TP2.2 - GENERADORES DE NÚMEROS PSEUDOALEATORIOS DE DISTINTAS DISTRIBUCIONES DE PROBABILIDAD

Carlucci, Gino
Universidad Tecnológica Nacional
Rosario, Santa fe
ginocarlucci@hotmail.com

Docampo, Juan Manuel
Universidad Tecnológica Nacional
Rosario, Santa fe
docampojuan@gmail.com

Menegozzi, Milton
Universidad Tecnológica Nacional
Rosario, Santa fe
miltonmenegozzi@gmail.com

1 de julio de 2020

ABSTRACT

El presente trabajo práctico tiene por objetivo la construcción de generadores de números pseudoaleatorios de distintas probabilidades.

1. Introducción

En nuestro trabajo anterior (Generadores pseudoaleatorios[1]), creamos un generador congruencial lineal y realizamos un estudio para determinar la calidad del mismo. Gracias a ello, logramos obtener sucesiones de números independientes que se puedan considerar como observaciones de una distribución uniforme en el intervalo (0, 1).
En el presente trabajo, lo que buscaremos es transformar esa sucesión de números generadas por nuestro generador GCL, a distintas distribuciones de probabilidad por medio de algoritmos ya definidos.
Existen varios métodos que nos permiten generar variables aleatorias. La mayoría de las técnicas utilizadas para la generación se pueden agrupar en:

- Método de la transformada inversa
- Método de aceptación-rechazo
- Método de composición
- Método de convolución

Para el presente estudio, se utilizará el método de la transformada inversa por su simplicidad. El mismo permite la generación de números aleatorios de cualquier distribución de probabilidad, aunque será aplicado solo a algunas de las distribuciones descriptas, ya que para otras, puede resultar muy complicado obtener una expresión analítica de la inversa de su distribución de probabilidad.
El método será explicado con mayor detalle en la siguiente sección, para conocer su funcionamiento y posteriormente ser aplicado a las distintas distribuciones.

2. Metodo de la Transformadan inversa

El método de la transformada inversa puede utilizarse para simular variables aleatorias continuas, como la uniforme, exponencial, normal, etc. Lo cual se logra mediante la función acumulada $F(x)$ y la generación de números pseudoaleatorios $r_i \sim U(0, 1)$.

También, el mismo método puede emplearse para simular variables aleatorias de tipo discreto, como en las distribuciones de Poisson, binomial, geométrica, etc. La generación se lleva a cabo a través de la probabilidad acumulada $P(x)$ y la generación de números pseudoaleatorios $r_i \sim U(0, 1)$.

Para el caso de las distribuciones continuas, el método se puede resumir en los siguientes pasos:

1. Definir la función de densidad f(x) que representa la variable a modelar.
2. Calcular la función acumulada F(x).
3. Puesto que $F(x)$ se define sobre el rango $(0, 1)$, podemos generar números aleatorios distribuidos uniformemente y además hacer $F(x) = r$. Entonces para cualquier valor particular de r que generemos, es posible encontrar el valor de x.
$$x_0 = F^{-1}(r_0) \tag{1}$$
Donde F^{-1} es el mapeo de r sobre el intervalo unitario en el dominio de x

3. Distribuciones continuas de probabilidad

Una distribución continua describe las probabilidades de los posibles valores de una variable aleatoria continua. Una variable aleatoria continua es una variable aleatoria con un conjunto de valores posibles (conocido como el rango) que es infinito y no se puede contar.

Las probabilidades de las variables aleatorias continuas (X) se definen como el área por debajo de la curva de su PDF (Función de densidad). Por lo tanto, solo los rangos de valores pueden tener una probabilidad diferente de cero. La probabilidad de que una variable aleatoria continua equivalga a algún valor siempre es cero.

3.1. Distribución uniforme

La distribución Uniforme es el modelo continuo más simple. Corresponde al caso de una variable aleatoria que sólo puede tomar valores comprendidos entre dos extremos a y b, de manera que todos los intervalos de una misma longitud (dentro de (a, b)) tienen la misma probabilidad. También puede expresarse como el modelo probabilístico correspondiente a tomar un número al azar dentro de un intervalo (a, b).

Su función de densidad **(FDP)** esta dada por:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in (a, b) \\ \\ 0 & x \notin (a, b) \end{cases} \tag{2}$$

La función de la distribución acumulada **(FDA)** se obtiene integrando la función de densidad (2):

$$F(x) = \int_a^{\infty} \frac{1}{b-a} dt = \frac{x-a}{b-a} \quad 0 \leq F(x) \leq 1 \tag{3}$$

El valor esperado y la varianza estan dadas por:

$$EX = \int_a^b \frac{1}{b-a} x \ dx = \frac{b+a}{2} \tag{4}$$

$$VX = \int_a^b \frac{(x-EX)^2}{b-a} \ dx = \frac{(b-a)^2}{12} \tag{5}$$

Para simular una distribución uniforme sobre cierto intervalo (a, b), debemos obtener la transformación inversa de (3). Aplicando (1) obtenemos:

$$\begin{aligned} r &= \frac{x-a}{b-a} \\ x &= a + (b-a)r \quad 0 \leq r \leq 1 \end{aligned} \tag{6}$$

Se adjunta a continuación el código para realizar la transformación de valores del intervalo $(0, 1)$ al intervalo $(0, 20)$ y su correspondiente gráfica obtenida.

```python
def distribucion_uniforme(a,b):
    for r in numerosGCL:
        x = a+(b-a)*r
        uniforme.append(x)
    return uniforme
```

Figura 1: Código python

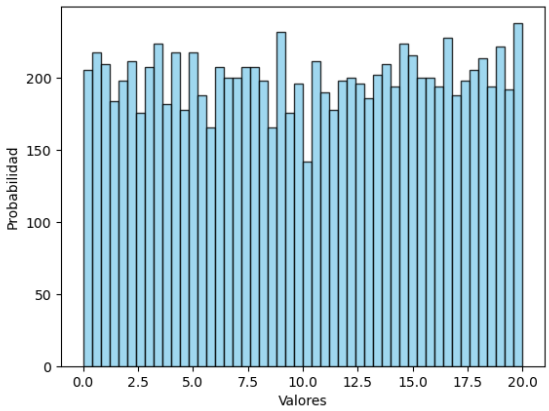

Figura 2: Gráfica distribución uniforme

Para probar que los datos provienen de una distribución uniforme, se aplicara la prueba de Kolmogorov-Smirnov. La misma, permite la medición del grado de concordancia existente entre la distribución que generamos y una distribución teorica específica, en nuestro caso, la uniforme.

Hipotesis:

H_0 Los datos pertenecen a una distribución uniforme
H_1 Los datos no pertenecen a una distribución uniforme

Resultados:

D^+	D^-	D_{max}	Dα
0.00107	-0.00087	0.00107	0.00366

Cuadro 1: Prueba de Kolmogorov-Smirnov

Al observar los resultados, podemos ver que el valor crítico obtenido de la tabla KS para un nivel de significancia $0{,}05$ y tamaño 10000, es mayor al valor máximo D obtenido. Por lo tanto, podemos aceptar la Hipótesis H_0 y afirmar que la suceción de números generada pertenecen a una distribución uniforme.

3.2. Distribución exponencial

La distribución exponencial suele utilizarse para variables que describen el tiempo hasta que se produce un determinado suceso. Este modelo depende de un único parámetro α, el cuál debe ser positivo ($\alpha > 0$).

Las distribuciones exponenciales muestran el intervalo de tiempo entre ocurrencias de eventos cuya probabilidad es baja en un intervalo corto, y si la ocurrencia es estadísticamente independiente respecto a otros eventos. El hecho de que un proceso de la vida real proporcione o no valores de variables aleatorias de tipo exponencial, está sujeto al grado en que se satisfagan las siguientes suposiciones:

1. La probabilidad de que ocurra un evento en el intervalo $[t, (t + \Delta t)]$ es $\alpha\Delta t$.
2. α es una constante que no depende de t o de algún otro factor.
3. La probabilidad de que durante un intervalo $[t, (t + \Delta t)]$ ocurra más de un evento, tiende a 0 a medida que $\Delta t \longrightarrow 0$, y su orden de magnitud deberá ser menor que el de $\alpha\Delta t$.

Su función de densidad (**FDP**) está dada por:

$$f(x) = \alpha e^{-\alpha x} \quad \alpha > 0 \text{ y } x \geq 0 \tag{7}$$

La función de la distribución acumulada **(FDA)** se obtiene integrando la función de densidad (7):

$$F(x) = \int_0^\infty \alpha e^{-\alpha t} dt = 1 - e^{-\alpha x} \tag{8}$$

El valor esperado y la varianza estan dadas por:

$$EX = \int_0^\infty x\alpha e^{-\alpha x} dx = \frac{1}{\alpha} \tag{9}$$

$$VX = \int_0^\infty (x - \frac{1}{\alpha})^2 \ \alpha e^{-\alpha x} dx = \frac{1}{\alpha^2} = (EX)^2 \tag{10}$$

Existen muchas maneras para lograr la generación de valores de variables aleatorias exponenciales, pero como mencionamos anteriormente, aplicaremos otra vez el método de la transformada inversa.
Puesto que $F(x)$ existe explícitamente, la técnica de la transformada inversa nos permite desarrollar métodos directos para dicha generación. Debido a la simetría que existe entre la distribución uniforme sigue que la intercambiabilidad de $F(x)$ y $1 - F(x)$. Por lo tanto:

$$\begin{aligned} r &= e^{\alpha x} \\ x &= -(\frac{1}{\alpha}) \log r \\ x &= -EX \log r \end{aligned} \tag{11}$$

Se adjunta a continuación el código para realizar la transformación de valores del intervalo $(0, 1)$ a valores que siguen una distribución exponencial y su correspondiente gráfica obtenida con el parámetro $ex = 0{,}5$.

```
24  def distribucion_exponencial(ex):
25      for r in numerosGCL:
26          x = -ex*np.log(r)
27          exponencial.append(x)
28      return exponencial
```

Figura 3: Código python

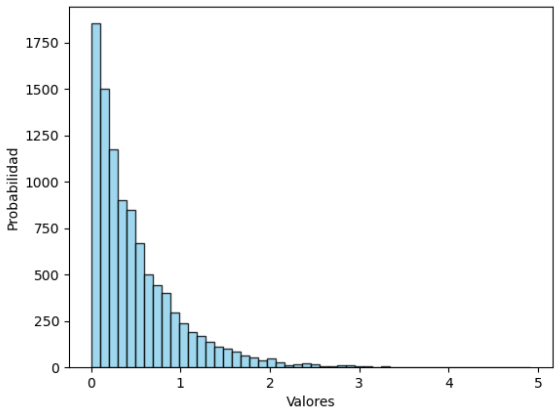

Figura 4: Gráfica distribución exponencial

Para determinar que los datos generados provienen de una distribución exponencial, se aplicará la prueba Anderson-Darling. La misma es una prueba no paramétrica para determinar si los datos de una muestra provienen de una distribución específica.

Hipotesis:

H_0 Los datos siguen una distribución exponencial

H_1 Los datos no siguen una distribución exponencial

Resultados:

Se obtuvieron los valores adjuntados en el siguiente Cuadro.

Valor estadístico = 0.903	
Nivel de significancia	Valores críticos
15 %	0.921
10 %	1.077
5 %	1.340
2 %	1.605
1 %	1.956

Cuadro 2: Resultados obtenidos para cada valor crítico calculado según nivel de significancia.

Como podemos observar, en todos los casos el valor estadístico calculado es menor al valor crítico, por lo que se acepta la hipótesis H_0 y concluimos que los datos generados provienen de una distribución exponencial.

3.3. Distribución gamma

Las distribuciones gamma, con parámetros α y k se dan cuando un determinado proceso consiste de k eventos sucesivos y si el total de tiempo transcurrido para dicho proceso se puede considerar igual a la suma de k valores independientes de la variable aleatoria con distribución exponencial, cada uno de los cuales tiene un parámetro a definido. Si la suma de los k (donde k es un entero positivo) valores de una variable aleatoria con distribución exponencial tiene un mismo parámetro α, se denomina distribución erlang.

Su función de densidad (**FDP**) está dada por:

$$f(x) = \frac{\alpha^k x^{k-1} e^{-\alpha x}}{(k-1)!} \tag{12}$$

Donde $\alpha > 0$, $k > 0$ y x se considera no negativo.
La distribución gamma no posee una forma explícita de escribir su función acumulativa pero se ha logrado presentar sus valor en un formato tabular. Con respecto a la media y la variancia, sus expresiones están formuladas como sigue:

$$EX = \frac{k}{\alpha} \tag{13}$$

$$VX = \frac{k}{\alpha^2} \tag{14}$$

Podemos notar que si k toma el valor de 1, la distribución gamma resulta ser idéntica a la exponencial.

Para generar valores de la variable aleatoria con distribución gamma y con un valor esperado y variancia dados, se pueden utilizar las siguientes ecuaciones a fin de determinar los parámetros de $f(x)$ de (12).

$$\alpha = \frac{EX}{VX} \tag{15}$$

$$k = \frac{EX^2}{VX} \tag{16}$$

Debido a que no contamos con la distribución acumulativa, para este caso no podremos usar el método de la transformada inversa, por lo que recurriremos a un método alternativo para la generación de valores de una variable aleatoria. Para realizarlo, se debe tomar la suma de los k valores de la variables aleatorias con distribución exponencial $x_1, x_2, x_3, ..., x_k$ cuyo valor esperado (o media) es el mismo e igual a $1/\alpha$. En consecuencia el valor de la variable aleatoria se puede expresar como:

$$x = \sum_{i=1}^{k} x_i$$

Como x_i se distribuye exponencialmente:

$$x = -\frac{1}{\alpha} \ln \left(\prod_{i=1}^{k} r_i \right)$$

Se adjunta a continuación el código para realizar la transformación de valores del intervalo $(0, 1)$ a valores que siguen una distribución gamma y su correspondiente gráfica obtenida con los parámetros: $k = 5$ y $\alpha = 1$.

```
30  def distribucion_gamma(k,alpha):
31      for i in range(10000):
32          tr = 1.0
33          for i in range(k):
34              r = random.choice(numerosGCL)
35              tr = tr *  r
36          x = -np.log(tr)/alpha
37          gamma.append(x)
38      return gamma
```

Figura 5: Código python

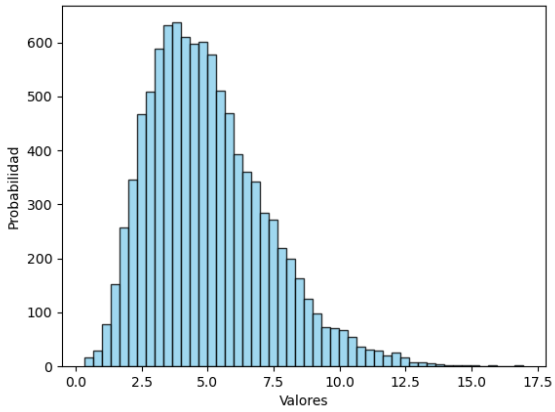

Figura 6: Gráfica distribución gamma

Para corroborar que los valores obtenidos siguen la distribucion en estudio se comparara la media y varianza obtenida en cinco corridas con la media y varianza esperada, reemplazando en las ecuaciones (13) y (14) con los parámetros utilizados.

Comparación Media y Varianza gamma				
-	Media obtenida	Varianza obtenida	Media esperada	Varianza esperada
Primera Corrida	4.998	5.237		
Segunda Corrida	5.038	5.186		
Tercera Corrida	4.974	5.011	5	5
Cuarta Corrida	4.999	5.111		
Quinta Corrida	4.993	5.178		

Cuadro 3: Gamma

3.4. Distribución normal

La distribución normal es un modelo teórico capaz de aproximar satisfactoriamente el valor de una variable aleatoria continua a una situación ideal. Basa su utilidad en el teoria del limite central. Este teorema postula que, la ditribucion de probabilidad de la suma de N valores de variable aleatoria Xi independientes pero identicamente distribuidos, con medias respectivas μ_i y variancias σ^2 se aproxima asintoticamente a una distribucion normal, a medida que N se hace muy grande, y que dicha distribucion tiene como media y variancia respectivamente a:

$$\begin{aligned} \mu &= \sum_{i=1}^{N} \mu_i \\ \sigma_i^2 &= \sum_{i=1}^{N} \sigma_i^2 \end{aligned} \tag{17}$$

Su función de densidad **(FDP)** esta dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{18}$$

El valor espereado y la variancia de la distribucion normal no estandar estan dados por:

$$EX = \mu_x \tag{19}$$

$$VX = \sigma_x^2 \tag{20}$$

Si la variable aleatoria x tiene una función de densidad $f(x)$ dada como:

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2}$$

Donde $-\infty \leq x \leq \infty$, con σ_x positiva, entonces se dice que x tiene una distribución normal o gaussiana, con parámetros μ_x y σ_x. La función de distribución acumulativa $F(x)$ no existe en forma explícita, sin embargo, se encuentra tabulada. El método para obtener valores de una variable aleatoria normal que se utiliza mayormente es el procedimiento llamado del límite central, de este se obtiene:

$$x = \sigma_x \left(\frac{12}{K}\right)^{\frac{1}{2}} \left(\left(\sum_{i=1}^{K} r_i\right) - \frac{K}{2}\right) + \mu_x$$

Para generar un solo valor de x (un valor de una variable aleatoria con distribución normal) bastará como sumar K números aleatorios definidos en el intervalo de 0 a 1. Es decir, K va a ser un valor que me va a indicar cuantos números aleatorios necesito para generar un único valor de x (se recomienda utilizar $K = 12$ de manera de simplificar los cálculos de la ecuación anterior).

Se adjunta a continuación el código para la obtencuin de valores que siguen una distribución normal y su correspondiente gráfica obtenida con los parámetros $\mu = 10$ y $\sigma = 20$.

```
40  def distribucion_normal(mu,sigma):
41      for i in range(10000):
42          sum = 0.0
43          for i in range(12):
44              r = random.choice(numerosGCL)
45              sum = sum + r
46          x = sigma * (sum - 6.0) + mu
47          normal.append(x)
48      return normal
```

Figura 7: Código python

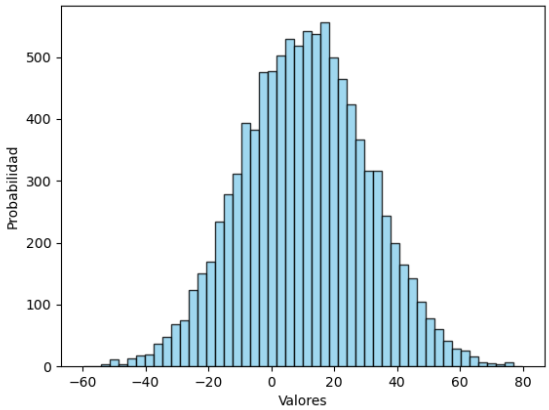

Figura 8: Gráfica distribución normal

Para determinar que los datos generados provienen de una distribución normal, se aplicará la prueba Anderson-Darling. La misma es una prueba no paramétrica para determinar si los datos de una muestra provienen de una distribución específica.

Hipotesis:

H_0 Los datos siguen una distribución normal
H_1 Los datos no siguen una distribución normal

Resultados:

Se obtuvieron los valores adjuntados en el siguiente Cuadro.

Valor estadístico = 0.400	
Nivel de significancia	Valores críticos
15 %	0.576
10 %	0.656
5 %	0.787
2 %	0.918
1 %	1.092

Cuadro 4: Resultados obtenidos para cada valor crítico calculado según nivel de significancia.

Como podemos observar, en todos los casos el valor estadístico calculado es menor al valor crítico, por lo que se acepta la hipótesis $H0$ y concluimos que los datos generados provienen de una distribución normal.

4. Distribuciones discretas de probabilidad

Una distribución discreta describe la probabilidad de ocurrencia de cada valor de una variable aleatoria discreta. Una variable aleatoria discreta es una variable aleatoria que tiene valores contables, tales como una lista de enteros no negativos.

Con una distribución de probabilidad discreta, cada valor posible de la variable aleatoria discreta puede estar asociado con una probabilidad distinta de cero.

4.1. Distribución Pascal

Cuando los procesos de ensayos de Bernoulli se repiten hasta lograr que ocurran k exitos ($k > 1$), la variable aleatoria que caracteriza al numero de fallas tendra una distribucion binomial negativa. Por consiguiente, los valores de variables aleatorias con distribucion binomial negativa coinciden esencialmente con la suma de k valores de variable aleatoria con distribucion geometrica; en este caso, k es un numero entero y la distribucion recibe el nombre de distribucion de Pascal. En consecuancia, la distribucion geometrica constituye un caso particular de Pascal, especificada para k igual a uno.

La funcion de distribucion de probabilidad para una distribucion binomial negativa esta dada por:

$$f(x) = \binom{k + x * 1}{x} p^k q^x \tag{21}$$

donde k es el numero total de exitos en una sucesion de $k + x$ ensayos, con x el numero de fallas que ocurren antes de obtener k exitos.El valor esperado y la variancia de X se representa con:

$$EX = \frac{kq}{p} \tag{22}$$

$$VX = \frac{kq}{p^2} \tag{23}$$

Para una media y una variancia dadas, se pueden determinar los parametros p y k de la siguiente manera:

$$p = \frac{EX}{VX} \tag{24}$$

$$k = \frac{(EX)^2}{VX - EX} \tag{25}$$

Cuando k es un entero, los valores de la variable aleatoria con distribución de Pascal se pueden generar con sólo considerar la suma de k va lores con distribución geométrica. En consecuencia:

$$x = \frac{\sum_{i=1}^{k} \log r_i}{\log q} = \frac{\log(\prod_{i=1}^{k} r_i)}{\log q} \tag{26}$$

viene a ser un valor de variable aleatoria con distribucion de Pascal, una vez que su magnitud se redondea con respecto al menor entero mas proximo al valor calculado.

Se adjunta a continuación el código para la obtención de valores que siguen una distribución pascal y su correspondiente gráfica obtenida con los parámetros $k = 1$ y $q = 0{,}5$.

```
50  def distribucion_pascal(k,q):
51      for i in range(10000):
52          tr = 1.0
53          qr = np.log(q)
54          for i in range(k):
55              r = random.choice(numerosGCL)
56              tr = tr * r
57          nx = np.log(tr)/qr
58          x = nx
59          pascal.append(x)
60      return pascal
```

Figura 9: Código python

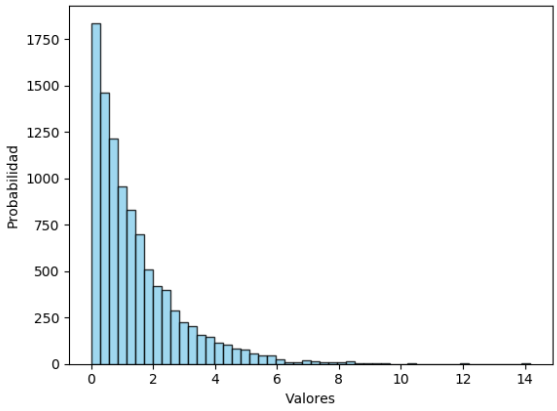

Figura 10: Gráfica distribución pascal

Para corroborar que los valores obtenidos siguen la distribucion en estudio se comparara la media y varianza obtenida en cinco corridas con la media y varianza esperada, aplicando las ecuaciones 22 y 23

Comparación Media y Varianza pascal				
-	Media obtenida	Varianza obtenida	Media esperada	Varianza esperada
Primera Corrida	1.476	2.229		
Segunda Corrida	1.451 3	2.171		
Tercera Corrida	1.446	2.158	1	2
Cuarta Corrida	1.433	2.074		
Quinta Corrida	1.452	2.141		

Cuadro 5: Pascal

4.2. Distribución Binomial

La distribución binomial proporciona la probabilidad de que un evento o acontecimiento tenga lugar x veces en un conjunto de n ensayo, donde la probabilidad de éxito esta dada por p.

En esta distribución las variables alatorias estan definidas por el numero de eventos exitosos en una sucesión de n ensayos independientes de Bernoulli para los cuales la probabilidad de éxito es p en cada ensayo.

Las variables al tomar el número de éxitos que ocurren en cada ensayo, toma valores discretos por lo tanto la distribución es una distribución de variables discretas.

La función de probabilidad para la distribución binomial

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad con \;\; x\,\varepsilon\mathbb{N}, x \leq n \;\; y \;\; q = (1-p) \tag{27}$$

Una variable binomialmente distribuida x se la nota:

$$x \sim B(n, p) \tag{28}$$

El valor esperado y la variancia de X se representa con:

$$EX = np \tag{29}$$

$$VX = npq \tag{30}$$

Se adjunta a continuación el código para la obtención de valores que siguen una distribución Binomial y su correspondiente gráfica obtenida con los parámetros $n = 20$ y $p = 0{,}5$.

```
62  def distribucion_binomial(n,p):
63      for i in range(10000):
64          x = 0.0
65          for i in range(n):
66              r = random.choice(numerosGCL)
67              if (r-p) < 0:
68                  x += 1
69          binomial.append(x)
70      return binomial
```

Figura 11: Código python

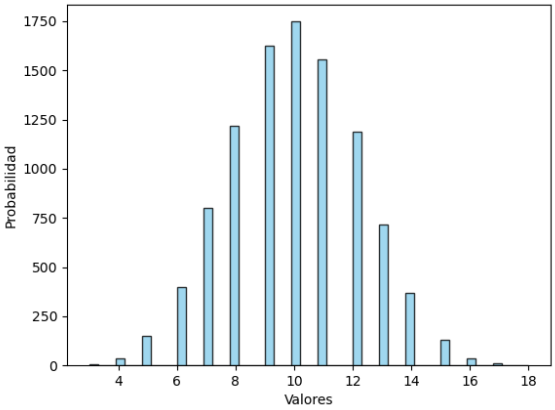

Figura 12: Gráfica distribución binomial

Para corroborar que los valores obtenidos siguen la distribucion en estudio se comparara la media y varianza obtenida en cinco corridas con la media y varianza esperada, aplicando (29) y (30) con los parámetros utilizados.

Comparación Media y Varianza binomial				
-	Media obtenida	Varianza obtenida	Media esperada	Varianza esperada
Primera Corrida	9.967	4.927		
Segunda Corrida	9.966 3	5.010		
Tercera Corrida	9.948	4.988	10	5
Cuarta Corrida	9.952	4.988		
Quinta Corrida	9.936	4.939		

Cuadro 6: Binomial

4.3. Distribución Hipergeométrica

Si en una poblacion de N elementos se toma una muestra aleatoria que conste de n elementos $(n < N)$ sin que tenga lugar algun reemplazo, entonces el numero de elementos x de la clase I en la muestra de n elemntos, tendra una distribucion de probabilidad hipergeometrica.

La distribucion hipergeometrica esta descrtita por la siguiente funcion de probabilidad:

$$f(x) = \frac{\binom{N_p}{x}\binom{N_q}{n-x}}{\binom{N}{n}} \begin{pmatrix} 0 < x < N_p \\ 0 < n - x < N_q \end{pmatrix} \tag{31}$$

donde x, n y N son enteros. El valor esperado y la variancia se caracterizan como sigue:

$$EX = np \tag{32}$$

$$VX = npq(\frac{N-n}{N-1}) \tag{33}$$

Se adjunta a continuación el código para la obtención de valores que siguen una distribución Hipergeométrica y su correspondiente gráfica obtenida con los parámetros $tn = 5000000$, $ns = 500$, $p = 0{,}4$.

```
72  def distribucion_hipergeometrica(tn,ns,p):
73      for i in range(10000):
74          x = 0.0
75          for i in range(ns):
76              r = random.choice(numerosGCL)
77              if (r-p) > 0:
78                  s = 0.0
79              else:
80                  s = 1.0
81                  x = x + 1.0
82              p = (tn*p-s) / (tn-1.0)
83              tn = tn - 1.0
84              if(tn<2):break
85          hipergeometrica.append(x)
86      return hipergeometrica
```

Figura 13: Código python

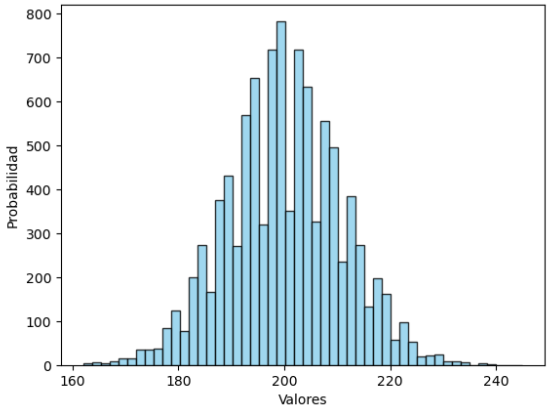

Figura 14: Gráfica distribución hipergeométrica

Para corroborar que los valores obtenidos siguen la distribucion en estudio se comparara la media y varianza obtenida en cinco corridas con la media y varianza esperada, aplicando (32) y (33) con los parámetros utilizados.

Comparación Media y Varianza Hipergeometrica				
-	Media obtenida	Varianza obtenida	Media esperada	Varianza esperada
Primera Corrida	199.999	120.104	200	120
Segunda Corrida	200.03	120.795		
Tercera Corrida	199.999	119.484		
Cuarta Corrida	200.0	118.891		
Quinta Corrida	200.0	121.002		

Cuadro 7: Hipergeomtrica

4.4. Distribución Poisson

Si tomamos una serie de n ensayos independientes de Bernoulli, en cada uno de los cuales se tenga una probabilidad p muy pequeña relativa a la ocurrencia de un cierto evento, a medida que n tiene al inifinito la probablidad de x ocurrencias esta dada por la ditribución de Poisson.

La relación que habíamos dicho con la distribución exponencial es que si en aquella distribución la x asumía los valores del tiempo entre dos eventos y por lo tanto era continua, en esta distribución la x asumirá los valores que corresponden al numero de éxitos en un determinado intervalo de tiempo, por lo tanto seran valores discretos.

La función de densidad esta dada por:

$$f(x) = e^{-\lambda}\frac{\lambda}{x!} \quad x = 0, 1, 2, .. \quad y \quad \lambda > 0 \tag{34}$$

Un variable x Distribuida con Poisson se nota:

$$x \sim P(\lambda) \tag{35}$$

El valor esperado y la variancia se caracterizan como sigue:

$$EX = \lambda \tag{36}$$

$$VX = \lambda \tag{37}$$

Se adjunta a continuación el código para la obtención de valores que siguen una distribución Poisson y su correspondiente gráfica obtenida con los parámetros $p = 10$.

```
88  def distribucion_poisson(p):
89      for i in range(10000):
90          x = 0.0
91          b = np.exp(-p)
92          tr = 1.0
93          while (tr-b) >= 0:
94              r = random.choice(numerosGCL)
95              tr = tr * r
96              if(tr-b >= 0):
97                  x = x + 1.0
98          poisson.append(x)
99      return poisson
```

Figura 15: Código python

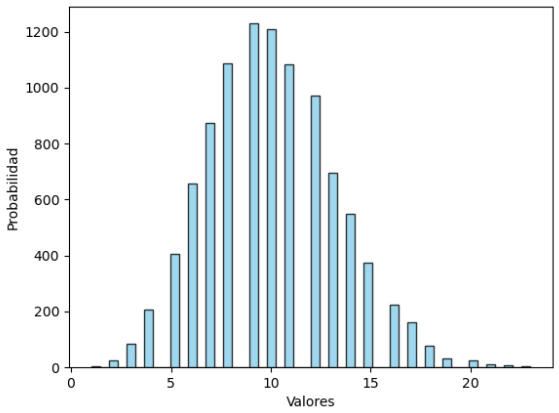

Figura 16: Gráfica distribución poisson

Para corroborar que los valores obtenidos siguen la distribucion en estudio se comparara la media y varianza obtenida en cinco corridas con la media y varianza esperada, aplicando las ecuaciones 36 y 37

Comparación Media y Varianza POISSON				
-	Media obtenida	Varianza obtenida	Media esperada	Varianza esperada
Primera Corrida	10.038	10.314		
Segunda Corrida	10.049 3	10.026		
Tercera Corrida	10.041	10.257	10	10
Cuarta Corrida	9.995	10.219		
Quinta Corrida	10.044	10.111		

Cuadro 8: Poisson

4.5. Distribución Empírica Discreta

Una idea a tener en cuenta es que cualquier distribución puede ser una destribución empírica. Es decir se conforma en base datos emípiricos que bien se podrian corresponder a cualquiera de las distribuciones dadas o bien a alguna distribución que no se corresponde a ninguna conocida.

Para poder generar una disribución empírica simplemente necesitamos armar una tabla de frecuencias con valores asignados a una variable alatoria que asuma estos valores en particular, en el caso de una distribución empírica discreta o bien valores dentro de un intervalo en el caso de una distribución empírica continua.

El ejemplo debajo es una gráfica que representa una distribución empírica continua en base a 10000 muestras de la siguiente tabla la cual muestra los diferentes eventos que pueden ocurrir si x toma un valor dentro del intevalo.

$[0{,}273, 0{,}037, 0{,}195, 0{,}009, 0{,}124, 0{,}058, 0{,}062, 0{,}151, 0{,}047, 0{,}044]$

Cuando x asume un valor se lo coloca en el intevalo correspondiente y se va contando el número de ocurrencias que tiene cada evento.

Se adjunta a continuación el código para la obtención de valores que siguen una distribución Empírica y su correspondiente gráfica obtenida.

```
101  def distribucion_empirica():
102      p=[0.273,0.037,0.195,0.009,0.124,0.058,0.062,0.151,
103      for i in range(10000):
104          r = random.choice(numerosGCL)
105          acum=0
106          cont=1
107          for j in p:
108              acum = acum + j
109              if (r<=acum):
110                  break
111              else:
112                  cont+=1
113          empirica.append(cont)
114      return empirica
```

Figura 17: Código python

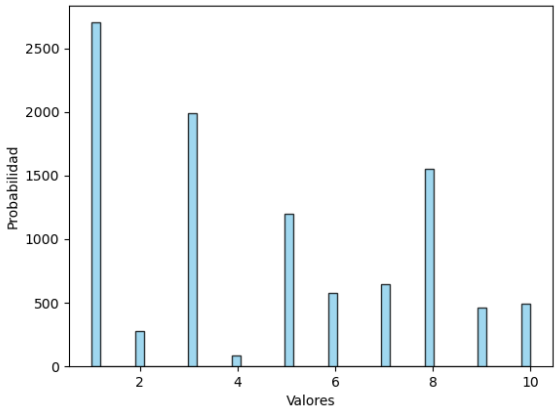

Figura 18: Gráfica distribución Empírica

Prueba Chi Cuadrado
En general un test de bondad de ajuste se utiliza para discriminar si una colección de datos o muestra se ajusta a una distribución teórica de una determinada población. En otras palabras, nos dice si la muestra disponible representa (o ajusta) razonablemente los datos que uno esperaría encontrar en la población.

Hipotesis:

H_0 Los datos son muestra de la distribución Empírica
H_1 Los datos no son muestra de la distribución Empírica

Resultados:

Se obtuvieron los siguentes valores adjuntados en la Cuadro 9

Intervalo	$\frac{(O_i-E_i)^2}{E_i}$
1	0.193
2	4.110
3	0.185
4	0.544
5	0.725
6	0.208
7	0.645
8	0.016
9	0.938
10	0.511
Total	$\sum_{i=1}^{10} \frac{(o_i-e_i)^2}{e_i} = 8{,}079$

Cuadro 9: Sumatoria estadísticas de prueba

Para un intervalo de confianza del 95 % y 9 grados de libertad, se obtuvo el valor de la tabla Chi Cuadrado = 16.918

Como $X^2 < X_{0,05,9}$, se acepta la hipotesis H_0 y podemos afirmar que no existe diferencia entre la distribución de números generada y la distribución Empírica.

5. Resumen trabajo

Distribución de Probabilidad	Tipo	Transformada inversa	Código	Testeo
Uniforme	Continua	Si	Si	Kolmogorov-Smirnov
Exponencial	Continua	Si	Si	Anderson-Darling
Gamma	Continua	No	Si	Comparacion Ex Vx
Normal	Continua	No	Si	Anderson-Darling
Pascal	Discreta	No	Si	Comparacion Ex Vx
Binomial	Discreta	No	Si	Comparacion Ex Vx
Hipergeometrica	Discreta	No	Si	Comparacion Ex Vx
Poisson	Discreta	No	Si	Comparacion Ex Vx
Empirica	Discreta	No	Si	Chi-Cuadrado

Cuadro 10: Resumen del trabajo realizado

6. Conclusiones

Luego del trabajo realizado, podemos concluir que a partir de una secuencia de números uniformes en el intervalo $(0, 1)$, se pueden generar cualquier tipo de distribución con los métodos y algoritmos empleados.
A su vez, todas las distribuciones generadas pasan las pruebas estádisticas realizadas, por lo que podemos concluir con total firmeza, de que las nuevas sucesiones obtenidas pertenecen a las distribuciones generadas.

Referencias

[1] Generadores pseudoaleatorios. In *https://github.com/ginocarlucci/Informes/blob/master/TP2*$_1$*.pdf* $//$.