Unidad 4. Distribuciones muestrales y grandes muestras

1. Sumas de variables aleatorias

Introducción general

La sumas de variables aleatorias y funciones que dependen de ellas son de especial importancia en la teoría y en las aplicaciones de la probabilidad y la estadística. Comenzamos enunciando tres problemas simples, tomados del libro de Baron (2014) con modificaciones.

Problema 1. *Un disco tiene 330 megabytes libres y queremos alojar 300 imágenes, de modo independiente una de otras. Si cada imagen tiene en promedio un tamaño de 1 megabyte con un desvío estándar de 0.5 megabytes, ¿será posible alojar la totalidad de las imágenes?*

Problema 2. *La gestión de memoria de un dispositivo informático en MS-DOS tuvo, en el pasado, una barrera de* 740 *KB. Cierta población de programas tiene un consumo medio de memoria de* 63.3 *KB y un desvío estándar de* 4 *KB. En una PC 10 personas dejan sus programas a un administrador para que los ejecute, ¿le dejarías al administrador tu programa para que lo corra en esa PC (sin correr el riesgo agotar la memoria) sabiendo que tu programa consume* 67 *KB? (asumir que la variable aleatoria "consumo" tiene distribución normal).*

Problema 3. *Un virus ataca una carpeta que contiene 200 archivos. Cada uno puede estar dañado con probabilidad 0.2, independientemente de los restantes archivos. ¿Cuál es la probabilidad de que haya menos de 50 archivos dañados?*

Abordaremos las primeras instancias de formalización. Dadas n variables aleatorias $X_1, \ldots, X_n$, interesa el problema de hallar la distribución de la variable

$$\sum_{i=1}^{n} X_i$$

.

Para algunas familias (modelos) de distribuciones la distribución de la suma pertenece a la misma familia, como se enuncia en la siguiente proposición.

Proposición 1. *Sean X e Y variables aleatorias independientes.*

a) Si $X \sim binomial(n, p)$ e $Y \sim binomial(m, p)$ entonces $X + Y \sim binomial(n + m, p)$.

b) Si $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ entonces $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

c) Si $X \sim gamma(\alpha_1, \lambda)$ e $Y \sim gamma(\alpha_2, \lambda)$ entonces $X + Y \sim gamma(\alpha_1 + \alpha_2, \lambda)$.

Demostración. Ver Apéndice.

Las siguientes afirmaciones son consecuencia de la proposición anterior y la demostración de su validez se propone como ejercicio.

Corolario 1. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes*

a) Si $X_i \sim$ *binomial*(m_i, p), $i = 1, \ldots, n$ *entonces* $\sum_{i=1}^{n} X_i \sim$ *binomial*(m, p) *con* $m = \sum_{i=1}^{n} m_i$

b) Si $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ *entonces* $\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2)$ *con* $\mu = \sum_{i=1}^{n} \mu_i$ *y* $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

c) Si $X_i \sim$ *gamma*(α_i, λ), *entonces* $\sum_{i=1}^{n} X_i \sim$ *gamma*(α, λ) *con* $\alpha = \sum_{i=1}^{n} \alpha_i$.

d) Si $X_i \sim$ *exponencial*(λ), *entonces* $\sum_{i=1}^{n} X_i \sim$ *gamma*(n, λ).

Se dice que las familias binomial y normal satisfacen la propiedad reproductiva; i.e., la suma de variables aleatorias pertenecientes a una familia de distribuciones pertenece a la familia.

Una pregunta que surge "naturalmente" es: ¿la distribución de cualquier suma de variables aleatorias independientes de una familia pertenece a la familia? La respuesta en general es no, tal como lo muestra el siguiente problema [1].

Problema 4. *Si* X *e* Y *son variables aleatorias independientes e idénticamente distribuidas con distribución uniforme* $[0, 1]$. *Entonces la suma* $Z = X + Y$ *tiene una distribución que viene dada por la densidad*

$$f_Z(z) = \begin{cases} z, & \text{if } 0 \leq z \leq 1 \\ 2 - z, & \text{if } 1 < z \leq 2 \\ 0, & \text{en caso contrario .} \end{cases}$$

Solución. En el Apéndice.

Volvamos ahora a los problemas planteados al inicio de esta unidad.

Solución al Problema 2.

El inciso b) del Corolario anterior nos permite arribar a una respuesta al *Problema 2* antes enunciado. La probabilidad que queremos computar es $P(\sum_{i=1}^{10} X_i + 67 > 740)$ donde X_i es el consumo del i-ésimo programa. Como $\sum_{i=1}^{10} X_i \sim N(10 \times 63.3, 10 \times 4^2) = N(633, 160)$ entonces

$$\sum_{i=1}^{10} X_i + 67 \sim N(700, 10 \times 4^2) = N(700, 160)$$

y así, calculando en R, obtenemos

```
 > 1-pnorm(740,mean=700,sd=sqrt(10*16))
[1] 0.0007827011
```

Es decir, que con probabilidad $0.9992 \approx 1$ la ejecución de los 11 programas no alcanzará el límite de gestión de memoria. □

[1] tomado de https://math.dartmouth.edu/ prob/prob/prob.pdf.

2. Distribución de la media y la varianza muestral

Si bien ya hemos utilizado la noción de idéntica distribución, la formalizamos en la siguiente definición.

Definición 1. *Dadas dos variables aleatorias X e Y se dicen idénticamente distribuidas (id) si tienen la misma función de distribución; i.e. $F_X(u) = F_Y(u), \forall u \in \mathbb{R}$. Utilizaremos la notación $X \stackrel{d}{=} Y$.*

A continuación introducimos los conceptos de *muestra aleatoria* y *media muestral*.

Definición 2. *Dada una variable aleatoria X diremos que la sucesión $X_1, \ldots, X_n$ es una muestra aleatoria de X si las variables $X_i, i = 1, \ldots, n$ son independientes e idénticamente distribuidas (iid) con $X_1 \stackrel{d}{=} X$. A la muestra aleatoria también se la denota con el vector $(X_1, \ldots, X_n)$.*

Dada una muestra aleatoria de X a la variable aleatoria

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

se le denomina *media muestral*.

La siguiente afirmación establece la distribución de la media muestral *bajo el supuesto de normalidad o Gaussianidad* de las variables aleatorias de la muestra aleatoria.

Proposición 2. ***Distribución de la media muestral***

Si $X_1, \ldots, X_n$ es una muestra aleatoria de una variable aleatoria $X \sim N(\mu, \sigma^2)$ entonces

$$\overline{X}_n \sim N(\mu, \sigma^2/n) \tag{1}$$

Demostración: ejercicio.

Observación 1. *Es importante notar que si la distribución de X_i de la muestra no es normal, la distribución exacta de la media no será normal (¿Por ejemplo?)*

En la figura 1 se muestran las densidades de $\overline{X}_n$ para los tamaños de muestra $n = 1, 10, 30, 50$ y $X \sim N(2, 1)$.

Otra función de una muestra aleatoria $X_1, \ldots, X_n$ que es de gran utilidad en estadística es la *varianza muestral* definida como

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2. \tag{2}$$

Para hallar la distribución de s_n^2 necesitamos introducir la siguiente familia de distribuciones.

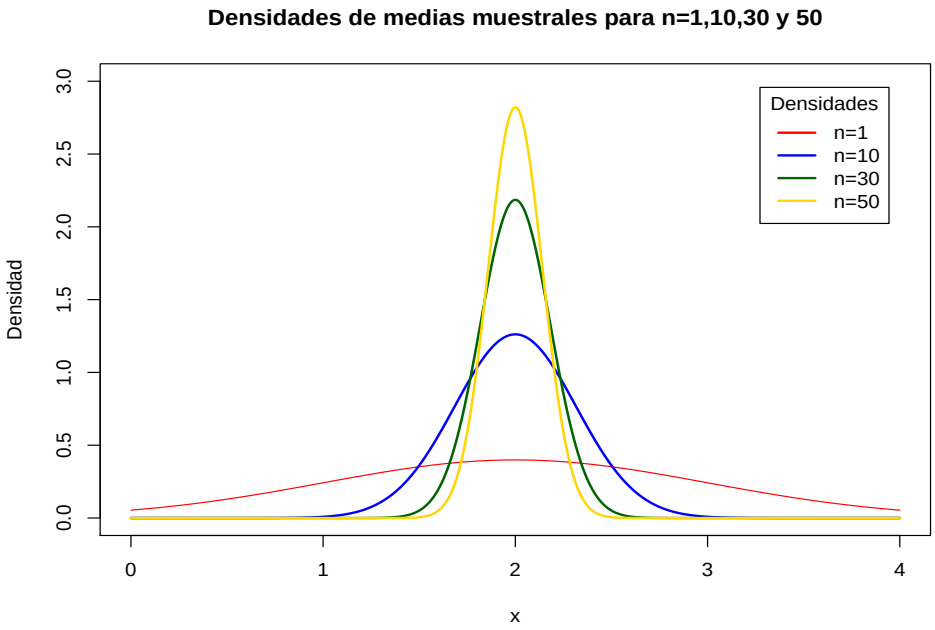

Figura 1: Densidades de $\overline{X}_n \sim N(\mu, \sigma^2/n)$ para $\mu = 2$, $\sigma^2 = 1$, $n = 1, 10, 30, 50$.

Una variable aleatoria X se dice que tiene distribución "chi-cuadrado con ν grados de libertad", $X \sim \chi^2_\nu$, si $X \sim \text{gamma}(\nu/2, 1/2)$ siendo ν un número real positivo.

Si recordamos la expresión de la función de densidad de la distribución gamma, la función de densidad de una variable aleatoria $X \sim \chi^2_\nu$ es

$$f(x) = \frac{1}{2^{v/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, \quad x > 0.$$

Proposición 3. ***Distribución de la varianza muestral***

Si $X_1, \ldots, X_n$ es una muestra aleatoria de una variable aleatoria $X \sim N(\mu, \sigma^2)$ entonces

$$\frac{(n-1)s_n^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{3}$$

La demostración de esta proposición cae fuera del alcance de los contenidos de este curso.

3. Teorema Central del Límite

Hasta ahora nos hemos preguntado sobre la distribución exacta de sumas de variables aleatorias. Hay un resultado relacionado con el anterior pero muy diferente en su significado y que afirma que la suma de variables aleatorias independientes es aproximadamente normal si el tamaño de la muestra es grande y sin asumir ninguna distribución para las variables aleatorias de la muestra. Este resultado se conoce como Teorema Central del Límite, tiene diferentes formulaciones y una extensa historia.

Teorema 1. ***Teorema Central del Límite*** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X con $E(X) = \mu$ y $\text{VAR}(X) = \sigma^2$. Sea $S_n = \sum_{i=1}^n X_i$ la suma de las variables aleatorias y Z_n la suma estandarizada*

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{\text{VAR}(S_n)}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$$

Entonces

$$\forall x \in I\!R : F_{Z_n}(x) \longrightarrow \Phi(x), \ \textit{si } n \longrightarrow \infty \tag{4}$$

donde Φ es la función de distribución de la $N(0,1)$; i.e. $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-\frac{1}{2}u^2} du$.

Una afirmación equivalente de la tesis del teorema es que;

la distribución de $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ es aproximadamente normal estándar si n es grande.

Solución al Problema 1. Si definimos X la variable "tamaño de la imagen", $S_n = \displaystyle\sum_{i=1}^{300} X_i$ con $X_i \overset{d}{=} X$, $\mu = 1$, $\sigma = 0.5$ entonces

$$\begin{aligned} P(\text{contar con espacio suficiente en el disco}) &= P(S_n \leq 330) \\ &= P\left(Z_n \leq \frac{330 - (300)1}{0.5\sqrt{300}}\right) \\ &\approxeq \Phi(3.46) = 0.9997 \end{aligned}$$

la cual es una probabilidad muy alta. □

Y ahora abordemos la resolución del problema restante.

Solución al Problema 3. Tengamos en cuenta que la variable S_n : "número de archivos dañados en un total de 200" archivos tiene una distribución binomial(n, p) con $n = 200$ y $p = 0.2$. Aplicando el Teorema Central del Límite, con

$$n\mu = np = 40 \text{ y } \sigma\sqrt{n} = \sqrt{p(1-p)}\sqrt{n} = 5.657$$

obtenemos $P(S_n < 50) \approxeq \Phi((50-40)/5.657) = 0.9614$. Es decir, la probabilidad de que haya menos de 50 archivos dañados es de 0.96. □

4. Apéndice

4.1. Demostraciones y soluciones complementarias

Para poder demostrar la Proposición 1 es necesario el siguiente teorema y su corolario.

Teorema 2.

a) Sean X e Y variables aleatorias discretas con función de probabilidad de masa conjunta $f(x, y)$. Entonces la función de probabilidad de masa de $X + Y$ tiene la siguiente expresión:

$$\forall z : \ f_{X+Y}(z) = P(X + Y = z) = \sum_x f(x, z - x)$$

b) Si X e Y son variables aleatorias continuas con función de densidad conjunta $f(x, y)$, entonces $X + Y$ tiene densidad dada por:

$$\forall z : \ f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z - x)dx$$

Demostración. Probaremos solo a). Teniendo en cuenta que

$$\forall z : \ \{X + Y = z\} = \bigcup_x (\{X = x\} \cap \{Y = z - x\})$$

de esta igualdad se deduce fácilmente la tesis. □

Dadas dos funciones de probabilidad de masa f_X y f_Y a la función tal que para cada z:

$$f_X * f_Y(z) = \sum_x f_X(x) f_Y(z - x) = \sum_y f_X(z - y) f_Y(y) \tag{5}$$

se le llama *convolución* de las funciones de probabilidad de masa. Y, cuando f_X y f_Y son funciones de densidad entonces la convolución de ellas es

$$f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)dx = \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y)dy. \tag{6}$$

El siguiente corolario es una consecuencia inmediata del Teorema 2.

Corolario 2. *Dadas dos variables aleatorias independientes, discretas o continuas, su suma tiene como función de probabilidad de masa o función de densidad a la convolución de las respectivas funciones de probabilidad de masa o de densidad de X e Y. Es decir,*

$$\forall z : f_{X+Y}(z) = f_X * f_Y(z).$$

Demostración de la Proposición 1. Demostraremos a), un caso particular de b) y c).

A partir de la expresión (5)) alcanzada para la convolución de dos funciones de probabilidad de masa obtenemos, para $z = 0, 1, \ldots, n+m$,

$$\begin{aligned} f_{X+Y}(z) &= f_X * f_Y(z) \\ &= \sum_y f_X(z-y) f_Y(y) \\ &= \sum_{y=0}^{z} \binom{n}{z-y} p^{z-y}(1-p)^{n-(z-y)} \binom{m}{y} p^y (1-p)^{m-y} \\ &= \binom{n+m}{z} p^z (1-p)^{n+m-z}, \end{aligned}$$

donde además hemos utilizado la igualdad

$$\sum_{y=0}^{z} \binom{n}{z-y}\binom{m}{y} = \binom{n+m}{z}.$$

Hemos probado la afirmación del inciso a). Demostremos el inciso b) para dos distribuciones normales estándares independientes con densidad común:

$$f_X(x) = f_Y(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Así, si Z denota la variable suma de X e Y queremos demostrar que $Z \sim \mathrm{N}(0, 2)$. Por la independencia y por (6) se tiene que:

$$\begin{aligned} f_Z(z) &= f_X * f_Y(z) \\ &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-(z-y)^2/2} e^{-y^2/2} dy \\ &= \frac{1}{2\pi} e^{-z^2/4} \int_{-\infty}^{+\infty} e^{-(y-z/2)^2} dy \\ &= \frac{1}{2\pi} e^{-z^2/4} \sqrt{\pi} \left[\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-(y-z/2)^2} dy \right] \end{aligned} \tag{7}$$

donde en la tercera igualdad hemos utilizado la identidad $(z-y)^2/2 + y^2/2 = z^2/4 + (y - z/2)^2$, válidad para cualesquiera z e y números reales. Por otro lado, haciendo el cambio de variable $u = 2y$

y reacomodando términos:

$$\begin{aligned}\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-(y-z/2)^2}dy &= \frac{1}{\sqrt{\pi}}\frac{1}{2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{u-z}{\sqrt{2}})^2}du \\ &= \frac{1}{\sqrt{\pi(\sqrt{2})^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{u-z}{\sqrt{2}})^2}du \\ &= 1\end{aligned}$$

ya que $h(u) = \frac{1}{\sqrt{\pi(\sqrt{2})^2}}e^{-\frac{1}{2}(\frac{u-z}{\sqrt{2}})^2}$ es la densidad de una distribución normal de media $\mu = z$ y varianza $\sigma^2 = 2$. Volviendo a (7), hemos probado que

$$f_Z(z) = \frac{1}{2\pi}e^{-z^2/4}\sqrt{\pi}$$

que es la densidad de una normal de media 0 y varianza 2.

Probemos ahora c). Queremos demostrar que la función de densidad de la suma, f_{X+Y}, satisface la siguiente igualdad

$$f_{X+Y}(z) = \frac{1}{\Gamma(\alpha_1+\alpha_2)}\lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\lambda z}.$$

Para ello tengamos en cuenta que

$$\begin{aligned}f_{X+Y}(z) &= f_X * f_Y(z) \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^z \lambda^{\alpha_1}(z-x)^{\alpha_1-1}\mathrm{e}^{-\lambda(z-x)}\lambda^{\alpha_2}x^{\alpha_2-1}\mathrm{e}^{-\lambda x}dx \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\lambda^{\alpha_1+\alpha_2}\mathrm{e}^{-\lambda z}\int_0^z (z-x)^{\alpha_1-1}x^{\alpha_2-1}dx \\ &= \lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\lambda z}\frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 (1-y)^{\alpha_1-1}y^{\alpha_2-1}dy \\ &= c\lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\lambda z}.\end{aligned}$$

Observar que para la cuarta igualdad hemos utilizado el cambio de de variable $y = x/z$ y hemos considerado que la constante c de la última expresión satisface

$$c = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\int_0^1 (1-y)^{\alpha_1-1}y^{\alpha_2-1}dy.$$

Dado que f_{X+Y} es una densidad, su integral satisface

$$1 = \int_0^\infty c\lambda^{\alpha_1+\alpha_2}z^{\alpha_1+\alpha_2-1}\mathrm{e}^{-\lambda z}dz \tag{8}$$

y, por la definición de la densidad de una distribución gamma$(\alpha_1+\alpha_2, \lambda)$, se tiene que

$$\int_0^\infty \lambda^{\alpha_1+\alpha_2} z^{\alpha_1+\alpha_2-1} \mathrm{e}^{-\lambda z} dz = \Gamma(\alpha_1+\alpha_2). \tag{9}$$

De (8) y (9), la constante c también satisface

$$c = \frac{1}{\Gamma(\alpha_1+\alpha_2)}.$$

En consecuencia

$$f_{X+Y}(z) = \frac{1}{\Gamma(\alpha_1+\alpha_2)} \lambda^{\alpha_1+\alpha_2} z^{\alpha_1+\alpha_2-1} \mathrm{e}^{-\lambda z}$$

como queríamos demostrar. □

Solución al Problema 4.

Por hipótesis

$$f_X(x) = f_Y(x) = \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{en caso contrario} \end{cases}$$

y la función de densidad de la suma, de acuerdo a (6), viene dada por $f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy$ con lo cual

$$f_Z(z) = \int_0^1 f_X(z-y) dy.$$

Ahora bien, el integrando se anula a menos que $0 \le z-y \le 1$ (i.e., $z-1 \le y \le z$) y en este caso es 1. Así, cuando $0 \le z \le 1$, tenemos

$$f_Z(z) = \int_0^z dy = z$$

mientras que si $1 < z \le 2$, se tiene

$$f_Z(z) = \int_{z-1}^1 dy = 2 - z$$

y si $z < 0$ o $z > 2$, tenemos $f_Z(z) = 0$. Resumiendo,

$$f_Z(z) = \begin{cases} z, & \text{if } 0 \le z \le 1 \\ 2-z, & \text{if } 1 < z \le 2 \\ 0, & \text{en caso contrario} \end{cases}$$

tal como se enunció en la tesis del problema. □

En este curso no abordaremos la demostración del Teorema Central del Límite, pero haremos una suerte de "comprobación empírica" utilizando *simulación*. En la próxima sección damos elementos introductorios a este problema.

4.2. Simulación

Las simulaciones por computadora refieren a la regeneración de un proceso a través de un código (programa) y a la observación de sus resultados.

Uno de los principales objetivos de la simulación es estimar aquellas cantidades cuyo cómputo directo es complicado, riesgoso, caro, o imposible de ser realizado.

Antes del surgimiento de la computadora se generaban números aleatorios a través de experimentos físicos, como por ejemplo el lanzamiento de una moneda, de un dado, bolilleros, etc. Estos métodos adolecen de varios defectos, entre ellos falta de repetibilidad y de velocidad.

En una computadora no se pueden generar números verdaderamente aleatorios pero sí pseudo aleatorios, que tienen la apariencia de ser aleatorios [2]

Un generador de números pseudo-aleatorios puede pensarse como una larga lista de números. Un usuario señala una semilla, que indica la posición a partir de la cual se comenzará a leer una secuencia de números de cierta longitud. Si la misma computadora utiliza la misma semilla entonces una próxima ejecución genera la misma secuencia.

Hay diferentes métodos para generar números pseudo-aleatorios, la mayoría se reduce a la generación de números aleatorios provenientes de una distribución uniforme y a partir de ello se generan números provenientes de otras distribuciones. Entre estos métodos se encuentran los generadores congruenciales y el de Mersenne-Twister que es el utilizado por R.

Los métodos Monte Carlo se refieren a las técnicas de cálculo basadas en una simulación por computadora (ver Baron, 2014, pág. 101)

Los métodos Monte Carlo son usualmente utilizados para estimar probabilidades, valores esperados y otras características asociadas a distribuciones.

4.2.1. Generación de números pseudo-aleatorios de una variable aleatoria discreta

La siguiente proposición es de simple demostración.

Proposición 4. *X una variable aleatoria con rango $R_X = \{x_0, x_1, \ldots\}$ y con función de probabilidad de masa $f(x_i) = p_i$, $i = 0, 1, 2, \ldots$. Sea U una variable aleatoria con distribución uniforme en $[0, 1]$ y definamos la variable aleatoria*

$$Y(\omega) = \begin{cases} x_0 & si\ 0 \leq U(\omega) < p_0 \\ x_1 & si\ p_0 \leq U(\omega) < p_0 + p_1 \\ & \ldots \\ x_i & si\ p_0 + \ldots + p_{i-1} \leq U(\omega) < p_0 + \ldots + p_i \\ & \ldots \end{cases} .$$

Entonces X e Y tienen la misma distribución.

Esta proposición nos permite formular el siguiente algoritmo de generación de números pseudo-

[2] Owen, J., Maillardet, R., Robinson, A. (2014). Introduction to Scientific Programming and Simulation Using R. CRC Press

aleatorios de X.

Algoritmo 4.1

1. Dividir el intervalo $[0,1]$ en los subintervalos

$$\begin{aligned} A_0 &= [0, p_0) \\ A_1 &= [p_0, p_0 + p_1) \\ A_2 &= [p_0 + p_1, p_0 + p_1 + p_2) \\ &\dots \\ A_i &= [p_0 + \dots + p_{i-1}, p_0 + \dots + p_i) \\ &\dots \end{aligned}$$

2. A partir de un generador de una $U \sim$ uniforme$[0,1]$ obtener una realización u de la variable aleatoria U.
3. Si i es tal que $u \in A_i$, entonces elegir x_i como el valor generado por X.

Analicemos a continuación algunos ejemplos.

Ejemplo 1.

Considerar una variable aleatoria X tal que $P(X = 1) = 0.4$, $P(X = 2) = 0.3$ y $P(X = 3) = 0.3$. Una observación de esta variable se puede generar con el siguiente programa

```
set.seed(1234);
P = c(0.4, 0.7, 1);
X = c(1, 2, 3);
counter = 1;
r = runif(1, min = 0, max = 1);
while(r > P[counter]){
 counter = counter + 1;
}
X[counter]
```

□

Ejemplo 2. *Generación de números pseudo-aleatorios de una variable aleatoria Bernoulli*

El siguiente programa permite generar una observación de una Bernoulli(p), con $p = 1/3$.

```
set.seed(1234);
p<- 1/3;
unif.sim  <- runif(n=1, min = 0, max = 1);
as.numeric(unif.sim < p)
```

□

Ejemplo 3. *Generación de números pseudo-aleatorios de una variable aleatoria binomial*(n, p)

El anterior programa nos permite implementar muy simplemente el algoritmo para generar una observación de, por ejemplo, una binomial(n, p) con $n = 4$ y $p = 1/2$[3]

```
set.seed(1234);
p<- 1/2;
unif.sim <- runif(n=4, min = 0, max = 1);
ber.sim <- as.numeric(unif.sim < p);
bin.sim <- sum(ber.sim)
```

Y con

```
set.seed(1234);
p<- 1/2;
bin.sim<- c()
for(i in 1:1000) {
 unif.sim <- runif(n=4, min = 0, max = 1);
 ber.sim <- as.numeric(unif.sim < p)
 bin.sim[i] <- sum(ber.sim)
}
```

podemos generar una muestra de 1000 observaciones de binomial$(12, 1/2)$, a la que podemos representar con un histograma

```
hist(bin.sim, nclass = 4, main="Histograma",
xlab = "Valores simulados", ylab = "Frecuencias", right=FALSE)
```

□

R permite generar valores de cuantiles, de distribuciones de probabilidad y de funciones de probabilidades de masa de un grupo grande de familias de distribución basándose en este tipo de algoritmos.

4.2.2. Generación de números pseudo-aleatorios de una variable aleatoria continua

Consideremos una variable aleatoria continua X. El siguiente teorema nos provee un insumo clave para la generación de números pseudo-aleatorios de X.

Teorema 3. *Sea X una variable aleatoria continua con función de distribución F estrictamente creciente, sea U una variable aleatoria con distribución uniforme en el $(0, 1)$ y definamos la variable aleatoria $Y = F^{-1}(U)$. Entonces $X \stackrel{d}{=} Y$.*

Demostración. Sólo probaremos el caso en que F es continua y estrictamente creciente. Denotemos

[3]H. Pishro-Nik, *Introduction to probability, statistics, and random processes*, disponible en https://www.probabilitycourse.com, Kappa Research LLC, 2014.

con F_Y la función de distribución de Y. Entonces, si y es cualquier número real

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F^{-1}(U) \leq y) \\ &= P(U \leq F(y)) \\ &= F(y) \end{aligned}$$

donde hemos utilizado que $F(y) \in (0,1)$ y U tiene como función de distribución a la identidad en el abierto $(0,1)$. □

Una definición más general de F^{-1} permite establecer la validez del teorema anterior para cualquier función de distribución F.

En el contexto del Teorema 3 damos el siguiente algoritmo, asumiendo que la inversa F^{-1} está bien definida.

Algoritmo 4.2

1. A partir de un generador de una variable aleatoria $U \sim \mathrm{U}(0,1)$ obtener una realización u de U.
2. Entonces $F^{-1}(u)$ es una realización de la variable aleatoria X.

Al igual que para las variables aleatorias discretas, con R es posible generar números pseudo-aleatorios de muchas familias de distribuciones de variables aleatorias continuas ya estudiadas.

Observación 2. *Los algoritmos 4.1 y 4.2 dicen esencialmente lo mismo. Consideremos una variable aleatoria discreta X con función de distribución F y definamos*

$$\forall y \in [0,1] : F^{-1}(y) =: min\{x : F(x) \geq y\}.$$

Entonces, el método de utilizar la "inversa" de la F es el mismo en variables aleatorias discretas y continuas.

4.3. Complementos de resultados asintóticos.

Los siguientes dos teoremas son de gran importancia y los utilizaremos, en conjunto con el Teorema Central del Límite, en simulación.

Teorema 4. *Ley fuerte de los grandes números.*

Sean $X_1, \ldots, X_n$ variables aleatorias iid con $E(|X_1|) < \infty$ y sea $\mu = E(X_1)$.

Entonces

$$P(\omega : \overline{X}_n(\omega) \to \mu \text{ cuando } n \to \infty) = 1.$$

Para resumir, decimos que la media muestral $\overline{X}_n$ converge a μ casi seguramente o con probabilidad 1.

Consideremos $X_1, \ldots, X_n$ una sucesión de variables aleatorias independientes con función de distritución común $F(x)$. A

$$F_n^*(x, \omega) = \frac{1}{n} \# \{i : X_i(\omega) \leq x\}$$

se le llama la función de distribución empírica de la muestra $X_1, \ldots, X_n$.

Por simplicidad, omitimos el argumento ω y escribimos $F_n^*(x)$.

El siguiente teorema nos dice que F_n^* y F están cercanas, si el tamaño de la muestra es grande.

Teorema 5. *Teorema de Glivenko-Cantelli.*

Sean $X_1, \ldots, X_n$ variables aleatorias iid con con función de distritución común $F(x)$.

Entonces

$$P(\omega : \forall x, F_n^*(x, \omega) \to F^*(x) \text{ cuando, } n \to \infty) = 1.$$

Análogamente, para resumir decimos que $F_n^*(x)$ converge a $F(x)$, para todo x casi seguramente o con probabilidad 1.