Aproximaciones de la distribución binomial. y teorema central del límite

Pablo L. De Nápoli

Departamento de Matemática
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Probabilidades y Estadística para Matemática
Segundo cuatrimestre de 2021

Parte I

La aproximación de Poisson a la distribución binomial

Bernoulli, siempre Bernoulli

Una vez más volvemos a los ensayos de Bernoulli, donde considerábamos un experimento aleatorio con dos resultados que convencionalmente se llaman

- éxito (1) con probabilidad p.
- fracaso(0) con probabilidad $q = 1 - p$.

Introdujimos las variables aleatorias de Bernoulli X_i dadas por

$$X_i = \begin{cases} 1 & \text{si la i-ésima realización del experimento es un éxito} \\ 0 & \text{si la i-ésima realización del experimento es un fracaso} \end{cases}$$

Las X_i son variables aleatorias discretas. También lo es el número de éxitos en n ensayos.

$$S_n = X_1 + X_2 + \ldots + X_n$$

Como ya vimos S_n tiene distribución binomial; $S_n \sim Bi(n, p)$

$$P\{S_n = k\} = b(k, n, p) = \binom{n}{k} p^k q^{n-k}, \quad 0 \leq k \leq n$$

La distribución de Poisson: definición

Esto significa que bajo estas hipótesis, la distribución binomial puede aproximarse por la distribución de Poisson. Recordamos su definición:

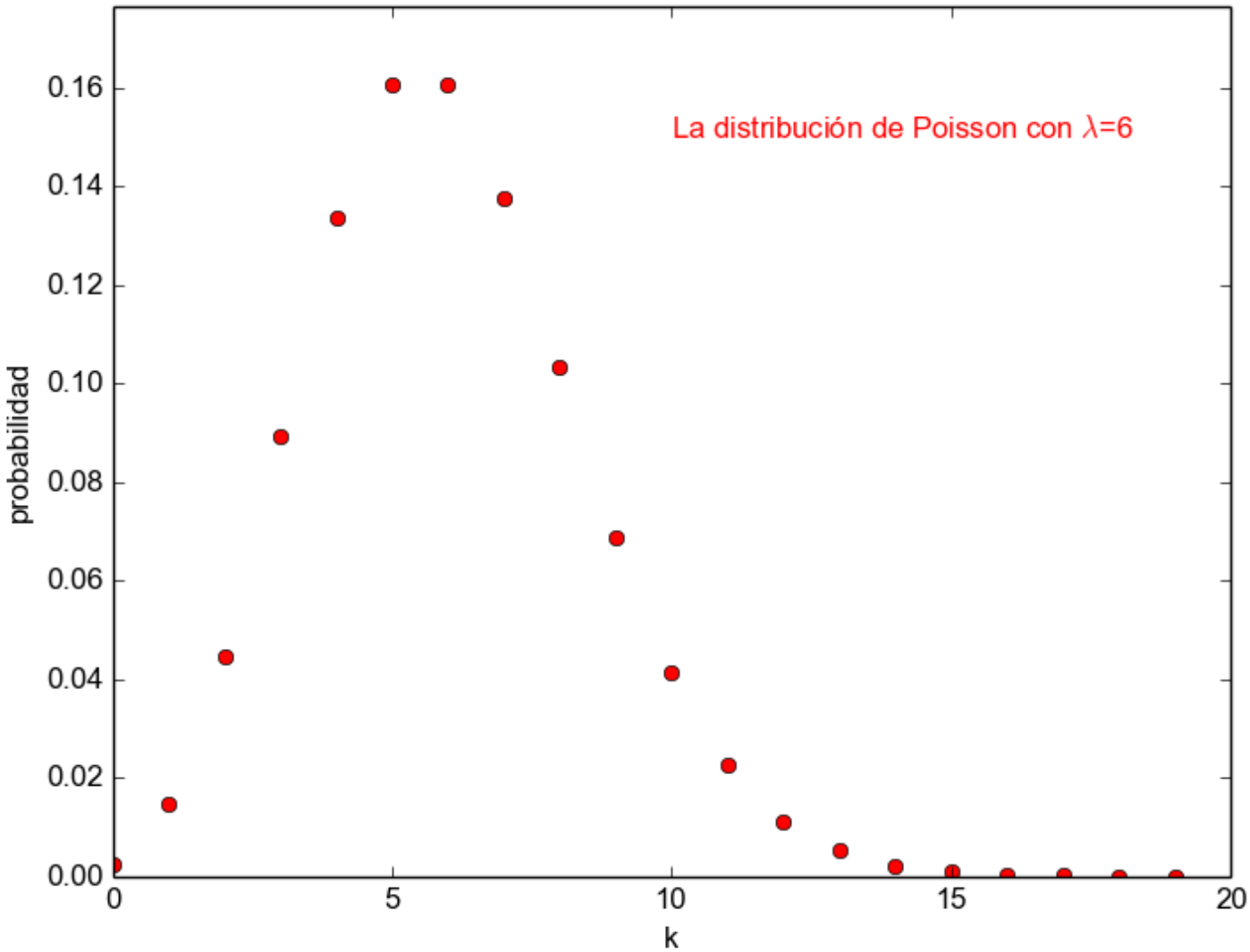
Definición

Sea $X : \Omega \to \mathbb{N}_0$ una variable aleatoria entera. Diremos que X tiene distribución de Poisson de parámetro $\lambda > 0$, si

$$P\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Notación: $X \sim \mathcal{P}(\lambda)$.

La distribución de Poisson: gráfico



La aproximación de Poisson a la distribución binomial

La aproximación de Poisson es una aproximación de la distribución binomial bajo las siguientes hipótesis:

- k es pequeño comparado con n.
- p es también pequeño.
- pero $\lambda = np$ es moderado.

Bajo estas hipótesis, la distribución binomial puede aproximarse por la distribución de Poisson:

Aproximación de Poisson

$$b(k, n, p) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

Parte II

Aproximación de la distribución binomial por la normal: El teorema de De Moivre-Laplace

La Fórmula de Stirling para aproximar $n!$ para n grande

Teorema (Fórmula de Stirling)

$$n! \sim \sqrt{2\pi}\, n^{n+1/2} e^{-n}$$

Idea de la prueba:

$$\log(n!) = \sum_{k=1}^{n} logk \approx \int_2^n \log\ x\ dx = n\log(n) - n - 2\log(2) + 2$$

Un ejemplo de la aproximación de Stirling

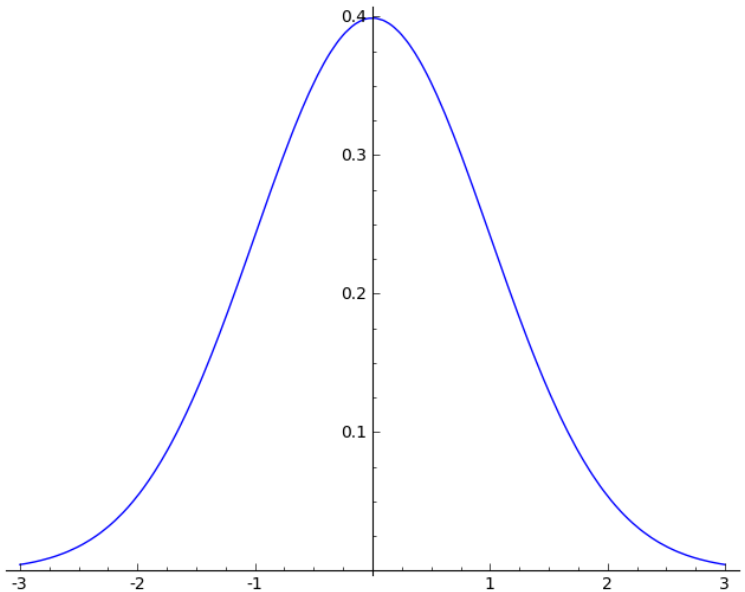
```
n= 15
factorial de n= 1307674368000
aproximación de Stirling= 1300430722199.4658
Error relativo = -0.0055393345451993 9
```

Recordamos una definición: la distribución normal

Decimos que N tiene **distribución normal**, y lo notaremos $N \sim N(\mu, \sigma^2)$, si su función de densidad de probabilidad viene dada por:

$$f_N(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/(2\sigma^2)}$$

donde μ, σ son dos parámetros reales con $\sigma > 0$. El caso $\mu = 0$, $\sigma = 1$, es decir $N(0, 1)$, se conoce como **distribución normal estándar**.

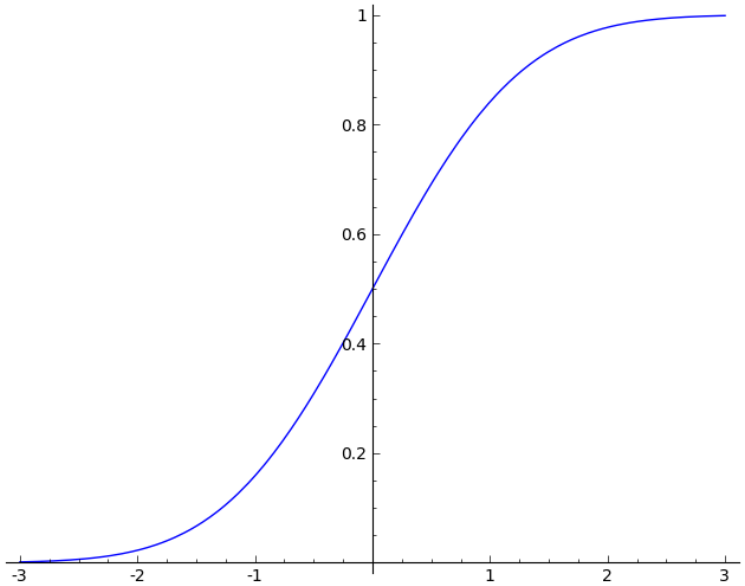


Reordamos que $E[N] = \mu$ y $\mathrm{Var}(N) = \sigma^2$.

Recordamos una definición: la distribución normal

Si $N \sim N(0,1)$, la función de distribución acumulada de N será la función:

$$F_N(x) = P\{N \leq x\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/(2\sigma^2)}\, dt \tag{1}$$



Teorema local de De Moivre-Laplace

Utilizando la fórmula de Stirling, podemos obtener otra aproximación de la distribución binomial, por medio de la distribución normal.

Teorema (Teorema local de De Moivre-Laplace)

$$b(k,n,p) = \frac{1}{\sqrt{2\pi npq}}\, e^{-x_k^2/2}\,(1+\beta_{n,k})$$

donde

$$x_k = \frac{k-np}{\sqrt{npq}}$$

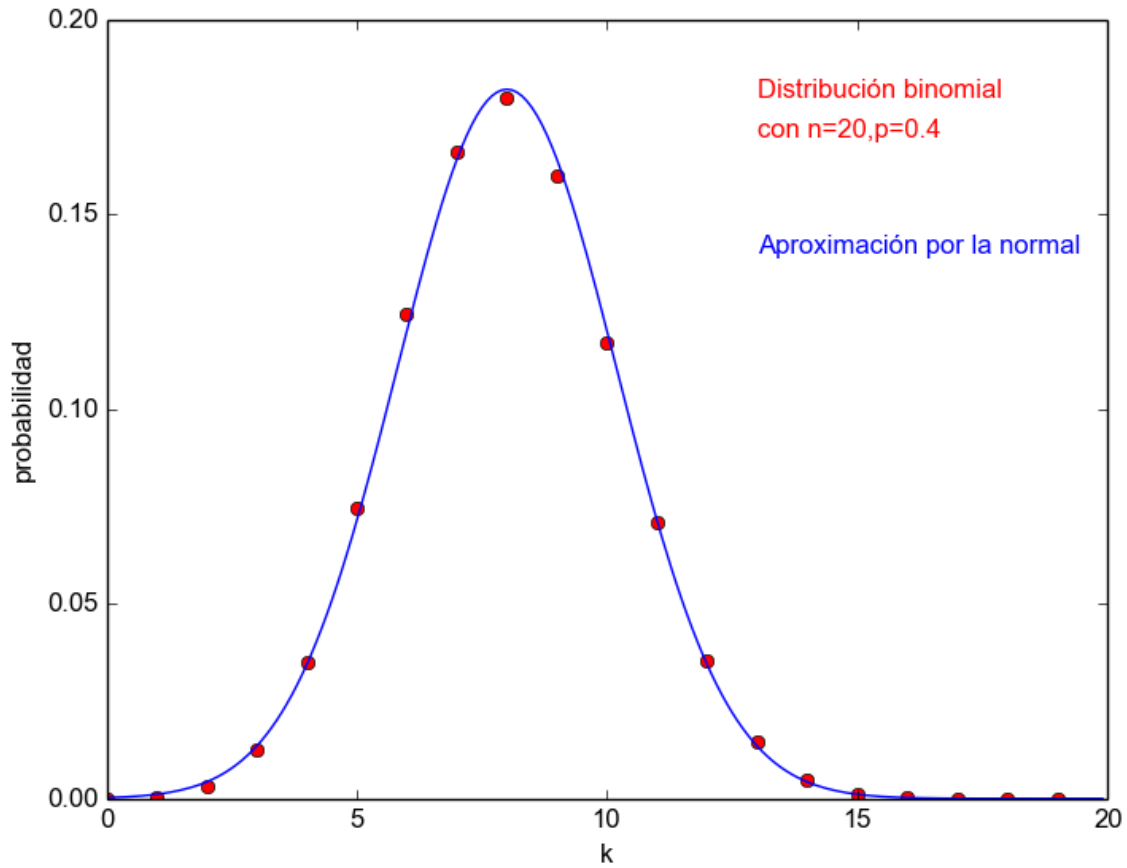
y para $M \geq 0$,

$$\max_{|x_k|\leq M} |\beta_{n,k}| \to 0 \text{ cuando } n \to \infty \tag{2}$$

Recordamos que $E[S_n] = np$ y $\text{Var}(S_n) = npq$. Entonces esto significa que podemos aproximar la distribución binomal por una distribución normal con la misma esperanza y varianza.

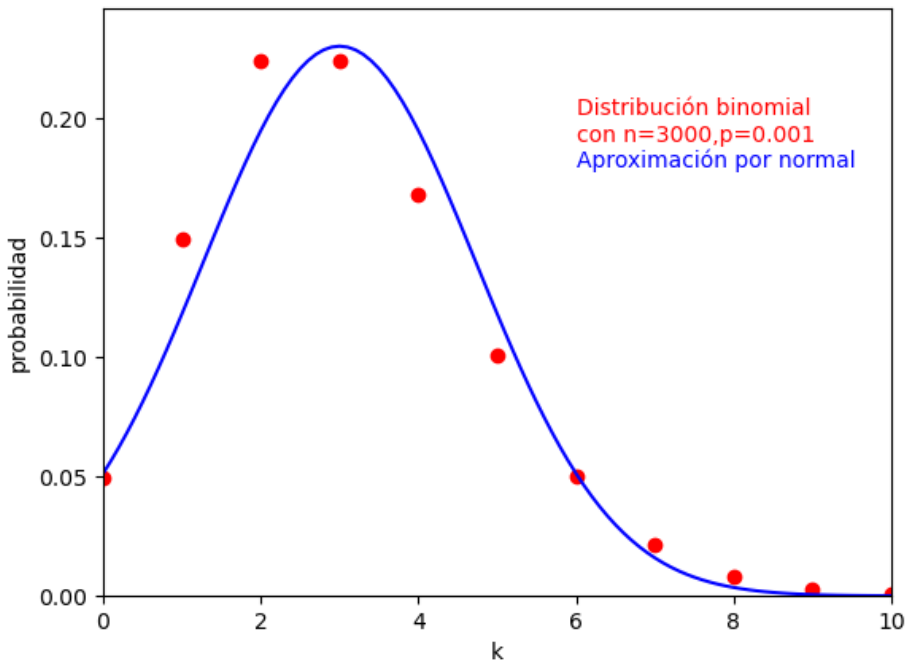
Ilustración gráfica

En el teorema Teorema local de De Moivre-Laplace estamos aproximando la distribución puntual de S_n por la función de densidad de la distribución normal (con parámetros adecuados)



Una Observación importante

La cota del error en el teorema significa la aproximación dada por el es buena en el centro de la distribución binomial, pero no en las colas de la misma. Por ejemplo, si n es grande y p es muy pequeño con np moderado. En esta situación es mejor la aproximación por la distribución de Poisson. Por simetría, tampoco es buena si p está muy cerca de 1.



¡Hagámoslo en la computadora!

Programita en Python, usando SciPy

```
from numpy import sqrt
from scipy.stats import binom, poisson, norm

def aproximar_normal(k, n, p):
        exacto = binom(n, p).pmf(k)
        mu = n * p
        q = 1 - p
        sigma = sqrt(n * p * q)
        aprox_poisson = poisson(mu).pmf(k)
        aprox_normal = norm(mu, sigma).pdf(k)

        err_rel_poisson = (aprox_poisson - exacto) / exacto
        err_rel_normal = (aprox_normal - exacto) / exacto
```

Pueden bajar mis programitas de
`https://pdenapo@bitbucket.org/pdenapo/programitas-proba.git`.

Ejemplo 1

Enunciado

Tiramos una moneda equilibrada 10 veces.¿Cuál es la probabilidad de obtener exactamente 3 caras?

Salida del programa

```
$python3 aproximaciones_normal.py  3 10 0.5
k= 3
n= 10
p= 0.5
Valor exacto =    0.11718750000000014
Aprox. de Poisson=   0.1403738958142805
Error rel. Poisson=   0.19785724428185886
Aprox. normal=   0.1133716522449791 4
Error relativo Normal= -0.03256190084284584
```

Vemos que la aproximación normal resulta buena (se equivoca en un $3,2\,\%$ mientras que la de Poisson tiene un error relativo de casi un $20\,\%$)

Ejemplo 2

Enunciado

Una vacuna tiene una probabilidad de producir una reacción alérgica de $0,001$. Se la aplica a 2000 individuos. ¿Cuál es la probabilidad de que exactamente 3 indivuos tengan dicha reacción alérgica?.

Salida del programa

```
$ python3 aproximaciones_normal.py  3 2000 0.001
k= 3
n= 2000
p= 0.001
Valor exacto =   0.1805373280324423 8
Aprox. de Poisson=  0.1804470443154835 6
Error rel. Poisson=  -0.0005000833785609052
Aprox. normal=  0.2197505754927657 2
Error relativo Normal=  0.2172029900280609
```

Vemos que ahora la aproximación de Poisson es muy buena, mientras que la aproximación normal tiene un error del 21 %.

Probabilidad de que la cantidad de éxitos en un cierto intervalo

En muchas situaciones no estamos interesados en las probabilidades individuales $P\{S_n = k\}$, sino en la probabilidad de que S_n caiga en un cierto intervalo $I = (a, b]$

$$P\{a < S_n \leq b\} = \sum_{a<k\leq b} b(k, n, p) = F_{S_n}(b) - F_{S_n}(a)$$

Esto podríamos aproximarlo reemplazando la función de distribución de S_n por la de la distribución de Poisson o la de la normal (según en que régimen estemos).

Normalización de una variable aleatoria

Sea X una variable aleatoria con segundo momento finito. Entonces la variable reescaldada (o "normalizada")

$$X^* = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

satisface que $E(X^*) = 0$ y $\text{Var}(X^*) = 1$.

Sea S_n el número de éxitos en n ensayos de Bernoulli con probabilidad $p \in (0, 1)$. Sabemos que S_n tiene distribución binomial que y que $E[S_n] = np$, $\text{Var}(S_n) = npq$. Consideramos entonces la variable normalizada:

$$S_n^* = \frac{S_n - np}{\sqrt{npq}}$$

Nuestro objetivo es estudiar el límite de la distribución de S_n^* cuando $n \to +\infty$ distribución de S_n^* cuando $n \to +\infty$:

El Teorema de De Moivre-Laplace

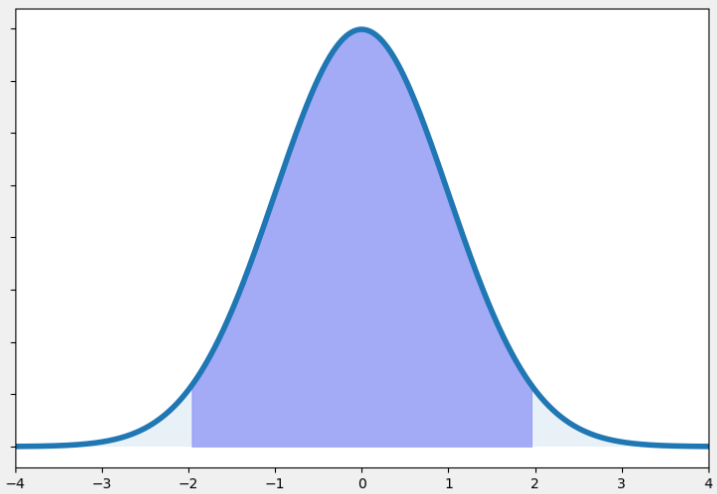
El siguiente teorema afirma que la distribución límite de la variable normalizada S_n^* está dada por la integral definida de $f_N(x)$:

Teorema (De Moivre-Laplace)

$$P\{a < S_n^* \leq b\} \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx = F_N(b) - F_N(a)$$

cuando $n \to +\infty$.

donde F_N era la función de distribución de una variable aleatoria con distribución normal estándar.



Idea de la Demostración

La idea básica de la demostración es la siguiente:

$$P_n(a,b) = P\{a < S_n^* \leq b\} = \sum_{a<x_k\leq b} b(k,n,p)$$

ya que si S_n^* toma el valor x_k, entonces S_n toma el valor k.
Los puntos x_k están cada vez más próximos a medida que $n \to +\infty$, ya que

$$\Delta x_k = x_{k+1} - x_k = \frac{1}{\sqrt{npq}} \to 0$$

y por el teorema anterior $b(k,n,p) \approx f_N(x_k)(x_{k+1} - x_k)$ entonces,

$$P_n(a,b) = P\{a < S_n^* \leq b\} \approx \sum_{a<x_k\leq b} f_N(x_k)\,\Delta x_k$$

y esta es una suma de Riemann para la integral $\int_a^b f_N(x)\,dx$. Por lo tanto, conforme $n \to +\infty$, es razonable que podamos aproximar $P_n(a,b)$ por dicha integral. La demostración consiste en una formalización de esta idea.

¡Hagámoslo en la computadora!

Programita en Python, usando SciPy

```
def aproximar_normal_acumulada(k, n, p):
        exacto = binom(n, p).cdf(k)
        mu = n * p

        q=1-p
        sigma = sqrt(n * p * q)
        aprox_poisson = poisson(mu).cdf(k)
        aprox_normal = norm(mu, sigma).cdf(k)

        err_rel_poisson = (aprox_poisson - exacto) / exacto
        err_rel_normal = (aprox_normal - exacto) / exacto
```

Ejemplo 1

Enunciado

Tiramos una moneda equilibrada 30 veces.¿Cómo poderíamos acotar la probabilidad de obtener a lo sumo 3 caras?

Salida del programa

```
python3 aproximaciones_normal_acumulada.py  3 10 0.5
k= 3
n= 30
p= 0.5
Valor exacto =              4.215165972709657e-06
Aprox. de Poisson=          0.0002113785034667617 4
Error rel. Poisson=         49.14713651497813
Aprox. normal=              5.885669548807499e-06
Error relativo Normal=      0.3963079003088421 6
```

Ejemplo 2

Enunciado

Una vacuna tiene una probabilidad de producir una reacción alérgica de $0,001$. Se la aplica a 2000 individuos. ¿Cuál es la probabilidad de que a lo sumo 3 indivuos tengan dicha reacción alérgica?.

Salida del programa

```
$ python3 aproximaciones_normal_acumulada.py  3 2000 0.001
k= 3
n= 2000
p= 0.001
Valor exacto =   0.8572137667929646
Aprox. de Poisson=  0.857123460498547
Error rel. Poisson=  -0.00010534862821369182
Aprox. normal=  0.7603598554307263
Error relativo Normal=  -0.1129868827522349
```

Vemos que ahora la aproximación de Poisson es muy buena, mientras que la aproximación normal tiene un error del 11 %.

Parte III

El teorema central del límite

Planteo del problema

Antes vimos que el número de éxitos en n ensayos de Bernoulli puede representarse como

$$S_n = X_1 + X_2 + \ldots + X_n$$

donde las (X_k) son variables aleatorias independientes e idénticamente distribuidas, con distribución Be(p), o sea:

$$X_i = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } q = 1 - p \end{cases}$$

El teorema de De Moivre-Laplace nos dice que la distribución de las variables estandarizadas S_n^* coverge a una distribución nomal cuando $n \to +\infty$.

Esto nos lleva a la pregunta, ¿qué pasa si las X_i fueran independientes e idénticamente distribuidas pero con otra distribución?

Enunciado del Teorema Central del Límite

Teorema (Teorema Central del Límite, versión sencilla)

Sea $(X_k)_{k\in\mathbb{N}} : \Omega \to \mathbb{R}$ *una sucesión de variables aleatorias independientes e idénticamente distribuidas con* $\mu = E[X_k]$ *y* $0 < \sigma^2 = Var(X_k)$ *ambas finitas (como suponemos que las* X_k *tienen todas la misma distribución, tendrán todas la misma esperanza y varianza). Notemos:*

$$S_n = X_1 + X_2 + \ldots + X_n$$

$$S_n^* = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n}\,\sigma}$$

Entonces fijados $a < b$,

$$P\{a < S_n^* \leq b\} \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx = F_N(b) - F_N(a)$$

cuando $n \to +\infty$.

Comentarios sobre el Teorema Central del Límite

- Un enunciado equivalente es que las funciones de distribución acumulada de una variable aleatoria N con distribución normal $N(0,1)$:

$$F_{S_n^*}(x) \to F_N(x) \text{ para todo } x$$

Esto se suele expresar diciendo que las variables aleatorias S_n^* convergen en distribución a N.

- Notemos que el teorema también puede formularse en términos de la media muestral

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

pues

$$S_n^* = \frac{\overline{X}_n - \mu}{\sigma} \cdot \sqrt{n}$$

- No vamos a demostrar este teorema en este curso (Una demostración aparece en el apunte de A. Bianco-E. Martínez.)

¡Hagámoslo en la computadora!

Programita en Python, usando SciPy - Definimos la distribución

```
import numpy as np
import scipy.stats as stats

# simulamos una moneda

distribucion = stats.bernoulli(0.5)

# simulamos un dado

xk = np.arange(start=1, stop=7)
p = 1 / 6
pk = (p, p, p, p, p, p)
distribucion = stats.rv_discrete(values=(xk, pk))
```

¡Hagámoslo en la computadora! (2)

Repetimos k veces el experimento de calcular S_n^*

```
def ensayar(distribucion, n, k):
    mu = distribucion.mean()
    sigma = distribucion.std()

    ensayos = np.array([])
    for i in range(k):
        Sn = np.sum(distribucion.rvs(size=n))
        Sn_star = (Sn - n * mu) / (np.sqrt(n) * sigma)
        ensayos = np.append(ensayos, Sn_star)

        return ensayos
```

¡Hagámoslo en la computadora! (3)

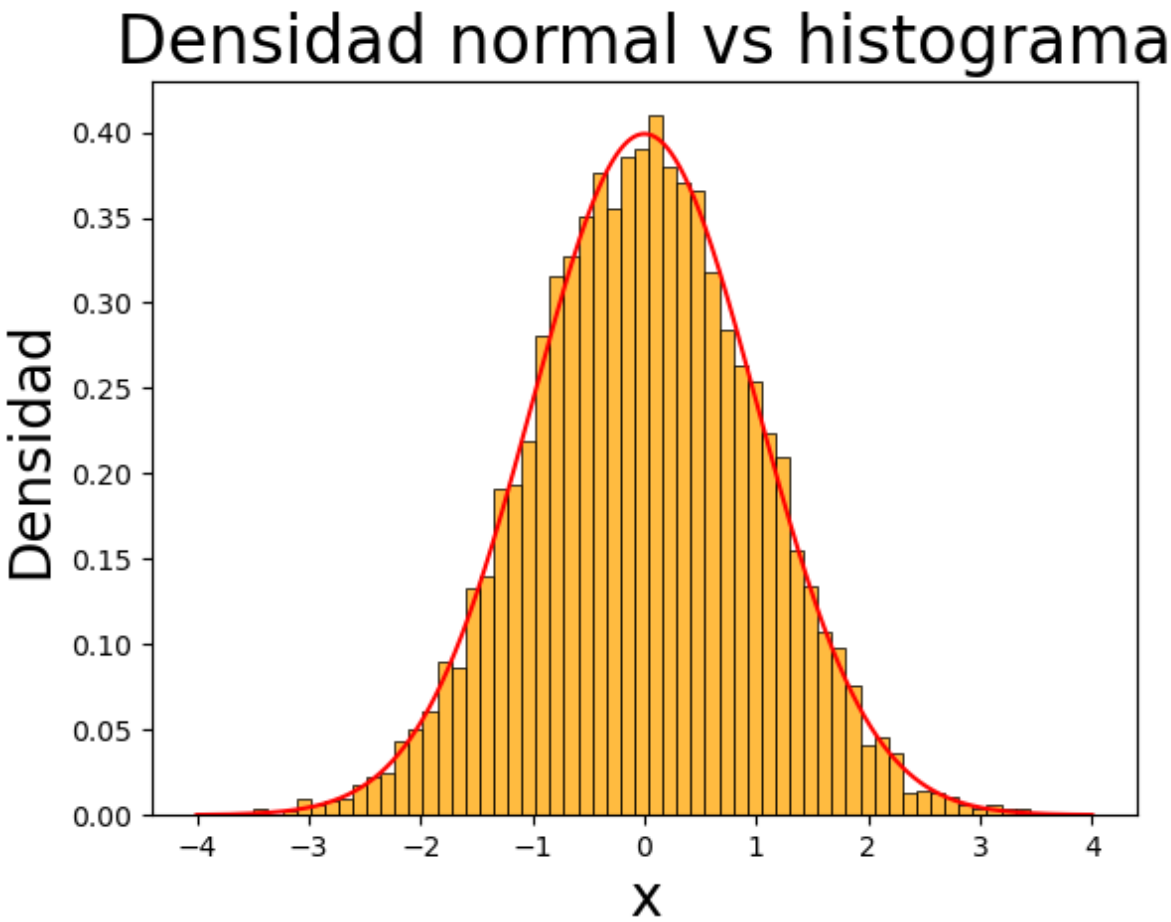
Ahora usamos esa función

```
n = 10000
k = 10000
ensayos = ensayar(distribucion, n, k)

import seaborn
seaborn.histplot(data=ensayos, stat='density', color='Orange')

import matplotlib.pyplot as plt
xs = np.linspace(-4, 4, 100001)
pdfs = stats.norm.pdf(xs)
plt.plot(xs, pdfs, color='red')
```

Histograma de S_n^* en el ejemplo del dado



Otro ejemplo: las distribuciones χ^2_n

- Cnsideremos las variables

$$Z_n = X_1^2 + X_2^2 + \ldots + X_n^2$$

donde las (X_k) son variables con distribución normal estándar independientes.

- Entonces, por definición Z_n tiene distribución χ^2_n (chi-cuadrado con n grados de libertad). Estas distribución la utilizaremos en la parte de estadística.
- Se puede ver que coincide con la distribución $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.
- Sabemos que $E[Z_n] = n$ y $\mathrm{Var}(Z_n) = 2n$.

Aplicación del Teorema Central del límite

Por el teorema del límite central, para n grande, la distribución normal proporciona una buena aproximación de la distribución χ^2_n en el sentido que las variables normalizadas

$$Z_n^* = \frac{Z_n - n}{\sqrt{2n}}$$

convergen en distribución a una normal estándar.

$$F_{Z_*^n}(x) \to F_N(x) \text{ para todo } x \text{ cuando } n \to \infty$$

Gráfico comparando las funciones de distribución de Z_n^*