Parte 1. Tema 3: Distribuciones Notables Parte I: Distribuciones Discretas

Probabilidad con R y python

27 abril, 2023

1. Distribuciones Notables I
2. Distribución Bernoulli
3. Distribución binomial
4. Distribución geométrica
5. Distribución binomial negativa
6. Distribución de Poisson
7. Distribución hipergeométrica

Lección 1

Distribuciones Notables I

Introducción

- En este tema estudiaremos diversos tipos de experimentos que son muy frecuentes y algunas de las variables aleatorias asociadas a ellos.
- Estas variables reciben distintos nombres que aplicaremos sin distinción al tipo de población del experimento a la variable o a su función de probabilidad, densidad o distribución.
- Empezaremos con las variables aleatorias discretas que se presentan con frecuencia ya que están relacionadas con situaciones muy comunes como el número de caras en varios lanzamiento de una moneda, el número de veces que una maquina funciona hasta que se estropea, el numero de clientes en una cola,...

Lección 2

Distribución Bernoulli

Distribución Bernoulli

Distribución Bernoulli

- Consideremos un experimento con dos resultados posibles éxito (E) y fracaso (F). El espacio de sucesos será $\Omega = \{E, F\}$.
- Supongamos que la probabilidad de éxito es $P(E) = p$, y naturalmente $P(F) = 1 - p = q$ con $0 < p < 1$.
- Consideremos la aplicación

$$X : \Omega = \{E, F\} \to \mathbb{R}$$

definida por

$$X(E) = 1,\ X(F) = 0.$$

Distribución Bernoulli

Su función de probabilidad es

$$P_X(x) = \begin{cases} 1 - p = q & \text{si } x = 0 \\ p & \text{si } x = 1 \\ 0 & \text{en cualquier otro caso} \end{cases}.$$

Su función de distribución es

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - p = q & \text{si } 0 \leq x < 1 \\ 1 & \text{si } 1 \leq x \end{cases}.$$

Distribución Bernoulli

- Bajo estas condiciones diremos que X **es una v.a. Bernoulli** o que sigue una ley de **distribución de probabilidad Bernoulli** de parámetro p.
- Lo denotaremos por
$$X \equiv Ber(p) \text{ o también } X \equiv B(1, p).$$
- A este tipo de experimentos (éxito/fracaso)se les denomina experimentos Bernoulli.
- Fue su descubridor un científico suizo Jacob Bernoulli, uno más de la de la conocida familia de científicos suizos Bernoulli.

Esperanza de una v.a. X $Ber(p)$

Su **valor esperado** es

$$E(X) = \sum_{x=0}^{1} x \cdot P(X = x) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Calculemos también $E(X^2)$

$$E(X^2) = \sum_{x=0}^{1} x^2 \cdot P(X = x) = 0^2 \cdot (1 - p) + 1^2 \cdot p = p.$$

Varianza de una v.a. X $Ber(p)$

Su **varianza** es

$$Var(X) = E(X^2) - (E(X))^2 = p - p^2 = p \cdot (1-p) = p \cdot q.$$

Su desviación típica es

$$\sqrt{Var(X)} = \sqrt{p \cdot (1-p)}.$$

Resumen v.a con distribución Bernoulli

X Bernoulli: $Ber(p)$
$D_X=\{0,1\}$
$P_X(x)=P(X=x)=\begin{cases} 1-p=q & \text{si } x=0 \\ p & \text{si } x=1 \\ 0 & \text{en otro caso.} \end{cases}$
$F_X(x)=P(X\leq X)=\begin{cases} 0 & \text{si } x<0 \\ (1-p) & \text{si } 0\leq x<1 \\ 1 & \text{si } x\geq 1 \end{cases}.$
$E(X)=p$; $Var(X)=p\cdot(1-p)$

Distribución Bernoulli. Ejemplo

Veamos los cálculos básicos $Ber(p = 0.25)$ en R.

```r
dbinom(0,size=1,prob=0.25)
```

```
## [1] 0.75
```

```r
dbinom(1,size=1,prob=0.25)
```

```
## [1] 0.25
```

```r
rbinom(n=20,size = 1,prob=0.25)
```

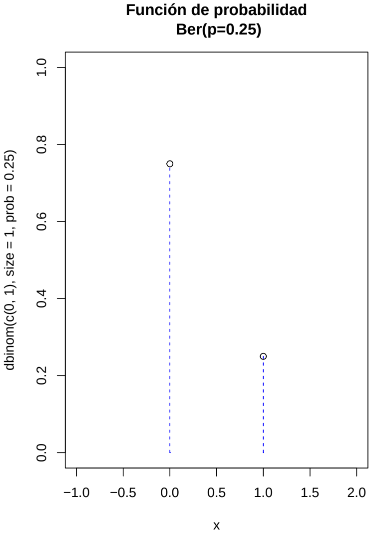
```
##  [1] 0 0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0
```

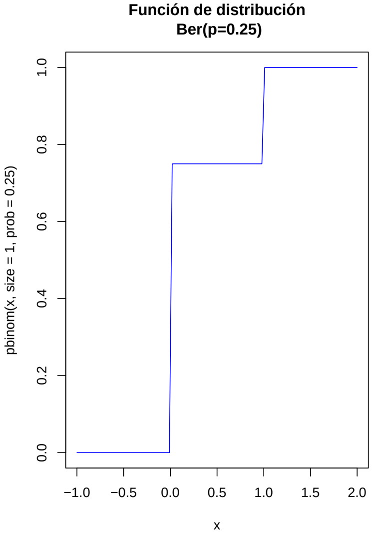
Distribución Bernoulli. Ejemplo

El siguiente código dibuja las función de probabilidad y la de distribución de una $Ber(p = 0.25)$

```
par(mfrow=c(1,2))
plot(x=c(0,1),y=dbinom(c(0,1),size=1,prob=0.25),
     ylim=c(0,1),xlim=c(-1,2),xlab="x",
     main="Función de probabilidad\n Ber(p=0.25)")
lines(x=c(0,0,1,1),y=c(0,0.75,0,0.25), type = "h", lty = 2,col="blue")
curve(pbinom(x,size=1,prob=0.25),
      xlim=c(-1,2),col="blue",
      main="Función de distribución\n Ber(p=0.25)")
par(mfrow=c(1,1))
```

Distribución Bernoulli. Ejemplo





Lección 3

Distribución binomial

Distribución binomial

Distribución binomial

Si repetimos n veces de forma independiente un experimento Bernoulli de parámetro p.

El espacio muestral Ω estará formado por cadenas de E's y F's de longitud n Consideremos la v.a.

$$X(\overbrace{EFFF\ldots EEF}^{n}) = \text{número de éxitos en la cadena.}$$

A la variable aleatoria anterior se le conoce como distribución binomial de parámetros n y p, y lo denotaremos por $X \equiv B(n,p)$.

Función de probabilidad de una binomial

Entonces su **función de probabilidad** es

$$P_X(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} & \text{si } x = 0, 1, \ldots, n \\ 0 & \text{en otro caso} \end{cases}.$$

Función de distribución de binomial

Su **función de distribución** no tiene una fórmula cerrada. Hay que acumular la función de probabilidad:

$$
\begin{aligned}
F_X(x) = P(X \leq x) &= \sum_{i=0}^{x} P_X(i) \\
&= \begin{cases} 0 & \text{si } x < 0 \\ \displaystyle\sum_{i=0}^{k} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} & \text{si } \begin{cases} k \leq x < k+1 \\ k = 0, 1, \ldots, n. \end{cases} \\ 1 & \text{si } n \leq x \end{cases}
\end{aligned}
$$

Números binomiales con R

Los números binomiales calculan el número de equipos de baloncesto distintos que ($k = 5$ jugadores) se pueden hacer con 6 jugadores ($n = 6$).

Es decir cuántas maneras distintas hay para elegir (*choose*) 5 jugadores en un conjunto de 6 jugadores. Todo el mundo diría ¡¡¡6!!!. Efectivamente con R es

```
choose(6,5)
```

```
## [1] 6
```

Números binomiales con R

Con 10 jugadores el número de equipos de 5 distintos es bastante más grande

```
choose(10,5)
```

```
## [1] 252
```

Y, por ejemplo, con un equipo de fútbol profesional que tiene en plantilla 22 jugadores (quitando los guardametas) se pueden formar ¡¡nada menos que!!

```
choose(22,10)
```

```
## [1] 646646
```

un bonito número capicúa que nos da el número de equipos distintos que se pueden formar.

Distribución Binomial

Obviamente se tiene que una v.a. Bernoulli es una binomial con $n = 1$

$$B(1, p) = Ber(p).$$

Ejercicio

Calculad las funciones de distribución de una binomial $B(n = 1, p = 0.3)$ y comprobar que coinciden con las distribuciones de una $Ber(p = 0.3)$.

Observaciones sobre la distribución binomial

- La probabilidad de fracaso se suele denotar con $q = 1 - p$, **sin ningún aviso adicional**, con el fin de acortar y agilizar la escritura de las fórmulas.
- Su **función de distribución no tienen una formula general**, hay que calcularla con una función de R o python... En el siglo pasado se tabulaban en los libros de papel :-).
- En el material adicional os pondremos unas tablas de esta distribución para distintos valores de n y p para que disfrutéis de tan ancestral método de cálculo.
- Cualquier paquete estadístico, hoja de cálculo dispone de funciones para el cálculo de estas probabilidades, así que el **uso de las tablas** queda **totalmente anticuado**.

Esperanza de una $B(n, p)$

Su **esperanza** es

$$E(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k} = n \cdot p.$$

La esperanza de X^2 es

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} \cdot p^k \cdot q^{n-k} \\ &= n \cdot p \cdot q + (n \cdot p)^2. \end{aligned}$$

Varianza de una $B(n,p)$

Su **varianza** es

$$Var(X) = E(X^2) - (E(X))^2 = n \cdot p \cdot q = n \cdot p \cdot (1-p).$$

Su desviación típica es

$$\sqrt{n \cdot p \cdot q} = \sqrt{n \cdot p \cdot (1-p)}.$$

En temas posteriores veremos una forma sencilla del cálculo de la esperanza y varianza de una $B(n,p)$ como las suma de n v.a. $Ber(p)$ independientes.

Ejercicio

Justificar de forma intuitiva que si X_i con $i = 1, 2, \ldots, n$ son v.a. $Ber(p)$ independientes entonces $X = \sum_{i=1}^{n} X_i$ sigue una distribución $B(n,p)$.

Resumen v.a con distribución binomial $B(n,p)$

X binomial: $B(n,p)$
$D_X=\{0,1,\ldots n\}$
$P_X(x)=P(X=x)=\begin{cases} \binom{n}{x}\cdot p^x\cdot(1-p)^{n-x} & \text{si } x=0,1,\ldots,n \\ 0 & \text{en otro caso.} \end{cases}$
$F_X(x)=P(X\leq X)=\begin{cases} 0 & \text{si } x<0 \\ \sum_{i=0}^{k}\binom{n}{i}\cdot p^i\cdot(1-p)^{n-i} & \text{si } k\leq x<k+1 \text{para } k=0,1,\ldots,n \\ 1 & \text{si } x\geq n \end{cases}.$
$E(X)=n\cdot p;\ Var(X)=n\cdot p\cdot(1-p)$

Cálculos binomial con R

Veamos los cálculos básicos con funciones de R para una v.a X con distribución binomial $B(n = 10, p = 0.25)$.

Si queremos calcular con `R` algún valor de la función de distribución como por ejemplo $F_X(0) = P(X \leq 0)$, tenemos que hacer:

```
pbinom(0,size=10,prob=0.25)
```

```
## [1] 0.05631351
```

y si queremos por ejemplo $F_X(4) = P(X \leq 4)$, tenemos que hacer:

```
pbinom(4,size=10,prob=0.25)
```

```
## [1] 0.9218731
```

Cálculos binomial con R

Sin embargo, si queremos calcular algún valor de la función de probabilidad como por ejemplo $P(X = 0)$, tenemos que hacer:

```
dbinom(0,size=10,prob=0.25)
```

```
## [1] 0.05631351
```

o por ejemplo para $P(X = 4)$:

```
dbinom(4,size=10,prob=0.25)
```

```
## [1] 0.145998
```

Generación de muestras aleatorias con R

Generaremos una muestra aleatoria de 100 valores de una población con distribución $B(20, 0.5)$

```
set.seed(2019)
rbinom(100,size = 20,prob=0.5)
```

```
##   [1] 12 11  9 11  6  6 12  5  7 11 12 11  8  8 11 11  7 11  9 10  9 10 14
##  [26]  5 11 14 11 10 11  5 12  8  6  7  9 10  5 12 11  9 12 11 12 10 13 13
##  [51]  9  7  6  9 10  9 16 13  6  6  8  8 11  9 12 15  9  7 12 11  9  8  9
##  [76] 15  7 10  9 12  6 13 14  8 10  8 10 11 11  9 10 11 12  8 10 12  9 13
```

Ejemplo

El ejemplo anterior correspondería a repetir 100 veces el experimento de lanzar una moneda 20 veces y contar el número de caras.

Cálculos distribución binomial con python

Veamos los cálculos básicos con funciones de python para una v.a X con distribución binomial $B(n=10, p=0.25)$.

Primero importamos la función `binom` de la librería `scipy.stat`

```
from scipy.stats import binom
```

En general en el paquete `scipy`, la función de probabilidad se invocará con el método `pmf`, la de distribución con el método `cdf` mientras que una muestra aleatoria que siga esta distribución con el método `rvs`. En todos ellos aparecerá siempre el parámetro `loc` que se utiliza para desplazar el dominio de la variable aleatoria. Por ejemplo, en este caso

```
binom.pmf(k, n, p, loc) =  binom.pmf(k - loc, n, p)
```

Cálculos distribución binomial con python

Para calcular los valores de la función de distribución como por ejemplo $F_X(0) = P(X \leq 0)$ y $F_X(4) = P(X \leq 4)$ utilizamos la función `cdf`

```
binom.cdf(0,n=10,p=0.25)
```

```
## 0.056313514709472684
```

```
binom.cdf(4,n=10,p=0.25)
```

```
## 0.9218730926513672
```

Notemos que al no indicar el valor de `loc`, se le asume que toma el valor 0.

Cálculos distribución binomial con python

Para calcular los valores de la función de probabilidad $P(X=0)$ y $P(X=4)$ utilizamos la función `pmf`:

```
binom.pmf(0,n=10,p=0.25)
```

```
## 0.056313514709472656
```

```
binom.pmf(4,n=10,p=0.25)
```

```
## 0.14599800109863284
```

Notemos que al no indicar el valor de `loc`, se le asume que toma el valor 0.

Cálculos distribución binomial con python

Si queremos generar una muestras aleatorias que siga una distribución binomial, podemos usar la función `rvs`. En este caso, generaremos una muestra aleatoria de 100 valores de una población $B(20, 0.5)$

```
binom.rvs(n=20,p=0.25,size = 100)
```

```
## array([6, 6, 3, 4, 5, 5, 5, 4, 7, 4, 2, 7, 1, 4, 4, 9, 6, 4, 3, 3, 7, 6,
##        5, 4, 0, 2, 6, 4, 7, 4, 5, 2, 4, 6, 7, 6, 7, 3, 6, 3, 8, 6, 4, 7,
##        7, 8, 4, 6, 6, 6, 3, 6, 1, 7, 2, 2, 5, 5, 7, 5, 5, 6, 8, 6, 3, 3,
##        7, 3, 5, 2, 4, 6, 4, 6, 4, 5, 3, 6, 4, 5, 2, 5, 3, 4, 7, 6, 6, 9,
##        2, 5, 4, 4, 3, 4, 4, 6, 4, 4, 3, 4], dtype=int64)
```

Cálculos distribución binomial con python

Observación Notemos que la secuencia aleatoria generada no es la misma que con R. De hecho, si volvemos a ejecutar esta función obtendremos una muestra aleatoria distinta.

```
binom.rvs(n=20,p=0.25,size = 100)
```

```
## array([ 5,  7,  6,  4,  5,  4,  6,  4,  7,  3,  4,  2,  5,  5,  8,  3,  3,
##          4,  3,  2,  7,  3,  5,  6,  5,  5,  7,  6,  6,  7,  4,  3,  5,  5,
##          3,  5,  8,  9,  4,  5,  9,  5,  9,  5,  6,  4,  4,  4,  2,  5,  3,
##          7,  5,  3, 10,  5,  3,  9,  5,  5,  7,  5,  1,  6,  5,  6,  5,  2,
##          5,  6,  6,  6,  2,  4,  6,  4,  3,  4, 10,  4,  8,  4,  4,  5,  6,
##          5,  6,  3,  2,  4,  7,  4,  9,  7,  4,  2,  3, 10,  4,  3],
##        dtype=int64)
```

Cálculos binomial con python

Veamos algunos cálculos básicos con funciones de python para la binomial $B(n = 10, p = 0.25)$.

```
binom.cdf(5,n=10,p=0.25)
```

```
## 0.9802722930908203
```

```
binom.pmf(1,n=10,p=0.25)
```

```
## 0.1877117156982421
```

```
binom.rvs(n=20,p=0.25,size=10)
```

```
## array([ 5,  5,  6,  8,  4,  2, 11,  6,  6,  3], dtype=int64)
```

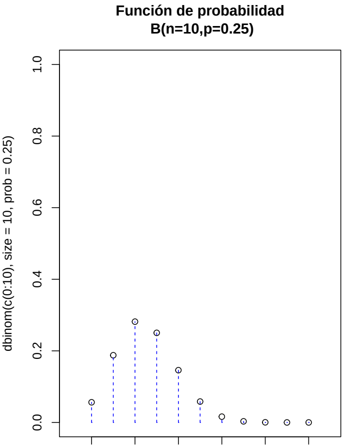
Gráficas de la distribución binomial con R

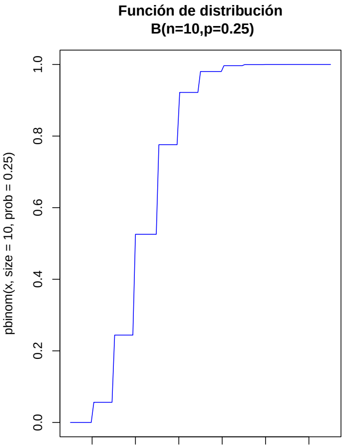
El siguiente código de R dibuja las función de probabilidad y la de distribución de una $B(n = 10, p = 0.25)$

```
par(mfrow=c(1,2))
aux=rep(0,22)
aux[seq(2,22,2)]=dbinom(c(0:10),size=10,prob=0.25)
plot(x=c(0:10),y=dbinom(c(0:10),size=10,prob=0.25),
  ylim=c(0,1),xlim=c(-1,11),xlab="x",
  main="Función de probabilidad\n B(n=10,p=0.25)")
lines(x=rep(0:10,each=2),y=aux, type = "h", lty = 2,col="blue")
curve(pbinom(x,size=10,prob=0.25),
  xlim=c(-1,11),col="blue",
  main="Función de distribución\n B(n=10,p=0.25)")
par(mfrow=c(1,1))
```

Gráficas de la distribución binomial con R

El siguiente código de R dibuja las función de probabilidad y la de distribución de una $B(n = 10, p = 0.25)$





Gráficos de la distribución binomial con python

Ejercicio

Buscad en la documentación de python cómo se dibuja la función de probabilidad y de distribución de una binomial y recread los gráficos anteriores.

Pista: Necesitaremos investigar más librerías:

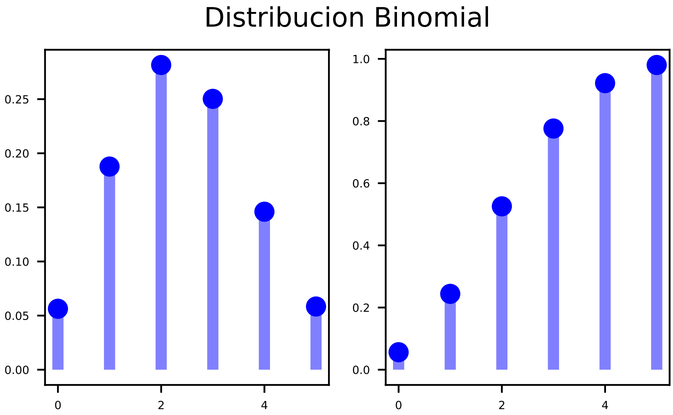
```
import numpy as np
import matplotlib.pyplot as plt
```

Gráficos de la distribución binomial con python

```
n, p = 10, 0.25
x = np.arange(binom.ppf(0.01, n, p),binom.ppf(0.99, n, p))
fig =plt.figure(figsize=(5, 2.7))
ax = fig.add_subplot(1,2,1)
ax.plot(x, binom.pmf(x, n, p), 'bo', ms=8, label='binom pmf')
ax.vlines(x, 0, binom.pmf(x, n, p), colors='b', lw=5, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  tick.label.set_fontsize(5)
ax = fig.add_subplot(1,2,2)
ax.plot(x, binom.cdf(x, n, p), 'bo', ms=8, label='binom pmf')
ax.vlines(x, 0, binom.cdf(x, n, p), colors='b', lw=5, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
```

Gráficos de la distribución binomial con python

```
## <string>:2: MatplotlibDeprecationWarning: The label function was deprecatec
```



Ejemplo distribución binomial

Ejemplo: número de bolas rojas extraídas de una urna con reposición

Tenemos una urna con 100 bolas de las cuales 40 son rojas y 60 son blancas. Extraemos al azar una bola, anotamos su color y la devolvemos a (reponemos en) la urna.

Supongamos que repetimos este proceso $n = 10$ reponiendo en cada ocasión la bola extraída.

Consideremos la variable aleatoria X como el número de bolas rojas extraídas (con reposición) en $n = 10$ repeticiones del mismo experimento de Bernoulli.

Bajo estas condiciones repetimos $n = 10$ veces el mismo experimento de Bernouilli con probabilidad de éxito (sacar bola roja)

$$P(Roja) = P(\text{É}xito) = p = \frac{40}{100} = 0.4.$$

Así que la variable X que es el número de bolas rojas extraídas de la urna (con reposición) en $n = 10$ ocasiones sigue una ley binomial $B(n = 10, p = 0.4)$.

Ejemplo $B(n = 10, p = 0.4)$.

Nos preguntamos:

1. ¿Cuál es la probabilidad de que saquemos exactamente 4 bolas rojas?
2. ¿Cuál es la probabilidad de que saquemos al menos 4 bolas rojas?
3. ¿Cuál es la probabilidad de que saquemos menos de 3 bolas rojas?
4. ¿Cuál es el valor esperado del número de bolas rojas?
5. ¿Cuál es la desviación típica del número de bolas rojas?

Ejemplo $B(n = 10, p = 0.4)$.

Solución 1. ¿Cuál es la probabilidad de que saquemos exactamente 4 rojas?

Utilizando la función de probabilidad, tenemos que:

$$\begin{aligned} P(X = 4) &= \binom{10}{4} \cdot 0.4^4 \cdot (1 - 0.4)^{10-4} = \frac{10!}{(10-4)! \cdot 4!} \cdot 0.4^4 \cdot 0.6^6 \\ &= \frac{7 \cdot 8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3 \cdot 4} \cdot 0.4^4 \cdot 0.6^6 = 0.2508227. \end{aligned}$$

Con R

```
dbinom(4,size=10,prob = 0.4)
```

```
## [1] 0.2508227
```

Ejemplo $B(n = 10, p = 0.4)$.

Solución 2. ¿Cuál es la probabilidad de que saquemos al menos 4 bolas rojas?

La probabilidad de sacar al menos 4 rojas se expresa como

$P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3)$:

$$\begin{aligned}
P(x \leq 3) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) \\
&= \binom{10}{0} \cdot 0.4^0 \cdot (1 - 0.4)^{10-0} + \binom{10}{1} \cdot 0.4^1 \cdot (1 - 0.4)^{10-1} \\
&+ \binom{10}{2} \cdot 0.4^2 \cdot (1 - 0.4)^{10-2} + \binom{10}{3} \cdot 0.4^3 \cdot (1 - 0.4)^{10-3} \\
&= 0.3822806.
\end{aligned}$$

Ejemplo $B(n = 10, p = 0.4)$.

Con R

```
pbinom(3,10,0.4)
```

```
## [1] 0.3822806
```

Así que

$$P(X \geq 4) = 1 - P(X < 4) = 1 - P(X \leq 3) = 1 - 0.3822806 = 0.6177194.$$

Ejemplo $B(n = 10, p = 0.4)$.

Otra manera usando R sería:

```
1-pbinom(3,10,0.4)
```

```
## [1] 0.6177194
```

Aunque en estos casos el parámetro `lower.tail = FALSE` es sin duda nuestra mejor opción:

```
pbinom(3,10,0.4,lower.tail = FALSE)
```

```
## [1] 0.6177194
```

Ejemplo $B(n = 10, p = 0.4)$.

Solución 3. ¿Cuál es la probabilidad de que saquemos menos de 3 bolas rojas?

$$
\begin{aligned}
P(X < 3) &= P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \\
&= \binom{10}{0} \cdot 0.4^0 \cdot (1 - 0.4)^{10-0} + \binom{10}{1} \cdot 0.4^1 \cdot (1 - 0.4)^{10-1} \\
&\quad + \binom{10}{2} \cdot 0.4^2 \cdot (1 - 0.4)^{10-2} \\
&= 0.1672898.
\end{aligned}
$$

En R:

```
dbinom(0,10,0.4)+dbinom(1,10,0.4)+dbinom(2,10,0.4)
```

```
## [1] 0.1672898
```

```
pbinom(2,10,0.4)
```

Ejemplo $B(n = 10, p = 0.4)$.

Solución 4. ¿Cuál es el valor esperado del número de bolas rojas?

Como X es una $B(n = 10, p = 0.4)$ sabemos que

$$E(X) = n \cdot p = 10 \cdot 0.4 = 4.$$

Aunque en python tenemos la función `stats` que nos lo calcula directamente:

```
print("E(X) = {m}".format(m=binom.stats(n = 10, p = 0.4, moments='m')))
```

Ejemplo $B(n = 10, p = 0.4)$.

Solución 5. ¿Cuál es la desviación típica del número de bolas rojas?

La varianza es:

$$Var(X) = n \cdot p \cdot (1 - p) = 10 \cdot 0.4 \cdot 0.6 = 2.4.$$

Por lo tanto la desviación típica es

$$\sqrt{Var(X)} = \sqrt{2.4} = 1.5491933.$$

Aunque en python tenemos la función `stats` que nos lo calcula directamente:

```
print("Var(X) = {v}".format(v=binom.stats(n = 10, p = 0.4, moments='v')))
```

Lección 4

Distribución geométrica

Distribución geométrica

- Todos hemos jugado a, por ejemplo, tirar una moneda hasta que obtengamos la primera cara.
- O también tirar una pelota a una canasta de baloncesto hasta obtener la primera canasta.
- Desde otro punto de vista también podemos intentar modelar el número de veces que accionamos una interruptor y la bombilla se ilumina hasta que falla.
- O también el número de veces que un cajero automático nos da dinero hasta que falla.

La **modelización de este tipo de problemas se consigue con la llamada distribución geométrica**.

Distribución geométrica

Distribución geométrica

- Repitamos un experimento Bernoulli, de parámetro p, de forma independiente hasta obtener el primer éxito.
- Sea X la v.a. que cuenta el número de fracasos antes del primer éxito. Por ejemplo que hayamos tenido x fracasos será una cadena de x fracasos culminada con un éxito. Más concretamente

$$P(\overbrace{FFF\ldots F}^{x}E) = P(F)^{x}\cdot P(E) = (1-p)^{x}\cdot p = q^{x}\cdot p.$$

Distribución geométrica

Su función de probabilidad es

$$P_X(x) = P(X = x) = \begin{cases} (1-p)^x \cdot p & \text{si } x = 0, 1, 2, \ldots \\ 0 & \text{en otro caso} \end{cases}.$$

- La v.a. definida anteriormente diremos que sigue una distribución geométrica de parámetro p.
- La denotaremos por $Ge(p)$.
- Su dominio es $D_X = \{0, 1, 2, \ldots\}$.

Función de distribución geométrica

Calculemos $\text{P}(X \leq 3)$.

Por la propiedad de la probabilidad del suceso complementario tenemos que

$$P(X \leq 3) = 1 - P(X > 3) = 1 - P(X \geq 4)$$

Efectivamente, el complementario del evento $X \leq 3$ nos dice que hemos fracasado más de tres veces hasta conseguir el primer éxito, es decir, **hemos fracasado 4 o más veces**. Podemos simbolizar dicho evento de la forma siguiente:

$$\{X > 3\} = \{X \geq 4\} = \{FFFF\}$$

Función de distribución geométrica

Ahora, al ser los intentos independientes, tenemos que:

$$\begin{aligned} P(X>3) &= P(\{FFFF\}) = P(F)\cdot P(F)\cdot P(F)\cdot P(F) \\ &= (1-p)\cdot(1-p)\cdot(1-p)\cdot(1-p) = (1-p)^{3+1} \\ &= (1-p)^4. \end{aligned}$$

El valor de la función de distribución de X en $x=3$ será, pues:

$$F_X(3) = P(X\leq 3) = 1-P(X>3) = 1-(1-p)^{3+1}.$$

Generalizando el resultado anterior a cualquier entero positivo $k=0,1,2,\ldots$, tenemos:

$$F_X(k) = P(X\leq k) = 1-(1-p)^{k+1},\ \text{si } k=0,1,2,\ldots$$

Función de distribución geométrica

En general, tendremos que:

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{si } x < 0, \\ 1-(1-p), & \text{si } k = 0 \leq x < 1, \\ 1-(1-p)^2, & \text{si } k = 1 \leq x < 2, \\ 1-(1-p)^3, & \text{si } k = 2 \leq x < 3, \\ 1-(1-p)^{k+1}, & \text{si } \begin{cases} k \leq x < k+1, \\ \text{para } k = 0, 1, 2, \ldots \end{cases} \end{cases}.$$

Función de distribución geométrica

De forma más compacta, tendremos que

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{si } x < 0, \\ 1-(1-p)^{k+1}, & \text{si } \begin{cases} k \leq x < k+1, \\ \text{para } k = 0, 1, 2, \ldots \end{cases} \end{cases}.$$

Notemos que el límite de la función de distribución es:

$$\lim_{k \to +\infty} F_X(k) = \lim_{k \to +\infty} 1-(1-p)^{k+1} = 1,$$

ya que $0 < 1-p < 1$.

Sumas derivadas series geométricas

Recordemos del tema de variables aleatorias que

Propiedades

- Si $|r| < 1$ también son convergentes las derivadas, respecto de r, de la serie geométrica y convergen a la derivada correspondiente. Así tenemos que

$$\begin{aligned}
\left(\sum_{k=0}^{+\infty} r^k\right)' &= \sum_{k=1}^{+\infty} k\cdot r^{k-1} &&= \left(\frac{1}{1-r}\right)' = \frac{1}{(1-r)^2}\\
\left(\sum_{k=0}^{+\infty} r^k\right)'' &= \sum_{k=2}^{+\infty} k\cdot(k-1)\cdot r^{k-2} &&= \left(\frac{1}{1-r}\right)'' = \frac{2}{(1-r)^3}
\end{aligned}$$

Esperanza de una v.a. $Ge(p)$

Recordemos que $P(X=x)=(1-p)^x\cdot p$ si $x=0,1,2,\ldots$ y aplicado la fórmula anterior con $r=1-p$

$$\begin{aligned}
E(X) &= \sum_{x=0}^{+\infty} x\cdot P_x(x) = \sum_{x=0}^{+\infty} x\cdot(1-p)^x\cdot p \\
&= p\cdot(1-p)\cdot\sum_{x=1}^{+\infty} x\cdot(1-p)^{x-1} \\
&= p\cdot(1-p)\cdot\frac{1}{(1-(1-p))^2} = p\cdot(1-p)\cdot\frac{1}{p^2} = \frac{1-p}{p}
\end{aligned}$$

Valor $E(X^2)$ de una v.a. $Ge(p)$

$$
\begin{aligned}
E(X^2) &= \sum_{x=0}^{+\infty} x^2 \cdot P_X(x) = \sum_{x=1}^{+\infty} x^2 \cdot (1-p)^x \cdot p \\
&= \sum_{x=1}^{+\infty} (x\cdot(x-1)+x)\cdot(1-p)^x \cdot p \\
&= \sum_{x=1}^{+\infty} x\cdot(x-1)\cdot(1-p)^x \cdot p + \sum_{x=1}^{+\infty} x\cdot(1-p)^x\cdot p \\
&= (1-p)^2 \cdot p \cdot \sum_{x=2}^{+\infty} x\cdot(x-1)\cdot(1-p)^{x-2} \\
&+ (1-p)\cdot p \sum_{x=1}^{+\infty} x\cdot(1-p)^{x-1} = \ldots
\end{aligned}
$$

.

Valor $E(X^2)$ de una v.a. $Ge(p)$

$$
\begin{aligned}
E(X^2) &= \ldots \\
&= (1-p)^2 \cdot p \cdot \sum_{x=2}^{+\infty} x \cdot (x-1) \cdot (1-p)^{x-2} \\
&+ (1-p) \cdot p \sum_{x=1}^{+\infty} x \cdot (1-p)^{x-1} \\
&= p \cdot (1-p)^2 \frac{2}{(1-(1-p))^3} + (1-p) \cdot p \frac{1}{(1-(1-p))^2} \\
&= p \cdot (1-p)^2 \frac{2}{p^3} + (1-p) \cdot p \frac{1}{p^2} \\
&= \frac{2 \cdot (1-p)^2}{p^2} + \frac{1-p}{p}.
\end{aligned}
$$

Varianza de una v.a. $Ge(p)$

$$
\begin{aligned}
Var(X) &= E(X^2) - E(X)^2 = \frac{2\cdot(1-p)^2}{p^2} + \frac{1-p}{p} - \left(\frac{1-p}{p}\right)^2 \\
&= \frac{2\cdot(1-p)^2 + p\cdot(1-p) - (1-p)^2}{p^2} = \frac{(1-p)^2 + p\cdot(1-p)}{p^2} \\
&= \frac{1 - 2\cdot p + p^2 + p - p^2}{p^2} \\
&= \frac{1-p}{p^2}.
\end{aligned}
$$

Y su desviación típica será

$$\sqrt{Var(X)} = \sqrt{\frac{1-p}{p^2}}.$$

Resumen distribución geométrica $Ge(p)$ empezando en 0

X = Geométrica (empieza en 0) número de fracasos para conseguir el primer éxito
$D_X = \{0, 1, \ldots n, \ldots\}$
$P_X(x) = P(X = x) = \begin{cases} (1-p)^x \cdot p & \text{si } x = 0, 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{cases}$
$F_X(x) = P(X \leq X) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - (1-p)^{k+1} & \text{si } \begin{cases} k \leq x < k+1 \\ \text{para } k = 0, 1, 2, \ldots \end{cases} \end{cases}$
$E(X) = \frac{1-p}{p}$; $Var(X) = \frac{1-p}{p^2}$

La variable geométrica que cuenta los intentos para obtener el primer éxito.

- Supongamos que sólo estamos interesados en el **número de intentos** para obtener el primer éxito.
- Si definimos $Y=$ número de intentos para obtener el primer éxito. Entonces $Y = X + 1$ donde $X \equiv Ge(p)$.
- Su dominio es $D_Y = \{1, 2, \ldots\}$
- La media se incrementa en un intento debido al éxito $E(Y) = E(X+1) = E(X) + 1 = \frac{1-p}{p} + 1 = \frac{1}{p}$.
- La varianza es la misma $Var(Y) = Var(X+1) = Var(X) = \frac{1-p}{p^2}$.

Resumen distribución geométrica $Ge(p)$ empezando en 1.

Y geométrica (que cuenta el éxito) número de INTENTOS para OBTENER el primer éxito
$D_Y = \{1, 2, \ldots n, \ldots\}$
$P_Y(y) = P(Y = y) = \begin{cases} (1-p)^{y-1} \cdot p & \text{si } y = 1, 2, 3, \ldots \\ 0 & \text{en otro caso.} \end{cases}$
$F_Y(y) = P(Y \leq y) = \begin{cases} 0 & \text{si } y < 1 \\ 1-(1-p)^k & \text{si } \begin{cases} k \leq y < k+1 \\ \text{para } k = 1, 2, 3, \ldots \end{cases} \end{cases}$
$E(X) = \frac{1}{p}$; $Var(X) = \frac{1-p}{p^2}$

Propiedad de la falta de memoria

Propiedad de la falta de memoria

Sea X una v.a. discreta con dominio $D_X = \{0, 1, 2, \ldots\}$, con $P(X = 0) = p$.

Entonces X sigue una ley $Ge(p)$ si, y sólo si,

$$P\left(X > k + j \middle| X \geq j\right) = P(X > k)$$

para todo $k, j = 0, 1, 2, 3 \ldots$.

Propiedad de la falta de memoria

Demostración

Si X es geométrica entonces el lado derecho de la igualdad es

$$P(X > k) = 1 - P(X \leq k) = 1 - \left(1 - (1-p)^{k+1}\right) = (1-p)^{k+1},$$

y el lado de izquierdo es

$$\begin{aligned} P\left(X > k + j \middle| X \geq j\right) &= \frac{P\left(\{X > k + j\} \cap \{X \geq j\}\right)}{P\left(X \geq j\right)} = \frac{P\left(X > k + j\right)}{P\left(X \geq j\right)} = \frac{1 - P(X \leq k + j)}{1 - P(X \leq j - 1)} \\ &= \frac{1 - \left(1 - (1-p)^{k+j+1}\right)}{1 - \left(1 - (1-p)^{j-1+1}\right)} = \frac{(1-p)^{k+j+1}}{(1-p)^{j}} = (1-p)^{k+1}, \end{aligned}$$

lo que demuestra la igualdad.

Propiedad de la falta de memoria

Para demostrar el recíproco, tomemos $j = 1$ y $k \geq 0$. Entonces, por la propiedad de la pérdida de memoria:

$$P\left(X > k + 1 \middle| X \geq 1\right) = P(X > k)$$

Como $P(X = 0) = p$, tenemos que $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - p$.

Combinado las igualdades, tenemos que:

$$P\left(X > k + 1 \middle| X \geq 1\right) = \frac{P(X > k + 1, X \geq 1)}{P(X \geq 1)} = \frac{P(X > k + 1)}{P(X \geq 1)} = P(X > k).$$

Así podemos poner que

$$\begin{aligned} P(X > k + 1) &= P(X \geq 1) \cdot P(X > k) = (1 - P(X < 1)) \cdot P(X > k) \\ &= (1 - P(X = 0)) \cdot P(X > k) = (1 - p) \cdot P(X > k). \end{aligned}$$

Propiedad de la falta de memoria

Es decir en general tenemos que

$$P(X > k + 1) = (1 - p) \cdot P(X > k)$$

Del mismo modo para $j = 2$

$$P(X > k + 2) = (1 - p) \cdot P(X > k + 1)$$

Restando la primera igualdad de la última obtenemos.

$$P(X > k + 1) - P(X > k + 2) = (1 - p) \cdot P(X > k) - (1 - p) \cdot P(X > k + 1)$$

de donde operando en cada lado de la igualdad obtenemos la recurrencia

$$[1 - P(X \leq k + 1)] - [1 - P(X \leq k + 2)] = (1 - p) \cdot [P(X > k) - P(X > k + 1)]$$

Propiedad de la falta de memoria

De forma similar obtenemos

$$P(X = k + 1) = (1 - p) \cdot P(X = k)$$

Utilizando la recurrencia anterior, podemos calcular todas las probabilidades $P(X = k)$ a partir de la $P(X = 0) = p$:

$$\begin{aligned}
P(X = 0) &= p, \\
P(X = 1) &= P(X = 0 + 1) = (1 - p) \cdot P(X = 0) = (1 - p) \cdot p, \\
P(X = 2) &= P(X = 1 + 1) = (1 - p) \cdot P(X = 1) = (1 - p) \cdot (1 - p) \cdot p = (1 - p)^2 \cdot p, \\
\vdots \; & \; \vdots \\
P(X = k) &= P(X = (k - 1) + 1) = (1 - p) \cdot P(X = k - 1) = (1 - p) \cdot (1 - p)^{k-1} \cdot p = (1 - p)^k \cdot p,
\end{aligned}$$

lo que demuestra el recíproco, es decir, que X es $Geom(p)$.

Falta de memoria

Observación: Interpretación de la propiedad

La propiedad de la falta de memoria

$$P(X > k + j \big| X \geq j) = P(X > k),$$

significa que, aunque **ya llevemos al menos** j **fracasos**, la probabilidad de **que fracasemos** k **veces más** no disminuye, es la misma que era cuando empezamos el experimento.

A este efecto se le suele etiquetar con la frase **el experimento carece de memoria** o es un **experimento sin memoria** (*Memoryless Property*).

Ejemplo falta de memoria

Un ejemplo muy sencillo nos aclarará el alcance de esta propiedad

Ejercicio: la llave que abre la puerta

Tenemos un llavero con 10 llaves, solo una de ellas abre una puerta. Cada vez que probamos una llave y falla olvidamos que llave hemos probado. ¿Cuál es la probabilidad de que si ya lo hemos intentado 5 veces necesitemos más de 4 intentos adicionales para abrir la puerta?

Tomemos $k = 4, j = 5$, aplicando la propiedad de la falta de memoria

$$P(X > 4 + 5/X \geq 5) = P(X > 4)$$

Después de 5 fracasos no estamos "más cerca" de abrir la puerta. La propiedad de la falta de memoria nos dice que en **después de cada intento es como si empezásemos de nuevo a abrir la puerta**. Tras 5 fracasos la probabilidad de que fallemos más de 4 veces más es la misma que cuando lo intentamos la primera vez.

Ejemplo falta de memoria

¿Cuál es el número esperado de fracasos hasta abrir la puerta?

$$E(X) = \frac{1-p}{p} = \frac{1-\frac{1}{10}}{\frac{1}{10}} = \frac{\frac{9}{10}}{\frac{1}{10}} = 9.$$

La varianza es

$$Var(X) = \frac{1-p}{p^2} = \frac{1-\frac{1}{10}}{\left(\frac{1}{10}\right)^2} = \frac{\frac{9}{10}}{\frac{1}{100}} = 90.$$

La desviación típica es $\sqrt{90} = 9.486833.$

Ejemplo: El clásico del fútbol

Ejemplo: partidos hasta que el Barça gana al Madrid

Los partidos Real Madrid vs FC Barcelona de **la liga** española se suelen denominar **El Clásico**, sean en el Bernabeu (estadio del Real Madrid) o en el Camp Nou (estadio del Barça)

Sea X la variable que cuenta el número de veces consecutivas que en un partido de fútbol de la liga el Barça no gana al Madrid sea en el Camp Nou o el Bernabeu.

Nuestra amiga Aina es muy culé (hincha del Barça) y quiere averiguar cuántos partidos consecutivos de **El Clásico** tiene que ver hasta ver ganar al Barça por primera vez.

Le interesa estimar cuánto le va a costar este capricho. Tendrá que comprar las entradas y pagar los viajes de Barcelona a Madrid.

En datos historicos de **El clásico** en la wikipedia están los datos hasta el 3 de marzo de 2019: se han jugado en total 178 **Clásicos** donde el Real Madrid ganó en 72 ocasiones, el Barça, en 72 y empataron 34 veces.

Ejemplo: El clásico del fútbol

Nos hacemos las siguientes preguntas:

- Si Aina solo tiene dinero para ir a ver 3 partidos, ¿cuál es la probabilidad de no ver ganar al Barça en al menos tres partidos consecutivos?
- ¿Cuántos partidos se tienen que jugar de media para ver ganar al Barça por primera vez?

Con los datos anteriores, podemos estimar que la probabilidad de que el Barça gane un clásico cualquiera es:

$$P(\text{Barça}) = \frac{72}{178} = 0.4045.$$

Por tanto, podemos modelar la variable X, que cuenta el número de veces consecutivas que en un partido de fútbol de la liga el Barça no gana al Madrid, con una ley geométrica empezando en cero con probabilidad de éxito $p = P(\text{Barça}) = \frac{72}{178}$,

Ejemplo: El clásico del fútbol

$$X = Ge\left(p = \frac{72}{178} = 0.4045\right)$$

Así que lo que nos pregunta Aina es la siguiente probabilidad

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - \left(1 - \frac{72}{178}\right)^{2+1} = 0.7888.$$

Así que Aina tiene una probabilidad del 78.88% de no ver ganar al Barça en al menos 3 partidos antes de ver uno en el sí que gane.

Variable geométrica: El clásico

Para responder a la segunda pregunta, usando que la distribución de X es:

$$X = Ge\left(p = \frac{72}{178} = 0.4045\right)$$

entonces

$$E(X) = \frac{1-p}{p} = \frac{1-0.4045}{0.4045} = 1.4722$$

y

$$Var(X) = \frac{1-p}{p^2} = \frac{1-0.4045}{0.4045^2} = 3.6397$$

La desviación típica es

$$\sqrt{3.6397} = 1.9078.$$

Cálculos con R

Veamos los cálculos básicos con R para la distribución geométrica $Ge(p = 0.25)$. R implementa la geométrica que cuenta el número de fracasos.

$P(X = 0) = (1 - 0.25)^0 \cdot 0.25^1 = 0.25$

```r
dgeom(0,prob=0.25)
```

```
## [1] 0.25
```

$P(X \leq 0) = 1 - (1 - 0.25)^{0+1} = 1 - 0.75 = 0.25$

```r
pgeom(0,prob=0.25)
```

```
## [1] 0.25
```

Cálculos con R

$P(X \leq 4) = 1 - (1-0.25)^{4+1} = 1 - 0.75 = 1 - 0.75^5 = 0.7626953.$

```
pgeom(4,prob=0.25)
```

```
## [1] 0.7626953
```

Una muestra aleatoria de tamaño 25 de una $Ge(0.25)$

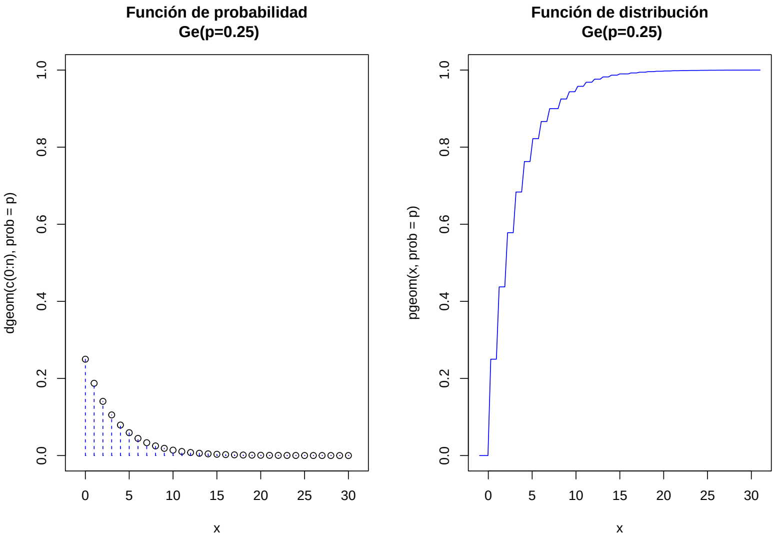
```
rgeom(n=25,prob=0.25)
```

```
##  [1]  5  4  1  6 10  0  0 10  7  0  6  2  1  3  0  2  5  0  0  5  5  3  3
```

Gráficos con R el código

```
par(mfrow=c(1,2))
x=c(0:10)
plot(x=x,y=dgeom(x,prob=0.25),
  ylim=c(0,1),xlim=c(-1,11),xlab="x",
  main="Función de probabilidad\n Ge(p=0.25)")
lines(x=rep(0:10,each=2),y=aux, type = "h", lty = 2,col="blue")
aux0=dgeom(c(0:10),prob=0.25)
ceros=rep(0,21)
ceros
aux=ceros
aux[2*(c(1:11))]<-aux0
curve(pgeom(x,prob=0.25),
  xlim=c(-1,10),col="blue",
  main="Función de distribución\n Ge(p=0.25)")
par(mfrow=c(1,1))
```

Los gráficos con R



Cálculos con python

Veamos los cálculos básicos con python para la distribución geométrica $Ge(p = 0.25)$. scipy.stats implementa la distribución geométrica que cuenta el número intentos así que empieza en 1

Cargamos la función de la librería

```
from scipy.stats import geom
```

Cálculos con python

La función de probabilidad es `geom.pmf(x,p,loc=0)=geom.pmf(x,p)` es un geométrica que cuenta el número de intentos para obtener el primer éxito el valor por defecto del último parámetro es `loc=0`.

Si queremos la que cuenta el número de fracasos para obtener el primer éxito (la geométrica que empieza en 0) tenemos que usar `geom.pmf(x,p,loc=-1)`.

Es decir `geom.pmf(x,p,loc=-1)=geom.pmf(x-1,p,loc=0)`

Veamos pues los cálculos para la $Ge(p)$ que empieza en 0.

$P(X=0)=(1-0.25)^0\cdot 0.25^1=0.25$

```
geom.pmf(0,p=0.25,loc=-1)
```

Cálculos con python

Ejercicio

Qué probabilidades son las que calcula el siguiente código y qué tipo de variables geométricas son?

```
geom.cdf(range(5),p=0.3,loc=0)
geom.cdf(range(5),p=0.3,loc=-1)
```

Cálculos con python esperanza y varianza

Con python también podemos calcular directamente algunos parámetros asociado a una función de distribución predefinida

```
geom.stats(p=0.25, loc=0, moments='mv')
geom.stats(p=0.25, loc=-1, moments='mv')
```

Cálculos con python esperanza y varianza

Ejercicio

Comprobad que las medias y las varianzas calculadas en el código anterior, corresponden a una $Ge(p = 0.3)$ empezando en 1 y a una $Ge(p = 0.3)$ empezando en 0.

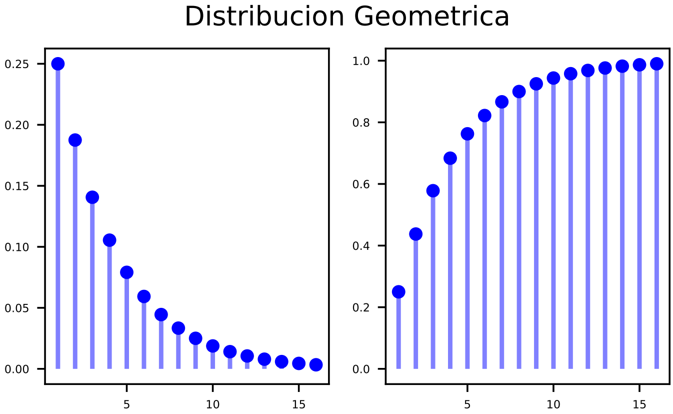
¿Son las varianzas siempre iguales?

Gráficos con python

```
p = 0.25
x = np.arange(geom.ppf(0.01, p),geom.ppf(0.99, p))
fig =plt.figure(figsize=(5, 2.7))
ax = fig.add_subplot(1,2,1)
ax.plot(x, geom.pmf(x, p), 'bo', ms=5, label='geom pmf')
ax.vlines(x, 0, geom.pmf(x, p), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  tick.label.set_fontsize(5)
ax = fig.add_subplot(1,2,2)
ax.plot(x, geom.cdf(x, p), 'bo', ms=5, label='geom pmf')
ax.vlines(x, 0, geom.cdf(x, p), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
```

Gráficos con python

```
## <string>:2: MatplotlibDeprecationWarning: The label function was deprecate
```



Lección 5

Distribución binomial negativa

El problema de la puerta con dos cerraduras

Supongamos que disponemos de 10 llaves distintas y tenemos que abrir una puerta con **dos cerraduras**.

Comenzamos por la primera cerradura, de tal forma que cada vez olvidamos qué llave hemos probado.

Una vez abierta la primera cerradura probamos de igual forma con la segunda hasta que también la abrimos.

Sea X = la v.a. que cuenta el número de fracasos hasta abrir la puerta.

Acertar una llave de la puerta es un experimento Bernoulli con probabilidad de éxito $p = 0.1$. Lo repetiremos hasta obtener 2 éxitos.

Distribución binomial negativa

En general tendremos un experimento de Bernoulli con probabilidad de éxito $0 < p < 1$ tal que:

- Repetimos el experimento hasta obtener el n-ésimo éxito ¡¡abrir la maldita puerta!!.
- Sea X la v.a. que cuenta el número fallos hasta abrir la puerta, es decir, hasta conseguir el n-ésimo éxito. Notemos que no contamos los éxitos, solo contamos los fracasos

Distribución binomial negativa

Si representamos como es habitual un suceso como una cadena de F's y E's, para $n = 2$, algunos sucesos elementales serán:

$$\{EE, FEE, EFE, FFEE, FEFE, EFFE, FFFEE, FFEFE, FEFFE, EFFFE\}.$$

Calculemos algunas probabilidades para $n = 2$:

$$\begin{aligned}
P(X = 0) &= P(\{EE\}) = p^2, \\
P(X = 1) &= P(\{FEE, EFE\}) = 2 \cdot (1 - p) \cdot p^2, \\
P(X = 2) &= P(\{FFEE, FEFE, EFFE\}) = 3 \cdot (1 - p)^2 \cdot p^2, \\
P(X = 3) &= P(\{FFFEE, FFEFE, FEFFE, EFFFE\}) = 4 \cdot (1 - p)^3 \cdot p^2.
\end{aligned}$$

Distribución binomial negativa

En general su función de probabilidad es

$$P_X(k) = P(X = k) = \begin{cases} \binom{k+n-1}{n-1} \cdot (1-p)^k \cdot p^n & \text{si } k = 0, 1, \ldots \\ 0 & \text{en otro caso} \end{cases}$$

Distribución binomial negativa

Una v.a. con este tipo de distribución recibe el nombre de **binomial negativa** y la denotaremos por $BN(n, p)$.

Notemos que $BN(1, p) = Ge(p)$.

Distribución binomial negativa

Demostración

Justifiquemos el resultado. Sea X una $BN(n,p)$ y sea $k = 0, 1, 2, \ldots$.

$$P(X = k) = P(\text{Todas las cadenas de E's y F' con } k \text{ F, con } n \text{ E y acabadas en E})$$

$$\overbrace{\overbrace{\underbrace{EFFF\ldots EEF}_{k \quad \text{Fracasos}}}^{n-1 \quad \text{Éxitos.}}}^{k+n-1 \text{ posiciones}} \quad E$$

De estas cadenas hay tantas como maneras de elegir de entre las $k+n-1$ primeras posiciones $n-1$ para colocar los éxitos. Esta cantidad es el número binomial $\binom{k+n-1}{n-1}$.

Números binomiales negativos

Números binomiales negativos

Dados dos enteros positivos n y k se define el número binomial negativo como

$$\binom{-n}{k} = \frac{(-n)(-n-1)\cdots(-n-k+1)}{k!}.$$

Los números binomiales negativos generalizan la fórmula de Newton para exponentes negativos:

$$(t+1)^{-n} = \sum_{k=0}^{+\infty} \left(\begin{array}{c} -n \\ k \end{array} \right) t^k$$

Números binomiales negativos

`R` usa la función `choose` para calcular números binomiales, sean negativos o no. Veámoslo con un ejemplo:

$$\binom{-6}{4} = \frac{-6\cdot(-6-1)\cdot\cdot(-6-2)\cdot(-6-3)}{4!} = \frac{-6\cdot(-7)\cdot(-8)\cdot(-9)}{24} = \frac{3024}{24} = 126.$$

Si realizamos el cálculo con `R` obtenemos el mismo resultado:

```
choose(-6,4)
```

```
## [1] 126
```

Esperanza de una $BN(n, p)$

Su **esperanza es**

$$E(X) = \sum_{k=0}^{+\infty} k \cdot \binom{k+n-1}{n-1} \cdot (1-p)^k \cdot p^n = n \cdot \frac{1-p}{p}.$$

La **esperanza de** X^2 **es**

$$E(X^2) = \sum_{k=0}^{+\infty} k^2 \cdot \binom{k+n-1}{n-1} \cdot (1-p)^k \cdot p^n = n \cdot \frac{1-p}{p^2} + \left(n \cdot \frac{1-p}{p}\right)^2.$$

Varianza de una $BN(n,p)$

Por último la **varianza es**

$$Var(X) = E(X^2) - E(X)^2 =$$

$$= n\cdot\frac{1-p}{p^2} + \left(n\cdot\frac{1-p}{p}\right)^2 - \left(n\cdot\frac{1-p}{p}\right)^2 = n\cdot\frac{1-p}{p^2}.$$

y por tanto la desviación típica es

$$\sqrt{Var(X)} = \frac{\sqrt{n(1-p)}}{p}$$

Resumen distribución Binomial Negativa $BN(n, p)$

X = Número de fracasos antes de conseguir el n-ésimo éxito, $P(\text{Éxito}) = p$. $BN(n, p)$
$D_X = \{0, 1, 2, 3 \ldots\}$
$P_X(k) = P(X = k) = \begin{cases} \binom{k+n-1}{n-1} \cdot (1-p)^k \cdot p^n, & \text{si } k = 0, 1, \ldots \\ 0, & \text{en otro caso.} \end{cases}$
$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{si } x < 0 \\ \displaystyle\sum_{i=0}^{k} P(X = i) & \text{si } \begin{cases} k \leq x < k+1, \\ k = 0, 1, 2, \ldots \end{cases} \end{cases}$
$E(X) = n \cdot \frac{1-p}{p}$; $Var(X) = n \cdot \frac{1-p}{p^2}$.

Ejemplo puerta dos cerraduras $BN(n = 2, p = 0.1)$.

Ejercicio: Puerta con dos cerraduras

Recordemos nuestra puerta con dos cerraduras que se abren secuencialmente. Tenemos un manojo de 10 llaves casi idénticas de manera que cada vez que probamos una llave olvidamos qué llave hemos usado.

Sea X la v.a que nos da el número de intentos fallidos hasta abrir abrir la puerta.

Ejemplo $BN(n, p)$

Estamos interesado en modelar este problema. La preguntas son:

1. ¿Cuál es la distribución de probabilidad de X la v.a que nos da el número fallos hasta abrir la puerta?
2. ¿Cuál es la función de probabilidad y de distribución de X?
3. ¿Cuál es la probabilidad de fallar exactamente 5 veces antes de abrir la puerta?
4. ¿Cuál es la probabilidad de fallar más de 4?
5. ¿Cuál es el número esperado de fallos? ¿Y su desviación típica?

Ejemplo dos cerraduras $BN(n = 2, p = 0.1)$.

Solución 1. ¿Cuál es la distribución de probabilidad de X la v.a que nos da el número fallos hasta abrir la puerta?

Bajo estados condiciones tenemos que la probabilidad de "éxito" de cada intento es $p = \frac{1}{10} = 0.1$. Como cada vez *olvidamos* qué llave hemos probado, cada intento será independiente del anterior.

Así que la variable X que queremos modelar cuenta el número fallos de repeticiones sucesivas e independientes de un experimento $Ber(p = 0.1)$ hasta conseguir 2 éxitos en un experimento.

Por lo tanto podemos asegurar que X sigue un distribución $BN(n = 2, p = 0.1)$.

Ejemplo $BN(n=2, p=0.1)$

Solución 2. ¿Cuál es la función de probabilidad y de distribución del X?

En general la función de probabilidad de una $BN(n,p)$ es

$$P_X(k) = P(X=k) = \left\{ \begin{array}{cl} \binom{k+n-1}{n-1} \cdot (1-p)^k \cdot p^n & \text{si } k = 0, 1, \ldots \\ 0 & \text{en otro caso.} \end{array} \right.$$

Si aplicamos la expresión anterior para $n=2$ y $p=0.1$, obtenemos:

$$P_X(k) = P(X=k) = \left\{ \begin{array}{cl} \binom{k+2-1}{2-1} \cdot 0.9^k \cdot 0.1^2 & \text{si } k = 0, 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{array} \right.$$

Ejemplo $BN(n = 2, p = 0.1)$

Simplificando

$$P_X(X = k) = P(X = k) = \begin{cases} 0.01 \cdot (k+1) \cdot 0.9^k, & \text{si } k = 0, 1, 2, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

La función de distribución en general es

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < 0 \\ \displaystyle\sum_{i=0}^{k} \binom{i+n-1}{n-1} \cdot (1-p)^{i+n-1} \cdot p^n & \text{si } \begin{cases} k \leq x < k+1 \\ k = 0, 1, 2, \ldots \end{cases} \end{cases}$$

Ejemplo $BN(n=2, p=0.1)$

Simplificando para $n=2,\ p=0.1$.

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{si } x < 0, \\ \displaystyle\sum_{i=0}^{k} 0.01 \cdot (i+1) \cdot 0.9^{i+1}, & \text{si } \begin{cases} k \leq x < k+1, \\ k = 0, 1, 2, \ldots \end{cases} \end{cases}$$

Solución 3. ¿Cuál es la probabilidad de fallar exactamente 5 veces antes de abrir la puerta?

$$P(X=5) = 0.01 \cdot (5+1) \cdot 0.9^5 = 0.06 \cdot 0.9^5 = 0.0354294.$$

Ejemplo $BN(n = 2, p = 0.1)$

Solución 4. ¿Cuál es la probabilidad de fallar más de 4?

Nos piden que

$$P(X > 4) = 1 - P(X \leq 4).$$

Calculemos primero $P(X \leq 4)$:

$$\begin{aligned} P(X \leq 4) &= \sum_{x=0}^{4} P(X = x) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= 0.01 \cdot (0 + 1) \cdot 0.9^0 + 0.01 \cdot (1 + 1) \cdot 0.9^1 + 0.01 \cdot (2 + 1) \cdot 0.9^2 \\ &\quad +0.01 \cdot (3 + 1) \cdot 0.9^3 + 0.01 \cdot (4 + 1) \cdot 0.9^4 \\ &= 0.01 + 0.018 + 0.0243 + 0.02916 + 0.032805 = 0.114265. \end{aligned}$$

Ejemplo $BN(n = 2, p = 0.1)$

Por lo tanto

$$P(X > 4) = 1 - P(X \leq 4) = 1 - 0.114265 = 0.885735.$$

Solución 5. ¿Cuál es el número esperado de fallos? ¿Y su desviación típica?

Como X sigue una ley $BN(n = 2, p = 0.1)$

$$E(X) = n \cdot \frac{1-p}{p} = 2 \cdot \frac{1-0.1}{0.1} = 18.$$

El número de fallos esperado es 18. La varianza es

$$Var(X) = n \cdot \frac{1-p}{p^2} = 2 \cdot \frac{1-0.1}{0.1^2} = 180,$$

y su desviación típica $\sqrt{180} = 13.41641$.

Cálculos con R

La función de `R` que calcula la función de probabilidad de la binomial negativa con sus parámetros básicos es:

`dnbinom(x, size, prob,...)`‘

donde `size` (n) es el número de éxitos y `prob` (p), la probabilidad de éxito.

Así en el ejemplo de la puerta con dos cerraduras, X es una $BN(n = size = 2, p = prob = 0.1)$. Por ejemplo, $P(X = 5)$ que hemos calculado en el ejemplo anterior, vale:

```
dnbinom(5,size=2,p=0.1)
```

```
## [1] 0.0354294
```

Cálculos con R

De forma similar calculamos calculamos $P(X \leq 4)$, $P(X > 4) = 1 - P(X \leq 4)$ y $P(X > 4)$.

```
pnbinom(4,size=2,p=0.1)
```

```
## [1] 0.114265
```

```
1-pnbinom(4,size=2,p=0.1)
```

```
## [1] 0.885735
```

```
pnbinom(4,size=2,p=0.1,lower.tail=FALSE)
```

```
## [1] 0.885735
```

Cálculos con python

La función con python es `nbinom.pmf(k, n, p, loc)`. Hay que cargarla desde `scpi.stats`

```
from scipy.stats import nbinom
```

Recordemos que de nuevo se cumple que

```
nbinom.pmf(k, n, p, loc) = nbinom.pmf(k-loc, n, p)`
```

Cálculos $BN(n, p)$ con python

```
nbinom.pmf(k=5,n=2,p=0.1)
```

```
## 0.0354294
```

```
nbinom.pmf(k=5,n=2,p=0.1,loc=0)
```

```
## 0.0354294
```

```
nbinom.cdf(k=4,n=2,p=0.1)
```

```
## 0.11426500000000002
```

```
1-nbinom.cdf(k=4,n=2,p=0.1)
```

```
## 0.8857349999999999
```

Cálculos $BN(n,p)$ con python

Generemos 100 observaciones aleatorias de una $BN(n=2,0.1)$. Es decir serán las veces que hemos fallado hasta abrir la puerta 100 veces.

```
nbinom.rvs(n=2, p=0.1, size=100)
```

```
## array([ 7, 16, 12, 91,  4, 12, 93, 13,  5, 28, 19, 23, 48, 12, 36, 14, 25,
##        19,  3, 25, 63, 25,  6,  9, 11, 39,  7, 24, 13, 12, 13, 13,  1,  9,
##         8, 32, 15, 23,  1, 10, 11, 39, 11, 24, 20, 22,  6, 17,  3,  1, 43,
##        39,  4,  2, 59,  9,  2, 12,  7, 15, 34,  9,  4, 38, 43,  5, 19,  5,
##        10, 12, 24,  9, 14,  8,  1, 52, 13, 11, 19, 26, 26, 11,  7, 26, 25,
##        79, 14, 14,  3,  9, 27, 17, 12, 13,  3,  8, 15, 16, 32, 24],
##       dtype=int64)
```

Cálculos $BN(n, p)$ con python

La **esperanza** y la **varianza**de una $BN(n = 2, 0.1)$ valen:

```
n, p=2,0.1
params = nbinom.stats(n,p,moments='mv')
print("E(X)={m}".format(m=params[0]))
```

```
## E(X)=18.0
```

```
print("Var(X)={v}".format(v=params[1]))
```

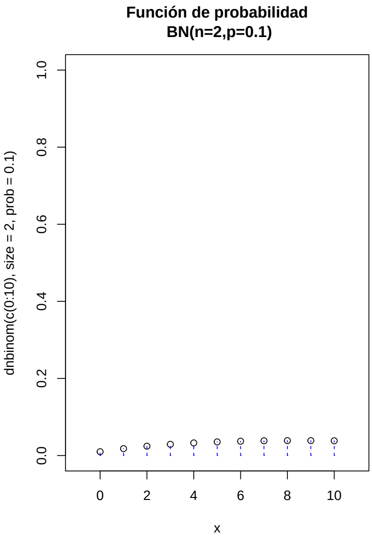
```
## Var(X)=179.99999999999997
```

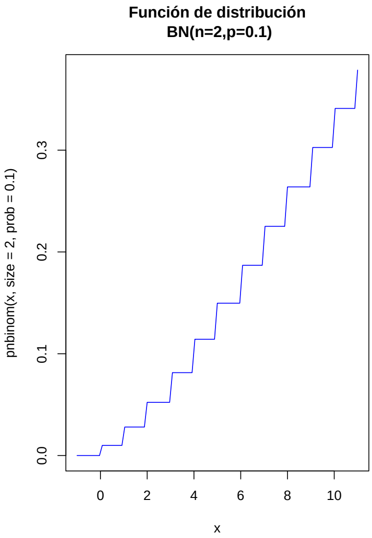
Gráficas de la binomial negativa con R

El siguiente código de R dibuja las función de probabilidad y la de distribución de una $BN(n = 2, p = 0.1)$

```
par(mfrow=c(1,2))
aux=rep(0,22)
aux[seq(2,22,2)]=dnbinom(c(0:10),size=2,prob=0.1)
plot(x=c(0:10),y=dnbinom(c(0:10),size=2,prob=0.1),
  ylim=c(0,1),xlim=c(-1,11),xlab="x",
  main="Función de probabilidad\n BN(n=2,p=0.1)")
lines(x=rep(0:10,each=2),y=aux, type = "h", lty = 2,col="blue")
curve(pnbinom(x,size=2,prob=0,1),
  xlim=c(-1,11),col="blue",
  main="Función de distribución\n BN(n=2,p=0.1)")
par(mfrow=c(1,1))
```

Gráficas de la binomial negativa con R





Gráficos de la binomial negativa con python

Ejercicio

Buscad en los manuales de python cómo se dibuja la función de probabilidad y de distribución de una binomial. negativa

Necesitamos de nuevo más librerías

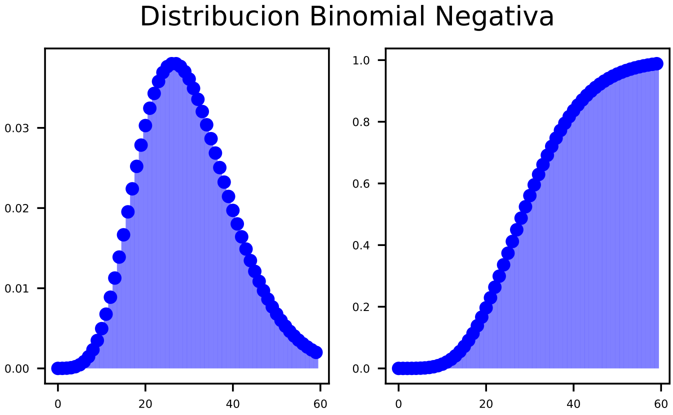
```
import numpy as np
from scipy.stats import nbinom
import matplotlib.pyplot as plt
```

Gráficos de la binomial negativa con python

```
n, p = 10, 0.25
x = np.arange(0,nbinom.ppf(0.99, n, p))
fig =plt.figure(figsize=(5, 2.7))
ax = fig.add_subplot(1,2,1)
ax.plot(x, nbinom.pmf(x, n, p), 'bo', ms=5, label='nbinom pmf')
ax.vlines(x, 0, nbinom.pmf(x, n, p), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  tick.label.set_fontsize(5)
ax = fig.add_subplot(1,2,2)
ax.plot(x, nbinom.cdf(x, n, p), 'bo', ms=5, label='nbinom pmf')
ax.vlines(x, 0, nbinom.cdf(x, n, p), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
```

Gráficos de la binomial negativa con python

```
## <string>:2: MatplotlibDeprecationWarning: The label function was deprecatec
```



Ejercicio: Acceso aleatorio a un sistema con triple clave.

Sistema con tres claves de acceso

Supongamos que tenemos un sistema informático tiene un programa de seguridad que genera accesos con claves de 3 dígitos $000, 001, \ldots 999$. En total 1000 posibilidades.

Como una clave de tres dígitos es fácil de romper proponemos considerar tres claves consecutivas de acceso al sistema, cada una de 3 dígitos.

Para acceder al sistema hay que dar las tres claves de forma consecutiva y por orden.

Es decir hasta que no averiguamos la primera clave no pasamos a la segunda clave.

Supongamos que cada vez que ponemos las dos claves olvidamos el resultado y seguimos poniendo claves al azar hasta adivinar la contraseña.

Así hasta conseguir entrar en el sistema.

Sea X la v.a que nos da el número de fallos antes de entrar en el sistema.

Ejercicio acceso aleatorio a un sistema con triple clave.

Estamos interesados en modelar este problema. La preguntas son:

1. ¿Cuál es la distribución de probabilidad de X, la v.a que nos da el número de fallos antes de acceder al sistema.
2. ¿Cuál es la función de probabilidad y de distribución del X?
3. ¿Cuál es la probabilidad de fallar 150 veces antes de acceder en el sistema?
4. ¿Cuál es la probabilidad de fallar más de 150 veces antes de entrar en el sistema?
5. ¿Cuál es el número esperado de fallos antes de acceder al sistema? ¿Y su varianza?

Ejemplo $BN(r,p)$

Solución 1. ¿Cuál es la distribución de probabilidad de X, la v.a que nos da el número de fallos antes de acceder al sistema?

Bajo estados condiciones tenemos que la probabilidad de “éxito” de cada intento es $p=\frac{1}{1000}=0.001$. Y como cada vez *olvidamos* en los dígitos cada intento será independiente del anterior.

Así que la variable X cuenta el número de fracasos independientes hasta conseguir 3 éxitos en un experimento $Ber(p=0.001)$ por lo tanto X sigue un distribución $BN(n=3,p=0.001)$.

Ejemplo $BN(r,p)$

Solución 2. ¿Cuál es la función de probabilidad y de distribución del X

En general la función de probabilidad de una $BN(n,p)$ es

$$P_X(X=x)=P(X=x)=\begin{cases} \binom{x+n-1}{n-1}\cdot(1-p)^x\cdot p^n & \text{si } x=0,1,\ldots \\ 0 & \text{en otro caso.}\end{cases}$$

En particular la función de probabilidad de una $BN(n=3,p=0.001)$ es

$$P_X(X=x)=P(X=x)=\begin{cases} \binom{x+2}{2}\cdot 0.999^x\cdot 0.001^3 & \text{si } x=0,1,2,\ldots \\ 0 & \text{en otro caso.}\end{cases}$$

Solución ejemplo $BN(n = 3, p = 0.001)$

Solución 3. ¿Cuál es la probabilidad de fallar 150 veces antes de acceder en el sistema?

Nos piden

$$P(X = 150) = \binom{152}{2} \cdot 0.999^{150} \cdot 0.001^3$$

Lo calcularemos operando con R

```
choose(152,2)*0.999^150*0.001^3
```

```
## [1] 9.876743e-06
```

con la función de R

```
dnbinom(150,size=3,p=0.001)
```

```
## [1] 9.876743e-06
```

Solución ejemplo $BN(n = 3, p = 0.001)$

Solución 3. ¿Cuál es la probabilidad de fallar 150 veces antes de acceder en el sistema?

Nos piden

$$P(X = 150) = \binom{152}{2} \cdot 0.999^{150} \cdot 0.001^3$$

Pero también lo podemos hacer con python

```
from  scipy.special import binom
binom(152,2)*0.999**150*0.001**3
```

```
## 9.876743459670526e-06
```

```
nbinom.pmf(150,n=3,p=0.001)
```

```
## 9.876743459670532e-06
```

Solución ejemplo $BN(n, p)$

Solución 4. ¿Cuál es la probabilidad de fallar más de 150 veces antes de entrar en el sistema?

$$P(X > 150) = 1 - P(X \leq 150)$$

Calculemos $P(X \leq 150)$

$$\begin{aligned} P(X \leq 150) &= P(X = 0) + P(X = 1) + P(X = 2) + \ldots + P(X = 150) \\ &= \sum_{k=0}^{150} \binom{k+3-1}{3-1} \cdot (0.999)^k \cdot 0.001^3 \ldots = \ldots = 5.2320035 \times 10^{-4} \end{aligned}$$

Solución ejemplo $BN(n, p)$

Con R

```
pnbinom(150,3,0.001)
```

```
## [1] 0.0005232003
```

Con python

```
nbinom.cdf(150,n=3,p=0.001)
```

```
## 0.0005232003490824064
```

El valor pedido será pues:

$$P(X > 150) = 1 - P(X \leq 150) = 1 - 5.2320035 \times 10^{-4} = 0.9994768.$$

Vemos que es muy probable que fallemos más de 150 veces antes de entrar en el sistema.

Solución ejemplo $BN(n, p)$

Solución 5. ¿Cuál es el número esperado de fallos antes de acceder al sistema? ¿Y su desviación típica?

$$E(X) = n \cdot \frac{1-p}{p} = 3 \cdot \frac{1-0.001}{0.001} = 2997.$$

$$Var(X) = n \cdot \frac{1-p}{p^2} = 3 \cdot \frac{1-0.001^2}{0.001^2} = 2.997 \times 10^6.$$

Con python

```
params = nbinom.stats(n=3,p=0.001,moments='mv')
print("E(X) = {m}".format(m=params[0]))
```

```
## E(X) = 2997.0
```

```
print("Var(X) = {v}".format(v=params[1]))
```

¿Tres claves de tres dígitos o una de 9 dígitos?

Ejercicio

Supongamos que ponemos una sola clave de 9 dígitos. Estudiemos en este caso la variable aleatoria que da el número de fallos antes de entrar en el sistema y comparemos los resultados.

Si seguimos suponiendo que cada vez ponemos la contraseña al azar pero esta vez con una clave de 9 dígitos. La probabilidad de éxito será ahora $p = \frac{1}{10^9}$.

Si llamamos X_9 a la variable aleatoria que nos da el número de fallos antes de entra en el sistema seguirá una distribución $Ge(p = \frac{1}{10^9} = 0.000000001)$.

Qué da más seguridad ¿tres claves de tres dígitos o una de 9 dígitos?

Su valor esperado es

$$E(X_9) = \frac{1-p}{p} = \frac{1-0.000000001}{0.000000001} = 10 \times 10^8.$$

1000000000 son 1000 millones de fallos esperados hasta abrir la puerta.

Recordemos que con tres contraseñas de 3 dígitos el valor esperado de fallos es

$$3 \cdot \frac{1-0.001}{0.001} = 2997.$$

Por lo tanto, desde el punto de vista de la seguridad, es mejor una clave larga de 9 dígitos que tres cortas si escribimos las contraseñas al azar.

Lección 6

Distribución de Poisson

Distribución Poisson

Diremos que una v.a. discreta X con $X(\Omega) = \mathbf{N}$ tiene distribución de Poisson con parámetro $\lambda > 0$, y lo denotaremos por $Po(\lambda)$ si su función de probabilidad es:

$$P_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{si } x = 0, 1, \ldots \\ 0 & \text{en otro caso} \end{cases}.$$

Distribución Poisson

Usando que el desarrollo en serie de Taylor de la función exponencial es

$$e^{\lambda} = \sum_{x=0}^{+\infty} \frac{\lambda^x}{x!},$$

es fácil comprobar que la suma de la función de probabilidad en todos los valores del dominio de X, o sea, los enteros positivos, vale 1.

Además recordemos que dado $x \in \mathbb{R} - \{0\}$ se tiene que

$$\lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

Distribución Poisson

Usando la expresión anterior para $x = -\lambda$, tenemos:

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \lim_{n\to\infty}\left(1+\frac{-\lambda}{n}\right)^n = e^{-\lambda}.$$

La distribución de Poisson como “límite” de una binomial.

La distribución de Poisson (Siméon Denis Poisson) aparece en el conteo de determinados eventos que se producen en un intervalo de tiempo o en el espacio.

Supongamos que nuestra variable de interés es X, el número de eventos en el intervalo de tiempo $(0, t]$, como por ejemplo el número de llamadas a un *call center* en una hora donde suponemos que se cumplen las siguientes condiciones:

La distribución Poisson como "límite" de una binomial.

1. El número promedio de eventos en el intervalo $(0, t]$ es $\lambda > 0$.
2. Es posible dividir el intervalo de tiempo en un gran número de subintervalos (denotemos por n al número de intervalos) de forma que:
   - La probabilidad de que se produzcan dos o más eventos en un subintervalo es despreciable.
   - El número de ocurrencias de eventos en un intervalo es independiente del número de ocurrencias en otro intervalo.
   - La probabilidad de que un evento ocurra en un subintervalo es $p_n = \frac{\lambda}{n}$ ·

La distribución Poisson como "límite" de una binomial.

Bajo estas condiciones, podemos considerar que el número de eventos en el intervalo $(0, t]$ será el número de "éxitos" en n repeticiones independientes de un proceso Bernoulli de parámetro p_n

Entonces si $n \to \infty$ y $p_n \cdot n$ se mantiene igual a λ resulta que la función de probabilidad de X se puede escribir como

La distribución Poisson como "límite" de una binomial.

$$
\begin{aligned}
P(X_n = k) &= \left(\begin{array}{c} n \\ k \end{array} \right) \cdot p_n^k \cdot (1 - p_n)^{n-k} \\
&= \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \cdot \frac{n!}{(n-k)! \cdot n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}.
\end{aligned}
$$

La distribución Poisson como "límite" de una binomial.

Si hacemos tender n hacia ∞, obtenemos:

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \frac{\lambda^k}{k!} \cdot \frac{n!}{(n-k)! \cdot n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}.$$

Calculemos el límite de algunos de los factores de la expresión

$$\lim_{n\to\infty} \frac{n!}{(n-k)! \cdot n^k} = \lim_{n\to\infty} \frac{n \cdot (n-1) \cdots (n-k-1)}{n^k} = \lim_{n\to\infty} \frac{n^k + \cdots}{n^k} = 1.$$

La distribución Poisson como "límite" de una binomial.

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=e^{-\lambda}$$

Y también teniendo en cuanta que k es constante.

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}=\lim_{n\to\infty}1^{-k}=\lim_{n\to\infty}1=1.$$

La distribución Poisson como “límite” de una binomial.

Para acabar

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} \left(\begin{array}{c} n \\ k \end{array}\right) \cdot p_n^k \cdot (1 - p_n)^{n-k} = \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^k}{k!} \cdot e^{-\lambda}.$$

Lo que confirma que límite de una serie de variables $B(n, p_n = \frac{\lambda}{n})$ sigue una ley $Po(\lambda)$.

Procesos de Poisson

Lo interesante de las variables Poisson es que podemos modificar (si el modelo lo permite) el intervalo de tiempo $(0,t]$ en el que contamos los eventos.

Claro que esto no tiene que poder ser así.

Pero en general si la variable es poisson en $(0,t]$ también lo será en cualquier subintervalo $(0,t']$ para todo t' tal que $0<t'<t$.

Así que podremos definir una serie de variables X_t de distribución $Po(\lambda\cdot t)$.

Procesos de Poisson

Definición procesos de Poisson

Consideremos un experimento *Poisson* con λ igual al promedio de eventos en una unidad de tiempo (u.t.).

Si t es una cantidad de tiempo en u.t., la v.a. X_t=numero de eventos en el intervalo $(0, t]$ es una $Po(\lambda \cdot t)$.

El conjunto de variables $\{X_t\}_{t>0}$ recibe el nombre de **proceso de Poisson**.

Resumen distribución Poisson $X \equiv Po(\lambda)$

X con distribución Poisson de media o promedio λ, $Po(\lambda)$
$D_X = \{0, 1, \ldots\}$
$P_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{si } x = 0, 1, \ldots \\ 0 & \text{en otro caso.} \end{cases}$
$F_X(x)=P(X\leq X)=\begin{cases} 0 & \text{si } x < 0 \\ \sum_{i=0}^{k} P(X=i)=\sum_{i=0}^{k} \frac{\lambda^i}{i!}\cdot e^{-\lambda} & \text{si } \begin{cases} k\leq x<k+1 \\ k=0,1,2,\ldots \end{cases} \end{cases}$
$E(X) = \lambda;\ Var(X) = \lambda$

Resumen proceso Poisson $X_t \equiv Po(\lambda \cdot t)$

X_t = número de eventos en el intervalo $(0, t]$ $Po(\lambda \cdot t)$ donde λ promedio por u.t.
$D_X = \{0, 1, \ldots\}$
$P_X(x) = P(X = x) = \begin{cases} \frac{(\lambda \cdot t)^x}{x!} e^{-\lambda \cdot t} & \text{si } x = 0, 1, \ldots \\ 0 & \text{en otro caso.} \end{cases}$
$F_X(x) = P(X \leq X) = \begin{cases} 0 & \text{si } x < 0 \\ \sum_{i=0}^{k} P(X=i) = \sum_{i=0}^{k} \frac{(\lambda \cdot t)^i}{i!} \cdot e^{-\lambda \cdot t} & \text{si } \begin{cases} k \leq x < k+1 \\ k = 0,1,2,\ldots \end{cases} \end{cases}$
$E(X) = \lambda \cdot t$; $Var(X) = \lambda \cdot t$

Aproximación de la distribución binomial por la Poisson

Bajo el punto de vista anterior y si p es pequeño y n suficientemente grande la distribución $B(n,p)$ se aproxima a una $Po(\lambda = n\cdot p)$.

Existen distintos criterios (ninguno perfecto) de cuando la aproximación es buena.

Por ejemplo si

$$n \geq 20 \text{ o mejor } n\geq 30, n\cdot p < 10 \text{ y } p \leq 0.05,$$

la aproximación de una $B(n,p)$ por una $Po(n\cdot p)$ es buena. Sobre todo para los valores cercanos a $E(X)=\lambda$.

Condición deseable $n\geq 20$, $n\cdot p<10$, $p\leq 0.05$.

Ejemplo $Po(\lambda)$

Ejemplo: Trampa insectos.

La conocida lámpara antiinsectos o insecticida eléctrico atrae a los insectos voladores con una luz ultravioleta y los mata por electrocución.

Consideremos la v.a. X que cuenta el número de insectos caídos en la trampa en una hora. Supongamos que el número promedio de insectos que captura la trampa en una hora es $E(X) = 20$ y que podemos admitir que X sigue una ley de probabilidad $Po(\lambda = 20)$.

Nos piden

1. Comentar de forma breve si se cumplen intuitivamente las condiciones para tener una distribución Poisson.
2. Escribir de forma explicita la función de probabilidad y de distribución de X.
3. Calculad la probabilidad de que en una hora caigan en la trampa exactamente 21 insectos.
4. Calculad la probabilidad de que en una hora caigan en la trampa al menos 6 insectos.
5. ¿Cuál es el valor esperando, la varianza y la desviación típica de X?

Ejemplo $Po(\lambda)$

Solución 1. Comentar de forma breve si se cumplen intuitivamente las condiciones para tener una distribución Poisson.

1. El número promedio de eventos en el intervalo $(0,1]$, una hora es $\lambda = 20 > 0$.
2. Es posible dividir el intervalo de tiempo de una hora en un gran número de subintervalos (denotemos por n al número de intervalos) de forma que:
   - La probabilidad de que se produzcan dos o más electrocuciones un subintervalo es despreciable. No es posible que dos mosquitos se electrocuten al mismo tiempo.
   - El número de ocurrencias, electrocuciones de insectos, en un intervalo es independiente del número de electrocuciones en otro intervalo.
   - La probabilidad de que un evento ocurra en un subintervalo es $p_n = \frac{\lambda}{n}\cdot$ Podemos dividir los 20 insectos promedio entre los n intervalos (trozo de hora) de forma que $p_n = \frac{\lambda}{n}$.
   - Por ejemplo si $n = 60$ tenemos que $p_n = \frac{20}{60} = \frac{1}{3}$. La probabilidad de que en un minuto la trampa chisporrotee es $\frac{1}{3}$.

Ejemplo $Po(\lambda)$

Solución 2. Escribid de forma explicita la función de probabilidad y de distribución de X.

La distribución de probabilidad de un $Po(\lambda)$ es

$$P_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{si } x = 0, 1, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

En nuestro caso, $\lambda = 20$:

$$P_X(x) = P(X = x) = \begin{cases} \frac{20^x}{x!} e^{-20} & \text{si } x = 0, 1, \ldots \\ 0 & \text{en otro caso.} \end{cases}$$

Ejemplo $Po(\lambda)$

La función de distribución es

$$F_X(x) = P(X \leq X) = \begin{cases} 0 & \text{si } x < 0 \\ \displaystyle\sum_{i=0}^{k} P(X = i) = \sum_{i=0}^{k} \frac{\lambda^i}{i!} \cdot e^{-\lambda} & \text{si } \begin{cases} k \leq x < k+1 \\ k = 0, 1, 2, \ldots \end{cases} \end{cases}$$

En nuestro caso

$$F_X(x) = P(X \leq X) = \begin{cases} 0 & \text{si } x < 0 \\ \displaystyle\sum_{i=0}^{k} P(X = i) = \sum_{i=0}^{k} \frac{20^i}{i!} \cdot e^{-20} & \text{si } \begin{cases} k \leq x < k+1 \\ k = 0, 1, 2, \ldots \end{cases} \end{cases}$$

Ejemplo $Po(\lambda)$

Solución 3. Calculad la probabilidad de que en una hora caigan en la trampa exactamente 21 insectos.

Nos piden la probabilidad siguiente:

$$P(X = 21) = \frac{20^{21}}{21!} e^{-20} = 0.0846051.$$

Para realizar el cálculo anterior, podemos usar `R` como calculadora o usar la función `dpois` que nos calcula la función de distribución de la variable de Poisson:

```
20^21/factorial(21)*exp(-20)
```

```
## [1] 0.08460506
```

```
dpois(21,lambda = 20)
```

```
## [1] 0.08460506
```

Ejemplo $Po(\lambda)$

Solución 4. Calculad la probabilidad de que en una hora caigan en la trampa al menos 6 insectos.

Nos piden la probabilidad siguiente:

$$
\begin{aligned}
P(X \geq 6) &= 1 - P(X < 6) = 1 - P(X \leq 5) = 1 - F_X(5) = 1 - \sum_{x=0}^{5} \frac{20^x}{x!} \cdot e^{-20} \\
&= 1 - \left(\frac{20^0}{0!} \cdot e^{-20} + \frac{20^1}{1!} \cdot e^{-20} + \frac{20^2}{2!} \cdot e^{-20} + \frac{20^3}{3!} \cdot e^{-20} + \frac{20^4}{4!} \cdot e^{-20} + \frac{20^5}{5!} \cdot e^{-20} \right) \\
&= 1 - e^{-20} \cdot \left(1 + 20 + \frac{400}{4} + \frac{8000}{6} + \frac{160000}{24} + \frac{3200000}{120}\right) \\
&= 1 - e^{-20} \cdot \left(\frac{1 \cdot 120 + 20 \cdot 120 + 400 \cdot 30 + 8000 \cdot 20 + 160000 \cdot 24 + 3200000 \cdot 1}{120} \right) \\
&= 1 - e^{-20} \cdot \left(\frac{4186520}{120} \right) = 1 - 7.1908841 \times 10^{-5} = 0.9999281.
\end{aligned}
$$

Ejemplo $Po(\lambda)$

Solución 5. ¿Cuál es el valor esperado, la varianza y la desviación típica de X?

El valor esperado del número de insectos caídos en la trampa en una hora es

$$E(X) = \lambda = 20$$

Su varianza es

$$Var(X) = \lambda = 20$$

y su desviación típica vale

$$\sqrt{Var(X)} = +\sqrt{\lambda} = +\sqrt{20} = 4.47214.$$

Cálculos con R

Consideremos por ejemplo una v.a. X con distribución $Po(\lambda = 3)$. Calculemos $P_X(0) = P(X = 0), P_X(1) = P(X = 1)$ con R:

```
dpois(0,lambda = 3)
```

```
## [1] 0.04978707
```

```
dpois(1,lambda = 3)
```

```
## [1] 0.1493612
```

Cálculos con R

Si quisiéramos hallar la función de distribución en los mismos valores anteriores, $F_X(0) = P(X \leq 0), F_X(1) = P(X \leq 1)$, haríamos lo siguiente:

```
ppois(0,lambda = 3)
```

```
## [1] 0.04978707
```

```
ppois(1,lambda = 3)
```

```
## [1] 0.1991483
```

```
dpois(0,lambda = 3)+dpois(1,lambda = 3) ## es igual a ppois(1,lambda=3)
```

```
## [1] 0.1991483
```

Cálculos con R

A continuación, comprobemos que $F_X(10) = \sum_{x=0}^{10} P_X(x)$:

```
dpois(0:10,3)
```

```
##  [1] 0.0497870684 0.1493612051 0.2240418077 0.2240418077 0.1680313557
##  [6] 0.1008188134 0.0504094067 0.0216040315 0.0081015118 0.0027005039
## [11] 0.0008101512
```

```
sum(dpois(0:10,3))
```

```
## [1] 0.9997077
```

```
ppois(10,3)
```

```
## [1] 0.9997077
```

Cálculos distribución Poisson con R

Si quisiéramos generar una secuencia de 100 observaciones para una distribución de Poisson de parámetro $\lambda = 3$, $Po(3)$, tendríamos que hacer:

```
rpois(n=100,lambda = 3)
```

```
##   [1] 2 5 3 3 2 2 5 2 4 4 2 3 2 2 2 2 2 3 3 5 3 3 2 4 2 3 2 1 1 3 4 6 2 5
##  [38] 1 6 3 4 1 4 3 4 3 0 2 1 4 3 0 2 4 2 3 5 2 1 3 3 4 2 5 0 3 1 1 4 6 4
##  [75] 0 3 3 3 4 1 2 6 2 2 2 2 1 2 5 2 5 3 7 3 5 2 3 2 1 3
```

Cálculos con R

Ejercicio de la trampa para insectos (continuación)

En el ejercicio de la trampa para insectos teníamos que X es una $Po(20)$. Responded con R a la preguntas 3 y 4 de este ejercicio

Pregunta 3. Calculad la probabilidad de que en una hora caigan en la trampa exactamente 21 insectos.

Recordemos que la probabilidad pedida es $P(X=21)$:

```
dpois(21,lambda=20)# P(X=21)
```

```
## [1] 0.08460506
```

Cálculos con R

Pregunta 4. Calculad la probabilidad de que en una hora caigan en la trampa al menos 6 insectos.

Recordemos que la probabilidad pedida es $P(X \geq 6) = 1 - P(X < 6) = 1 - P(X \leq 5)$:

```
ppois(5,lambda=20)
```

```
## [1] 7.190884e-05
```

```
1-ppois(5,lambda=20) # es 1-P(X<=5)=P(X>=6)
```

```
## [1] 0.9999281
```

```
ppois(5,lambda=20,lower.tail =FALSE ) # acumula hacia arriba
```

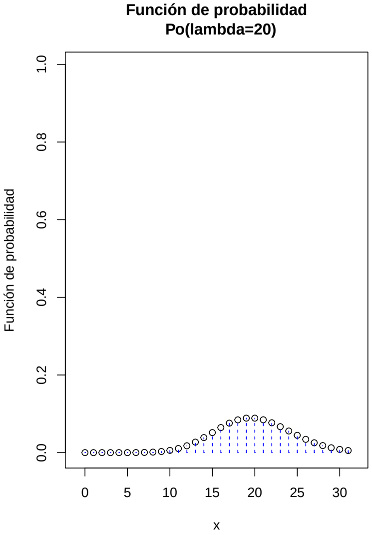
```
## [1] 0.9999281
```

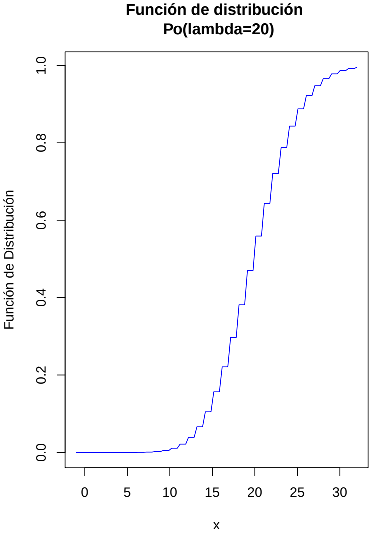
```
# P(X>5)=P(X>=6)=P(X=6)+P(X=7)+
```

Gráficos de la distribución Poisson con R

```
lambda=20; par(mfrow=c(1,2)); n=qpois(0.99,lambda=lambda)
aux=rep(0,(n+1)*2); aux[seq(2,(n+1)*2,2)]=dpois(c(0:n),lambda=lambda)
ymax=max(ppois(0:n,lambda=lambda))
plot(x=c(0:n),y=dpois(c(0:n),lambda=lambda),
     ylim=c(0,ymax),xlim=c(-1,n+1),xlab="x", ylab="Función de probabilidad",
     main=paste0(c("Función de probabilidad\n  Po(lambda=",lambda,")")
                 collapse = ""))
lines(x=rep(0:n,each=2),y=aux,pch=21, type = "h", lty = 2,col="blue")
curve(ppois(x,lambda=lambda),
      xlim=c(-1,n+1),col="blue",ylab="Función de Distribución",
      main=paste0(c("Función de distribución \n Po(lambda=",lambda,")"),
                  collapse = ""))
par(mfrow=c(1,1))
```

Gráficos de la distribución Poisson con R





Cálculos con python

Sea X un una v.a. $Po(\lambda = 3)$. Entonces

$P_X(0) = P(X = 0), P_X(1) = P(X = 1)$ en este orden son

```
from scipy.stats import poisson
poisson.pmf(0,mu = 3)
```

```
## 0.049787068367863944
```

```
poisson.pmf(1,mu = 3)
```

```
## 0.14936120510359185
```

Cálculos con python

Sea X un una v.a. $Po(\lambda = 3)$. Entonces

$F_X(0) = P(X \leq 0), F_X(1) = P(X \leq 1)$ en este orden son

```
poisson.cdf(0,mu = 3)
```

```
## 0.04978706836786395
```

```
poisson.cdf(1,mu = 3)
```

```
## 0.1991482734714558
```

```
poisson.pmf(0,mu = 3)+poisson.pmf(1,mu= 3)
## es igual a poisson.cdf(1,lambda=3)
```

```
## 0.1991482734714558
```

Cálculos con python

Por ejemplo podemos comprobar que $F_X(10) = \sum_{0}^{10} P_X(x)$

```
poisson.pmf(range(0,10),mu=3)
```

```
## array([0.04978707, 0.14936121, 0.22404181, 0.22404181, 0.16803136,
##        0.10081881, 0.05040941, 0.02160403, 0.00810151, 0.0027005 ])
```

```
sum(poisson.pmf(range(0,10),mu=3))
```

```
## 0.9988975118698846
```

```
poisson.cdf(10,mu=3)
```

```
## 0.9997076630493527
```

Cálculos con python

En el ejercicio de la trampa para insectos teníamos que X es una $Po(20)$. Responded con python a la preguntas 3 y 4 de este ejercicio

Pregunta 3. Calculad la probabilidad de que en una hora caigan en la trampa exactamente 21 insectos.

La respuesta a la pregunta 3 es calcular $P(X = 21)$

```
poisson.pmf(21,mu=20)
# P(X=21)
```

```
## 0.08460506418293791
```

Cálculos con python

Pregunta 4. Calculad la probabilidad de que en una hora caigan en la trampa al menos 6 insectos.

La pregunta 4 nos pide calcular $P(X \geq 6) = 1 - P(X \leq 5)$

```
1-poisson.cdf(5,mu=20)
# es 1-P(X<=5)=P(X>=6)
```

```
## 0.9999280911594716
```

Cálculos con python

Como ya hemos visto con `scipy.stats` podemos pedir los momentos de una variable aleatoria $Po(3)$

```
poisson.stats(mu=3, moments='mv')
```

```
## (3.0, 3.0)
```

Y también generar secuencias de observaciones aleatorias de una población $Po(3)$

```
poisson.rvs(mu=3,size=40)
```

```
## array([2, 2, 3, 2, 3, 2, 3, 5, 2, 3, 6, 5, 3, 9, 4, 4, 4, 1, 2, 3, 8, 3,
##        5, 1, 4, 3, 5, 6, 5, 2, 3, 2, 3, 2, 4, 0, 1, 1, 5, 0], dtype=int64)
```

Gráficos con python

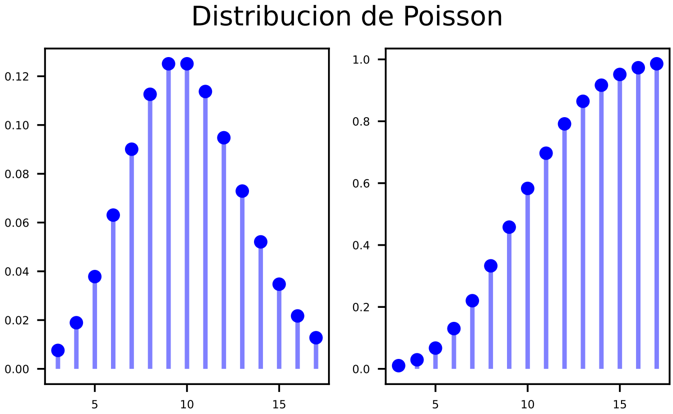
```
from scipy.stats import poisson
mu = 10 # mu = lambda
x = np.arange(poisson.ppf(0.01, mu),poisson.ppf(0.99, mu))
fig =plt.figure(figsize=(5, 2.7))
ax = fig.add_subplot(1,2,1)
ax.plot(x, poisson.pmf(x, mu), 'bo', ms=5, label='poisson pmf')
ax.vlines(x, 0, poisson.pmf(x, mu), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  tick.label.set_fontsize(5)
```

Gráficos con python

```
ax = fig.add_subplot(1,2,2)
ax.plot(x, poisson.cdf(x, mu), 'bo', ms=5, label='poisson cdf')
ax.vlines(x, 0, poisson.cdf(x, mu), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
  tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  tick.label.set_fontsize(5)
fig.suptitle('Distribucion de Poisson')
plt.show()
```

Gráficos con python

```
## <string>:2: MatplotlibDeprecationWarning: The label function was deprecated
```



Ejemplo proceso Poisson

Número de impactos de insectos en la visera de un casco

Un colega de trabajo, al que llamaremos JG, es muy aficionado a los grandes premios de velocidad tanto en coches como en motos.

Como es tan aficionado está obsesionado con muchas de las más extravagantes estadísticas de estos deportes. En particular le propusimos que estudiara el número de insectos que chocan contra la visera de un casco de un motorista GP o de un conductor de fórmula 1 .

Ejemplo proceso Poisson

La idea es que el número de insectos está igualmente repartido por todo el circuito y de promedio impactan $\lambda > 0$ insectos por minuto. También es razonable suponer que:

- podemos dividir la superficie de la visera en cuadrados suficientemente pequeños de forma que la probabilidad de que caigan dos insectos en la misma zona es prácticamente 0.
- la probabilidad de que un insecto impacte en un cuadrado cualquiera de la visera es independiente de cualquier otro cuadrado.
- si hemos dividido la visera en n cuadrados la probabilidad p_n de impacto de un cuadrado vale $p_n = \frac{\lambda}{n}$.

Bajo estas condiciones, si denotamos por X_t como el número de insectos que ha impactado en la visera en el intervalo $(0, t]$ (en t minutos), podemos afirmar que X_t es un proceso de Poisson $Po(\lambda \cdot t)$.

Ejemplo proceso Poisson

Supongamos que nos dicen que $\lambda = 3$ insectos por minuto. Entonces el proceso de poisson X_t seguirá un ley $Po(3 \cdot t)$.

Ahora estamos en condiciones de preguntar al proceso de Poisson.

¿Cuál es la probabilidad de que en 10 minutos impacten más de 25 insectos?

En este caso $t = 10$ $X_{10} =$ número de insectos que impactan en 10 minutos, el intervalo $[0, 10)$ que sigue una $P(3 \cdot 10 = 30)$. Por lo tanto

$$P(X > 25) = 1 - P(X \leq 25)$$

lo resolvemos con R

```
1-ppois(25,lambda=30)
```

```
## [1] 0.7916426
```

Ejemplo proceso Poisson

Otra pregunta interesante es que tengamos que esperar más de 2 minutos para observar el primer impacto

$$P(X_2 = 0) = \frac{(3 \cdot 2)^0}{0!} \cdot e^{-3 \cdot 2} = e^{-6} = 0.002479.$$

Con R

```
6^0/factorial(0)*exp(-6)
```

```
## [1] 0.002478752
```

```
ppois(0,lambda=3*2)
```

```
## [1] 0.002478752
```

Lección 7

Distribución hipergeométrica

Modelo de la distribución hipergeométrica

Supongamos que disponemos de una urna de de sorteos que contiene m bolas blancas y n bolas rojas.

En total en esta urna hay $m+n$ bolas, m blancas y n rojas. Si extraemos dos bolas de la urna lo podemos hacer de dos formas:

- Extraer una anotar su color y reponerla. Sacar otra y anotar su color. Hemos extraído la bola con reposición.
- Extraer simultáneamente dos bolas (sin reposición) y contar el número de bolas blancas.

Modelo de la distribución hipergeométrica

Sea X es la v.a. que cuenta el número de bolas blancas extraídas.

- En el primer caso, X es una $B(n = 2, p = \frac{m}{m+n})$ ya que consiste en repetir dos veces el mismo experimento de Bernoulli.
- En el segundo caso, X sigue una distribución hipergeométrica que estudiaremos en esta sección.

Modelo de la distribución hipergeométrica

Distribución hipergeométrica

Sean n, m y k tres número enteros positivos y tales que $k < m + n$.

Consideremos una urna que contiene $m + n$ bolas de las que m son blancas y las restantes n no (son no blancas).

El número total de bolas es $m + n$. Extraemos de forma aleatoria k bolas de la urna sin reemplazarlas.

Modelo de la distribución hipergeométrica

Sea X la v.a. que cuenta el número de bolas blancas extraídas. Diremos que la distribución de X es hipergeométrica de parámetros m, n y k y la denotaremos por $H(m,n,k)$.

Su dominio es

$$D_X = \{x \in \mathbf{N} \mid \max\{0, k-n\} \leq x \leq \min\{m,k\}\}$$

Para explicarlo, veamos varios ejemplos:

- $H(m = 5, n = 2, k = 3)$. Tenemos $m = 5$ bolas blancas, $n = 2$ no blancas y sacamos $k = 3$ bolas sin reposición.
  - En este caso el mínimo de bolas blancas extraídas es $1 = k - n = 3 - 2$, ya que sólo hay dos no blancas.
  - En cambio, el máximo si es $k = 3$, ya que tenemos bolas blancas de "sobra".

Modelo de la distribución hipergeométrica

$$D_X = \{x \in \mathbf{N} \mid \max\{0, k-n\} \leq x \leq \min\{m, k\}\}$$

- $H(m=2, n=5, k=3)$. Tenemos $m=2$ bolas blancas, $n=5$ no blancas y sacamos $k=3$ bolas sin reposición.
    - En este caso el mínimo de bolas blancas es 0 ya que puedo sacar 3 no blancas.
    - En cambio, el máximo si es $m=2$, ya que aunque saquemos $k=3$ bolas, al llegar a 2 ya hemos extraído todas las bolas blancas de la urna.
- $H(m=10, n=10, k=3)$. Tenemos $m=10$ bolas blancas, $n=10$ no blancas y sacamos $k=3$ bolas sin reposición.
    - En este caso podemos obtener desde 0 blancas hasta $k=3$ blancas.

Modelo de la distribución hipergeométrica

Su función de probabilidad es:

Su función de probabilidad es:

$$P_X(x) = \begin{cases} \frac{\binom{m}{x}\cdot\binom{n}{k-x}}{\binom{m+n}{k}}, & \text{si } \max\{0, k-n\} \leq x \leq \min\{m, k\}, \text{ para } x \in \mathbf{N}, \\ 0, & \text{en otro caso.} \end{cases}$$

Distribución hipergeométrica

Observación: otras parametrizaciones

En ocasiones se parametriza una v.a. hipergeométrica mediante $N = m + n$, número total de bolas, k, número de extracciones y p, probabilidad de extraer una bola blanca.

Así podemos **parametrizar alternativamente** la distribución hipergeométrica así

$$H(N, k, p) \text{ donde } p = \frac{m}{N}.$$

Resumen distribución Hipergeométrica $H(m, n, k)$.

$X = \begin{cases} \text{número de bolas blancas en } k \text{ extracciones} \\ \text{sin reposición de una urna conm} \\ \text{bolas blancas y n negras.} \end{cases}$	$; H(m, n, k)$
$D_X = \{x \in \mathbb{N} \mid \max\{0, k-n\} \leq x \leq \min\{m, k\}\}$	
$P_X(x) = P(X = x) = \begin{cases} \frac{\binom{m}{x} \cdot \binom{n}{k-x}}{\binom{m+n}{k}}, & \text{si } \max\{0, k-n\} \leq x \leq \min\{m, k\}, \\ 0, & \text{en otro caso.} \end{cases}$	
$F_X(x) = P(X \leq x)$.	
$E(X) = \frac{k \cdot m}{m+n}$; $Var(X) = k \cdot \frac{m}{m+n} \cdot \left(1 - \frac{m}{m+n}\right) \cdot \frac{m+n-k}{m+n-1}$	

Ejemplo clásico urna $m = 15$ blancas, $n = 10$ rojas y $k = 3$ extracciones sin reposición.

Urna con bolas blancas y rojas

Tenemos una urna con 15 bolas blancas y 10 bolas rojas. Extraemos al azar tres bolas de la urna sin reposición. Sea X el número de bolas **blancas** extraídas. Bajo esta condiciones, la v.a. X sigue una ley de distribución $H(m = 15, n = 10, k = 3)$.

La función de probabilidad es

$$P_X(x) = P(X = x) = \begin{cases} \frac{\binom{m}{x}\cdot\binom{n}{k-x}}{\binom{m+n}{k}} & \text{si } \max\{0, k-n\} \leq x \leq \min\{m, k\} \text{ para } x \in \mathbf{N} \\ 0 & \text{en otro caso} \end{cases},$$

$$\text{sustituyendo } P_X(x) = P(X = x) = \begin{cases} \frac{\binom{15}{x}\cdot\binom{10}{3-x}}{\binom{25}{3}} & \text{si } 0 \leq x \leq 3 \text{ para } x \in \mathbf{N} \\ 0 & \text{en otro caso} \end{cases}$$

Ejemplo clásico urna $m = 15$ blancas, $n = 10$ rojas y $k = 3$ extracciones sin reposición.

La probabilidad de sacar 2 blancas será

$$P(X = 2) = \frac{\binom{15}{2} \cdot \binom{10}{3-2}}{\binom{25}{3}}$$

```
c(choose(15,2), choose(10,1), choose(25,3))
```

```
## [1]  105   10 2300
```

$P(X = 2) = \frac{105 \cdot 10}{2300} = 0.4565217.$

Ejemplo clásico urna $m = 15$ blancas, $n = 10$ rojas y $k = 3$ extracciones sin reposición.

La probabilidad de que saquemos más de 1 bola blanca es

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) = 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - \left(\frac{\binom{15}{0} \cdot \binom{10}{3}}{\binom{25}{3}} + \frac{\binom{15}{1} \cdot \binom{10}{2}}{\binom{25}{3}} \right) \\ &= 1 - \left(\frac{1 \cdot 120}{2300} + \frac{15 \cdot 45}{2300} \right) = 1 - \frac{120 + 15 \cdot 45}{2300} = 0.6543478. \end{aligned}$$

Ejemplo clásico urna $m = 15$ blancas, $n = 10$ rojas y $k = 3$ extracciones sin reposición.

El número esperado de bolas blancas extraídas para una v.a. X $H(m = 15, n = 10, k = 3)$ es

$$E(X) = \frac{k \cdot m}{m + n} = \frac{3 \cdot 15}{15 + 10} = \frac{45}{35} = 1.285714.$$

La varianza vale:

$$\begin{aligned} Var(X) &= k \cdot \frac{m}{m+n} \cdot \left(1 - \frac{m}{m+n}\right) \cdot \frac{m+n-k}{m+n-1} \\ &= 3 \cdot \frac{15}{15+10} \cdot \left(1 - \frac{15}{15+10}\right) \cdot \frac{15+10-3}{15+10-1} \\ &= 3 \cdot \frac{15}{25} \cdot \left(1 - \frac{15}{25}\right) \cdot \frac{22}{24} = 3 \cdot \frac{15}{25} \cdot \frac{25-15}{25} \cdot \frac{22}{24} \\ &= 3 \cdot \frac{15}{25} \cdot \frac{10}{25} \cdot \frac{22}{24} = 0.66. \end{aligned}$$

Y por lo tanto su desviación típica es

$$+\sqrt{Var(X)} = +\sqrt{0.66} = 0.812404.$$

Cálculos con R

Sea X una v.a. $H(m, n, k)$. La función de `R` para calcular la función de probabilidad en un valor x, $P(X = x)$, es `dhyper(x,m,n,k)` y para calcular la función de distribución en un valor q, $P(X \leq q)$, es `phyper(q,m,n,k)`. Para generar una muestra de valores que siga la distribución $H(m, n, k)$, hay que usar la función `rhyper(nn,m,n,k)` donde `nn` es el número de observaciones aleatorias deseado de la muestra.

Por ejemplo, si X es una $H(m = 15, n = 10, k = 3)$, los valores de $P(X = 2)$ y que $P(X > 1) = 1 - P(X \leq 1)$ son:

Cálculos con R

```
dhyper(x=2,m=15,10,k=3)
```

```
## [1] 0.4565217
```

```
phyper(q=1,m=15,n=10,k=3)# sí, le han puesto q ya veremos el porqué
```

```
## [1] 0.3456522
```

```
1-phyper(q=1,m=15,n=10,k=3)
```

```
## [1] 0.6543478
```

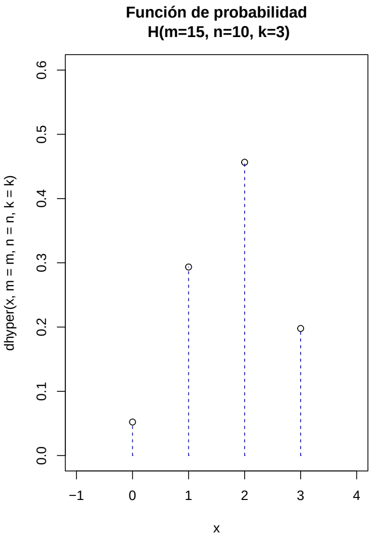
Cálculos con R

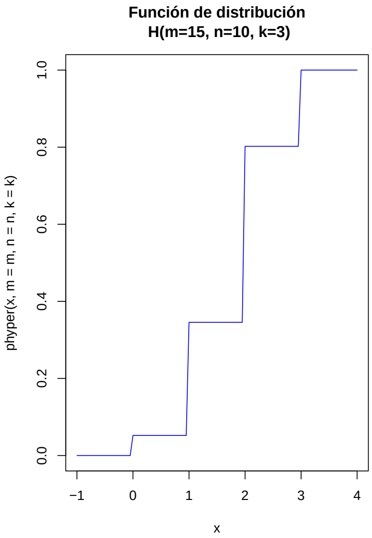
Una muestra aleatoria de este experimento de tamaño 200 sería:

```
rhyper(nn=200,m=15,n=10,k=3)
```

```
##   [1] 2 3 1 3 1 2 2 3 2 2 1 2 1 2 2 3 3 1 1 1 1 0 2 3 2 1 3 2 2 2 2 3 2 3
##  [38] 1 2 1 3 2 2 3 2 3 2 2 3 2 3 1 2 2 2 2 3 2 2 1 3 2 2 3 1 2 2 2 2 2 3
##  [75] 3 2 2 2 1 2 2 3 1 1 1 2 2 2 2 1 1 3 2 2 3 2 2 1 1 1 3 3 2 2 2 1 3 2
## [112] 1 2 3 2 2 1 2 2 2 2 2 2 3 1 2 3 3 1 1 2 2 1 1 3 2 1 1 2 2 3 1 1 1 2
## [149] 1 2 2 3 3 2 3 1 2 1 2 2 2 1 2 3 1 3 3 3 2 2 1 3 3 1 1 2 2 2 2 2 3 2
## [186] 1 1 1 2 1 1 2 2 2 2 3 3 1 0 2
```

Gráficas con R





Cálculos con python

Sea X una $H(m, n, k)$, las funciones de `scipy.stats` cambian los parámetros

- M es el número total de bolas. Con nuestra parametrización $M = m + n$.
- n es el número de bolas blancas. Con nuestra parametrización $n = m$.
- N es el número de extracciones. Con nuestra parametrización $N = k$.

```
from scipy.stats import hypergeom
```

Cálculos con python

```
hypergeom.pmf(1,M=15+10,n=15,N=3)
```

```
## 0.2934782608695652
```

```
hypergeom.cdf(1,M=15+10,n=15,N=3)
```

```
## 0.3456521739130434
```

```
1-hypergeom.cdf(1,M=15+10,n=15,N=3)
```

```
## 0.6543478260869566
```

Cálculos con python

Una muestra aleatoria de este experimento sería. . .

```
hypergeom.rvs(M=15+10,n=15,N=3,size=100)
```

```
## array([3, 2, 1, 3, 2, 3, 3, 1, 3, 2, 2, 1, 1, 0, 1, 3, 3, 2, 1, 1, 2, 2,
##        0, 2, 1, 2, 1, 2, 1, 0, 2, 3, 2, 2, 2, 2, 3, 2, 3, 1, 2, 3, 2, 2,
##        1, 2, 3, 2, 1, 1, 1, 0, 1, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 3, 2, 2,
##        1, 2, 2, 3, 1, 1, 3, 2, 3, 0, 2, 2, 1, 3, 3, 1, 2, 3, 3, 1, 2, 3,
##        3, 1, 2, 2, 1, 1, 1, 1, 2, 2, 0, 1], dtype=int64)
```

Gráficos con python

```
from scipy.stats import hypergeom
[M, n, N] = [20, 7, 12] ##20 elementos, 7 del tipo, extraemos 12
x = np.arange(max(0, N-M+n),min(n, N))
fig =plt.figure(figsize=(5, 2.7))
 =ax = fig.add_subplot(1,2,1)
 =ax.plot(x, hypergeom.pmf(x, M, n, N), 'bo', ms=5, label='hypergeom pmf')
 =ax.vlines(x, 0, hypergeom.pmf(x, M, n, N), colors='b', lw=2, alpha=0.5)
 =ax.set_ylim([0, max(hypergeom.pmf(x, M, n, N))*1.1])
```

Gráficos con python

```
for tick in ax.xaxis.get_major_ticks():
   =tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
  =tick.label.set_fontsize(5)
ax = fig.add_subplot(1,2,2)
 =ax.plot(x, hypergeom.cdf(x, M, n, N), 'bo', ms=5, label='hypergeom cdf')
 =ax.vlines(x, 0, hypergeom.cdf(x, M, n, N), colors='b', lw=2, alpha=0.5)
for tick in ax.xaxis.get_major_ticks():
   =tick.label.set_fontsize(5)
for tick in ax.yaxis.get_major_ticks():
   =tick.label.set_fontsize(5)
 =fig.suptitle('Distribucion Hipergeometrica')
 =plt.show()
```

Gráficos con python

```
## <string>:2: MatplotlibDeprecationWarning: The label function was deprecatec
```