1. Distribución binomial en RStudio Cloud

Enrique Lessa

2024-08-09

Introducción

Vamos a usar R, un lenguaje de programación de uso libre para estadística y un amplio conjunto de aplicaciones relacionadas (análisis de datos, gráficas, presentación de resultados). R permite cargar diferentes bibliotecas o paquetes que aportan distintas funciones (ej. hacer gráficos, distintos test estadísticos, etc). Como interfase usamos Rstudio. R y Rstudio pueden usarse en casi cualquier computadora, así como en la nube (posit.cloud).

El sitio de Posit.cloud y muchos otros tienen guías introductorias, incluyendo la interfase y sus distintos sectores. No vamos a reiterar esos materiales. Solamente noten que este mismo panel corresponde a un archivo *.Rmd* (en formato Rmarkdown). Notas mínimas sobre el entorno RStudio: - Abajo está la consola, donde se pueden escribir comandos directamente (se pueden copiar y pegar de este archivo). Ese espacio se puede limpiar con el comando Ctrl + L.

- Abajo a la derecha está un panel que incluye "Packages" (para activar o desactivar paquetes de R), "Files" (los archivos), y "Help", extremadamente útil para ver los formatos de los distintos comandos.
- En el panel "Environment" (ambiente) se pueden ver los valores definidos a lo largo de la sesión. Los mismos se pueden limpiar (con el ícono de la escoba) para volver a empezar el proceso.

Para producir los archivos de salida usamos *Rmarkdown*. En Rmarkdown, intercalamos texto común (estos párrafos son un ejemplo), en los que podemos también incluir ecuaciones en Latex, como se verá más abajo, con bloques ("chunks") de código (las líneas de código en R propiamente dichas). Si producimos una salida (en nuestro caso en html), esta intercalará texto, código y resultados del código.

El primer bloque de código activa "knitr" y "markdown", necesarios para las salidas en html, así como "lattice", un paquete gráfico.

```
library("knitr")
knitr::opts_chunk$set(echo = TRUE)
library("markdown")
library("lattice")
```

Prueba: cliquear en el ícono de *Knit* arriba de esta ventana para ver cómo se produce una salida de html para este mismo archivo.

Distribución binomial

Consideramos un evento con una probabilidad conocida p de ocurrir en una prueba. Por convención, calificamos a dicho evento como un "éxito" o resultado "favorable" en la prueba, y a complemento como "fracaso". El evento de referencia (éxito) está asociado a

la probabilidad p de observarlo. Naturalmente, el complemento del resultado favorable tiene una probabilidad complementaria de $1 - p$.

Por ejemplo, la probabilidad de sacar un valor en particular (4, por ejemplo) al tirar un dado es $p = 1/6$. En este ejemplo el complemento, la probabilidad de sacar algún otro número, es $1 - p = 5/6$. La distribución binomial nos permite calcular la probabilidad de obtener un resultado cualquiera, para un número n de pruebas independientes (esto significa que el resultado de cada prueba no depende de los de las restantes). Respetando la condición de independencia, no importa si las pruebas son simultáneas o sucesivas.

La probabilidad de obtener dos 4 en una tirada de dos dados (o, de nuevo, dos tiradas sucesivas) es $p*p = p^2$. La probabilidad de obtener un 4 en el primer dado y cualquier otro número en el segundo es $p\left(1 - p\right)$. El mismo resultado global (un 4 y un número diferente) puede obtenerse de dos maneras, cada una con la probabilidad antes mencionada, de modo que ese resultado global tiene probabilidad $2p\left(1 - p\right)$.

Generalizando, si llamamos $P(i)$ a la probabilidad de observar i veces nuestro evento de referencia (en el ejemplo, un 4), que ocurre en cada prueba con probabilidad p, en una muestra de tamaño n (en el ejemplo, una tirada de dos dados), tenemos que:

$$P(i) = \binom{n}{i} p^i \left(1 - p\right)^{n-i},$$

$$i = 0, 1, \ldots, n$$

Notamos que:

p^i es la probabilidad de obtener i veces el evento de referencia.

$\left(1 - p\right)^{n-i}$ es la probabilidad de obtener $n - i$ veces el evento complementario (es decir,de no obtener el de referencia).

$\binom{n}{i} = \frac{n!}{i!(n-i)!}$ es el llamado coeficiente binomial, que expresa el número de formas de obtener la combinación de eventos anteriores. Por una explicación concisa del coeficiente binomial, ver https://en.wikipedia.org/wiki/Binomial_coefficient.

Ejemplo En una tirada de 3 dados, calculamos la probabilidad de obtener un único 4 (sin importar el valor de los restantes dados) del modo siguiente:

- probabilidad de obtener 4 en un dado: $p^i = \left(1/6\right)^1 = 1/6$
- probabilidad de obtener cualquier otro valor en dos dados: $\left(1 - p\right)^{n-i} = \left(5/6\right)^2 = 0.69444$
- Hay 3 formas de obtener un 4 y otros dos valores cualesquiera, ya que el 4 puede ser el primero, el segundo o el tercer dado. Usando el coeficiente binomial:

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} = \frac{3!}{1!(2-i)!} = 3$$

Para una muestra de tamaño 3, es más fácil enumerar los casos que usar el coeficiente binomial. Obviamente, la situación es la inversa con tamaños muestrales mayores.

Distribución binomial y frecuencias alélicas

Observamos que esta sección es simplemente una aplicación de lo anterior. Lo único nuevo es verificar precisamente eso, del modo que sigue:

Si conocemos la frecuencia real del alelo A $p = f(A)$ en la población, podemos aplicar la binomial para calcular la probabilidad de observar i alelos de tipo A en una muestra de tamaño n. Como es lógico, dicha probabilidad depende de la frecuencia del alelo y del tamaño de la muestra. En concreto, toma exactamente la forma de la binomial, es decir:

$$P(i) = \binom{n}{i} p^i (1-p)^{n-i}, i = 0, 1, \ldots, n$$

Notamos que:

p^i es la probabilidad de muestrear i veces el alelo A.

$(1-p)^{n-i}$ es la probabilidad de muestrear $n - i$ veces otro alelo (por ejemplo, para un sistema de 2 alelos, el alelo a).

$\binom{n}{i}$ es el llamado coeficiente binomial, que corresponde al número de formas de obtener los resultados anteriores.

Aplicación

Este es un ejercicio deductivo: dada una frecuencia alélica conocida (p) y un tamaño de muestra, deducimos con qué probabilidad podemos obtener todos los resultados posibles, desde $i = 0$ hasta $i = n$.

Para explorar estas ideas, - Consideramos una muestra de $n = 10$ alelos tomada de una población en la cual la frecuencia de A es $p = 0,3$. Usando la función *dbinom*, calculamos la probabilidad de observar 0, 1, ... 10 copias de A en la muestra. Como estamos evaluando todos los resultados posibles, verificamos que la suma de los $P(i)$ es 1.

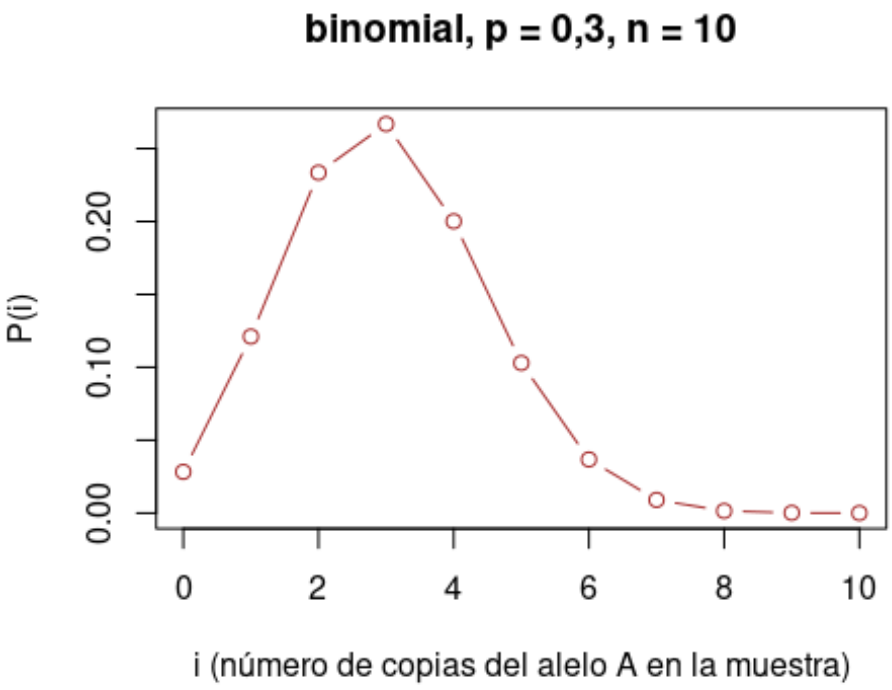
```
n = 10  # tamaño de la muestra
pr = 0.3 # valor de p (frecuencia del alelo A en la población)

#A. Probabilidad de observar 0, 1, ... n copias del alelo A en una
muestra de tamaño n, dado que la frecuencia del
#   alelo en la población es pr

  dist = dbinom(c(0:n), n, pr) # c(0:n) es un vector de números de 0
a n, el rango de resultados cuyas probabilidades queremos calcular);
si queremos calcular la probabilidad de un único resultado, entonces
usamos un valor entero x (entre 0 y n) en lugar de dicho vector

  print(dist)
```

```
##  [1] 0.0282475249 0.1210608210 0.2334744405 0.2668279320
0.2001209490
##  [6] 0.1029193452 0.0367569090 0.0090016920 0.0014467005
0.0001377810
## [11] 0.0000059049
```

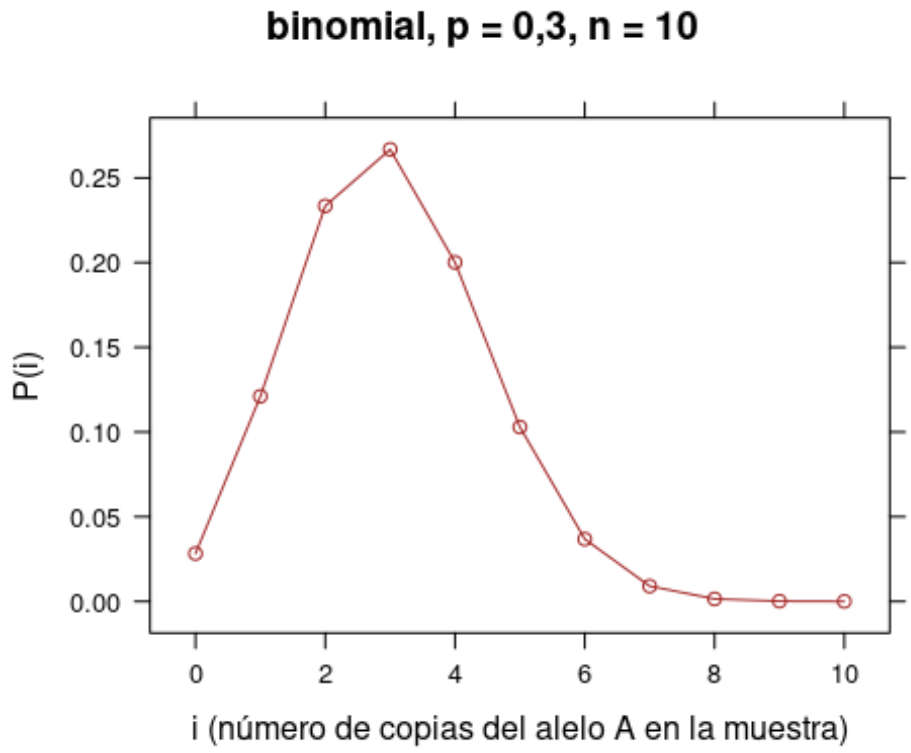
```
  sum(dist) # como calculamos las probabilidades para todos los
resultados posibles, verificamos que sumen 1
```

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

```
#B. Graficamos los valores obtenidos
 #B1. con gráficas básicas
   plot(c(0:10), dist, type = "b",
        main = "binomial, p = 0,3, n = 10",
        xlab = "i (número de copias del alelo A en la muestra)",
        ylab = "P(i)",
        col = "brown",
        )
```




```
 #B2. con lattice
 xyplot (as.vector(dist) ~ (c(0:10)),
         type = "b",
       main = "binomial, p = 0,3, n = 10",
       xlab = "i (número de copias del alelo A en la muestra)",
       ylab = "P(i)",
       col = "brown"
       )
```




```
  #xyplot(y ~ x)
```

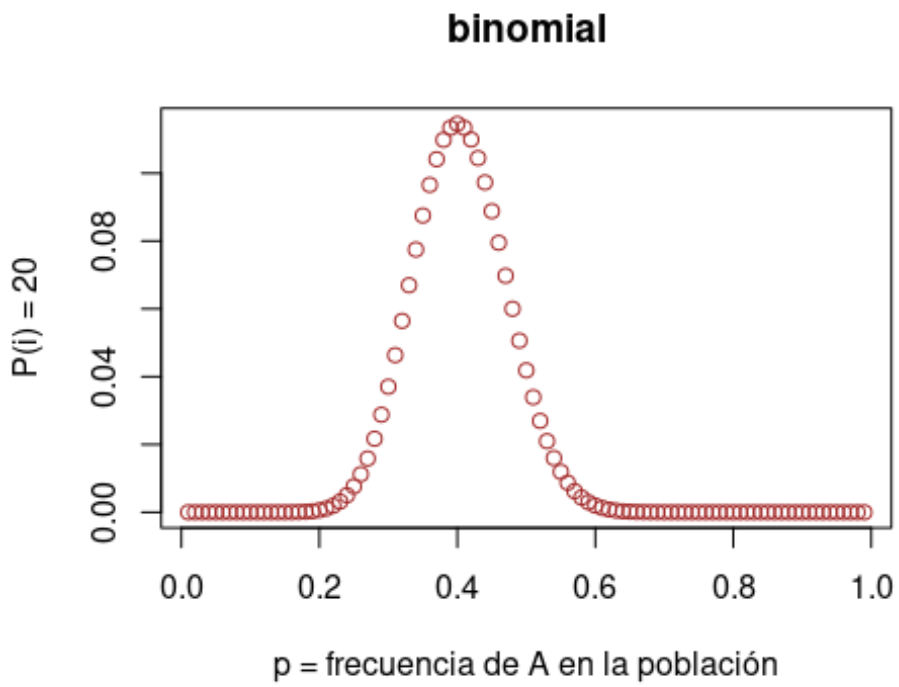
Probabilidad de observar un resultado en función de distintos valores de p

- Creamos un vector de frecuencias de interés; en el ejemplo que sigue, definimos 99 valores uniformemente distribuidos en el intervalo 0.01-0.99. [Nota: podríamos incluir las frecuencias extremas de p=0 y p=1, aunque sabemos a priori que en los dos casos la probabilidad de observar 20 alelos de tipo A en una muestra de 50 es cero.]

- Usando la función *dbinom*, calculamos la probabilidad de observar exactamente $n = 20$ alelos A en una muestra total de N= 50 alelos (nuestras observaciones) si la frecuencia real del alelo de interés A fuese, sucesivamente, $p = 0.01, 0.02, \ldots 0.99$, es decir nuestro vector de valores de interés.

```
# Observaciones:
  n  = 50  # tamaño de la muestra (número total de alelos en la
muestra)
  i = 20   # número observado de alelos de clase A


# Probabilidad de observar i alelos en una muestra de n alelos, en
función de la frecuencia del alelo A en la población (en el rango
0.01-0.99):
  pvector = c(1:99)/100 # vector de frecuencias de interés (rango
0.01-0.99)
  probs = dbinom(i, n, pvector)
  plot(pvector, probs, type = "b",
       main = "binomial",
       xlab = "p = frecuencia de A en la población",
```

```
      ylab = "P(i) = 20",
      col = "brown")
```



Ejercicio 1.

En una población, la frecuencia del alelo A es $p = 1/4$.

a) Completar la siguiente tabla, que enumera algunos de los resultados posibles para una muestra al azar de 4 alelos de dicha población:

i = número de alelos A	número de alelos a	Arreglos	Prob. de cada arreglo	$P(i)$
0	4	aaaa	0.3164	0.3164
1	3	Aaaa	0.1055	
		aAaa		
		aaAa		
		aaaA		
2	2	...		

b) Usando la binomial, ¿cuál es la probabilidad de observar 1 alelo A? Usar este resultado para verificar el obtenido por enumeración en "a)".

Muestras al azar

Hasta ahora, usamos *dbinom* ("distribución binomial") en ejercicios deductivos, calculando las probabilidades de observar ciertos resultados (combinaciones de eventos) tomando *p* como un parámetro conocido. Entendemos que cada resultado tiene una cierta probabilidad de ocurrir, y podemos calcularla usando la binomial. Ahora vamos a muestrear al azar uno o más resultados, usando las mismas reglas. Un concepto importante es que el resultado de una muestra particular es una *realización* al azar de un proceso estocástico. En otras palabras, dos muestras obtenidas bajo las mismas reglas pueden dar idénticos o distintos resultados (como lo sabe cualquiera que haya observado un juego de azar).

- Usando la función *rbinom* ("random binomial"), simulamos *x1* réplicas de observaciones, cada una de ellas consistente en muestras de *n1* alelos, asumiendo una frecuencia alélica en la población conocida (*p1*).

```
# A. Probando unas pocas realizaciones de la función binomial:
  # valores a definir para las pruebas:
    x1 = 40 # número de réplicas
    n1 = 10 # tamaño de la muestra en cada réplica
    p1  = 0.3 # frecuencia del alelo en la población
  # pruebas
    Pruebas = rbinom(x1, n1, p1)
  # resultado: cantidad de alelos A en cada una de las pruebas
    print(Pruebas)
```

```
##  [1] 1 1 3 1 1 2 2 4 4 4 4 3 4 2 3 3 4 3 2 5 1 2 4 3 1 6 4 4 3 1
3 3 1 5 5 2 4 4
## [39] 5 6
```

```
  # resultado: cantidad media de alelos A en las pruebas
    print(mean(Pruebas))
```

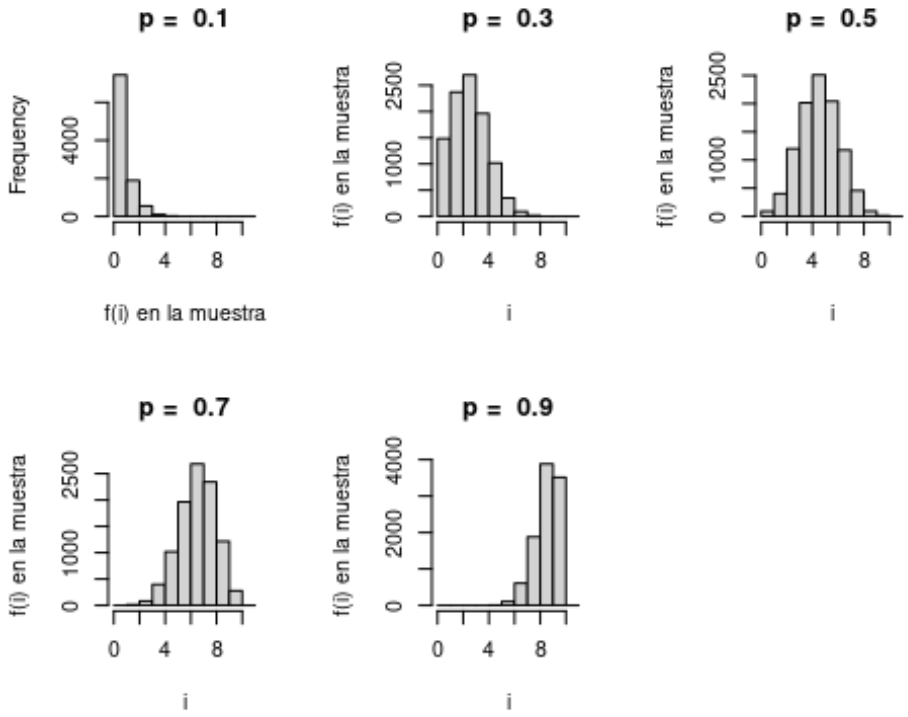
```
## [1] 3.075
```

Observamos, como sugiere la intuición, que los valores observados suelen estar "en el entorno" de la frecuencia real, pero que cada realización resulta en un número de alelos de tipo A particular, entre 0 y *n1*. [Nota: verificar más arriba que ya calculamos la probabilidad de obtener cada uno de estos valores]. También observamos que la cantidad media de alelos por prueba se aproxima a la frecuencia *p*1 en la población.

A continuación repetimos el ejercicio anterior, pero acumulando un gran número de réplicas (*x2*) para luego graficar los valores obtenidos.

```
 op = par(mfrow=c(2,3),pty="s")


# B. Acumulando un número grande de réplicas réplicas al azar de la
binomial:
  # valores a definir para las pruebas:
    x2 = 10000 # variar x2 para explorar el comportamiento de la
función
    n2 = 10 # tamaño de la muestra en cada prueba
    p2  = 0.1 # frecuencia del alelo en la población
```

```
  # B1. Réplicas:
    Replicas = rbinom(x2, n2, p2)
    hist(Replicas, 100, main = paste("p = ", p2),
           xlab = "f(i) en la muestra",
           breaks = as.vector(c(0:11)),
           include.lowest = TRUE,
         )
  # B2. Réplicas como las anteriores, pero incluidas en un "loop"
para variar la frecuencia p2 y observar las consecuencias
    for(j in rep(1:4)) {
      p2 = p2+0.2
      Replicas = rbinom(x2, n2, p2)
      hist(Replicas, 100,
           xlab = "i",
           ylab = "f(i) en la muestra",
           breaks = as.vector(c(0:11)),
           include.lowest = TRUE,
           main = paste("p = ", p2)
      )
     }
```



Ejercicio 2

a) Usando la binomial, obtener 20 muestras al azar, cada una con 10 alelos, de una población en la que la frecuencia del alelo A es $p = 0.4$. Calcular el promedio de dichas muestras y compararlo con la probabilidad de observar 1 copia del alelo A en la muestra.

b) Repetir el ejercicio y comparar los resultados con los obtenidos en el primer ensayo. Explicar similitudes y diferencias.

Comentarios finales

Para un gen con dos alelos, la binomial nos permite saber cuál es la probabilidad de observar un número determinado de cada uno de ellos dado que las frecuencias de los alelos son conocidas. Como $p + q = 1$, nos basta el valor de p. Notar además que la binomial es un caso particular de la distribución multinomial, que puede aplicarse a casos con más de dos clases de alelos.

Ejercicios resueltos

Ejercicio 1.

En una población, la frecuencia del alelo A es $p = 1/4$.

a) Completar la siguiente tabla, que enumera algunos de los resultados posibles para una muestra al azar de 4 alelos de dicha población:

i = número de alelos A	número de alelos a	Arreglos	Prob. de cada arreglo	$P(i)$
0	4	aaaa	0.3164	0.3164
1	3	Aaaa	0.1055	X2
		aAaa	X1	
		aaAa	X1	
		aaaA	X1	
2	2	…		

Cada uno de los arreglos con 1 copia de A y 3 copias de a tiene la misma probabilidad, de modo que todos los casilleros marcados como X1 se pueden completar copiando el valor dado para Aaaa, o sea $X1 = 0.1055$*. El casillero marcado como X2 debería completarse una única vez para el conjunto de arreglos de este tipo, ya que el encabezado de la tabla pide anotar* $P(i)$*. El resultado se obtiene sumando las probabilidades de cada uno de los arreglos con 1 A y 3 a. Obtenemos así* $X2 = 0.422$

Para el caso de 2 alelos de cada tipo, el número de arreglos es mayor. Se puede obtener usando la binomial (ver la parte "b)". Por enumeración, podemos razonarlo en términos de las posiciones de A entre los 4 alelos, puesto que los otros dos lugares son necesariamente ocupados por el alelo a: así, A puede ocupar los lugares 1 y 2, 1 y 3, 1 y 4, 2 y 3, 2 y 4, o 3 y 4, para un total de 6 arreglos. No vamos a hacerlo, pero se puede calcular la probabilidad de cada arreglo y la total para el caso de dos alelos de cada clase siguiendo el razonamiento del párrafo anterior

b) Usando la binomial, ¿cuál es la probabilidad de observar 1 alelo A? Usar este resultado para verificar el obtenido por enumeración en "a)".

La función dbinom, que nos permite obtener el resultado, tiene la siguiente forma:

$dbinom(x, size, prob)$,

donde "x" corresponde a nuestro "i", el número de copias del alelo A, "size" es el tamaño de la muestra, y prob es, en nuestro caso, la probabilidad de obtener un alelo A, es decir la frecuencia de A en la población.

En la consola, usamos

$$dbinom(1, 4, 0.25)$$

El resultado es 0.421875*, que verifica lo obtenido en "a)" (las diferencias se deben únicamente a que el valor* 0.1055 *está redondeado).*

Ejercicio 2

a) Usando la binomial, obtener 20 muestras al azar, cada una con 10 alelos, de una población en la que la frecuencia del alelo A es $p = 0.4$. Calcular el promedio de dichas muestras y compararlo con la probabilidad de observar 1 copia del alelo A en la muestra.

Hay que usar la función $rbinom$*, siendo la "r" de "random" la palabra clave que señala que buscamos obtener muestras al azar. La forma general de la función es:*

$rbinom(n, size, prob)$*, donde n es el número de muestras, size es el tamaño de cada una, y prob es, en nuestro caso, la frecuencia del alelo A. De modo que podemos usar la función en la consola como sigue:* $rbinom(20, 10, 0.40)$

Para calcular el promedio vamos a modificar el procedimiento ligeramente. Guardar las muestras en un vector (que llamamos "muestras", línea 1) y luego visualizamos tanto las muestras (línea 2) como su promedio (línea 3):

$$muestras = rbinom(20, 10, 0.40)$$

$$muestras$$

$$mean(muestras)$$

b) Repetir el ejercicio y comparar los resultados con los obtenidos en el primer ensayo. Explicar similitudes y diferencias.