular el valor de la función de distribución en el valor 5, utilizamos el comando `ppois` indicando el valor 5 y a continuación el valor del parámetro:

```
ppois(5,6)
[1] 0.44568
```

Por último, hemos de encontrar el menor valor x de manera que $P(X \leq x) \geq 0{,}10$. Utilizaremos el comando `qpois`, indicando el valor de la probabilidad 0.10 y a continuación el parámetro:

```
qpois(0.10,6)
[1]  3
```

Por lo tanto, si solamente hay tres días, o menos, con lluvia, se producirán cortes de agua; por contra, si hay cuatro o más días con lluvia, los cortes de agua no serán necesarios.

Ejemplo 1.2. Consideramos una variable X que sigue distribución exponencial de media 2. Eso significa que el parámetro es $\lambda = 0{,}5$, puesto que el parámetro de una exponencial es el inverso de la media. Vamos a resolver los siguientes apartados:

a) ¿Cuál es el valor de la densidad en el punto 1?

b) ¿Cuál es la probabilidad de que el valor de la variable no supere el valor 2.5?

c) ¿Cuál es el menor valor que deja por debajo suyo el 80 % de la probabilidad?

Para resolver el apartado a), simplemente hemos de utilizar el comando `exppdf`, indicando primero el valor en el cual queremos calcular la función de densidad y a continuación la media:

```
dexp(1,0.5)
[1] 0.30327
```

En el segundo apartado nos piden calcular el valor de la función de distribución F_X en el punto 2,5, esto es, $F_X(2{,}5) = P(X \leq 2{,}5)$. Para calcularlo, utilizamos el comando `pexp` indicando primero el valor en el cual queremos conocer la función de distribución y a continuación la media:

```
pexp(2.5,0.5)
[1]  0.71350
```

Por último, en el apartado c) nos piden calcular el valor de la inversa de la función de distribución en el punto 0.8. Para ello haremos uso del comando `qexp`, indicando primero el valor de la probabilidad y a continuación la media:

```
qexp(0.8,0.5)
[1] 3.2189
```

1.4. Simulación de valores aleatorios

Si queremos generar valores aleatorios que provienen de alguna de las distribuciones de probabilidad mencionadas en la Sección 1.2, podemos usar las siguientes órdenes:

Distribución	Comando	Ejemplo
Normal	`rnorm`	`rnorm`(m,mu,sigma)
Uniforme	`runif`	`runif`(m,a,b)
Exponencial	`rexp`	`rexp`(m,lambda)
Gamma*	`rgamma`	`gamrnd`(m,k,beta)
Beta	`rbeta`	`rbeta`(m,a,b)
Chi-cuadrado	`rchisq`	`rchisq`(m,k)
F de Snedecor	`rf`	`rf`(m,d1,d2)
t de Student	`rt`	`rt`(m,k)
Binomial	`rbinom`	`rbinom`(m,n,p)
Poisson	`rpois`	`rpois`(m,lambda)
Geométrica	`rgeom`	`rgeom`(m,p)
Distribución discreta	`sample`	`sample`(k,m,p,replace=TRUE)

Como se puede ver, basta añadir la letra `r` (abreviatura de *random*) al código de cada distribución, indicando el número de datos deseados `m` así como el/los parámetro(s) de la distribución.

Ejemplo 1.3. Si queremos generar 12 valores aleatorios que provienen de una distribución uniforme en el intervalo $(0, 1)$, podemos utilizar la instrucción:

```
runif(12,0,1)
```

Por otra parte, si queremos generar esos 12 valores aleatorios en forma de una matriz con 2 filas y 6 columnas, podemos utilizar la instrucción:

```
matrix(runif(12,0,1),nrow=2)
```

```
          [,1]      [,2]      [,3]       [,4]      [,5]      [,6]
[1,] 0.5988617 0.8291817 0.5520412 0.01242983 0.6438471 0.5357555
[2,] 0.6184497 0.6270486 0.2618494 0.01195311 0.1458818 0.2754560
```

Ejemplo 1.4. Si tenemos una variable aleatoria con distribución binomial con 5 repeticiones y probabilidad de éxito 0,4, podemos generar 100 valores aleatorios utilizando la instrucción:

```
x<-rbinom(100,5,0.4)
```

Utilizando estos datos generados aleatoriamente, podemos estimar las probabilidades de que una distribución $\mathcal{B}(5, 0{,}4)$ tome los valores 0, 1, 2, 3, 4, 5. Por ejemplo, podríamos utilizar las siguientes órdenes:

```
p<-c()
for(i in 0:5){p[i+1]<-sum(x==i)/100}
p
[1] 0.08 0.24 0.46 0.15 0.06 0.01
```

También podemos estimar la media o la desviación típica de $\mathcal{B}(5, 0{,}4)$:

```
mean(x)
[1] 1.9
sd(x)
[1] 1.020002
```

Si comparamos los valores estimados tanto de las probabilidades como de la media y desviación típica, obtenemos los siguientes resultados:

$X = \mathcal{B}(5,0{,}4)$	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$	$P(X=5)$	$E(X)$	$SD(X)$
Exacto	0.07776	0.25920	0.34560	0.23040	0.07680	0.01024	2	$\sqrt{1{,}2} \approx 1{,}0954$
Estimado	0.09000	0.28000	0.26000	0.27000	0.10000	0.00000	1.9	1.020002

Evidentemente, cuanto mayor sea el tamaño muestral, más cercanos tenderán a estar los valores estimados de los valores exactos.

Ejemplo 1.5. Vamos a estimar la función de distribución de una distribución $\mathcal{N}(0,1)$. Para ello, vamos a generar 10000 valores aleatorios:

```
x<-rnorm(10000,0,1)
```

A continuación vamos a estimar los valores de la función de distribución:

```
xcdf<-ecdf(x)
plot(xcdf)
```

En la vector `xcdf` tenemos la estimación de la función de distribución. Por ejemplo, si ahora queremos calcular el valor estimado de la función de distribución en el valor 0, que sabemos que toma el valor 0, lo podemos estimar como:

```
xcdf(0)
[1] 0.5007
```

Esta función de distribución empírica se puede dibujar de dos formas distintas. La primera de ellas sería mediante el comando `plot`, directamente aplicado sobre el vector `xcdf`. La segunda manera, claramente más elegante, sería mediante el paquete `ggplot2`. Para ello, utilizamos las siguientes instrucciones:

```
library(ggplot2)  # Instrucción para cargar el paquete ggplot2
ggplot(data=NULL, aes(x))+
  stat_ecdf(col="red",size=1.2)+  # Instrucción para dibujar la ECDF
  labs(x="", y="Función de distribución empírica")+
    # Añadir etiquetas a los ejes
  ggtitle("Estimación de la función de distribución de una N(0,1)")
    # Añadir un título al gráfico
```

Con la primera línea iniciamos el gráfico, y con la segunda parte añadimos la función de distribución empírica (`stat_ecdf`) en color rojo y con un ancho de línea de 1.2. Por último, en las líneas 3 y 4 cambiamos las etiquetas de los ejes y añadimos un título al gráfico. La representación obtenida se puede ver en la Figura 1.1. Si queremos estimar el valor del cuantil 95, podemos hacerlo de la siguiente manera:

```
quantile(x,0.95)
     95%
1.656144
```

De todas las distribuciones vistas, la única que tiene un comportamiento diferente es la distribución discreta. En este caso, como argumentos de entrada hay que indicar el vector de valores que toma la variable (o soporte), denotado por `k`, el número de valores deseados, denotado por `m`, así como el vector `p` de probabilidades.

Ejemplo 1.6. Consideremos una variable aleatoria X con distribución discreta, cuyos valores y probabilidades se resumen a continuación:

X	1	2	3
p_X	0,2	0,5	0,3

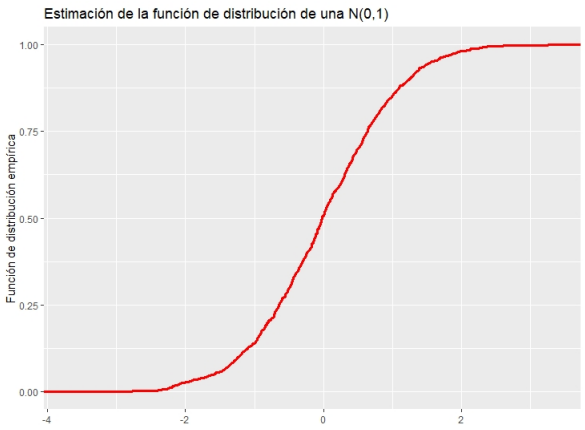

Figura 1.1: Representación gráfica de la función de distribución empírica del Ejemplo 1.5.

Si queremos simular 100 valores aleatorios, podemos utilizar el comando `sample` como se indica a continuación:

```
n<-100
k<-c(1,2,3)
prob<-c(0.2,0.5,0.3)
x<-sample(k,n,prob,replace=TRUE)
```

Una vez simulados los datos, es posible realizar esimaciones de parámetros, probabilidades o de la función de distribución como se ha mostrado en los ejemplos anteriores.

1.5. Ejercicios propuestos

Ejercicio 1.1. Sea X una variable aleatoria con distribución $\mathcal{N}(10,3)$. Genera 1000 valores aleatorios de dicha distribución y realiza los siguientes cálculos:

1. Calcula la máxima diferencia, en valor absoluto, entre la función de distribución "real" (utilizando el comando `pnorm`) y la función de distribución empírica.
2. Estima el valor de la media y de la desviación típica.
3. Estima el valor de la probabilidad $P(8 \leq \mathcal{N}(10,3) \leq 11)$ y calcula la diferencia, en valor absoluto, entre el valor obtenido con el comando `pnorm`.

Ejercicio 1.2. Rellena la información que falta en la siguiente tabla. Para hacer las estimaciones, utiliza una muestra de tamaño 200:

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X>4)$	$E(X)$	$DT(X)$
$X=\mathcal{B}(10,0{,}4)$							
$X=\mathcal{P}(7)$							
$X=\mathcal{G}(0{,}7)$							

Ejercicio 1.3. Rellena la siguiente tabla con los datos perdidos. En la columna de "exacto", indica el valor proporcionado por los comando `pnorm`, y en las columnas siguientes estima las probabilidades pedidas con una muestra del tamaño pedido.

$X = \mathcal{N}(0,1)$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 10000$	"Exacto"
$P(X \leq 0)$							
$P(-1 \leq X \leq 1)$							
$E(X)$							
$P(X > 2)$							