Distribución binomial en RStudio Cloud

Enrique Lessa

8/30/2019

Introducción

Vamos a usar R, un paquete de programas de uso libre para estadística y un amplio conjunto de aplicaciones relacionadas (análisis de datos, gráficas, presentación de resultados). Como interfase usamos RStudio, ampliamente utilizado, y para producir los archivos de salida usamos Rmarkdown. En Rmarkdown, intercalamos texto común (incluyendo ecuaciones en Latex) 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.

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

Distribución binomial

Consideramos un evento con una probabilidad conocida de ocurrir en una prueba. Por ejemplo, la probabilidad de sacar un 4 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*(1-p)*. *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* **2*p***(1-p)*.

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:

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

Notamos que:

p^i es la probabilidad de obtener *i* veces el evento de referencia. $(1-p)^{n-i}$ es la probabilidad de obtener *n-i* veces el evento complementario (es decir,de no obtener el de referencia). $\binom{n}{i}$ es el número de formas de obtener la combinación de eventos anteriores.

Distribución binomial y frecuencias alélicas

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

Detalles técnicos sobre el coeficiente binomial [Sección optativa en desarrollo]

$\binom{n}{i} = n!/[i!(n-i)!$ Por una explicación concisa, ver https://en.wikipedia.org/wiki/Binomial_coefficient (https://en.wikipedia.org/wiki/Binomial_coefficient).

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.

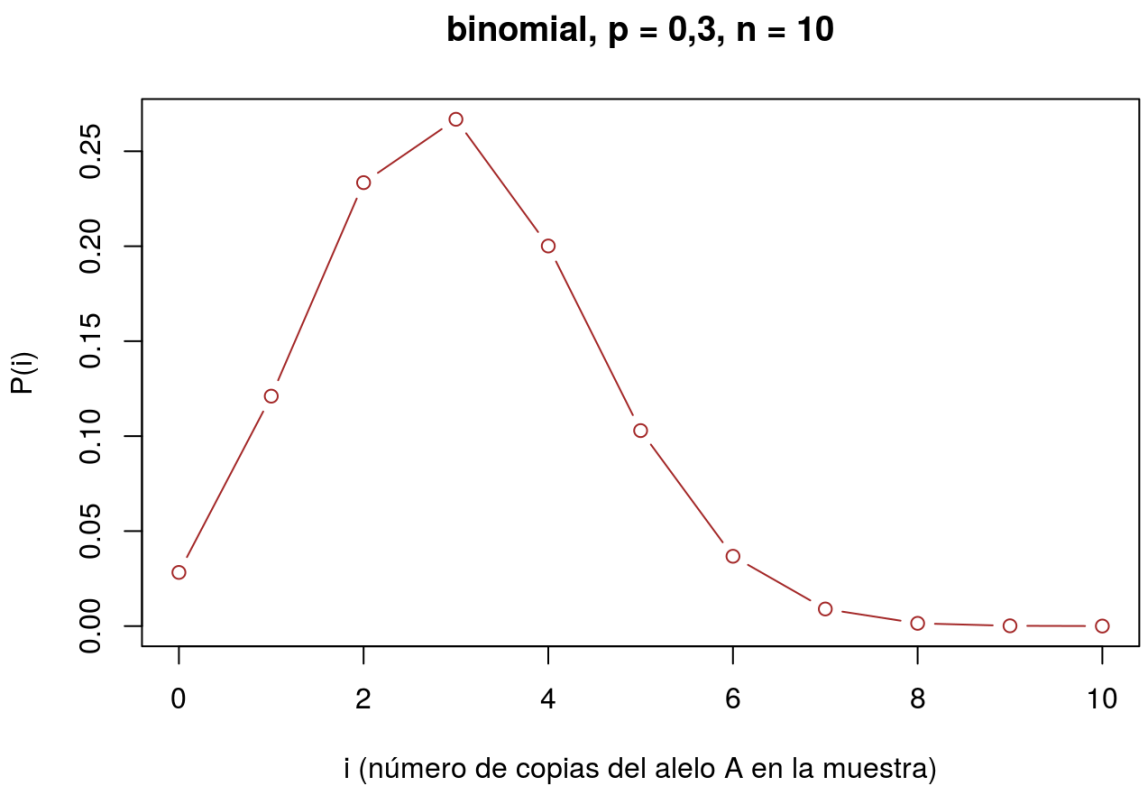
```
n = 10  # tamaño de la muestra
pr = 0.3 # valor de p (frecuencia del alelo A en la población)

#A. Probabilidad de observar 0, 1, ... n copias del alelo A en una muestra de n=10, dado que la frecuenc
ia del
#   alelo en la población es pr

  dist = dbinom(c(0:10), n, pr) # c(0:10 es el rango de resultados cuyas probabilidades queremos calcula
r)
  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 para todos los resultados posibles, verificamos que la suma da 1
```

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

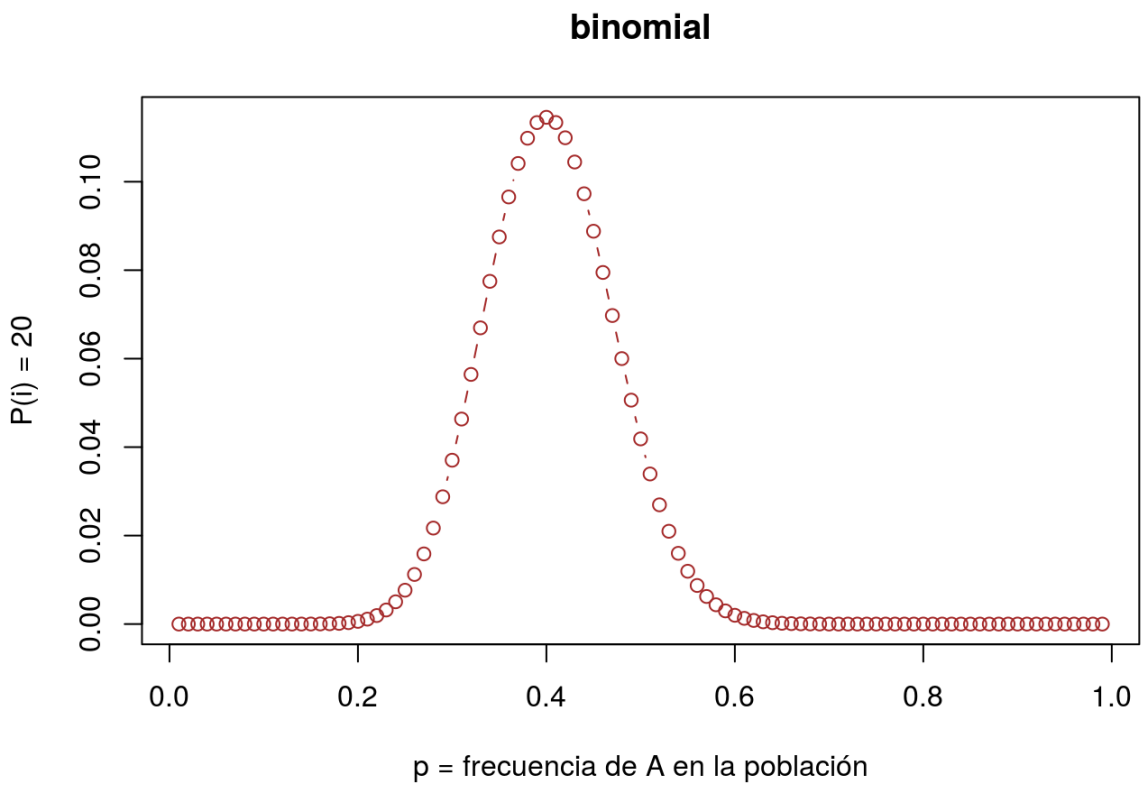
```
#B. Graficamos los valores obtenidos
  plot(c(0:10), dist, type = "b",
       main = "binomial, p = 0,3, n = 10",
       xlab = "i (número de copias del alelo A en la muestra)",
       ylab = "P(i)",
       col = "brown",
       )
```



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, … 0.99*, es decir nuestro vector de valores de interés.

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

# 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")
```



Muestras al azar

Hasta ahora, usamos *dbinom* en ejercicios predictivos, 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*, simulamos *x1* réplicas de observaciones consistentes en *n1* alelos, asumiendo una frecuencia alélica en la población conocida (*p1*).

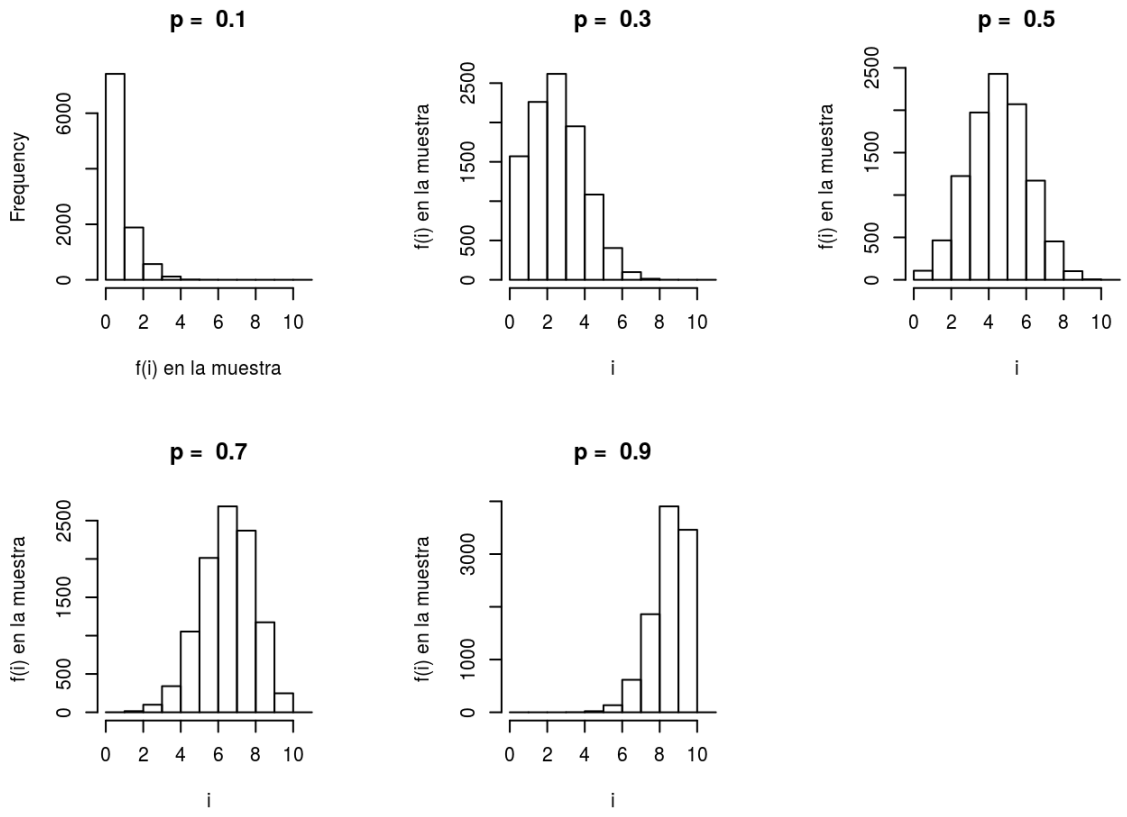
```
# A. Probando unas pocas realizaciones de la función binomial:
  # valores a definir para las pruebas:
    x1 = 40 # número de réplicas
    n1 = 10 # tamaño de la muestra en cada réplica
    p1  = 0.3 # frecuencia del alelo en la población
  # pruebas
    Pruebas = rbinom(x1, n1, p1)
    print(Pruebas)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

```
##  [1] 5 2 3 0 3 4 2 4 1 3 3 3 4 3 3 4 2 2 3 1 3 3 4 2 4 2 2 2 1 3 4 4 7 6 2
## [36] 4 1 5 2 3
```

Observamos lo que sugiere la intuición, es decir que los valores observados están “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]

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 observa
r 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)
      )
     }
```



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 no tiene la restricción de solamente dos clases de alelos.