Sesión 1

Introducción a la simulación de procesos estocásticos.

Un **proceso estocástico** es cualquier familia de variables aleatorias $\{X_t\}_{t\in T}$ definidas sobre un mismo espacio probabilístico (Ω, σ, P), donde T denota un conjunto de índices. En función de las relaciones existentes entre las distintas variables que componen la sucesión, el proceso estocástico recibe diferentes nombres: cadena de Markov, movimiento Browniano, etc. En esta práctica nos centraremos en ver cómo podemos simular un número determinado de observaciones de un proceso estocástico utilizando RCommander.

1.1. Simulación de variables aleatorias

El menú *Distribuciones* de R-Commander contiene un amplio conjunto de distribuciones de probabilidad, agrupadas en discretas y continuas. Para cada una de ellas, se tienen las siguientes posibilidades:

Cuantiles. Es el menor valor c tal que $\Pr[X \leq c] \geq p$ (cola inferior o izquierda) o que $\Pr[X > c] \leq p$ (cola superior o derecha).

Probabilidades. En variables discretas, da los valores de la función de probabilidad, es decir, $\Pr[X = k]$ para cierto k. En variables continuas, da la probabilidad acumulada (véase la siguiente entrada).

Probabilidades acumuladas. Dado un valor k, calcula la probabilidad $\Pr[X \leq k]$ (cola izquierda) o $\Pr[X > k]$ (cola derecha).

Gráfica. Dibuja la gráfica de la función de densidad (para variables continuas) o de probabilidad (para discretas), o bien la función de distribución.

Muestra. Permite generar un nuevo conjunto de datos aleatorio indicando los parámetros de la distribución y las cantidades de filas y columnas deseadas. Dos advertencias:

a) En algunas versiones de R-Commander, en *Introducir el nombre del conjunto de datos* aparece por omisión un nombre con un espacio en blanco. En general, los objetos de R-Commander no permiten nombres con espacios, por lo que si el usuario no lo cambia se producirá un error.

b) Según la ventana de diálogo, las filas corresponden a muestras y las columnas a observaciones. El usuario puede estar interesado en hacerlo al revés, de forma que las filas se correspondan con los individuos de cada muestra (la interpretación habitual). En tal caso, simplemente debemos ignorar las etiquetas *muestras* y *observaciones* y pensar sólo en filas y columnas.

La opción **muestra** será la más interesante para este curso. Otra posibilidad a la hora de generar una muestra de observaciones de una determinada distribución consiste en escribir la instrucción correspondiente en la `Ventana de Instrucciones` de RCommander. La estructura es siempre la siguiente:

r prefijo de la distribución (número de observaciones, parámetros de la distribución).

Los prefijos de las distribuciones más usuales son los siguientes:

Nombre		Parámetros		
en teoría	en R	en teoría	en R	Comentario
binomial	`binom`	(n, p)	`size, prob`	n=`size`, p=`prob`
geométrica	`geom`	(p)	`prob`	número de fallos hasta el primer éxito
Poisson	`pois`	(λ)	`lambda`	
uniforme	`unif`	(a, b)	`a, b`	por defecto, $a = 0$ y $b = 1$
normal	`norm`	(μ, σ)	`mean, sd`	μ=`mean`, σ=`sd` por omisión, $\mu = 0$ y $\sigma = 1$
exponencial	`exp`	(λ)	`rate`	λ=`rate`; por defecto, $\lambda = 1$
χ^2 (ji-cuadrado)	`chisq`	(n)	`df`	grados de libertad
t de Student	`t`	(n)	`df`	grados de libertad

Así, la instrucción

```
rbinom(50,10,0.3)
```

genera una muestra aleatoria de 50 observaciones de una variables con distribución binomial B(10,0.3).

Por otro lado, en ocasiones será interesante simular muestras de una distribución discreta que no pertenezca a ninguna de las familias anteriores. A continuación vemos a través de un ejemplo cómo hacerlo:

```
x=c(-1,0,-2,5)

prob=c(0.3,0.2,0.4,0.1)

y=sample(x,100,prob,replace=TRUE)
```

En estas instrucciones, creamos el vector `x` con los distintos valores de la distribución, el vector `prob` con las probabilidades de los distintos valores, y generamos un vector con 100 observaciones aleatorias de dicha distribución. Para ello, usamos la instrucción `sample`, y añadimos replace=TRUE (literalmente, haría un muestreo con reemplazamiento de un modelo con

las probabilidades anteriores; el reemplazamiento garantiza que siempre se mantiene la misma distribución).

En esta práctica, vamos a considerar el caso más sencillo de lo que llamaremos proceso estocástico: aquél en que las variables $\{X_t\}_{t\in T}$ que lo componen son independientes e idénticamente distribuidas. En ese caso, si queremos generar una observación del proceso formado por las n primeras variables, $(X_1, \ldots, X_n)$, resulta equivalente generar n observaciones de la distribución común a todas estas variables, de acuerdo con las instrucciones resumidas en esta sección.

La justificación teórica del uso de la simulación para la estimación de probabilidades se encuentra en resultados como la *ley fuerte de los grandes números*, que garantiza que la media muestral de una variable converge casi seguro hacia la media poblacional cuando el tamaño muestral converge hacia infinito. Como consecuencia de esto, se deduce que, con probabilidad uno, la frecuencia de cualquier suceso en una muestra converge hacia la probabilidad teórica de dicho suceso según aumentamos el tamaño muestral.

1.2. Comprobación del Teorema Central del Límite

En esta sección, vamos a comprobar de manera experimental el llamado Teorema Central del Límite, el cual nos dice que, si $\{X_n\}_{n\geq 1}$ es una sucesión de variables independientes e idénticamente distribuidas, con $E(X_n) = \mu$ y $Var(X_n) = \sigma^2$, entonces

$$\frac{X_1 + \cdots + X_n}{n} \to N(\mu, \frac{\sigma}{\sqrt{n}}),$$

y análogamente

$$X_1 + \cdots + X_n \to N(n\mu, \sqrt{n}\sigma).$$

Normalmente se considera que la aproximación que proporciona este resultado es válida para valores de n iguales o superiores a 30.

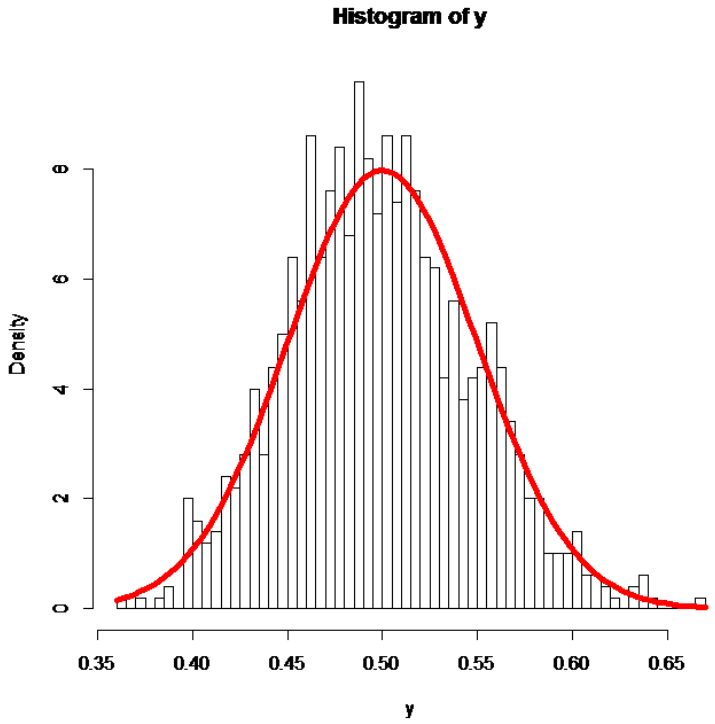
Vamos a comprobar el resultado cuando las variables X_i siguen una distribución exponencial con parámetro 2. Teniendo en cuenta lo visto en la sección anterior, es fácil obtener una observación de la variable $\frac{X_1+\cdots+X_{100}}{100}$:

```
x=rexp(100,2)
y=mean(x)
```

Para comprobar el resultado, vamos a representar una muestra de la variable $\frac{X_1+\cdots+X_{100}}{100}$, vamos a representar la misma en un histograma, y sobre ella superpondremos la función de densidad de la distribución normal a la que converge, que es una $N(0.5,0.05)$:

```
for (i in 1:1000)
{x=rexp(100,2)
y[i]=mean(x)}
hist(y,breaks=100,freq=FALSE)
curve(dnorm (x, 0.5, 0.05), col="red", lwd= 5, add=TRUE)
```

Obtenemos así lo siguiente:

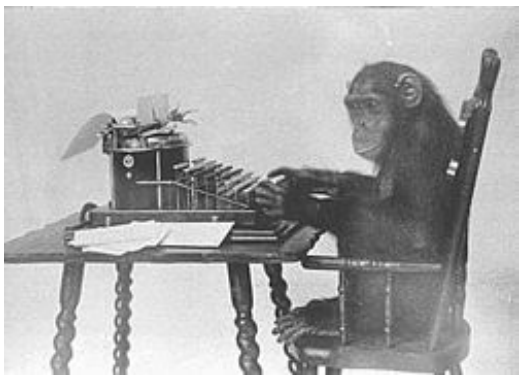


De manera análoga se podría comprobar el Teorema Central del Límite para distribuciones discretas, y para aproximar la suma de variables aleatorias independientes e idénticamente distribuidas.

1.3. Ejemplo: el teorema de los infinitos monos

El *teorema de los infinitos monos* es un ejemplo creado por Boole para explicar que todo suceso de probabilidad positiva (no importa cuán pequeña) sucede con probabilidad 1 cuando hacemos un número suficientemente grande de observaciones del experimento. En su formulación más conocida, dice lo siguiente:

> "Un mono pulsando teclas al azar sobre un teclado durante un periodo de tiempo infinito casi seguramente podrá escribir las obras de William Shakespeare."



La justificación matemática del resultado se basa en que, cuando la probabilidad p de un modelo binomial es mayor que 0, el valor $P(B(n,p) \neq 0) = 1 - (1-p)^n$ converge a 1 cuando n tiende a infinito. De hecho puede comprobarse de manera análoga que cualquier suceso de probabilidad positiva ocurre infinitas veces en infinitas repeticiones del experimento (es decir, el mono escribiría las obras de Shakespeare infinitas veces).

Veamos una comprobación experimental de este resultado en un caso más sencillo: vamos a comprobar que en sucesivos lanzamientos de una moneda, la probabilidad de que salgan en algún momento n caras consecutivas converge a 1 según aumentamos el número m de lanzamientos. Para ello, vamos a generar muestras de un proceso estocástico $\{X_1, \ldots, X_m\}$, donde X_i toma el valor 1 cuando el resultado del lanzamiento es una cara y 0 cuando es una cruz.

En cada muestra, comprobamos si existe alguna secuencia de n unos consecutivos:

```
racha=function(m,n){ #creamos una función que depende del número de
                     #lanzamientos m y la longitud de la racha n

x=rbinom(m,1,0.5) #generamos una muestra de m lanzamientos de una
                  #moneda equilibrada

z=rle(x)$length[rle(x)$values == 1] #almacenamos en un vector las
                                    #rachas de caras consecutivas

return(max(z)>=n)} #determinamos si hay alguna racha de longitud
                   #mayor o igual que n
```

Repetimos el proceso un número grande de veces (en este caso, 1000), y calculamos el porcentaje de veces en que se cumple la propiedad:

```
y=rep(0,1000)

prob=function(m,n){

for (i in 1:1000)

y[i]=racha(m,n)

return(mean(y))}
```

Se puede observar cómo, si fijamos un valor de n (por ejemplo, n=10) y aumentamos el valor de m, la probabilidad de que haya una secuencia de n caras consecutivas, que estimamos a partir del porcentaje de muestras en las que esto ocurre, va convergiendo hacia 1:

```
> prob(500,10)
[1] 0.22
```

```
> prob(5000,10)
[1] 0.91

> prob(50000,10)
[1] 1
```

Observación 1.1. *En las instrucciones anteriores, es importante el comando* `rle`*, que aplicado a un vector x nos devuelve una matriz de dos filas: en la primera se incluyen las longitudes de las siguientes rachas de números iguales consecutivos, mientras que en la segunda se incluye el valor del número asociado a la racha (en este ejemplo,* 0 *o* 1*).*

1.4. Ejercicios

1. (a) Genera $m = 100$ observaciones de una distribución binomial con parámetros $n = 10, p = 0{,}4$. Calcula la media de las observaciones simuladas.
   (b) Repite el apartado anterior $k = 100$ veces, y determina una estimación de la media de una distribución $B(10,0.4)$.

2. (a) Genera $m = 100$ observaciones del lanzamiento de un dado equilibrado, y determina si existe alguna racha de 4 unos consecutivos.
   (b) Repite el apartado anterior $k = 1000$ veces, y determina la proporción de muestras en las que aparecen al menos 4 unos consecutivos.
   (c) Repite los dos pasos anteriores para $m = 1000$, $m = 10000$ y $m = 100000$. Comenta los resultados.

3. Se considera que el estado de cuentas de una compañía de seguros al cabo de un año es una variable aleatoria X que puede viene determinada por la siguiente fórmula:

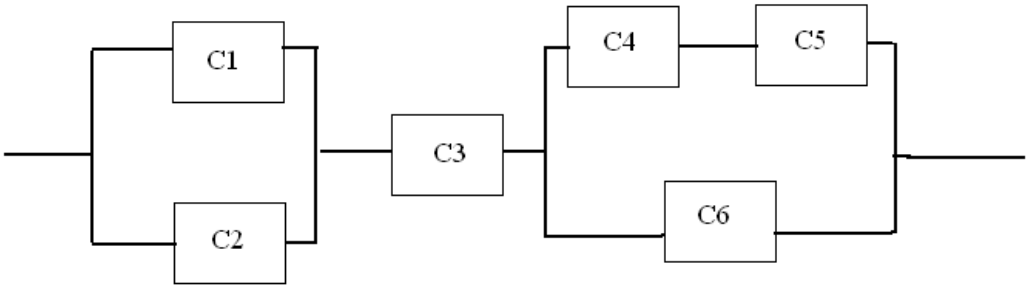
$$X = A \cdot (1 - B)^C,$$

   siendo:

   - A un coeficiente de volatidad, el cual se distribuye de acuerdo con una uniforme en el intervalo [0.5,1];
   - B es el estado de las reclamaciones por parte de los asegurados, el cual sigue una distribución exponencial con parámetro 6;
   - C es un coeficiente de deriva, el cual puede tomar el valor 1 con probabilidad 0.6 y el valor 0.5 con probabilidad 0.4.

   (a) Utiliza la simulación para estimar la media y la desviación típica de la variable X.
   (b) Usa las simulaciones anteriores y comprueba el teorema central del límite con la suma de 50 réplicas independientes de la variable X.

4. Un sistema está formado por componentes en serie y en paralelo, de acuerdo con la distribución de la siguiente figura:



Se supone que los bloques en serie funcionan cuando todas las componentes funcionan, mientras que cuando tenemos componentes en paralelo únicamente es necesario que funcione una de las componentes. Si el tiempo de vida de cada componente sigue una distribución exponencial con parámetro 5, estima mediante simulación el tiempo medio de vida del sistema.