TP 2.2 - GENERADORES DE NÚMEROS PSEUDOALEATORIOS DE DISTINTAS DISTRIBUCIONES DE PROBABILIDAD

Bruno A. Dolce
UTN - FRRo
Zeballos 1341, S2000
bdolce@frro.utn.edu.ar

Juan Pablo Castelli
UTN - FRRo
Zeballos 1341, S2000
juancastelli58@gmail.com

July 3, 2020

ABSTRACT

La intención de la simulación es imitar situaciones de la naturaleza, para eso es necesario replicar eventos de carácter aleatorio. En el trabajo anterior, desarrollamos programas que generan una variable aleatoria con una distribución uniforme. Muchas de estas situaciones que se intentan imitar, pueden tener distribuciones diferentes, por eso es necesario poder crear generadores pseudo-aleatorios que cumplan cada una de estas. En el siguiente trabajo, nos dedicaremos a generarlos.

1 Introducción.

Teniendo una variable X : Cantidad de llamadas telefónicas que recibe un centro de atención al cliente por minuto, con una distribución de Poisson $X \sim P(\lambda = 5)$, al querer hacer una simulación para esta situación necesitaríamos generar valores para esa variable que se correspondan con esa distribución, pero el generador creado en el trabajo anterior devuelve valores que con una distribución $Y \sim U(0; 1)$. En esta trabajo nos dedicaremos a producir generadores para las distribuciones continuas Uniforme, Gamma, Exponencial y Normal; y Binomial, Geométrica, Hipergeométrica y Poisson para las discretas.

2 Marco teórico.

A continuación detallaremos algunos conceptos que se utilizarán en el siguiente trabajo:

Variable aleatoria: en adelante abreviada como v.a., es una variable que surge como resultado de un evento o experimento y su naturaleza es aleatoria. Matemáticamente, denominando S como el espacio muestral asociado con un experimento aleatorio, una v.a. X se interpreta como una función que asocia a cada elemento wS un número real $X(w) = x$. Esto es:

$$X : S \to \mathbb{R} \tag{1}$$

Una v.a. se considera **discreta** si toma un número finito o infinito numerable de valores.
En cambio, una v.a. es **continua** si toma valores fijos dentro de un intervalo continuo determinado. Y siempre entre dos valores observables va a existir un tercer valor intermedio que también podría tomar la variable.

Función de densidad de probabilidad: también llamada **densidad** de una v.a. continua, es una función cuyo valor en cualquier punto del espacio muestral puede interpretarse como la *probabilidad relativa* que posee la v.a. para ese valor. Se utiliza para especificar la probabilidad que una v.a. caiga en cierto rango determinado de valores. La probabilidad de un valor en la función se obtiene calculando la integral de la función para ese rango, esto es, hallando el área bajo la curva de la función sobre el eje horizontal y acotada entre el rango especificado. La función es no negativa, y su integral sobre el espacio entero es 1.

Distribución de probabilidad: es una función sobre una v.a. que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra. Una distribución de probabilidades puede interpretarse como una frecuencia teórica, ya que describe cómo se espera que varíen los resultados.

Función de distribución acumulada: es la función que asocia a cada valor x la probabilidad acumulada para todos los valores menores o iguales a x. Esto es: $F(x) = P(X \leq x)$.

Función de distribución acumulada inversa: es la función inversa de la Función de distribución acumulada, por lo que provee el valor asociado con una probabilidad acumulada específica. Se utiliza para determinar el valor de la variable asociado con una cierta probabilidad.

2.1 Distribuciones

3 Metodología de trabajo.

Estudiaremos las distribuciones dadas a continuación, y encontraremos la mejor forma de generar números pseudoaleatorios, en caso de ser posible se utilizará el método de la transformada inversa, si no buscaremos formas alternas.

3.1 Método de transformada inversa.

Teniendo una función de densidad $f(x)$ de cierta variable aleatoria, debemos obtener la función de distribución acumulativa $F(x) = \int_{-\infty}^{x} f(t)\,dt$ que tiene como imagen el intervalo [0,1], de esta forma podemos generar números aleatorios r_i uniformes en el mismo intervalo, $F(x) = r$ y aplicarle la función inversa, para encontrar valores que se comporten en base a la distribución buscada, $F^{-1}(r_0) = x_0$. Posteriormente generaremos la gráfica de las listas de cada distribución generada para observar la similitud con las de ejemplo. Y por último aplicaremos ciertos test para comprobar de manera más fehaciente que el comportamiento del generador se acerca al esperado. [1]

3.2 Generador uniforme base.

Utilizamos el generador GLC, *Generador Lineal congruencial* producido en el trabajo anterior. Los parámetros elegidos para el generador son los utilizados en el generador de Borland C y C++.

Código Python

```python
class GLC():
    m = 2**32
    a = 22695477
    c = 1
    x = 0
    seed = 0
    def lcg(self):
        """Linear congruential generator. X is the seed (X0)."""
        self.x = (self.a * self.x + self.c) % self.m
        return (self.x / self.m)

    def get_seed():
        while True:
            string = str(int(time.time()*1000000))
            string = string[-8:]
            if string[0] != "0":
                break
        self.seed = int(string)
```

4 Estudio de distribuciones.

Detallaremos a continuación las distintas distribuciones sobre las cuales realizaremos el estudio [2]. Se brinda además información del cálculo de los parámetros estadísticos y el cálculo de su función inversa.

4.1 Distribución Uniforme.

Es una distribución, en este caso de estudio nos enfocaremos en su forma continua, que tiene como parámetros los extremos de un intervalo [a,b], intervalo donde la probabilidad de ocurrencia es constante y es 0 fuera de él.
Su función de probabilidad se define entonces como:

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

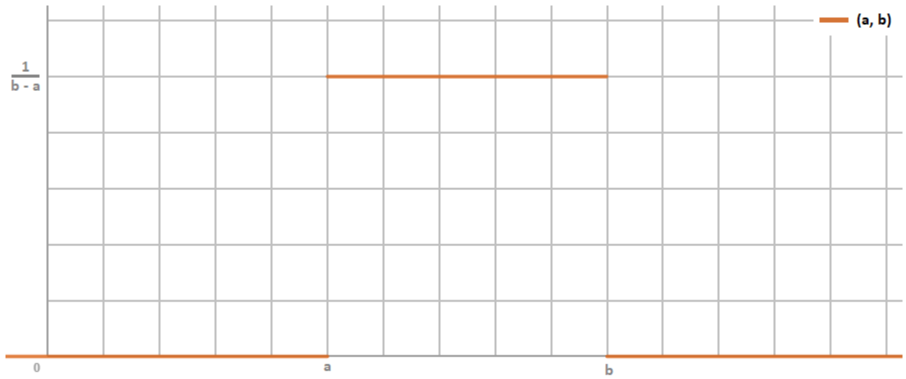

Figure 1: Probabilidad de variable X con distribución uniforme con parámetros a y b.

Parámetros estadísticos que nos serán útiles posteriormente son la media poblacional, la varianza y el desvío:

$$E(x) = \int_a^b \frac{1}{b-a} x\, dx = \frac{b-a}{2} \tag{3}$$

$$V(x) = \int_a^b \frac{(x-E(x))^2}{b-a}\, dx = \frac{(b-a)^2}{12} \tag{4}$$

$$D(x) = \sqrt{V(x)} \tag{5}$$

Al integrar la función de probabilidad (Formula 2) en cierto intervalo se obtiene la siguiente función de distribución acumulativa:

$$F(x) = \int_a^x \frac{1}{b-a}\, dt = \frac{x-a}{b-a} \tag{6}$$

A partir de esto podemos encontrar la función de mapeo o inversa de la siguiente manera:

$$F(x) = r = \frac{x-a}{b-a} \Rightarrow (b-a)r + a = x \Rightarrow F^{-1}(x) = (b-a)x + a = y : y \sim U(a,b) \tag{7}$$

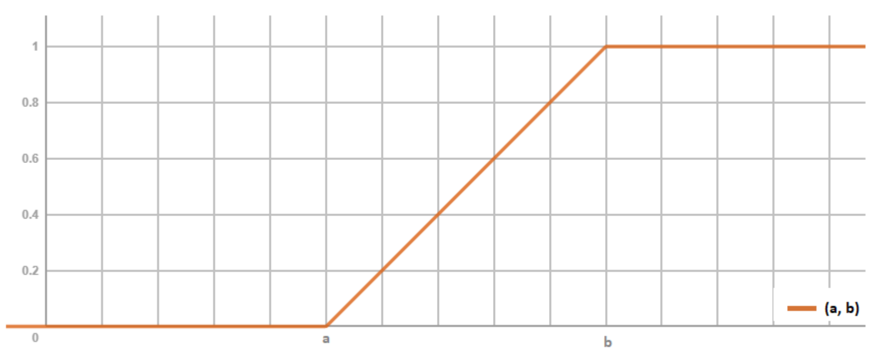

Figure 2: Función de frecuencia acumulada de la variable X con distribución uniforme con parámetros a y b.

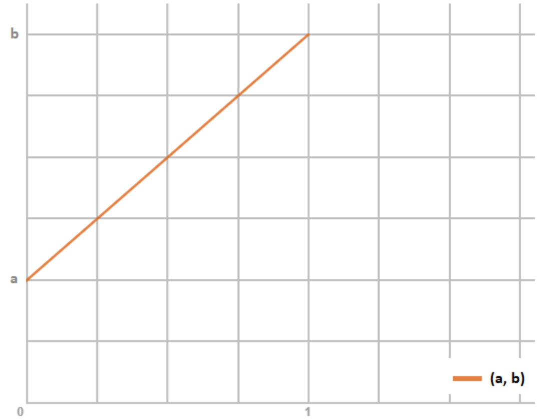

Figure 3: Función inversa de frecuencia acumulada o función de mapeo de la variable X con distribución uniforme con parámetros a y b.

4.2 Distribución exponencial.

esta distribución se refiere al período entre la ocurrencia de dos eventos, en la que a mayor intervalo entre ellos menor probabilidad, siguiendo el comportamiento de la exponencial con exponente negativo. Recibe como parámetro: λ, siendo éste el tiempo promedio esperado. La función de densidad es la siguiente. 4

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \text{ con } \lambda > 0 \tag{8}$$

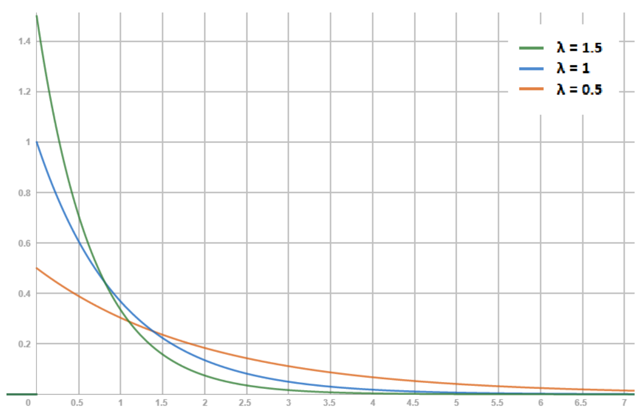

Figure 4: Probabilidad de variable X con distribución exponencial con parámetro λ.

Integramos para obtener su función de frecuencia acumulada:

$$F(x) = \int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \tag{9}$$

Procedemos a encontrar su función inversa:

$$F(x) = r = 1 - e^{-\lambda x} \Rightarrow \frac{ln(-r+1)}{-\lambda} = x \Rightarrow F^{-1}(x) = \frac{ln(-x+1)}{-\lambda} = y : y \sim Exp(\lambda) \tag{10}$$

Existe una forma un poco más simple y consiste en que considerar que:

$$1 - F(x) = e^{-\lambda x} = y \tag{11}$$

$$Im\left[e^{-\lambda x}\right] = \{y \in \mathbb{R} \wedge y \in (0;1)\} \tag{12}$$

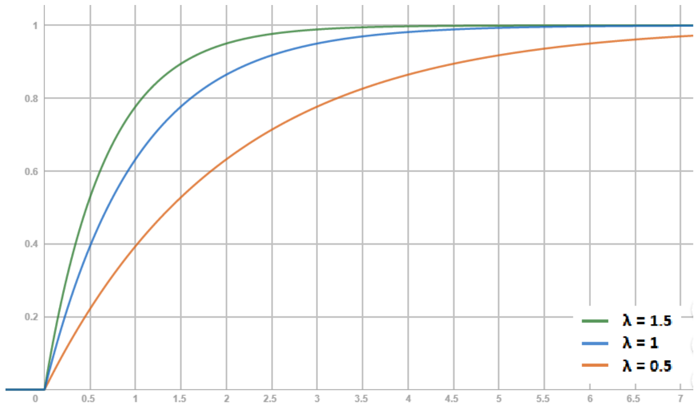

Figure 5: Función de frecuencia acumulada de variable x con distribución exponencial con parámetro λ.

$$e^{-\lambda x} = y = r \Rightarrow \frac{lg(r)}{-\lambda} = x \Rightarrow F^{-1}(x) = \frac{lg(x)}{-\lambda} = y : y \sim Exp(\lambda) \tag{13}$$

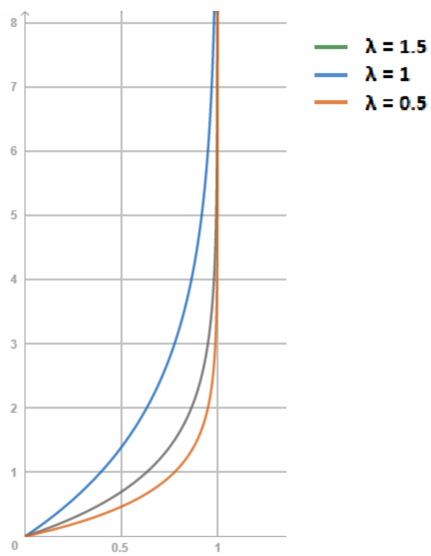

Figure 6: Función de frecuencia acumulada inversa de variable X con distribución exponencial con parámetro λ.

4.3 Distribución Gamma. [3]

Esta distribución se basa en un proceso formado por *k* eventos y el tiempo total de ese proceso se puede considerar como la suma de *k* valores de una variable x con distribución exponencial. De esta forma, los parámetros de esta distribución son la cantidad de eventos *k* y el tiempo promedio λ de la distribución exponencial. Para la función de densidad es necesario definir, primero, la *Función Gamma* $\Gamma(z)$ [10]:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \tag{14}$$

Y así nuestra función de probabilidad es:

$$f(x) = \frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x} \tag{15}$$

Considerando que

$$\text{si } n \in \mathbb{N} \Rightarrow \Gamma(n) = (n-1)! \tag{16}$$

simplificamos nuestra función de densidad a:

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \tag{17}$$

Si integramos para obtener la función de frecuencia acumulada:

$$F(x) = \frac{\lambda^k}{\Gamma(k)}\int_0^x t^{k-1}e^{-\lambda t}dt = \frac{\lambda^k}{(k-1)!)}\int_0^x t^{k-1}e^{-\lambda t}dt \tag{18}$$

Considerando la dificultad de calcular la función inversa de dicha función de densidad acumulada, para encontrar la función de mapeo hacemos uso del concepto enunciado al comienzo de esta sección. Consideramos la variable aleatoria de esta distribución como la sumatoria de *k* variables con distribución exponencial.

$$x \sim \Gamma(k,\lambda) \Leftrightarrow x = \sum_1^k y \text{ con } y \sim Exp(\lambda) \tag{19}$$

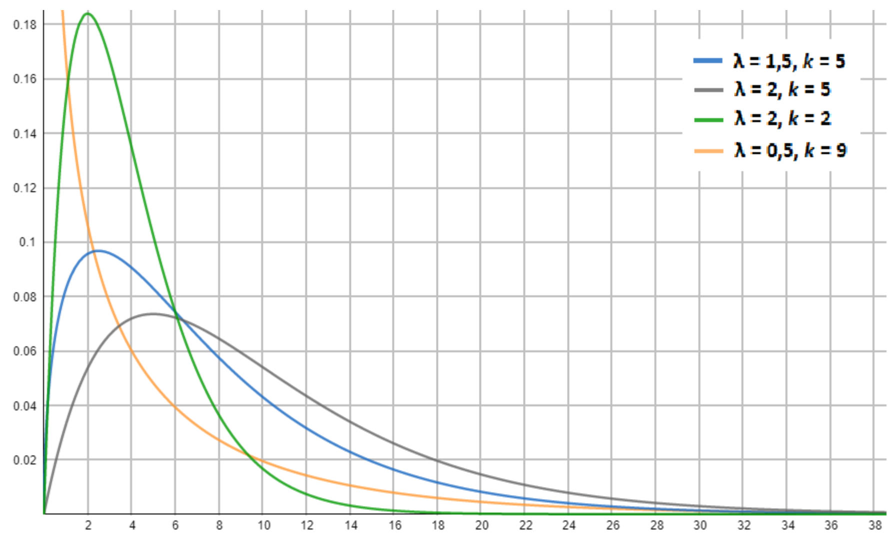

Figure 7: Probabilidad de variable X con distribución Gamma con parámetros k y λ.

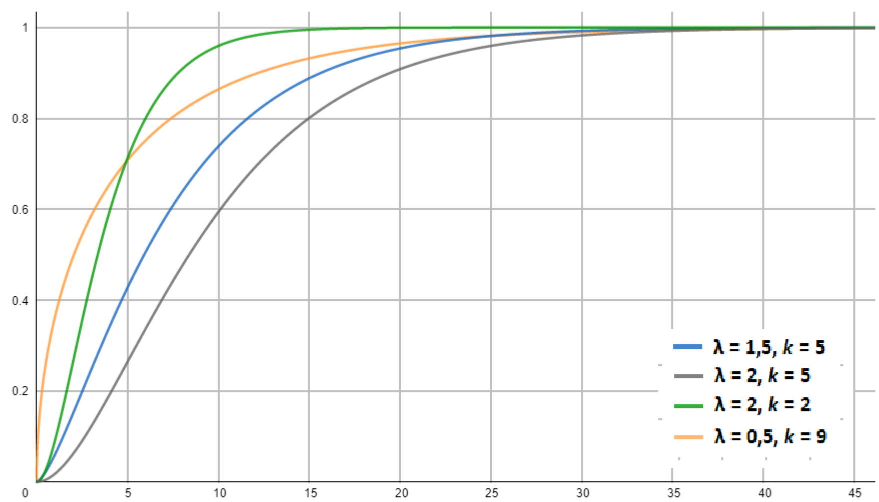

Figure 8: Función de frecuencia acumulada de la variable X con distribución Gamma con parámetros k y λ.

De esta manera, definimos la función de frecuencia acumulada inversa para la distribución Gamma como la sumatoria de funciones de frecuencia acumulada de la distribución Exponencial correspondiente:

$$F^{-1}(x) = \sum_{1}^{k} F^{-1}(y) = \sum_{1}^{k} \frac{ln(-x+1)}{-\lambda} \tag{20}$$

4.4 Distribución Normal. [4]

Esta distribución consiste en que la mayor densidad se encuentra exactamente en su media. Es muy común debido a que otras distribuciones, bajo ciertas condiciones tienen a la normalidad como lo demuestra el TCL, *Teorema Central del Límite*. Su parámetros son su media μ y su desvío estándar σ. Un caso particular y muy útil de esta distribución es la llamada Estándar, teniendo una media de 0 y una desviación de 1: $z \sim N(0,1)$.
Su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \text{ con } \mu \in \mathbb{R} \text{ y } \sigma \geq 0 \tag{21}$$

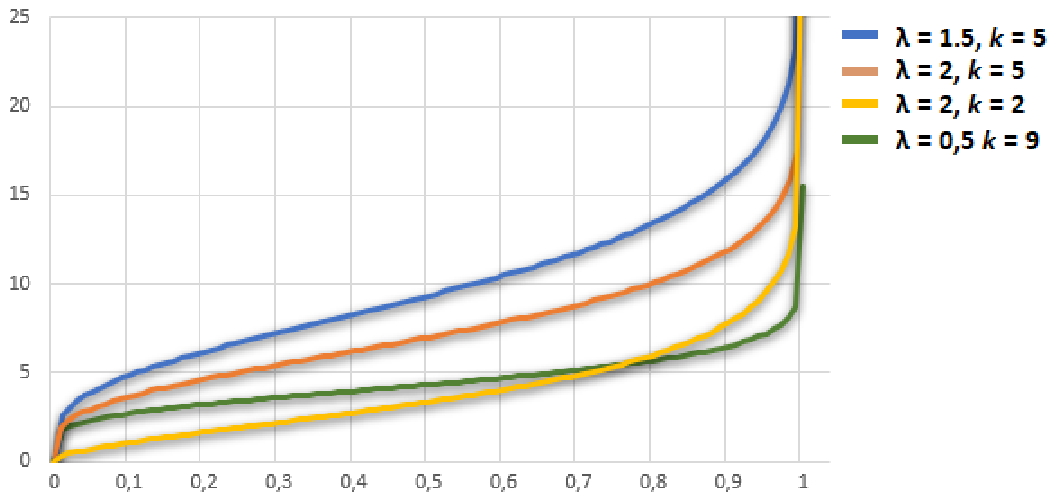

Figure 9: Función de mapeo de la variable X con distribución Gamma con parámetros k y λ.

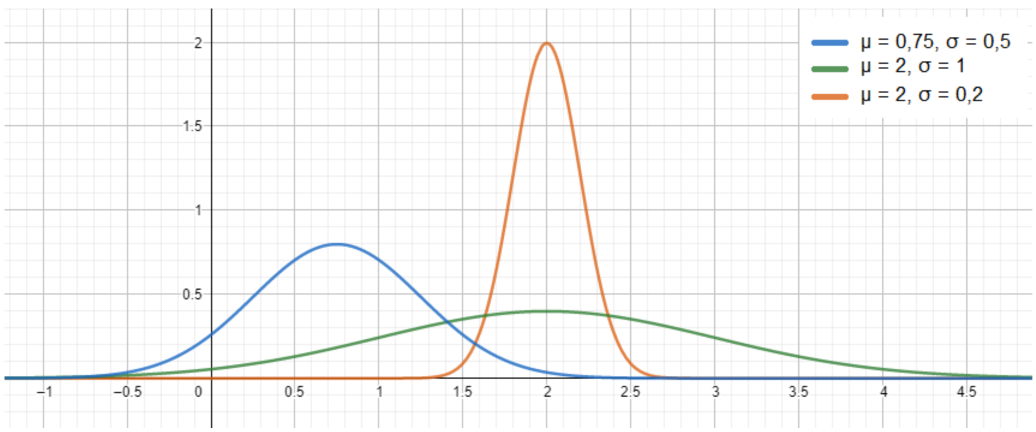

Figure 10: Función de densidad de la variable X con distribución Normal con parámetros μ y σ.

Es posible estandarizar cualquier variable con distribución normal, esto es muy útil, ya que las probabilidades de una variable y su forma estandarizada es la misma y en dichas probabilidades ya se encuentran calculadas para esta última, lo cual muchas veces facilita el cálculo. La fórmula es la siguiente:

$$z_{(0,1)} = \frac{x_{(\mu,\sigma)} - \mu_x}{\sigma_x} \tag{22}$$

Al integrar para obtener su función de frecuencia acumulada obtenemos:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-(t-\mu)^2}{2\sigma^2}} dt \tag{23}$$

También puede ser descrita utilizando la *Función de error* o *Función de error de Gauss*: [6]

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_{0}^{z} e^{-t^2} dt \tag{24}$$

de la siguiente manera:

$$F(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right] \tag{25}$$

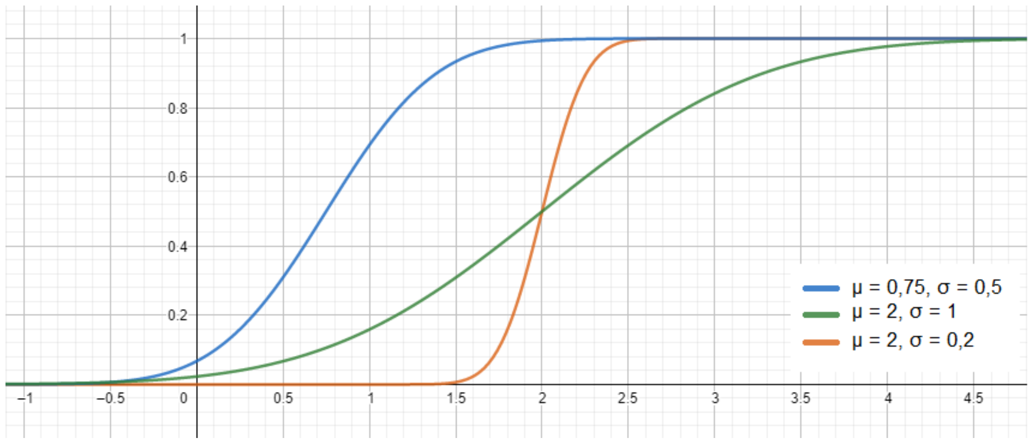

Figure 11: Función de densidad de la variable X con distribución Normal con parámetros μ y σ.

esta definición nos servirá para encontrar la función de frecuencia acumulativa inversa. Para eso primero obtenemos la función de error inversa:

$$erf^{-1}(z) = \sum_{k=0}^{\infty} \frac{c_k}{2k+1} \left(\frac{\sqrt{\pi}}{2} z \right)^{2k+1} \tag{26}$$

siendo,

$$c_0 = 1 \text{ y } c_k = \sum_{m=0}^{k-1} \frac{c_m c_{k-1} - m}{(m+1)(2m+1)} \tag{27}$$

Habiendo ya descrito las funciones necesarias podemos ahora hacer lo mismo con la Función de mapeo:[5]

$$F^{-1}(x) = \mu + \sigma\sqrt{2}\, erf^{-1}(2x-1) = y \tag{28}$$

Siendo y ya un valor de la variable x con distribución Normal, con parámetros μ y σ.

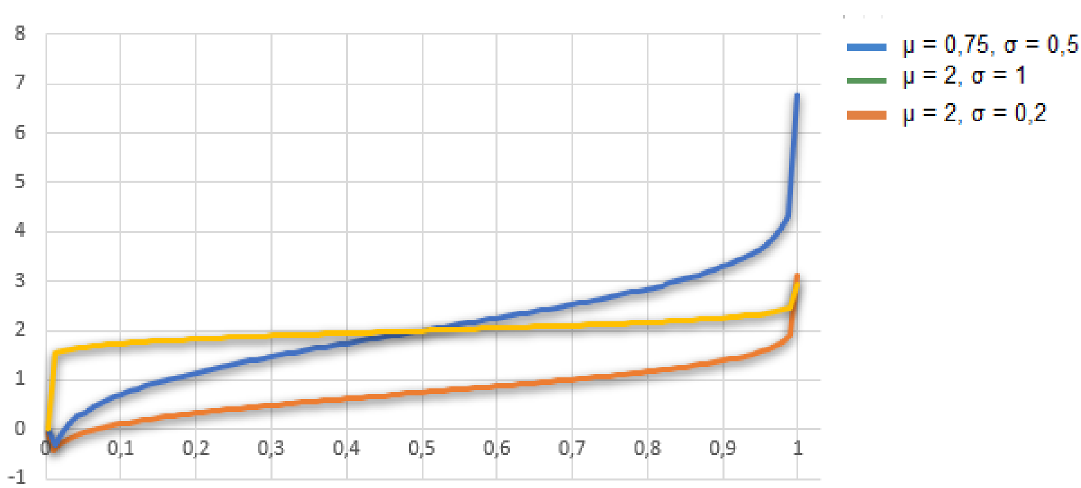

Figure 12: Función de mapeo de la variable X con distribución Normal con parámetros μ y σ.

Considerando la gran dificultad para conseguir valores de dicha función de mapeo, encontraremos los valores para esta distribución utilizando el ya mencionado TCL aplicado a éste caso, procedemos a enunciarlo:

Teorema Central del Límite. [7] Teniendo r_i variables independientes distribuidas uniformemente con parámetros (0,1), definimos $E(r_i) = \mu$ y $V(r_i) = \sigma^2$, entonces:

$$\lim_{N\to\infty} P\left[a < \frac{\sum_{i=1}^{N} r_i - r_i\mu}{\sqrt{N}\sigma} < b\right] = \frac{1}{\sqrt{2\pi}}\int_a^b e^{\frac{1}{2}z^2}dz \tag{29}$$

esta ecuación nos permite obtener un valor de z, es decir un valor estandarizado. De esta forma podemos generar K valores uniformemente distribuidos mediante nuestro generador GLC y aplicarle esta transformación. Para eso primero obtenemos los parámetros estadísticos de la distribución uniforme de dicho generador. $E(x) = \frac{1}{2}$, $Desv(x) = \frac{1}{\sqrt{12}}$
De esta forma nuestro valor de z queda definido por:

$$z = \frac{\sum_{i=1}^{K} r_i - \frac{K}{2}}{\sqrt{\frac{K}{12}}} \tag{30}$$

Aplicando la definición de Estandarización de una variable distribuida normalmente y despejando obtenemos:

$$z = \frac{x-\mu}{\sigma} = \frac{\sum_{i=1}^{K} r_i - \frac{K}{2}}{\sqrt{\frac{K}{12}}} \Rightarrow x = \frac{\sum_{i=1}^{K} r_i - \frac{K}{2}}{\sqrt{\frac{K}{12}}}\sigma + \mu \tag{31}$$

4.5 Distribución binomial.

La distribución discreta más sencilla, se basa en el *Experimento de Bernoulli*, en el que se registra para varias iteraciones de un evento sin memoria, uno de los dos posibles resultados, éxito o fracaso. En ésta distribución se expresa la probabilidad de obtener x éxitos en un experimento de Bernoulli con n ensayos, con una probabilidad de éxito p, siendo éstos últimos dos los parámetros de la distribución. De p obtenemos la probabilidad de fracasos que es $q = 1 - p$.
La función de distribución está dada por:

$$f(x) = \binom{n}{x} p^x q^{n-x} \tag{32}$$

Parámetros estadísticos de esta distribución son:

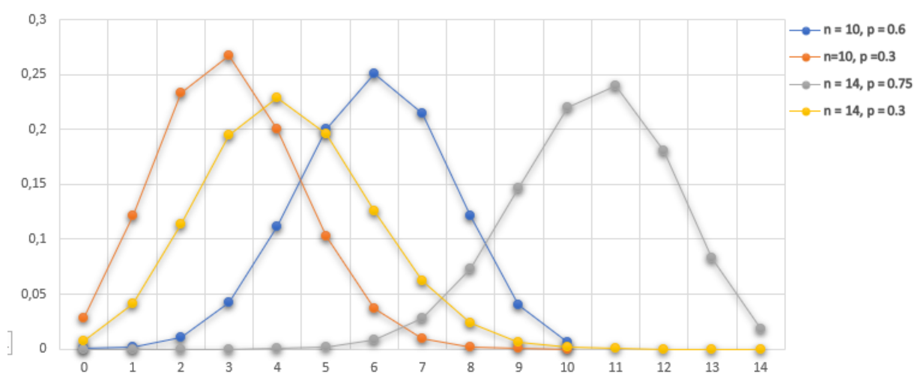

Figure 13: Función de densidad de la variable x con distribución Binomial con parámetros n y p.

$$E(x) = n\,p,\ V(x) = n\,p\,q \text{ y } D(x) = \sqrt{V(x)} \tag{33}$$

En las funciones discretas, las funciones acumulativas son sumatorias, ya no integrales, al tomar la variable, justamente, valores discretos.

$$F(x) = \sum_{i=0}^{x} \binom{n}{i} p^i q^{n-i} \tag{34}$$

Para las distribuciones discretas encontrar las funciones de frecuencia acumulativa inversa se vuelve más complejo, se puede observar en el siguiente ejemplo. Si tenemos una variable $x \sim Bi(5,\ 0,4)$, no existe x cuya $F(x) = 0.5$.

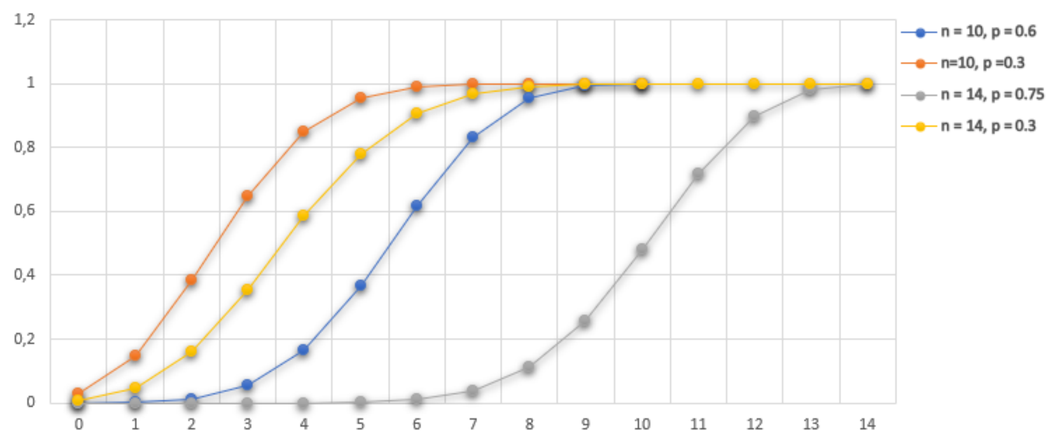

Figure 14: Función de frecuencia acumulada de la variable x con distribución Binomial con parámetros n y p.

$F(x = 1) = 0.3370$ y $F(x = 2) = 0.6826$. De ésta manera visualizamos como $F^{-1}(0,5) \nexists$ y nuestra tarea se convierte en encontrar el entero x más pequeño que satisfaga la siguiente inecuación: [8]

$$r \leq \sum_{i=0}^{x} \binom{n}{i} p^i q^{n-i} \tag{35}$$

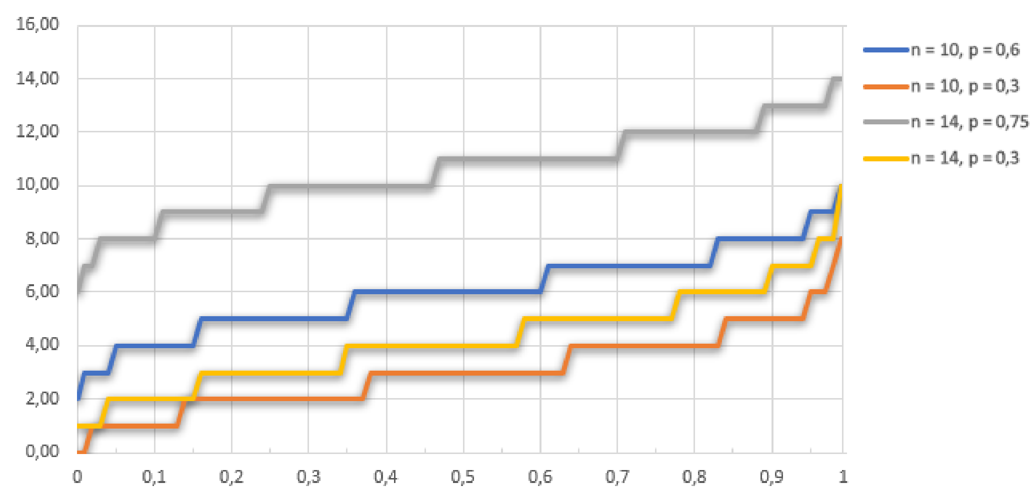

Figure 15: Función de mapeo de la variable x con distribución Binomial con parámetros n y p.

Para eso se podría iterar utilizando la función de frecuencia acumulada hasta encontrar dicho valor x pero eso conllevaría mucho tiempo de procesamiento. La forma más sencilla de generar valores para ésta distribución es simular un experimento de Bernoulli. Con nuestro generador GLC, generamos n valores y dependiendo si el valor pertenece a p aumentamos un uno el valor asignado al éxito, en caso contrario al fracaso.

$$x_i = x_i + 1 \text{ si } r_i \leq p \tag{36}$$

$$x_i = x_i \text{ si } r_i > p \tag{37}$$

4.6 Distribución Geométrica.

Ésta distribución calcula la probabilidad de que en un experimento de Bernoulli se obtenga el primer éxito en el ensayo número x. El parámetro es el mismo que el de la distribución binomial, la probabilidad p. De igual manera $q = 1 - p$ Su función de densidad está dada por:

$$f(x) = pq^{x-1} \tag{38}$$

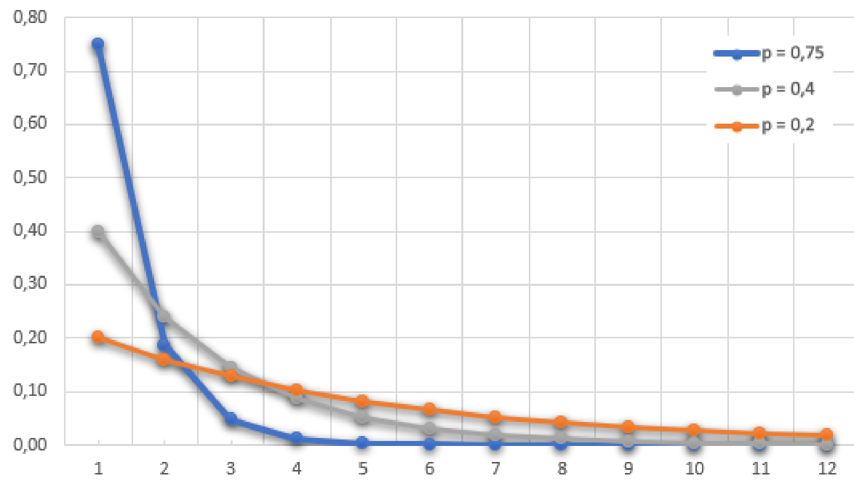

Figure 16: Función de densidad de la variable x con distribución Geométrica con parámetro p.

Para la función acumulativa realizamos la sumatoria:

$$F(x) = p\sum_{i=1}^{x} q^{i-1} = P(X \leq x) = 1 - q^{x+1} \tag{39}$$

Obtenemos la inversa de la Función de frecuencia acumulada:

$$F(x) = r = 1 - q^{x+1} \Rightarrow \left\lfloor \frac{log(1-r)}{log(q)} \right\rfloor = x \Rightarrow F^{-1}(x) = \left\lfloor \frac{log(1-x)}{log(q)} \right\rfloor = y \tag{40}$$

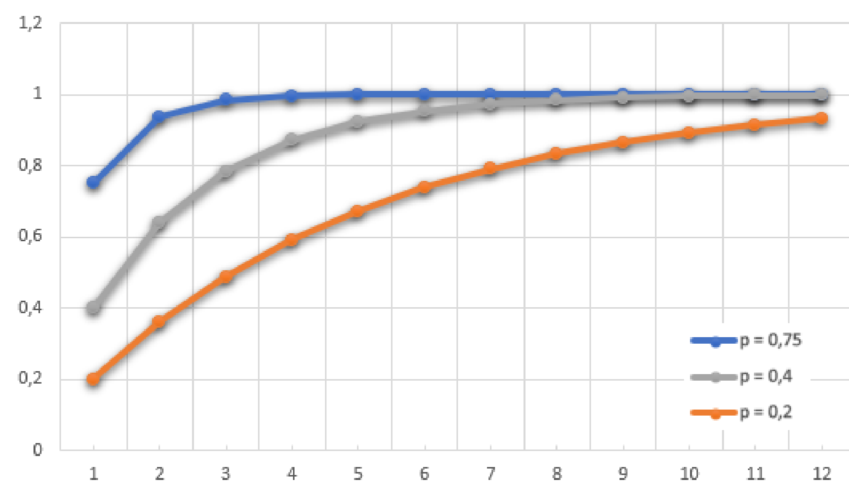

Figure 17: Función de frecuencia acumulada de la variable x con distribución Geométrica con parámetro p.

Podemos de todas formas, simplificar un poco más la función haciendo lo siguiente:

$$1 - F(x) = q^{x+1} = q^x \cdot q \Rightarrow \frac{1-F(x)}{q} = q^x \tag{41}$$

Teniendo en cuenta que:

$$Im\left(\frac{1-F(x)}{q}\right) = \{y \in \mathbb{R} \ \wedge y \in (0;1)\} \therefore y \sim U(0,1) \tag{42}$$

Consideramos:

$$\frac{1 - F(x)}{q} = q^x = y = r \tag{43}$$

De ahí, nuestra función de mapeo queda de la siguiente manera:

$$r = q^x \Rightarrow x = \frac{log(r)}{log(q)} \Rightarrow F^{-1}(x) = \left\lfloor \frac{log(x)}{log(q)} \right\rfloor = y \tag{44}$$

4.7 Distribución Pascal.[11]

Ésta distribución consiste en emular un ensayo de Bernoulli con p probabilidad de éxito hasta que se consigan k éxitos, siendo la cantidad de fracasos hasta entonces la variable con distribución de Pascal. La distribución geométrica es un caso particular de esta distribución, en la que $k = 1$, es decir, la cantidad de fracasos hasta el primer éxito. La distribución binomial negativa es una extensión de ésta distribución donde k pertenece a $\mathbb{R}$. La función de probabilidad puntual es:

$$f(x) = \binom{k + x - 1}{x} p^k q^x \tag{45}$$

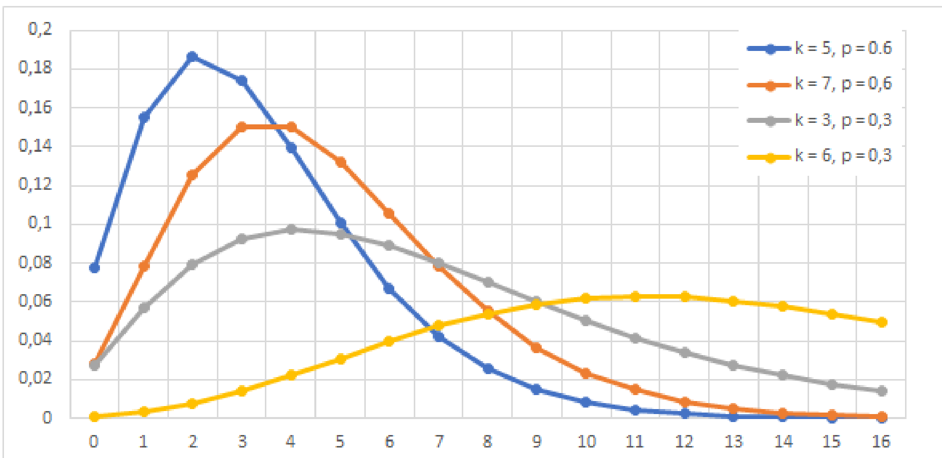

Figure 18: Función de densidad de la variable x con distribución Pascal con parámetros k y p.

La función de frecuencia acumulada se puede expresar como:

$$F(x) = \sum_{i=0}^{x} \binom{k + i - 1}{i} p^k q^i \tag{46}$$

Que puede también ser descrita usando la función beta incompleta regularizada: [12]

$$I_p(k, x + 1) = \frac{B(p; k, x + 1)}{B(k, x + 1)} = \frac{\int_0^p t^{k+1}(1 - t)^x dt}{\int_0^1 t^{k+1}(1 - t)^x dt} \tag{47}$$

Considerando la dificultad de encontrar una función inversa de dicha función de densidad acumulativa, se procede a considerar a la función Pascal como una suma de k elementos de una variable aleatoria con distribución Geométrica.

$$x = \sum_{i=1}^{k} \frac{log(r_i)}{log(q)} \tag{48}$$

Manipulando podemos llegar a:

$$\frac{\sum_{i=1}^{k} log(r_i)}{log(q)} = \frac{log(\prod_{i=i}^{k} r_i)}{log(q)} \tag{49}$$

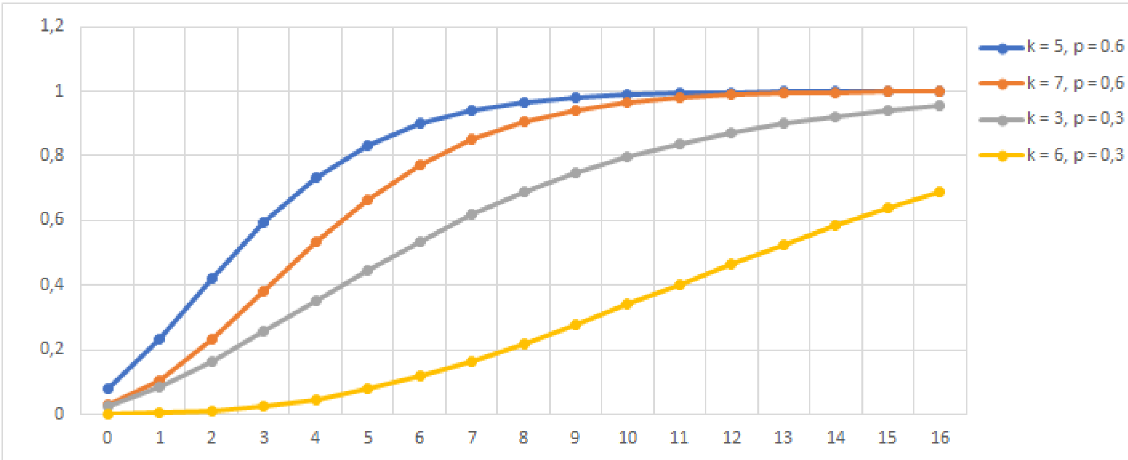

Figure 19: Función de densidad de la variable x con distribución Pascal con parámetros k y p.

4.8 Distribución Hipergeométrica.

Consiste en, teniendo una población de N elementos con una frecuencia absoluta de éxito p y una de fracaso $q = N - p$ y se toma una muestra sin reemplazo de n elementos, $n < N$ La variable x : *cantidad de elementos exitosos pertenecientes a* n se comporta de manera Hipergeométrica con parámetros, N, p y n.
La función de probabilidad es:

$$f(x) = \frac{\binom{p}{x}\binom{N-p}{n-x}}{\binom{N}{n}} \tag{50}$$

La función de frecuencia acumulada es:

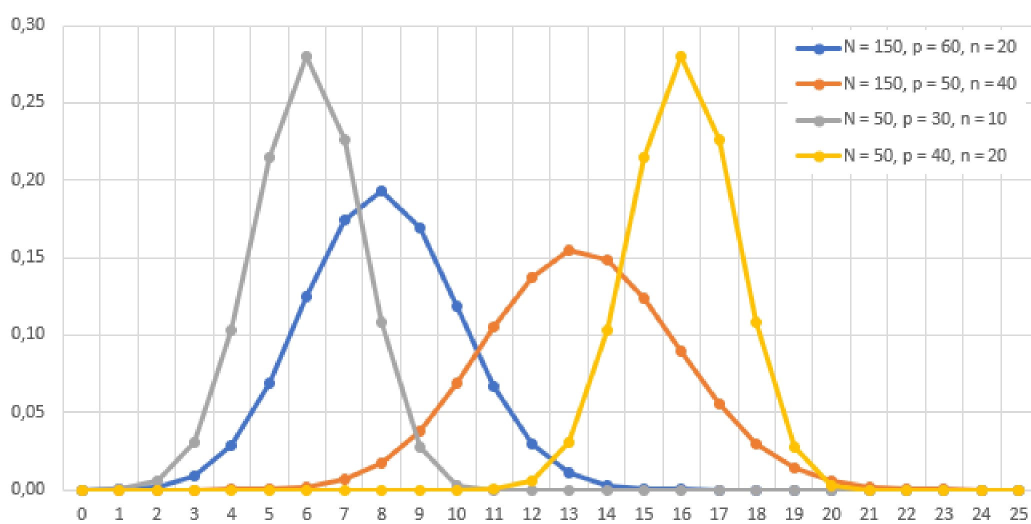

Figure 20: Función de distribución de la variable x con distribución Hipergeométrica con parámetros N, p y n.

$$F(x) = \frac{1}{\binom{N}{n}} \sum_{i=0}^{x} \binom{p}{i}\binom{N-p}{n-i} \tag{51}$$

Debido a la dificultad de obtener la función de mapeo para esta distribución, para generar valores computacionales se realiza igual que en la distribución Binomial, el ensayo de Bernoulli, asignando el valor que le corresponda en caso de éxito o fracaso, a la vez que se van reduciendo la cantidad de elementos disponibles en N y por ende en p.

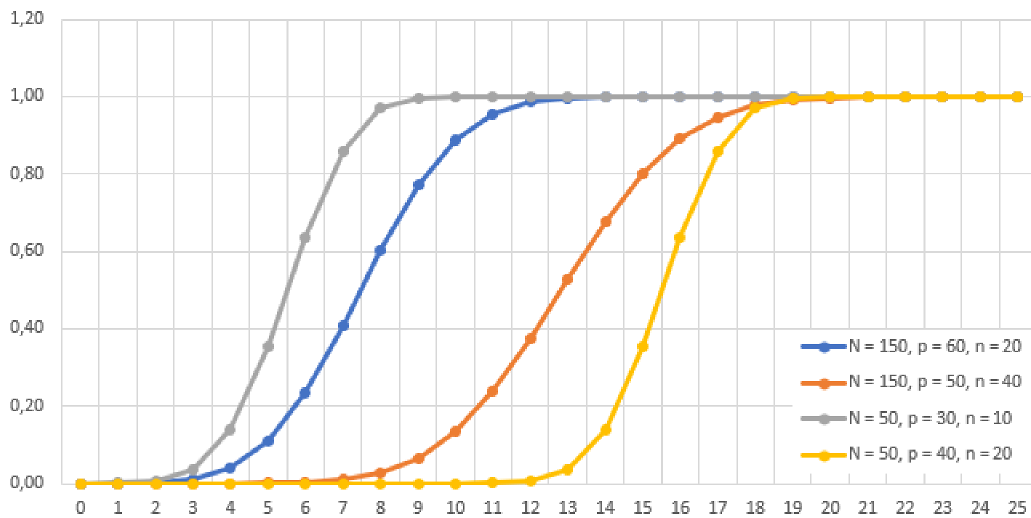

Figure 21: Función de frecuencia acumulada de la variable x con distribución Hipergeométrica con parámetros N, p y n.

4.9 Distribución de Poisson.

Ésta distribución describe la probabilidad de que en un intervalo de tiempo esperado λ (el mismo que en la distribución exponencial) ocurran x eventos independientes. Su función de densidad está dada por:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{52}$$

La función de frecuencia acumulada es:

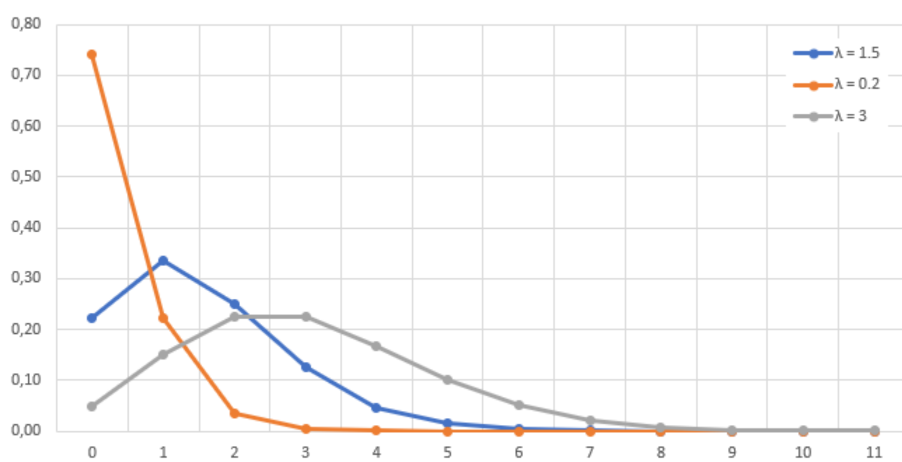

Figure 22: Función de distribución de la variable x con distribución Poisson con parámetro λ.

$$F(x) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} \tag{53}$$

Para generar valores con ésta distribución, se generan primero duraciones de intervalos con distribución exponencial con un parámetro $\lambda = 1$ hasta que la suma de éstos alcance el valor de λ

$$\sum_{i=0}^{x} t_i \leq \lambda < \sum_{i=0}^{x+1} t_i \tag{54}$$

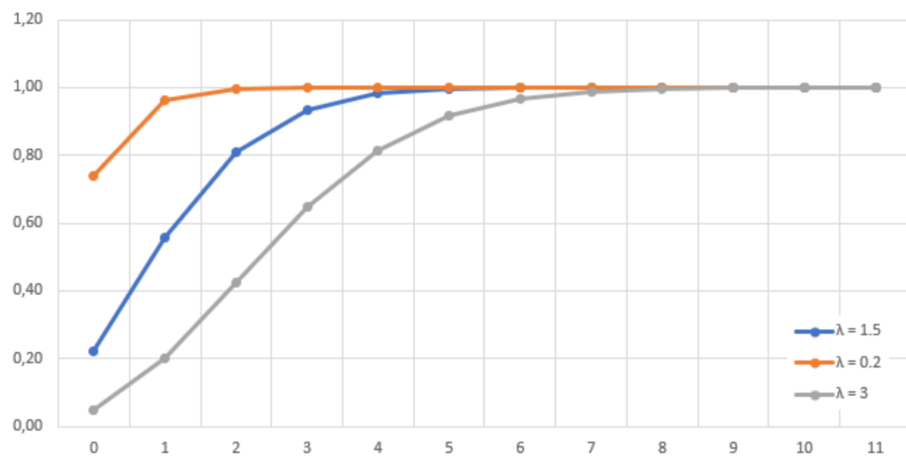

Figure 23: Función de frecuencia acumulada de la variable x con distribución Poisson con parámetro λ.

siendo que $t_i = -log(r_i)$ la fórmula (49) se puede reescribir de la siguiente manera:

$$\sum_{i=0}^{x} -log(r_i) \leq \lambda < \sum_{i=0}^{x+1} -log(r_i) \Rightarrow -log(\prod_{i=0}^{x} r_i) \leq \lambda < -log(\prod_{i=0}^{x+1} r_i) \Rightarrow log(\prod_{i=0}^{x} r_i) \geq -\lambda > log(\prod_{i=0}^{x+1} r_i) \quad (55)$$

y por último

$$\prod_{i=0}^{x} r_i \geq e^{-\lambda} > \prod_{i=0}^{x+1} r_i \quad (56)$$

De esta manera para generar nuestro números aleatorios con distribución de Poisson, nos basamos en la fórmula (51) multiplicamos tantos r_i distribuidos uniformemente hasta que sea menor que $e^{-\lambda}$, aumentando el posible valor Poisson en 1 cada vez.

4.10 Distribución empírica discreta.

En algunas circunstancias es necesario replicar la distribución de una muestra observada, para extenderla a una población. En este caso nos enfocaremos en una distribución muestral discreta, haciendo la aclaración de que existen de igual manera distribuciones empíricas continuas. Teniendo:

Table 1: Tabla de probabilidad puntual de una muestra de una variable aleatoria b

b_i	$P(b_i)$
b_1	p_1
b_2	p_2
b_3	p_3
.	.
.	.
b_n	p_n

No tenemos su función de probabilidad, pero sí podemos encontrar su función de frecuencia acumulativa.

$$F(b_i) = \sum_{k=1}^{i} p_k \quad (57)$$

Con esto es suficiente para poder generar números aleatorios que se comporten con ésta distribución. Para eso le asignamos a cada b_i un intervalo de la siguiente manera:

$$\text{para } b_i \rightarrow I_i = \{r \in \mathbb{R} | F(b_{i-1}) < r \leq F(b_i)\} \quad (58)$$

Podemos ahora, de manera similar al procedimiento utilizado en la distribución binomial, generar valores uniformes r_i y asignarle el valor b_i según el intervalo I_i en el que pertenezca dicho r_i.

5 Implementación en Python de distribuciones.

5.1 Distribución Uniforme.

```
def uniforme(self, a,b):
    r = self.gen.lcg()
    x = a + (b-a)*r
    return x
```

5.2 Distribución exponencial.

```
def exponencial(self, a):
    ex = 1/a
    r = self.gen.lcg()
    x = -ex * log(r)
    return x
```

5.3 Distribución Gamma.

```
def gamma(self, k, a):
    tr = 1.0
    for i in range(k):
        r = self.gen.lcg()
        tr = tr * r
    x = -log(tr) / a
    return x
```

5.4 Distribución Normal.

```
def normal(self, mu, std, K):
    suma = 0
    for i in range(K):
        r = self.gen.lcg()
        suma = suma + r
    x = std * (suma - K/2) / sqrt(K/12) + mu
    return x
```

5.5 Distribución binomial.

```
def binomial(self, N, p):
    x = 0
    for i in range(N):
        r = self.gen.lcg()
        if ((r-p) <= 0):
            x = x + 1
    return x
```

5.6 Distribución Pascal.

```
def pascal(self,k,q):
    tr = 1.0
    qr = log(q)
    for i in range(k):
        r = self.gen.lcg()
```

```
        tr = tr * r
    x = log(tr)/qr
    return x
```

5.7 Distribución Hipergeométrica.

```
def hipergeometrica(self, N, p, m):
    x = 0.0
    for i in range(m):
        r = self.gen.lcg()
        if (r-p)<= 0:
            s = 1.0
            x += 1.0
        else:
            s = 0.0
        p = (N * p - s) / (N - 1.0)
        N -= 1.0
    return x
```

5.8 Distribución de Poisson.

```
def poisson(self, l):
    x = 0.0
    b = exp(-l)
    t = 1.0
    while True:
        r = self.gen.lcg()
        t = t * r
        if (t - b)<= 0:
            break
        else:
            x += 1.0
    return x
```

5.9 Distribución empírica discreta.

```
def crear_casillas_empirica(self,muestra):
    casillas=[]
    casillas.append(-0.0001)
    pos = 0
    for i in muestra:
        pos +=  i[1]
        casillas.append(pos)
    return casillas

def empirica(self, muestra,casillas):
    r = self.gen.lcg()
    for i in range(1,len(casillas)):
        if casillas[i-1] < r <= casillas[i]:
            return muestra[i-1][0]
```

6 Métodos de testeo.

El método principal que utilizamos para medir la precisión de nuestros generadores es el método de bondad de ajuste utilizando Chi cuadrado desarrollado en el trabajo anterior, que consiste en medir la probabilidad de obtener el resultado obtenido suponiendo que la hipótesis nula es cierta, en nuestro caso, teniendo una variable aleatoria V con una distribución D, la hipótesis a comprobar será:

$$H_0 : \text{La variable } X \text{ se comporta con distribución } D. \tag{59}$$

En el caso de la distribución Normal utilizamos el método de *K cuadrado de D'Agostino.*[13] el cuál es un método de bondad de ajuste aplicado a distribuciones normales que mide la asimetría y la curtosis (distribución de los elementos con respecto a la media) de la muestra. Este método se desarrolla en el Anexo 1.

Para la distribución Exponencial utilizamos el método de *Anderson-Darling* [14] que evalúa si una se comporta con cierta distribución. Se ordena la muestra y se analiza si se comporta con cierta función de frecuencia acumulada teórica. Éste procedimiento se desarrolla en el Anexo 2.

7 Herramientas utilizadas.

7.1 Lenguaje de programación.

Se utilizó el lenguaje de programación **Python** en su versión 3 para el desarrollo de la simulación. Se utilizaron además módulos de la biblioteca estándar de Python, asi como bibliotecas externas para añadir más funcionalidad.

7.1.1 Bibliotecas y módulos.

Para realizar las gráficas de las distribuciones se utilizó la biblioteca ***Matplotlib 3.1.1***, en particular el método *pyplot* perteneciente a dicha biblioteca.

De la biblioteca ***math*** se utilizaron varios métodos para operaciones matemáticas. En particular, se emplearon los métodos de *log()* para el cálculo de logaritmos, *exp()* para el cálculo de e^x, *trunc()* para el truncado de números, *ceil()* para la función techo de un número, *floor* para la función piso de un número, *factorial()* para el cálculo del factorial de un número, y *sqrt()* para el cálculo de la raíz cuadrada de un número.

Se utilizaron varios métodos del módulo **stats**, perteneciente a la biblioteca **scipy**, para facilitar los cálculos estadísticos. En particular se utilizaron: el método *chisquare()* para el cálculo de χ^2; los métodos *norm()* y *expon()* para generar la distribución de referencia en los test de la distribución normal y exponencial respectivamente. También se utilizaron los métodos *anderson()* para el test de Anderson-Darling [15], y *normaltest()* para el método de K^2 de D'agostino [16].

Para obtener el tiempo actual y utilizarlo en la semilla del generador de números pseudoaleatorios, se utilizó el módulo **time**.

7.2 Herramientas externas

Para graficar las funciones de densidad, funciones de frecuencia acumulada y funciones de mapeo de ejemplo se utilizó el sitio web GeoGebra [17] y el programa Microsoft Office Excel.

8 Resultados obtenidos.

8.1 Parámetros utilizados.

Para realizar el experimento se utilizaron los siguientes. Siendo los parámetros de la distribución Empírica, los elementos de una muestra arbitraria M de 7 elementos.

Table 2: Parámetros utilizados para cada distribución

Distribución	Parámetro	Valor
Uniforme	a	1
	b	2
Exponencial	λ	1.3
Gamma	k	5
	α	0.5
Normal	μ	2
	σ	0.75
	K	102
Pascal	k	2
	p	0.2
Binomial	n	30
	p	0.45
Hipergeométrica	N	130
	p	0.2
	n	30
Poisson	λ	5

Table 3: Clases y probabilidades para la distribución Empírica

Distribución	Clase	Probabilidad
Empírica	0	0.15
	1	0.2
	2	0.1
	3	0.05
	4	0.1
	5	0.15
	6	0.25

8.2 Distribuciones Continuas

8.2.1 Distribución Uniforme

Se muestran la gráfica como el resultado del test para la distribución uniforme generada.

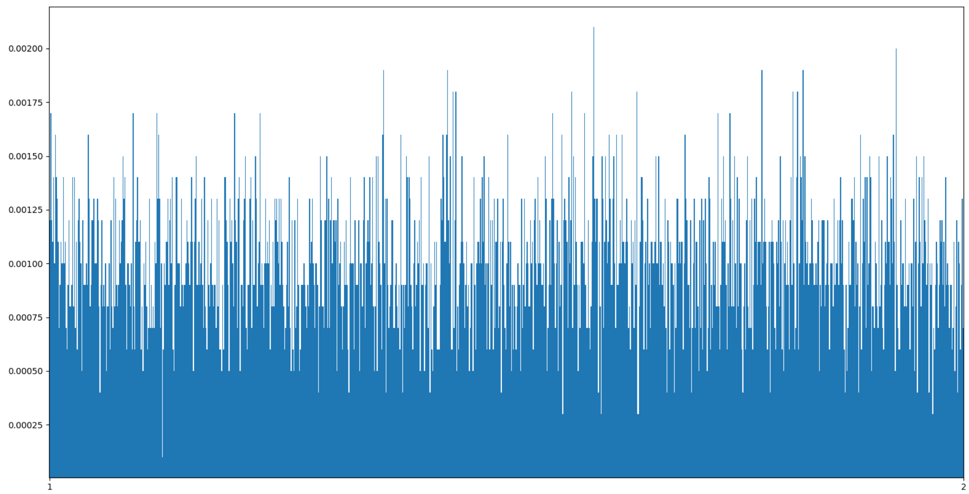

Figure 24: Distribución observada de la variable x con distribución Uniforme con parámetros $[a; b] = [1; 2]$.

Table 4: Resultados del test para distribución Uniforme

Tipo de test	P-valor	Resultado
Test de bondad de ajuste de χ^2	0.5877606886860093	Pasa

8.2.2 Distribución Gamma

Se muestra la gráfica para la distribución Gamma generada.

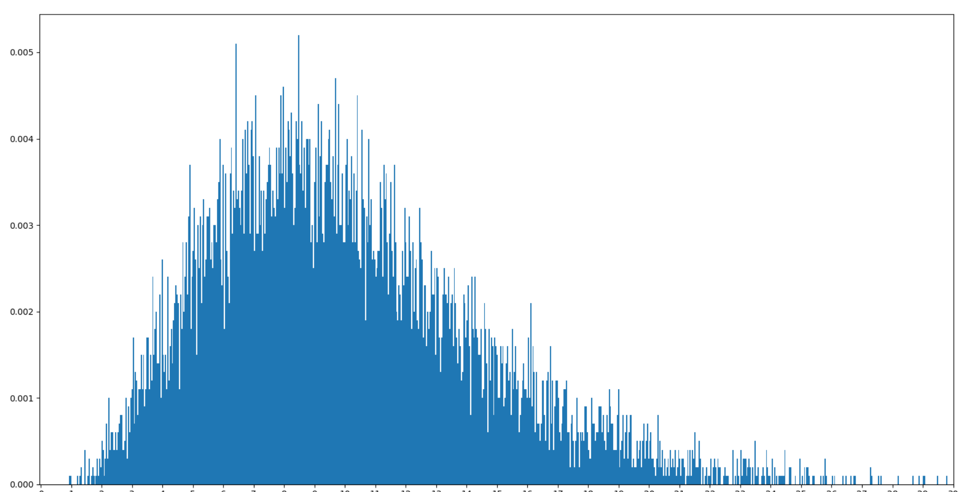

Figure 25: Distribución observada de la variable x con distribución Gamma con parámetros $k = 5$ y $\alpha = 0,5$.

8.2.3 Distribución Exponencial

Se muestran la gráfica como el resultado del test para la distribución Exponencial generada.

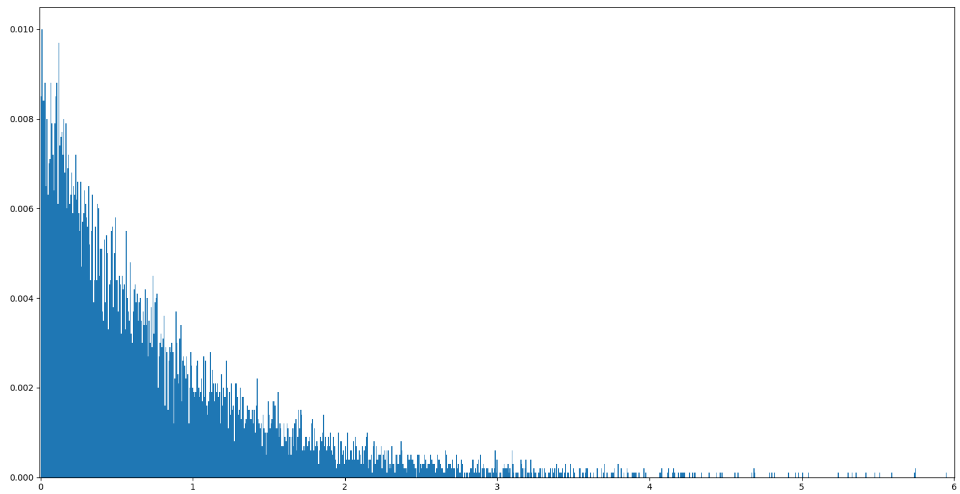

Figure 26: Distribución observada de la variable x con distribución Exponencial con parámetro $\lambda = 1,3$.

Table 5: Resultados del test para distribución Exponencial

Tipo de test	**Estadístico**	**Valor obtenido**	**Resultado**
Test de Anderson-Darling	A^2	0.2883518459912011	Pasa
Test de Anderson-Darling	Valor crítico (5% nivel de riesgo)	1.341	Pasa

8.2.4 Distribución Normal

Se muestran la gráfica como el resultado del test para la distribución Normal generada.

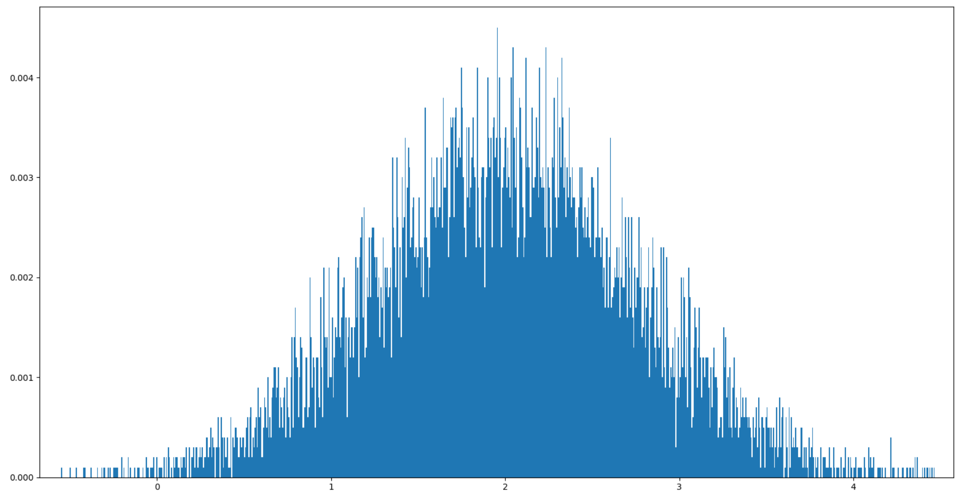

Figure 27: Distribución observada de la variable x con distribución Normal con parámetros $\mu = 2$ y $\sigma = 0,75$.

Table 6: Resultados del test para distribución Normal

Tipo de test	**Estadístico**	**Valor obtenido**	**Resultado**
Método de K^2 de D'agostino	P-valor	0.23547099748319072	Pasa

8.3 Distribuciones Discretas

8.3.1 Distribución Binomial

Se muestran la gráfica como el resultado del test para la distribución Binomial generada.

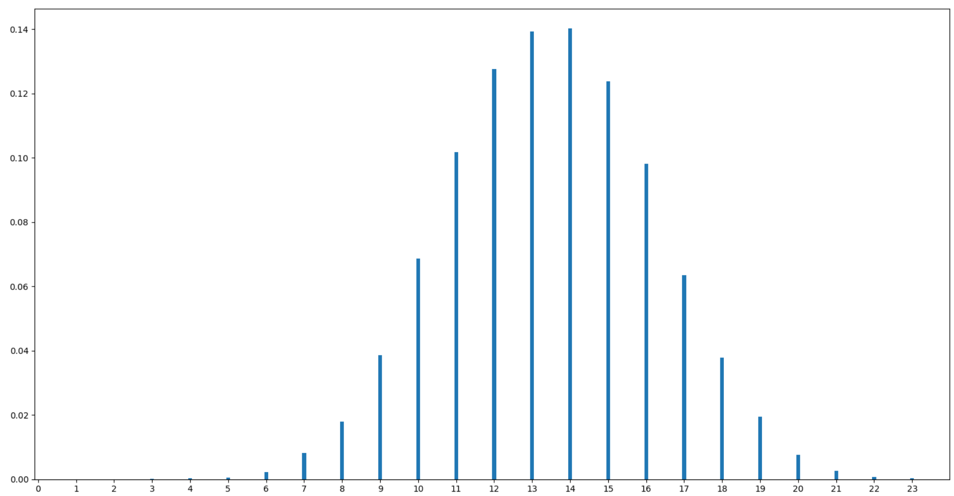

Figure 28: Función de distribución de la variable x con distribución Binomial con parámetros $n = 30$ y $p = 0,45$.

Table 7: Resultados del test para distribución Binomial

Tipo de test	**Estadístico**	**Valor obtenido**	**Resultado**
Test de bondad de ajuste de χ^2	P-valor	0.9745913219440376	Pasa

8.3.2 Distribución Poisson

Se muestran la gráfica como el resultado del test para la distribución Poisson generada.

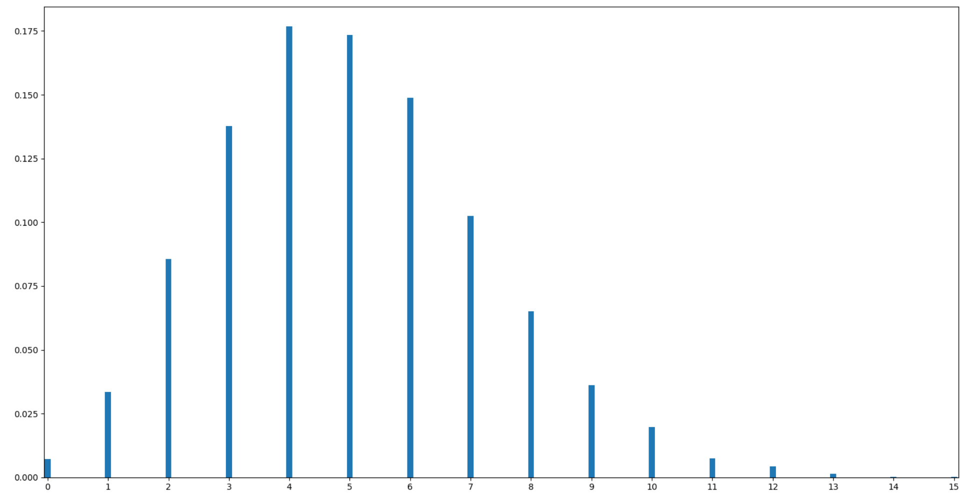

Figure 29: Función de distribución de la variable x con distribución Poisson con parámetro $\lambda = 5$.

Table 8: Resultados del test para distribución Poisson

Tipo de test	**Estadístico**	**Valor obtenido**	**Resultado**
Test de bondad de ajuste de χ^2	P-valor	0.9107564047803506	Pasa

8.3.3 Distribución Pascal

Se muestran la gráfica para la distribución Pascal generada.

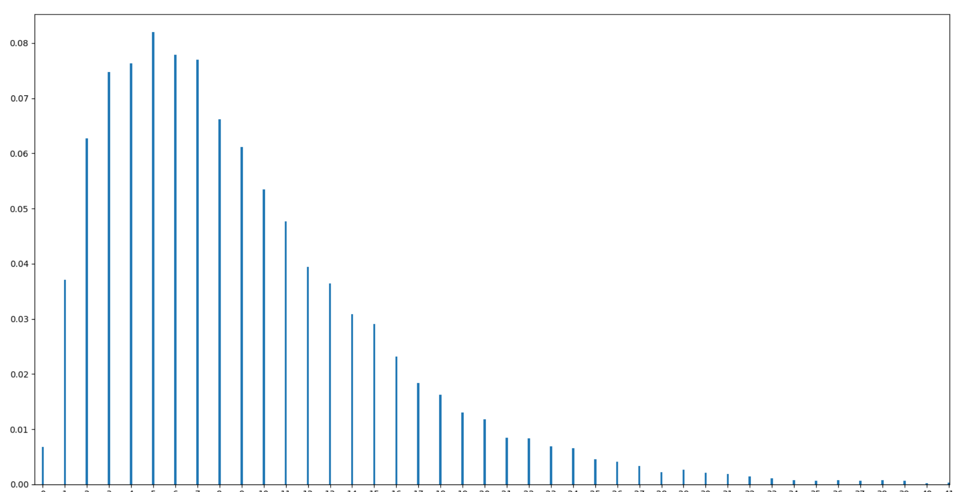

Figure 30: Función de distribución de la variable x con distribución Pascal con parámetros $k = 2$ y $p = 0,2$.

8.3.4 Distribución Hipergeométrica

Se muestran la gráfica para la distribución Hipergeométrica generada.

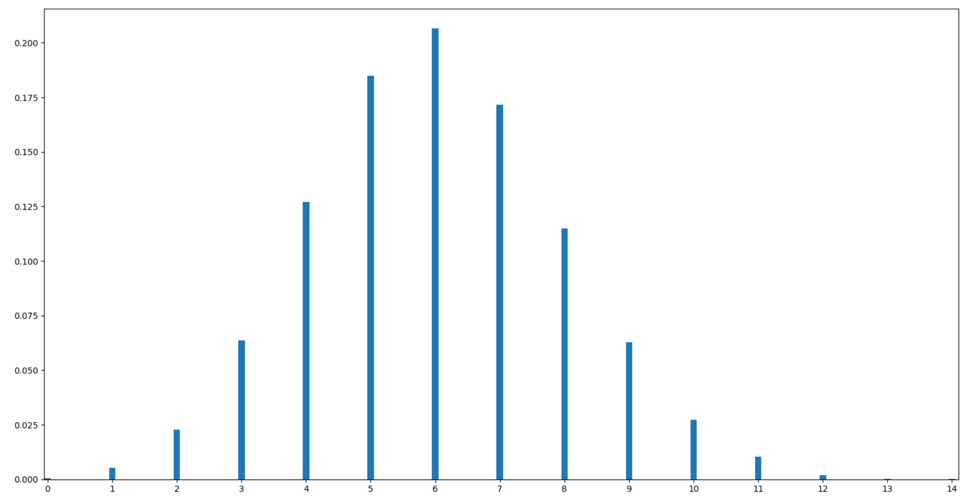

Figure 31: Función de distribución de la variable x con distribución Hipergeométrica con parámetros $N = 130$, $p = 0,2$ y $n = 30$.

8.3.5 Distribución Empírica discreta

Se muestran la gráfica como el resultado del test para la distribución Empírica discreta generada.

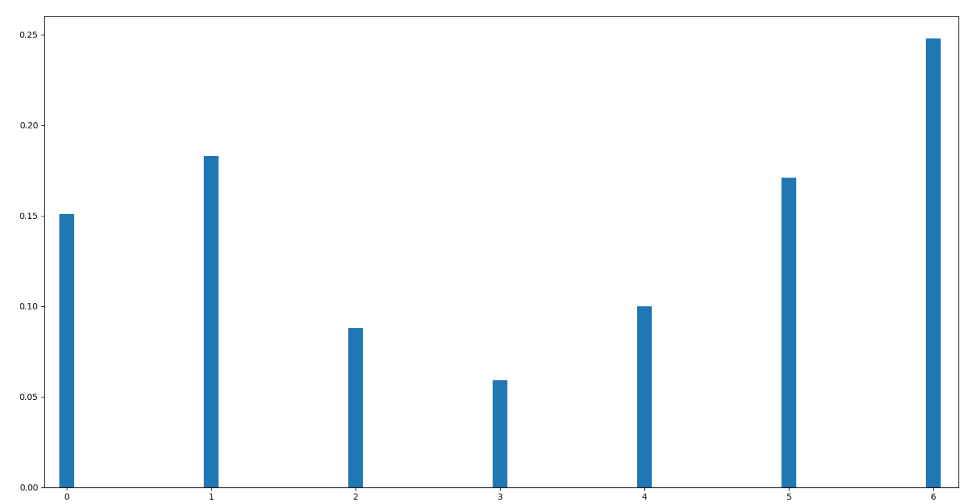

Figure 32: Función de distribución de la variable x con distribución Empírica discreta de la muestra M.

Table 9: Resultados del test para distribución Empírica discreta

Tipo de test	**Estadístico**	**Valor obtenido**	**Resultado**
Test de bondad de ajuste de χ^2	P-valor	0.6037788189702087	Pasa

9 Conclusiones.

Una vez obtenidos estos resultados, concluimos que generar números aleatorios con distribuciones específicas es posible y viable haciendo uso de los algoritmos adecuados. Esto dependerá del correcto análisis sobre la distribución en cuestión, así como de sus funciones asociadas, parámetros, estadísticos, sus relaciones con otras distribuciones, entre otras características de estudio. El resultado de este análisis, y emparejado con un generador de números pseudoaleatorios uniforme con imagen [0,1] abren paso a la posibilidad de generar muestras de números pseudoaleatorios que se comporten como si se trataran de una distribución de probabilidad en concreto, proveyendo así una potente herramienta para la modelización y la simulación de eventos.

References

[1] Methods and formulas for Inverse Cumulative Distribution Function (ICDF) En *https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/probability-distributions/how-to/inverse-cumulative-distribution-function-icdf/methods-and-formulas/methods-and-formulas/*.

[2] Generación de valores de las variables estocásticas empleadas en simulación En *Thomas H. Naylor*, *Técnicas de simulación en computadoras*, capítulo 4.

[3] Distribución Gamma En *https://la.mathworks.com/help/stats/gamma-distribution.html*.

[4] Distribución Normal En *https://la.mathworks.com/help/stats/normal-distribution.html?lang=en*.

[5] Quantile function § Normal distribution En *https://en.wikipedia.org/wiki/Normal_distribution#Quantile_function*.

[6] Error function En *https://en.wikipedia.org/wiki/Error_function#Inverse_functions*.

[7] Teorema central del límite En *https://es.wikipedia.org/wiki/Teorema_del_límite_central*.

[8] Efficient evaluation of the inverse Binomial cumulative distribution function where the number of trials is large En *https://www.lancaster.ac.uk/pg/moorhead/Work/SummerProject13.pdf*.

[9] Geometric distribution En *http://www.math.wm.edu/~leemis/chart/UDR/PDFs/Geometric.pdf*.

[10] Función gamma En *https://es.wikipedia.org/wiki/Función_gamma*.

[11] Negative binomial distribution En *https://en.wikipedia.org/wiki/Negative_binomial_distribution*.

[12] Incomplete beta function En *https://en.wikipedia.org/wiki/Beta_function#Incomplete_beta_function*.

[13] D'Agostino's K-squared test En *https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test*.

[14] Anderson–Darling test En *https://en.wikipedia.org/wiki/Anderson–Darling_test*.

[15] scipy.stats.anderson En *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html*.

[16] scipy.stats.normaltest En *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html*.

[17] GeoGebra En *https://www.geogebra.org/graphing*

[18] Curtosis En *https://es.wikipedia.org/wiki/Curtosis*.

[19] Skewness En *https://en.wikipedia.org/wiki/Skewness*.

[20] Momento central En *https://es.wikipedia.org/wiki/Momento_central*.

[21] Standardized moment En *https://en.wikipedia.org/wiki/Standardized_momen*.

Anexo I - Método de K^2 de D'agostino. [13]

Siendo M una muestra de n elementos, la muestra de asimetría M está dada por:

$$g_1(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^{\frac{3}{2}}} \tag{60}$$

La muestra de curtosis de la misma muestra se defino como:

$$g_2(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} - 3 \tag{61}$$

Se definen las siguientes funciones:

$$Z_1(g_1) = \delta \operatorname{asinh}\left(\frac{g_i}{\alpha\sqrt{\sigma^2(g_1)}}\right) \tag{62}$$

siendo $\sigma^2(M)$ la desviación estándar de M y

$$\delta = \frac{1}{\sqrt{ln(W)}} \tag{63}$$

$$\alpha^2 = \frac{2}{W^2 - 1} \tag{64}$$

$$W^2 = \sqrt{\gamma(g_1) + 4} - 1 \tag{65}$$

$$\gamma(M) = \frac{\mu_4(M)}{\sigma^4(M)} = \text{Curtosis de M[18]} \tag{66}$$

$$\mu_4(M) = \mathrm{E}\left[(X - \mathrm{E}[M])^4\right] = \int_{-\infty}^{+\infty}(x - \mu(M))^4 f(x)\,dx = \text{Momento centrado de orden 4 [20]} \tag{67}$$

Y a su vez la transformación

$$Z_2(g_2) = \sqrt{\frac{9A}{2}}\left[1 - \frac{2}{9A} - \left(\frac{1 - \frac{2}{A}}{1 + \frac{g_2 - \mu(g_2)}{\sigma^2(g_2)\sqrt{\frac{2}{A-4}}}}\right)^{\frac{1}{3}}\right] \tag{68}$$

siendo

$$A = 6 + \frac{8}{\beta(g_2)}\left(\frac{2}{\beta(g_2)} + \sqrt{1 + \frac{4}{\beta^2(g_2)}}\right) \tag{69}$$

y $\beta(M)$ la asimetría de M [19] o el momento centrado estandarizado de orden 3 de M [21]:

$$\beta(M) = \tilde{\mu}_3(M) = \mathrm{E}\left[\left(\frac{X - \mu(M)}{\sigma(M)}\right)^3\right] = \frac{\mu_3(M)}{\sigma^3(M)} = \frac{\mathrm{E}\left[(X - \mu(M))^3\right]}{(\mathrm{E}\left[(X - \mu(M))^2\right])^{3/2}} \tag{70}$$

A partir de éstas dos transformaciones Z_1 y Z_2 se define:

$$K^2 = Z_1(g_1)^2 + Z_2(g_2)^2 \tag{71}$$

Por último si el p-valor de $\chi_2^2(K^2(M)) > \alpha$, siendo α el nivel de riesgo elegido, no existen pruebas que refuten la hipótesis nula

$$H_0 : \text{La muestra } M \text{ se comporta con distribución } \textit{Normal.} \tag{72}$$

Anexo II - Método de Anderson-Darling. [14]

Teniendo una muestra M de n elementos, presentamos la hipótesis nula:

$$H_0 : \text{La muestra } M \text{ se comporta con distribución } D. \tag{73}$$

Se procede a ordenar M.
El estadístico A^2 viene dado por:

$$A^2 = -n - S \tag{74}$$

siendo

$$S = \sum_{i=1}^{n} s_i \tag{75}$$

$$s_i = \frac{2i-1}{n}\left[ln(F(x_i)) + ln(1 - F(n+1-i))\right] \tag{76}$$

, donde $F(x)$ es la función de frecuencia acumulada de $x \in M$.
Una vez definido, se calcula A^2 para nuestra muestra M y para la distribución teórica $T \sim D$ (valor crítico), existen tablas que presentan el valor crítico de distribuciones estándares, que permiten ahorrar el cálculo. Si $A^2_M < A^2_T$ no hay pruebas suficientes para rechazar H_0 y se puede suponer que $M \sim D$