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

Abinal Facundo

Ing. en Sistemas de Información
Universidad Tecnológica Nacional Rosario
Zeballos,1341
denovaabinal@gmail.com

Bruschi Federico

Ing. en Sistemas de Información
Universidad Tecnológica Nacional Rosario
Zeballos,1341
fedeJbruschi@gmail.com

Jeriha Alexis

Ing. en Sistemas de Información
Universidad Tecnológica Nacional Rosario
Zeballos,1341
alexis.jeriha@gmail.com

Schirmer Julián

Ing. en Sistemas de Información
Universidad Tecnológica Nacional Rosario
Zeballos,1341
julischirmer2@gmail.com

May 1, 2021

ABSTRACT

En el siguiente documento desarrollaremos las distribuciones estadísticas más conocidas hasta la actualidad, veremos como se comportan, su funcionamiento y como cada una de ellas nos ayudaran a generar números pseudoaleatorios. Luego, testaremos cada generador para corroborar y determinar cuales distribuciones pasan o no las pruebas. Dichas pruebas determinaran si los algoritmos de generación son o no aleatorios.

1 Introducción

La distribución de probabilidad permite establecer toda la gama de resultados probables de ocurrir en un experimento determinado. Es decir, describe la probabilidad de que un evento se realice en el futuro. Cuando hablamos de distribución de probabilidad, tenemos que tener presente que ésta se encuentra completamente especificada por la función de distribución, cuyo valor en cada x real es la probabilidad de que la variable aleatoria sea menor o igual que x . Por supuesto, esta función dependerá del tipo de distribución empleada. A su vez, la distribución puede ser de dos clases: continua o discreta. Es indispensable aclarar que estas características serán de suma importancia en la concreción de nuestros experimentos. El objetivo de nuestro trabajo es desarrollar las distribuciones estadísticas más conocidas hasta la actualidad, como así también su comportamiento, su funcionamiento, y la manera en que cada una de ellas nos ayuda a generar números pseudoaleatorios. Finalmente, testaremos cada generador utilizando los test que describimos y codificamos en el trabajo anterior para corroborar y determinar cuáles distribuciones pasan o no las pruebas. Dichas pruebas determinaran si los algoritmos de generación son o no aleatorios.

2 Definiciones

Comencemos por definir algunos de los conceptos que utilizaremos durante el desarrollo del estudio.

2.1 Variable aleatoria

En probabilidad y estadística, una variable aleatoria es una función que asigna un valor numérico al resultado de un experimento aleatorio.

Una variable aleatoria puede concebirse como un valor numérico que está afectado por el azar. Dada una variable aleatoria no es posible conocer con certeza el valor que tomará esta al ser medida o determinada, aunque sí se conoce que existe una distribución de probabilidad asociada al conjunto de valores posibles. Por ejemplo, en una pandemia, se sabe que una persona cualquiera puede enfermarse o no (suceso), pero no se sabe cuál de los dos sucesos va a ocurrir. Solamente se puede decir que existe una probabilidad de que la persona enferme.

Variable aleatoria discreta Es aquella que solo toma ciertos valores (frecuentemente enteros) y que resulta principalmente del conteo realizado.

Variable aleatoria continua Es aquella que resulta generalmente de la medición y puede tomar cualquier valor dentro de un intervalo dado.

2.2 Función de densidad

La función de densidad de probabilidad, función de densidad, o simplemente densidad de una variable aleatoria continua describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor. La probabilidad de que la variable aleatoria caiga en una región específica del espacio de posibilidades estará dada por la integral de la densidad de esta variable entre uno y otro límite de dicha región. La función de densidad de probabilidad (FDP) es positiva a lo largo de todo su dominio para toda x

$$f_X(x) \geq 0$$

y su integral sobre todo el espacio es de valor unitario

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

2.3 Función de distribución acumulada

La función de distribución acumulada (FDA, designada también a veces simplemente como función de distribución o FD) o función de probabilidad acumulada asociada a una variable aleatoria real X sujeta a cierta ley de distribución de probabilidad, es una función matemática de la variable real X que describe la probabilidad de que X tenga un valor menor o igual que x .

2.3.1 Caso discreto

Si X es una variable aleatoria discreta con función de probabilidad $f(x)$ entonces la función de distribución acumulada se calcula como

$$F(x) = P[X \leq x] = \sum_{u \leq x} f(u)$$

2.3.2 Caso continuo

Si X es una variable aleatoria continua con función de densidad $f(x)$ entonces la función de distribución acumulada se calcula como

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(u) du$$

2.4 Función cuantil

La función cuantil, asociada con la función de distribución de una variable aleatoria, indica el valor de la variable aleatoria para el cual la probabilidad de que esa variable aleatoria sea menor o igual a dicho valor sea la probabilidad dada. También es conocida como inversa de la función de distribución.

Dada una función de distribución continua estrictamente monótona, $F : \mathbb{R} \rightarrow [0, 1]$, la función cuantil, denotada por F^{-1} , devuelve el valor x tal que

$$F_x(x) = P[X \leq x] = p$$

Si la distribución de probabilidad es discreta, en lugar de continua, entonces puede haber saltos entre los valores en el dominio de su función de distribución, mientras que si la función de distribución es monótona no estricta, puede haber

"zonas llanas" (intervalos en los que el valor de la función se mantiene constante) en su rango. En cualquiera de los casos, la función no estaría bien definida, por lo que se establece la siguiente definición alternativa:

$$F^{-1}(p) = \inf \{x \in \mathbb{R} : p \leq F(x)\}$$

para una probabilidad $0 < p < 1$, devolviendo la función cuantil el valor mínimo de x para el cual se mantiene la probabilidad anterior.

2.5 Función de probabilidad

La función de probabilidad es una función que asocia a cada punto de su espacio muestral X la probabilidad de que esta lo asuma. En concreto, si el espacio muestral, E de la variable aleatoria X consta de los puntos x_1, x_2, \dots, x_k , la función de probabilidad P asociada a X es

$$P(x_i) = p_i,$$

donde p_i es la probabilidad del suceso $X = x_i$.

Por definición de probabilidad,

$$\sum_{i=1}^k P(x_i) = 1$$

2.6 Método de la transformada inversa

El método de la transformada (o transformación) inversa, también conocido como método de la transformada integral de probabilidad inversa, es un método para la generación de números aleatorios de cualquier distribución de probabilidad continua cuando se conoce la inversa de su función de distribución (cdf). Este método es en general aplicable, pero puede resultar muy complicado obtener una expresión analítica de la inversa para algunas distribuciones de probabilidad.

2.6.1 Método

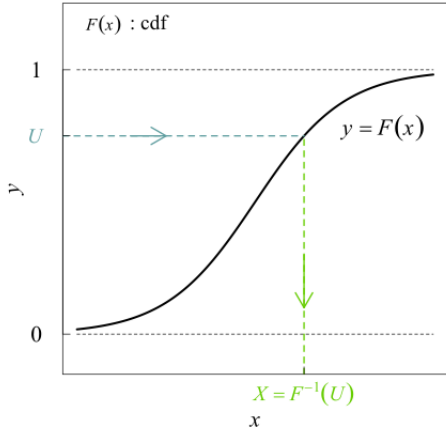
El problema que resuelve el método de la transformada inversa es el siguiente:

- Sea X una variable aleatoria cuya distribución puede ser descrita por la función de distribución F_X .
- Se desea generar valores de X que están distribuidos según dicha distribución.

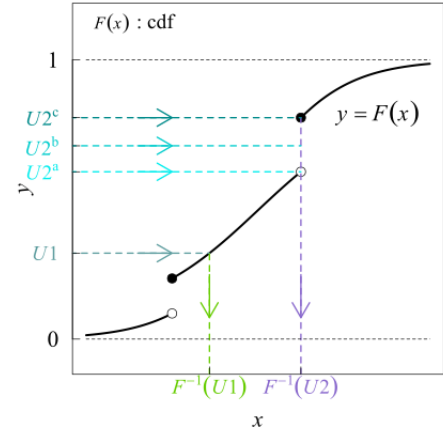
El método de la transformada inversa funciona de la siguiente manera:

1. Se genera un número aleatorio u a partir de la distribución uniforme en el intervalo $(0, 1)$, esto es $U \sim U(0, 1)$
2. Se halla la inversa de la función de distribución, esto es, $F_X^{-1}(x)$.
3. Calcular $X = F_X^{-1}(u)$, esta variable aleatoria X tiene distribución F_X .

Expresado de manera diferente, dada una variable aleatoria continua U en $(0, 1)$ y una función de distribución invertible F_X , la variable aleatoria $X = F_X^{-1}(U)$ tiene distribución F_X .



(a) Método de la transformada inversa.

(b) $y = F_X(x)$ escrita como $F_X^{-1}(y) = \inf \{x \mid F_X(x) \geq y\}$

2.7 Método del rechazo

Si $f(x)$ es una función acotada y x tiene además un rango finito, con $a \leq x \leq b$, entonces se puede usar la técnica de rechazos para generar los valores de variables aleatorias.

2.7.1 Método

1. Normalizar el rango de f mediante un factor de escala c tal que

$$c \cdot f(x) \leq 1 \quad \text{si } a \leq x \leq b$$

2. Definir a x como una función lineal de r , o sea

$$x = a + (b - a) \cdot u$$

3. Generar parejas de números aleatorios (u_1, u_2)
4. Siempre que se encuentre una pareja de números aleatorios que satisfagan la relación

$$u_2 \leq c \cdot f[a + (b - a) \cdot u_1].$$

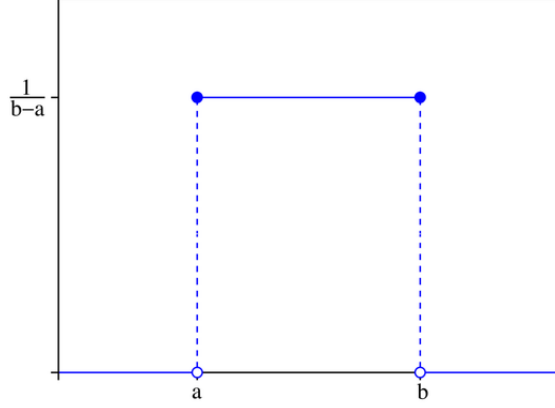
Dicho par será aceptado y se utilizará a $x = a + (b - a) \cdot u_1$ como el valor generado de la variable aleatoria.

3 Distribuciones de probabilidad

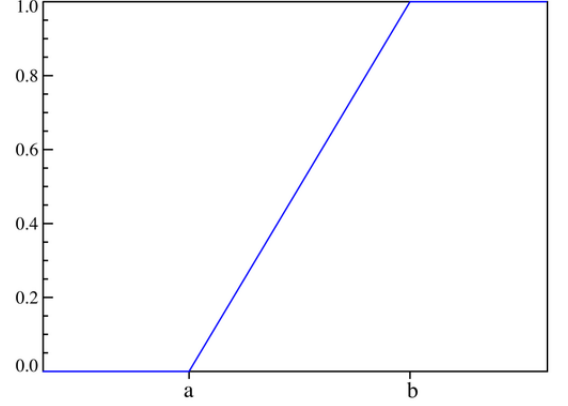
3.1 Distribución Uniforme Continua

La distribución uniforme es, posiblemente, la más sencilla entre las continuas: su densidad es 0 salvo en un determinado rango $[a, b]$ donde es constante y, como consecuencia, toma el valor $\frac{1}{b-a}$. Es decir, solo puede tomar valores en ese rango y, dentro de él, todos son equiprobables. Por lo que se dice que una variable aleatoria es uniforme con parámetros a y b denotado $X \sim U(a, b)$ y su función de densidad es

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in (a, b) \\ 0 & \text{si } x \notin (a, b) \end{cases}$$



(a) Función densidad



(b) Función de distribución

3.1.1 Obtención de la función inversa

En este caso la función inversa puede obtenerse mediante el método de la transformada inversa.

Paso 1 Identificar la función de densidad que describe la distribución.

$$f(x) = \frac{1}{b-a}$$

Paso 2 Obtener la función de distribución acumulada. Esto se logra integrando la función de densidad

$$F_X(x) = \int_a^x \frac{1}{b-a} du = \frac{1}{b-a} u \Big|_a^x = \frac{x-a}{b-a}$$

por lo que la función de densidad es

$$f(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x < b \\ 1 & \text{si } x \geq b \end{cases}$$

Paso 3 Reemplazar el $F(x)$ por un valor u muestreado de una distribución uniforme estándar

$$u = \frac{x-a}{b-a}$$

Paso 4 Despejar x para hallar la función inversa

$$u \cdot (b-a) = x-a$$

$$x = (b-a) \cdot u + a$$

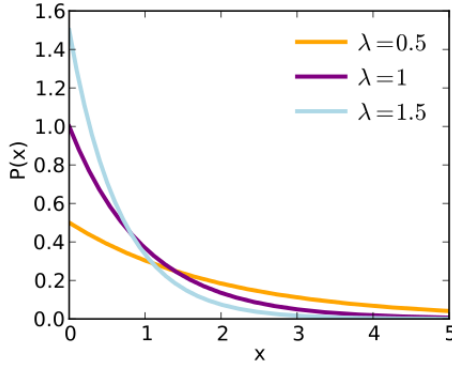
Si u es un valor muestreado de una distribución uniforme estándar, entonces el valor $a + (b-a) \cdot u$ posee una distribución uniforme parametrizada por a y b .

3.2 Distribución Exponencial Continua

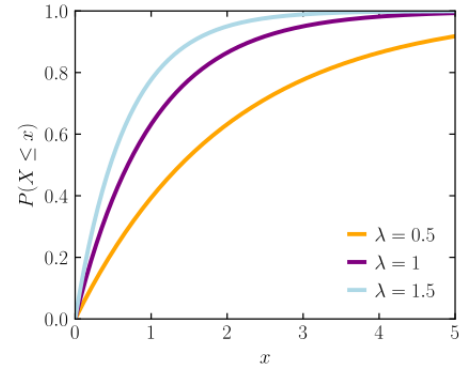
La distribución exponencial es una distribución continua que se utiliza para modelar tiempos de espera para la ocurrencia de un cierto evento. Decimos que una variable aleatoria X tiene una distribución exponencial de parámetro $\lambda > 0$ y notamos $X \sim \varepsilon(\lambda)$ cuando su función densidad de probabilidad es:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0, \\ 0 & \text{en otro caso.} \end{cases} \quad (1)$$

Si $x > 0$, entonces $F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$. Consecuentemente $P(X > x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$



(a) Función densidad



(b) Función de distribución

3.2.1 Obtención de la función inversa

En este caso la función inversa puede obtenerse mediante el método de la transformada inversa.

Paso 1 Identificar la función de densidad que describe la distribución.

$$f(x) = \lambda e^{-\lambda x}$$

Paso 2 Obtener la función de distribución acumulada. Esto se logra integrando la función de densidad

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \text{ si } x \geq 0$$

Paso 3 Reemplazar el $F(x)$ por un valor u muestreado de una distribución uniforme estándar

$$u = 1 - e^{-\lambda x}$$

Paso 4 Despejar x para hallar la función inversa

$$x = -\frac{1}{\lambda} \ln(1 - u)$$

Utilizando el hecho de que $U \sim (0, 1)$ entonces $1 - U \sim U(0, 1)$ por lo que una versión más eficiente del algoritmo es

$$x = -\frac{1}{\lambda} \ln(U)$$

3.3 Distribución gamma

La distribución gamma es una distribución con dos parámetros que pertenece a las distribuciones de probabilidad continuas. La distribución exponencial, distribución de Erlang y la distribución chi cuadrada son casos particulares de la distribución gamma. Hay dos diferentes parametrizaciones que suelen usarse:

1. Con parámetro de forma k y parámetro de escala θ .
2. Con parámetro de forma $\alpha = k$ y parámetro inverso de escala $\lambda = 1/\theta$

3.3.1 Funciones de densidad y distribución

Si una variable aleatoria continua X tiene distribución gamma con parámetros $\alpha > 0$ y $\lambda > 0$ entonces escribiremos $X \sim \Gamma(\alpha, \lambda)$

Si $X \sim \Gamma(\alpha, \lambda)$ entonces su función de densidad es:

$$F_X(x) = \frac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$$

Para $x > 0$ donde:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \cdot dt$$

Es la llamada **función gamma** y satisface lo siguiente:

1. $\Gamma(2) = \Gamma(1) = 1$
2. Para cualquier $\alpha > 0$ se cumple que $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
3. Si n pertenece a los **enteros positivos** entonces $\Gamma(n + 1) = n!$
4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$
5. Si n pertenece a los **enteros positivos** entonces:

$$\Gamma(\frac{n}{2}) = \frac{\sqrt{\pi}(n-1)!}{2^{n-1}(\frac{n-1}{2})!}$$

3.3.2 Propiedades

Si X es una variable aleatoria tal que $X \sim \Gamma(\alpha, \lambda)$ entonces X satisface algunas propiedades, de las cuales mencionaremos:

Varianza: La varianza de la variable aleatoria X es:

$$Var[X] = \frac{\alpha}{\lambda^2}$$

Media: La media de la variable aleatoria X es

$$E[X] = \frac{\alpha}{\lambda}$$

Suma de gammas: Si $X_i \sim \Gamma(\alpha_i, \lambda)$ para $i = 1, 2, \dots, n$ son variables aleatorias independientes entonces:

$$\sum_{i=1}^n X_i \sim \Gamma(\sum_{i=1}^n \alpha_i, \lambda)$$

3.4 Distribución normal

La distribución normal fue introducida en 1733 por Abraham De Moivre, al obtenerla como una forma límite de la distribución binomial. Medio siglo después fue redescubierta por Laplace y Gauss para describir el comportamiento de los errores en las mediciones astronómicas, que seguían una distribución simétrica en forma de campana. Decimos que una variable aleatoria X tiene una distribución normal de parámetros μ, σ , donde $\sigma > 0$, y notamos $X \sim N(\mu, \sigma)$ cuando su función de probabilidad es:

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

1. La función $f(x) > 0$, para todo x real y además $\int_{-\infty}^{+\infty} f(x)dx = 1$.
2. La gráfica de $f(x)$ es simétrica respecto a la recta de la ecuación $x = \mu$. En símbolos $f(\mu - x) = f(\mu + x), \forall x \in \mathbb{R}$
3. La función $f(x)$ tiene un máximo absoluto para $x = \mu$ y presenta puntos de inflexión en los puntos de abscisa $x = \mu \pm \sigma$

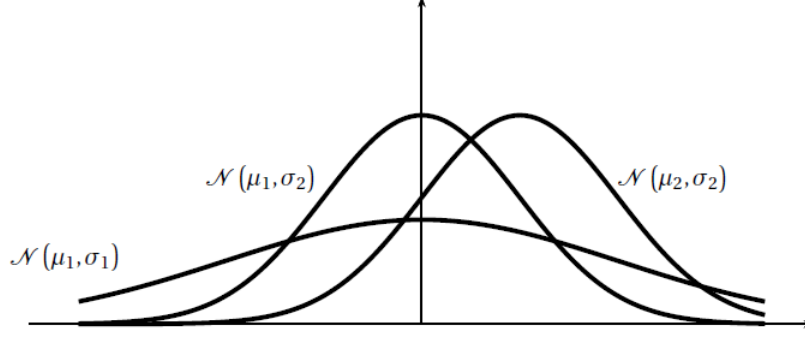
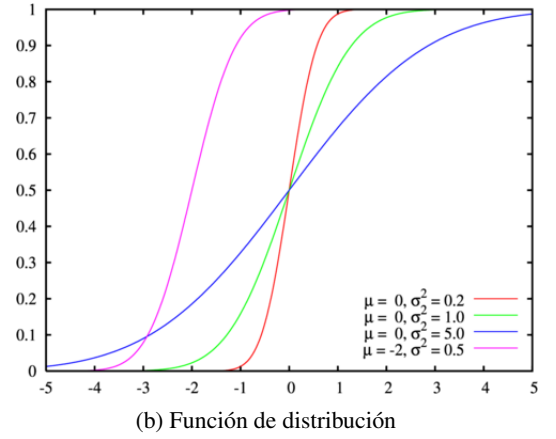
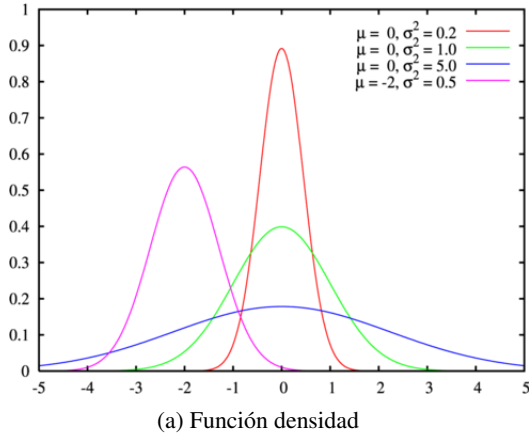


Fig. 4: Tres gráficas de curvas normales con diferentes medias y variancias. En este gráfico es $\mu_1 < \mu_2$ y $\sigma_1 > \sigma_2$



3.4.1 Método de Aceptación-Rechazo

Como ya sabemos, encontrar una fórmula explícita para $F^{-1}(x)$ de la función de distribución, no es siempre posible. Además, de ser posible de aplicar otros métodos, buscaremos aquellos que resultes más eficientes. Aquí presentaremos un método más inteligente, llamado *método de aceptación-rechazo*

Comenzamos asumiendo que la F que deseamos simular tiene una función de densidad de probabilidad $f(x)$; es decir, el caso continuo. Más adelante también daremos una versión discreta, que es muy similar. La idea básica es encontrar una distribución de probabilidad alternativa G , con función de densidad $g(x)$, a partir del cual ya tenemos un algoritmo eficiente para generar (por ejemplo, transformación inversa método o lo que sea), pero también tal que la función $g(x)$ este "cerca" a $f(x)$. En particular, suponemos que la razón $\frac{f(x)}{g(x)}$ está limitado por una constante $c > 0$;

$\sup_x \left\{ \frac{f(x)}{g(x)} \right\} \leq c$ (En la práctica queremos que c se acerque a 1 lo más posible.)

Aplicando a la distribución, si deseamos una $X \sim N(\mu, \sigma)$, podemos expresarla en función de $X = Z\sigma + \mu$, donde Z la denotamos como la reversa con una distribución $N(0, 1)$. Por tanto, basta con encontrar un algoritmo para generar $Z \sim (0, 1)$. Esto puede ser generado desde el valor absoluto, $|Z|$ entonces por simetría podemos obtener nuestro Z generando independientemente una reversa dada como S eso es ± 1 con probabilidad $1/2$ y ajustando a $Z = S|Z|$. En otras palabras, generamos una U y establecemos $Z = |Z|$ si $U \leq 0.5$ y establecer $Z = -|Z|$ si $U > 0.5$. $|Z|$ es no negativo y tiene densidad:

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, x \geq 0.$$

Para nuestro caso elegiremos $g(x) = e^{-x}, x \geq 0$, la densidad exponencial con tasa 1, algo que ya sabemos cómo simular fácilmente (método de transformación inversa, por ejemplo). Nada que $h(x) = f(x)/g(x) = e^{x-x^2/2} \sqrt{2/\pi}$, simplemente usamos para calcular su máximo (solucion $h'(x) = 0$) el cual debe ocurrir en el valor de x donde se

maximiza el exponente $x - x^2/2$, con valor $x = 1$. Por lo tanto $c = \sqrt{2e/\pi} \approx 1.32$ y entonces $f(y)/cg(y) = e^{-(y-1)^2/2}$.

Algoritmo para generar $Z \sim N(0, 1)$

1. Generar dos exponenciales independientes con tasa 1; $Y_1 = -\ln(U_1)$ y $Y_2 = -\ln(U_2)$
2. Si $Y_2 \geq (Y_1 - 1)^2/2$ guardar $|Z| = Y_1$, en otro caso, volver a 1.
3. Generar U . Guardar $Z = |Z|$ si $U \leq 0.5$; guardar $Z = -|Z|$ si $U > 0.5$

3.4.2 Obtención de la función inversa

[En este caso la función inversa puede obtenerse mediante el método de la transformada inversa.] [Paso a paso del método]

[En este caso la función inversa no puede obtenerse mediante el método de la transformada inversa, por lo que se debe recurrir al método del rechazo] [Paso a paso del método]

3.5 Distribución binomial

La distribución binomial es una distribución discreta muy importante que surge en muchas aplicaciones bioestadísticas. Fue obtenida por Jakob Bernoulli (1654-1705).

Esta distribución aparece de forma natural al realizar repeticiones independientes de un experimento que tenga respuesta binaria, generalmente clasificada como “éxito” o “fracaso”; este experimento recibe el nombre de experimento de Bernoulli. Ejemplos de respuesta binaria pueden ser el hábito de fumar (sí/no), si un paciente hospitalizado desarrolla o no una infección, o si un artículo de un lote es o no defectuoso. La variable discreta que cuenta el número de éxitos en n pruebas independientes de ese experimento, cada una de ellas con la misma probabilidad de “éxito” igual a p , sigue una distribución binomial de parámetros n y p , que se denota por $(Bi(n, p))$. Este modelo se aplica a poblaciones finitas de las que se toman elementos al azar con reemplazo, y también a poblaciones conceptualmente infinitas, como por ejemplo las piezas que produce una máquina, siempre que el proceso de producción sea estable (la proporción de piezas defectuosas se mantiene constante a largo plazo) y sin memoria (el resultado de cada pieza no depende de las anteriores).

Un ejemplo de variable binomial puede ser el número de pacientes con cáncer de pulmón ingresados en una unidad hospitalaria.

Un caso particular se tiene cuando $n=1$, que da lugar a la distribución de Bernoulli.

3.6 Distribución Binomial Negativa

La distribución binomial negativa es un modelo adecuado para tratar aquellos procesos en los que se repite un determinado ensayo o prueba hasta conseguir un número determinado de resultados favorables (por vez primera). Es por tanto de gran utilidad para aquellos muestreos que procedan de esta manera. Si el número de resultados favorables buscados fuera 1 estaríamos en el caso de la distribución geométrica. Está implicada también la existencia de una dicotomía de resultados posibles en cada prueba y la independencia de cada prueba o ensayo, o la reposición de los individuos muestreados.

Esta distribución o modelo puede hacerse derivar de un proceso experimental puro o de Bernoulli en el que se presenten las siguientes condiciones:

- El proceso consta de un número no definido de pruebas separadas o separables. El proceso concluirá cuando se obtenga un determinado número de resultados favorables K
- Cada prueba puede dar dos resultados posibles mutuamente excluyentes A y no A
- La probabilidad de obtener un resultado A en cada una de las pruebas es p siendo la probabilidad de no A, q . Lo que nos lleva a que $p+q=1$
- Las probabilidades p y q son constantes en todas las pruebas. Todas las pruebas son independientes. Si se trata de un experimento de extracción éste se llevará cabo con devolución del individuo extraído, a no ser que se trate de una población en la que el número de individuos tenga de carácter infinito.
- (Derivación de la distribución) Si, en estas circunstancias aleatorizamos de forma que la variable aleatoria x sea "el número de pruebas necesarias para conseguir K éxitos o resultados A"; entonces la variable aleatoria x

seguirá una distribución binomial negativa con parámetros p y k

$$\text{Entonces } x \Rightarrow BN(p, k)$$

La variable aleatoria x podrá tomar sólo valores superiores a k $x \in (k, k+1, k+2, \dots)$

El suceso del que se trata podría verse como:

$$\bar{A} \cap \bar{A} \cap \bar{A} \cap \dots \bar{A} \cap \bar{A} \cap \bar{A} (\text{veces})$$

Dado que las pruebas son independientes y conocemos que $P(A)=p$ y $P(\text{no } A)=q$

$$q^{x-k} * p^k$$

Que sería la probabilidad de x si el suceso fuera precisamente con los resultados en ese orden. Dado que pueden darse otros órdenes, en concreto $\binom{x-1}{x-k}$ formas u órdenes distintos. La función de cuantía de la distribución binomial negativa quedará como:

$$P(x) = \binom{x-1}{x-k} q^{x-k} * p^k$$

3.7 Distribución hipergeométrica

La distribución hipergeométrica es una distribución de probabilidad discreta relacionada con muestreos aleatorios y sin reemplazo. Suponga que se tiene una población de N elementos de los cuales, K pertenecen a la categoría A y $N - K$ pertenecen a la categoría B . La distribución hipergeométrica mide la probabilidad de obtener x ($0 \leq x \leq K$) elementos de la categoría A en una muestra sin reemplazo de n elementos de la población original.

3.7.1 Función de Probabilidad

Una variable aleatoria discreta X tiene una distribución hipergeométrica con parámetros $N = 0, 1, \dots, K = 0, 1, \dots, N$ y $n = 0, 1, \dots, N$ y escribimos $X \sim HG(N, K, n)$ si su función de probabilidad es:

$$P[X = x] = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}}$$

para $x = 0, 1, \dots, \min\{K, N - K\}$; donde N es el tamaño de población, n es el tamaño de la muestra extraída, K es el número de elementos en la población original que pertenecen a la categoría deseada y x es el número de elementos en la muestra que pertenecen a dicha categoría.

La notación

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

hace referencia al coeficiente binomial, es decir, el número de combinaciones posibles al seleccionar a elementos de un total b .

3.7.2 Propiedades

Si $X \sim HG(N, K, n)$ entonces X cumple algunas propiedades:

- El valor esperado de la variable aleatoria X es:

$$E[X] = \frac{nK}{N}$$

- La varianza está dada por:

$$Var[X] = \frac{nK}{N} \left(\frac{N-K}{N} \right) \left(\frac{N-n}{N-1} \right)$$

La distribución hipergeométrica es aplicable a muestreos sin reposición, a diferencia de la binomial que se aplica a muestreos con reposición. En situaciones en las que el número esperado de repeticiones en el muestreo es presumiblemente bajo, puede aproximarse la primera por medio de la segunda. Esto es así cuando N es grande y el tamaño relativo de la muestra extraída, n/N , es pequeño.

3.8 Distribución de Poisson

La distribución de Poisson debe su nombre al matemático francés Simeón Denis Poisson (1781-1840), aunque ya había sido introducida en 1718 por Abraham De Moivre (1667-1754) como una forma límite de la distribución binomial que surge cuando se observa un evento raro después de un número grande de repeticiones. En general, la distribución de Poisson de parámetro λ se puede utilizar como una aproximación de la binomial, $\text{Bin}(n, p)$, si el número de pruebas n es grande, pero la probabilidad de éxito p es pequeña, siendo $\lambda = np$; podemos considerar que la aproximación Poisson-binomial es “buena” si $n \geq 20$ y $p \leq 0,05$ y “muy buena” si $n \geq 100$ y $p \leq 0,01$.

La distribución de Poisson también surge cuando un evento o suceso “raro” ocurre aleatoriamente en el espacio o el tiempo. La variable asociada es el número de ocurrencias del evento en un intervalo o espacio continuo, por tanto, es una variable aleatoria discreta que toma valores enteros de 0 en adelante (0, 1, 2,...). Así, el número de pacientes que llegan a un consultorio en un lapso dado, el número de llamadas que recibe un servicio de atención a urgencias durante 1 hora, el número de células anormales en una superficie histológica o el número de glóbulos blancos en un milímetro cúbico de sangre son ejemplos de variables que siguen una distribución de Poisson. En general, es una distribución muy utilizada en diversas áreas de la investigación médica y, en particular, en epidemiología.

El concepto de evento “raro” o poco frecuente debe ser entendido en el sentido de que la probabilidad de observar k eventos decrece rápidamente a medida que k aumenta. Supóngase, por ejemplo, que el número de reacciones adversas tras la administración de un fármaco sigue una distribución de Poisson de media $\lambda = 2$. Si se administra este fármaco a 1.000 individuos, la probabilidad de que se produzca una reacción adversa ($k = 1$) es 0,27; los valores de dicha probabilidad para $k = 2, 3, 4, 5, 6$ reacciones, respectivamente, son: 0,27; 0,18; 0,09; 0,03 y 0,01. Para $k = 10$ o mayor, la probabilidad es virtualmente 0. El rápido descenso de la probabilidad de que se produzcan k reacciones adversas a medida que k aumenta.

Para que una variable recuento siga una distribución de Poisson deben cumplirse varias condiciones:

1. En un intervalo muy pequeño (p. e. de un milisegundo) la probabilidad de que ocurra un evento es proporcional al tamaño del intervalo.
2. La probabilidad de que ocurran dos o más eventos en un intervalo muy pequeño es tan reducida que, a efectos prácticos, se puede considerar nula.
3. El número de ocurrencias en un intervalo pequeño no depende de lo que ocurra en cualquier otro intervalo pequeño que no se solape con aquél.

Estas propiedades pueden resumirse en que el proceso que genera una distribución de Poisson es estable (produce, a largo plazo, un número medio de sucesos constante por unidad de observación) y no tiene memoria (conocer el número de sucesos en un intervalo no ayuda a predecir el número de sucesos en el siguiente).

El parámetro de la distribución, λ , representa el número promedio de eventos esperados por unidad de tiempo o de espacio, por lo que también se suele hablar de λ como “la tasa de ocurrencia” del fenómeno que se observa.

A veces se usan variables de Poisson con “intervalos” que no son espaciales ni temporales, sino de otro tipo. Por ejemplo, para medir la frecuencia de una enfermedad se puede contar, en un período dado, el número de enfermos en cierta población dividida en “intervalos” de, por ejemplo, 10.000 habitantes. Al número de personas enfermas en una población de tamaño prefijado, en un instante dado, se le denomina prevalencia de la enfermedad en ese instante y es una variable que sigue una distribución de Poisson. Otra medida para la frecuencia de una enfermedad es la incidencia, que es el número de personas que enferman en una población en un periodo determinado. En este caso, el intervalo es de personatiempo, generalmente personas-año, y es también una variable con distribución de Poisson. Habitualmente, ambas medidas se expresan para intervalos de tamaño unidad o, dicho de otro modo, en lugar de la variable número de enfermos, se usa el parámetro λ .

La distribución de Poisson tiene iguales la media y la varianza. Si la variación de los casos observados en una población excede a la variación esperada por la Poisson, se está ante la presencia de un problema conocido como sobredispersión y, en tal caso, la distribución binomial negativa es más adecuada.

Para valores de λ mayores de 20 la distribución de Poisson se aproxima a una distribución normal de media y varianza iguales a λ .

Valores:

k : 0, 1, 2, ...

Parámetros:

λ : tasa de ocurrencia, $\lambda > 0$.

3.9 Distribución empírica discreta

La función de distribución empírica es una manera alternativa de representar la información completa de un conjunto de datos numéricos. La *función de densidad* se especifica mediante una lista de pares de datos, (k_i, p_i) , donde cada par de datos consta de un punto de datos entero, k_i , y su correspondiente probabilidad p_i . La función de distribución esta dada por:

$$F(k_j) = \sum_{i=1}^j p_i = P_j$$

La entrada de esta distribución son pares de datos (k_i, p_i) donde $i = 1, 2, \dots, n$. Los puntos de datos debe estar ordenados de forma ascendente pero no es necesario que estén igualmente espaciados y las probabilidades deben sumar uno:

$k_i < k_j$ y solo si $i < j$ y $\sum_{i=1}^n p_i = 1$.

La salida sera $x \in \{k_1, k_2, \dots, k_n\}$

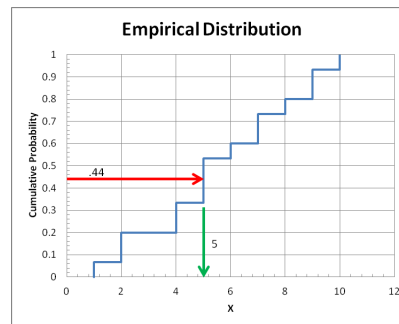


Fig. 6: En azul la gráfica de la función de distribución acumulada de la distribución empírica discreta

4 Realización de los test y análisis de lo esperado

4.1 Distribución uniforme

```

1  # Importe de librerías
2  import random
3  import matplotlib.pyplot as plt
4
5  #Definición de funciones
6
7  def uniform(a, b, u):
8      uniform_random = a + (b - a) * u
9      return uniform_random
10
11  #INICIA PROGRAMA PRINCIPAL
12
13  #Definición de variables y listas
14
15  n = 50 #cantidad de números a generar
16
17  a = 10 #cota inferior
18  b = 20 #cota superior
19
20  for i in range(n):
21      random_number = random.random ()
22      uniform_number = uniform(a, b, random_number)
23      #Imprimo valores
24      print(uniform_number)

```

Listing 1: Programa que genera e imprime 50 números uniformemente distribuidos entre 10 y 20

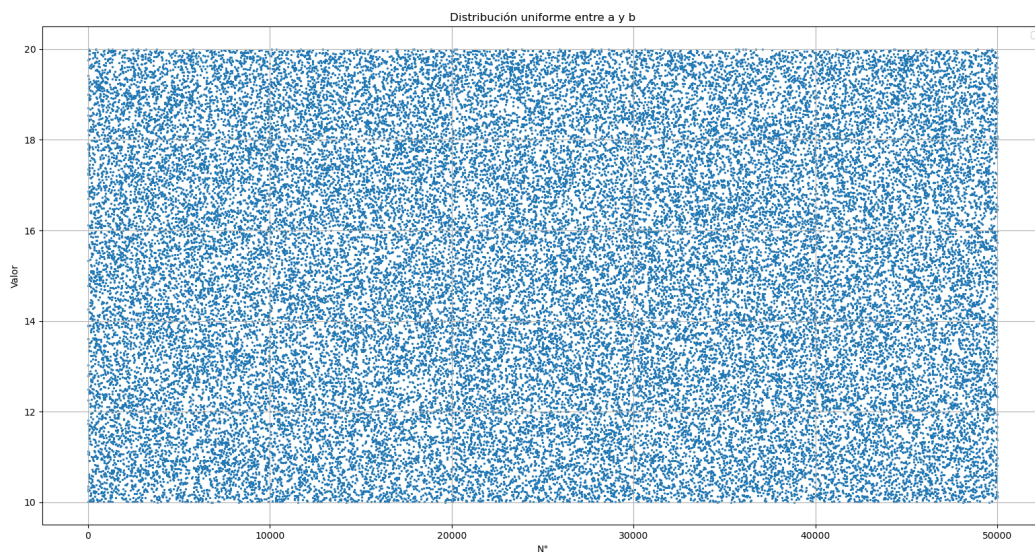


Fig. 7: Representación de 50.000 valores resultantes del generador uniforme en el intervalo [10,20]

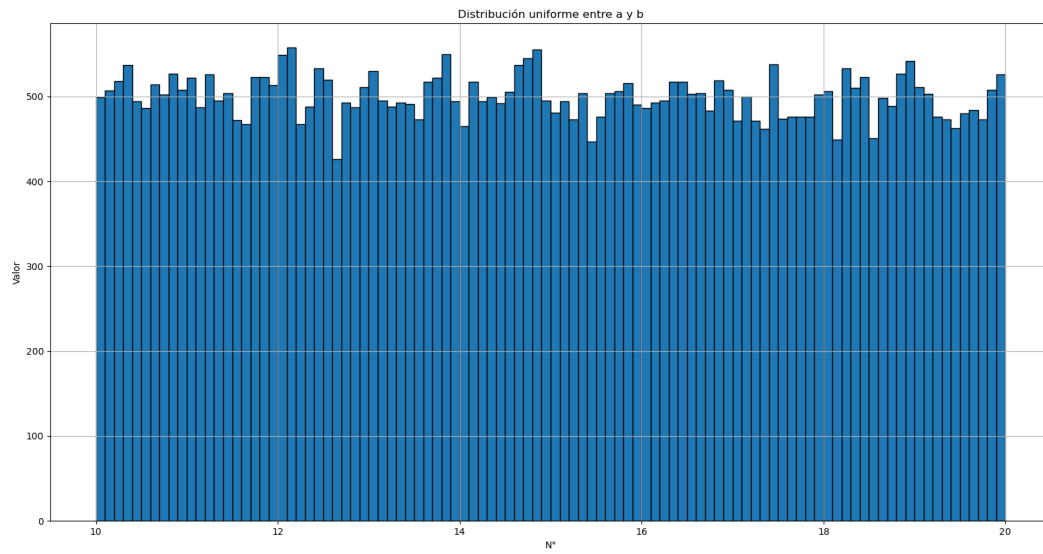


Fig. 8: Frecuencia absoluta valores anteriores - 100 intervalos

Gráfica esperada de la Distribución Uniforme:

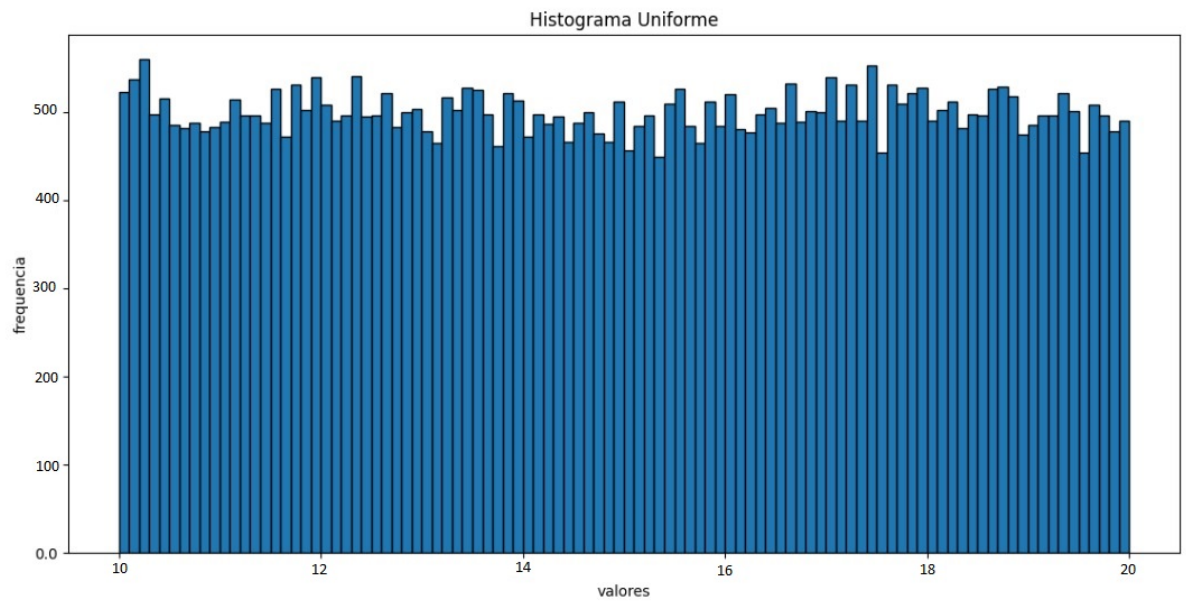


Fig. 9: Representación de 50.000 valores resultantes del generador uniforme en el intervalo [10,20]

4.2 Distribución exponencial

```

1  # Importe de librerías
2  import random
3  import math
4
5  #Definición de funciones
6  def exponential(param, u):
7      exponential_random = -(1 / param) * math.log(u)
8      return exponential_random
9
10 #INICIA PROGRAMA PRINCIPAL
11
12 #Definición de variables
13 n = 50 #cantidad de números a generar
14 lambda_param = 1 #parámetro Lambda
15 for i in range(n):
16     random_number = random.random ()
17     uniform_number = exponential(lambda_param, random_number)
18     #Imprime los valores
19     print(uniform_number)

```

Listing 2: Programa que genera e imprime 50 números distribuidos en forma exponencial con parámetro $\lambda = 1$

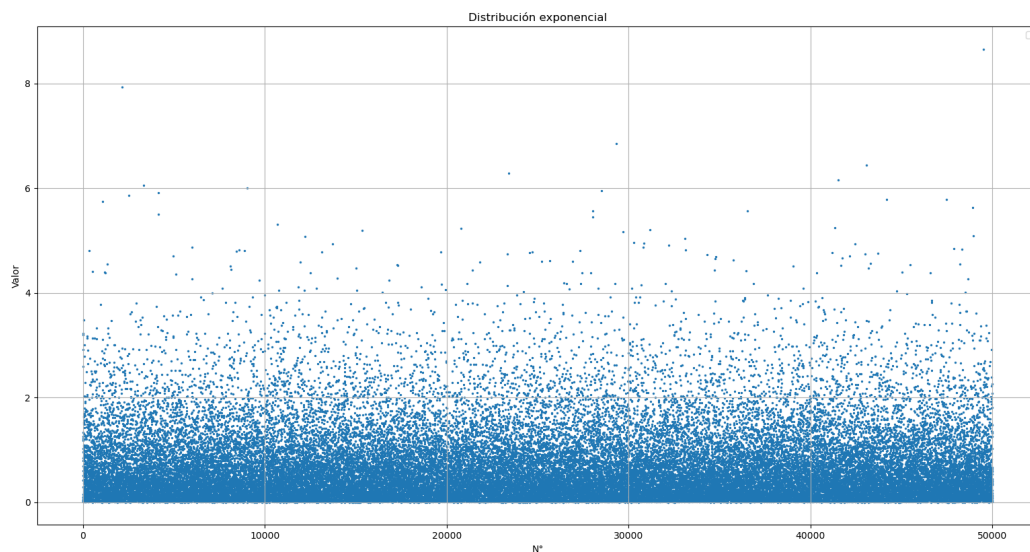


Fig. 10: Representación de 50.000 valores resultantes del generador exponencial con $\lambda = 1.5$

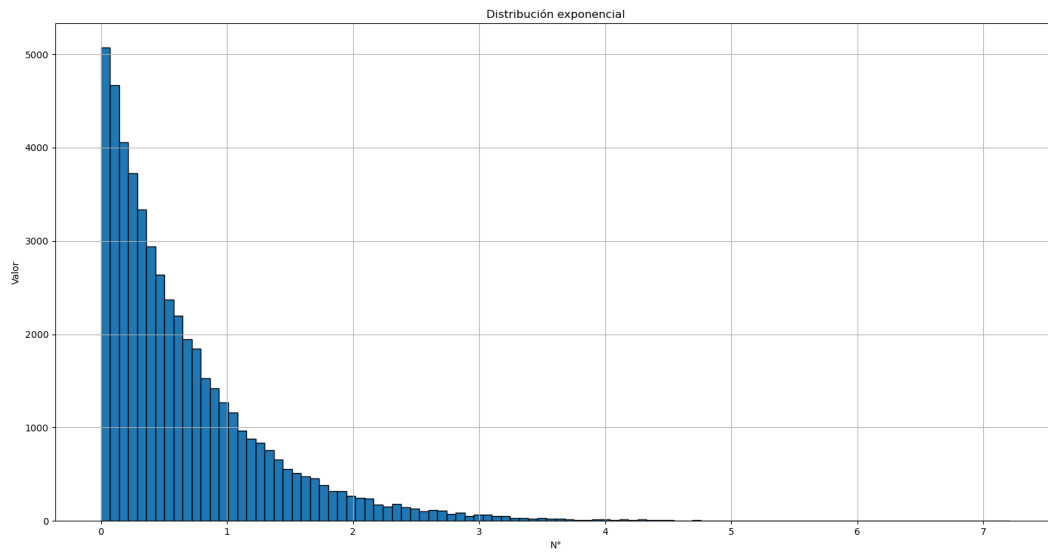


Fig. 11: Frecuencia absoluta valores anteriores - 100 intervalos

Gráfica esperada de la distribución exponencial:

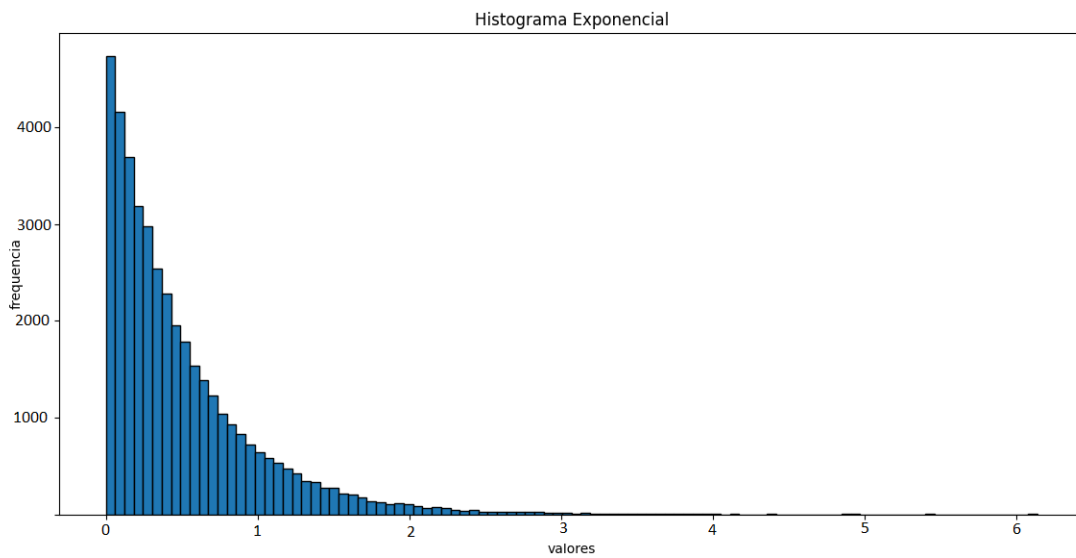


Fig. 12: Representación de 50.000 valores resultantes del generador exponencial

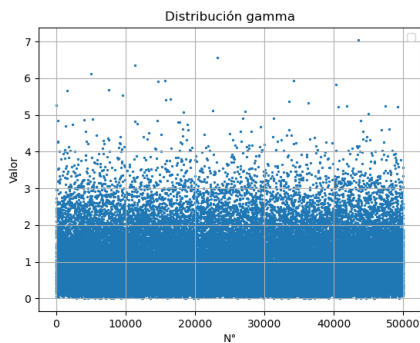
4.3 Distribución gamma

```

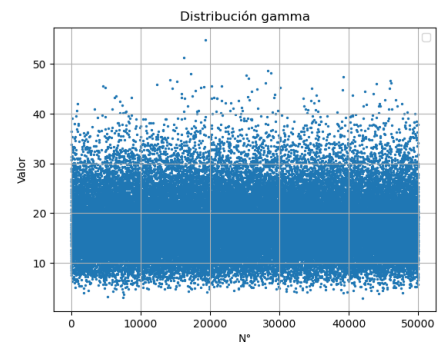
1  import random
2  import math
3  import matplotlib.pyplot as plt
4
5  def gamma(K, A):
6      tr = 1
7      for _ in range(K):
8          r = random.random()
9          tr *= r
10         x = -math.log(tr)/A
11         return x
12
13     #INICIA PROGRAMA PRINCIPAL
14
15     #Defino variables a utilizar
16     n = 50
17
18     #Parámetros de la distribución gamma
19     k = 2
20     a = 2
21
22     for i in range(n):
23         x = gamma(k, a)
24         #Imprimo valores
25         print(x)

```

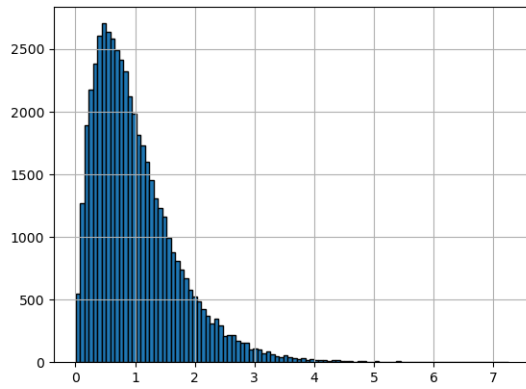
Listing 3: Programa que genera e imprime 50 números con una distribución $X \sim \Gamma(k = 2, \theta = 2)$



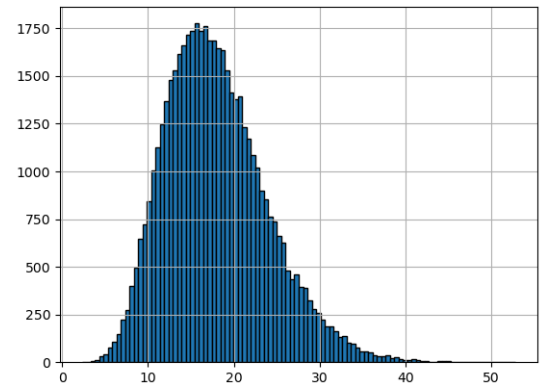
(a) Representación de los valores generados a partir de una distribución $X \sim \Gamma(k = 2, \theta = 2)$



(b) Representación de los valores generados a partir de una distribución $X \sim \Gamma(k = 9, \theta = 0.5)$

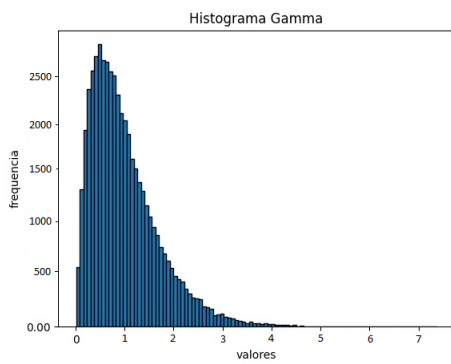


(a) Frecuencia absoluta de los valores - 100 intervalos

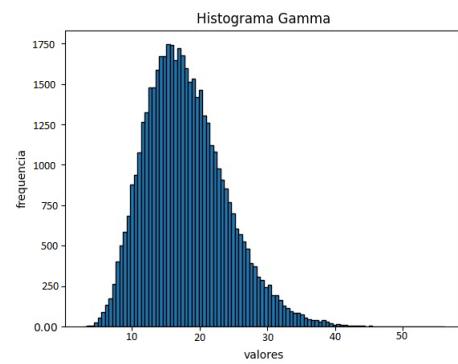


(b) Frecuencia absoluta de los valores - 100 intervalos

Gráfica esperada de la Distribución Gamma:



(a) Representación de los valores generados a partir de una distribución $X \sim \Gamma(k = 2, \theta = 2)$



(b) Representación de los valores generados a partir de una distribución $X \sim \Gamma(k = 9, \theta = 0.5)$

4.4 Distribución normal

```

1 import random
2 import math
3 import matplotlib.pyplot as plt
4
5 a1, a3, a5, a7, a9 = (3.949846138,
6                      0.252408784,
7                      0.076542912,
8                      0.008355968,
9                      0.029899776)
10
11 def normal():
12     sum = 0
13     for i in range(12):
14         sum += random.random()
15     y = (sum - 6) / 4
16     z = a1*y + a3*y**3 + a5*y**5 + a7*y**7 + a9*y**9
17     return z
18
19 #INICIA PROGRAMA PRINCIPAL
20
21 #Defino variables a utilizar
22 n = 50
23
24 #Parámetros de la distribución normal
25 mean, variance = (0, 1) # caso Normal Estándar
26 stddev = math.sqrt(variance)
27
28 for i in range(n):
29     x = normal() * stddev + mean
30     #Imprimo valores
31     print(x)

```

Listing 4: Programa que genera e imprime 50 números con una distribución normal estándar

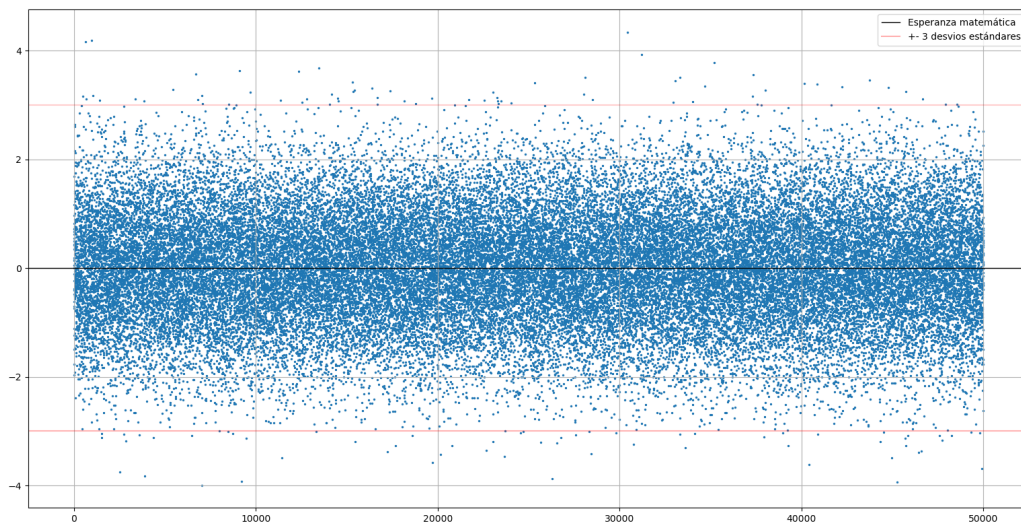


Fig. 16: Representación de 50.000 valores resultantes del generador de $X \sim N(\mu = 0, \sigma = 1)$

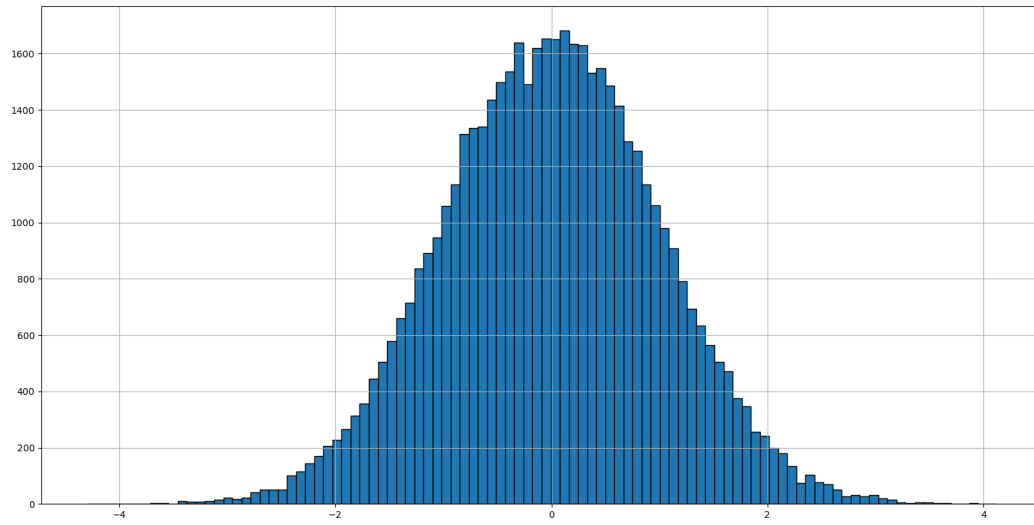


Fig. 17: Frecuencia absoluta valores anteriores - 100 intervalos

Gráfica esperada de la distribución normal:

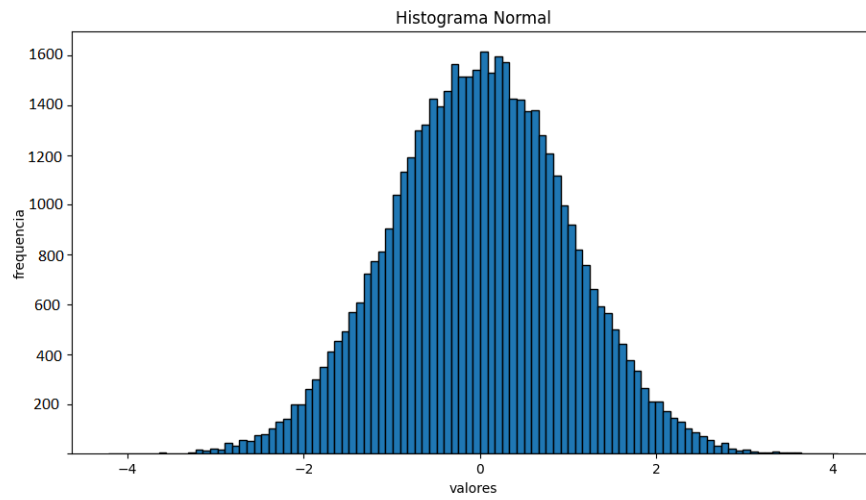


Fig. 18: Representación de 50.000 valores resultantes del generador normal con $X \sim N(\mu = 0, \sigma = 1)$

4.5 Distribución binomial

```

1  import random
2  import math
3  import matplotlib.pyplot as plt
4
5  def binomial(n, p):
6      x = 0
7      for _ in range(n):
8          r = random.random()
9          if (r - p) < 0:
10             x += 1
11      return x
12
13  #INICIA PROGRAMA PRINCIPAL
14
15  #Defino variables a utilizar
16  n = 50
17
18  #Parámetros de la distribución binomial
19  n_essay = 10 #ensayos de Bernoulli
20  p = 0.2 #probabilidad de "éxito"
21
22  for i in range(n):
23      x = binomial(n, p)
24      #Imprimo valores
25      print(x)

```

Listing 5: Programa que genera e imprime 50 números con una distribución $X \sim \text{Bin}(n = 50, p = 0.2)$

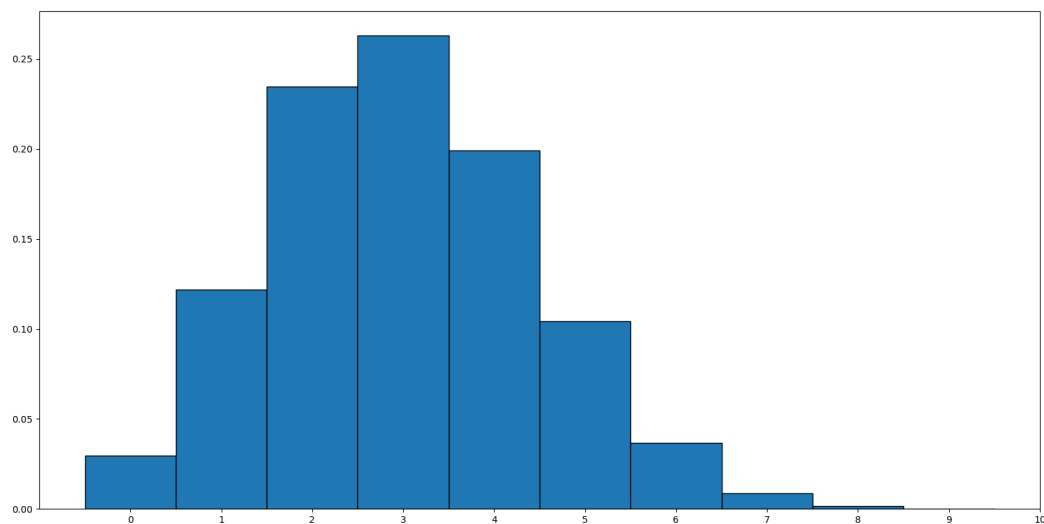


Fig. 19: Frecuencia relativa de 50.000 valores generados con distribución $X \sim \text{Bin}(n = 10, p = 0.2)$

Gráfica esperada de la Distribución Binomial:

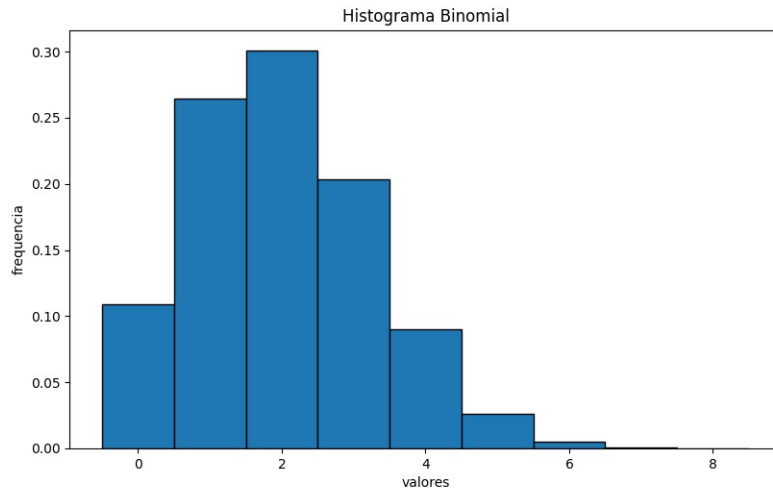


Fig. 20: Representación de 50.000 valores resultantes del generador exponencial

4.6 Distribución Binomial Negativa

```

1 import random
2 import matplotlib.pyplot as plt
3 import math
4
5 #Definición de funciones
6 def pascal(k, q):
7     tr = 1
8     qr = math.log(q)
9     for _ in range(k):
10         r = random.random()
11         tr = r * tr
12     nx = math.ceil((math.log(tr) / qr))
13     x = nx
14     return x
15
16 #INICIA PROGRAMA PRINCIPAL
17
18
19
20 #Definición de variables y listas
21
22
23 n = 50 #cantidad de números a generar
24 #Parámetros de la distribución de Pascal
25 r = 1 #número de éxitos
26 p = 0.4 #probabilidad de "éxito"
27
28 for i in range(n):
29     x = pascal(r, p)
30     print(x)

```

Listing 6: Programa que genera e imprime 50 números con una distribución $X \sim BN(n = 50, p = 0.4)$

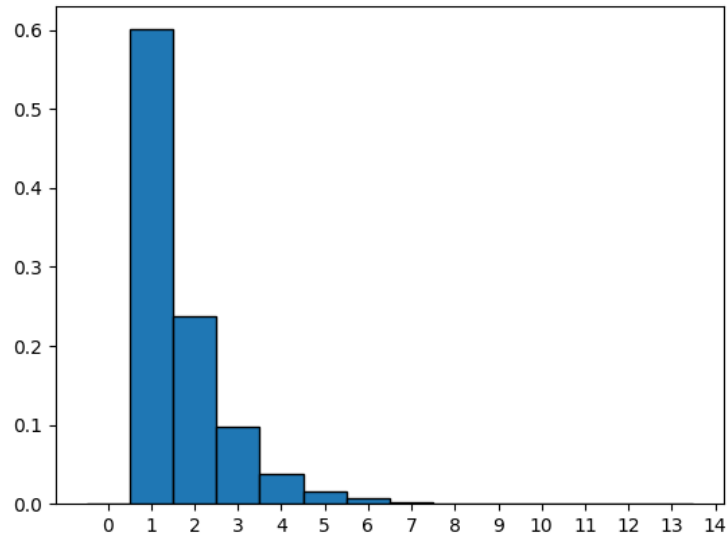


Fig. 21: Frecuencia relativa de 50.000 valores $X \sim BN(n = 10, p = 0.3)$

Gráfica esperada de la Distribución Binomial Negativa:

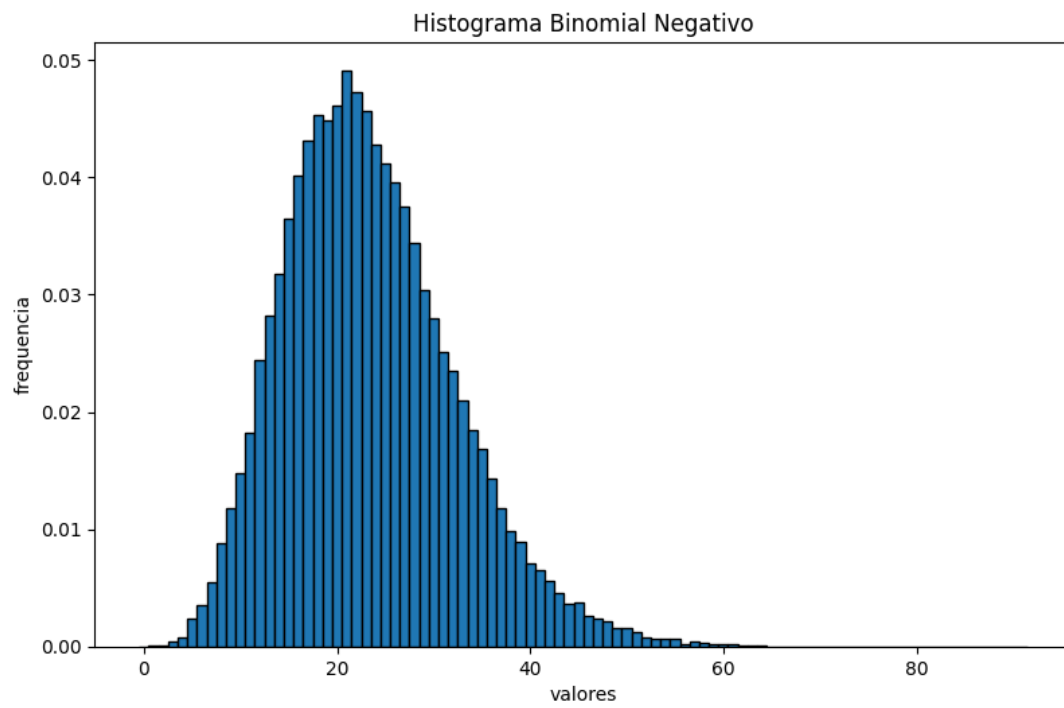


Fig. 22: Representación de 50.000 valores resultantes del generador binomial negativo $X \sim BN(n = 10, p = 0.3)$

4.7 Distribución hipergeométrica

```

1  import random
2  import math
3  import matplotlib.pyplot as plt
4
5  def hypergeometric(tn, ns, p):
6      x = 0
7      for _ in range(ns):
8          r = random.random()
9          if (r-p) >= 0:
10             s = 0
11          else:
12             s = 1
13             x += 1
14             p = (tn*p - s) / (tn - 1)
15             tn -= 1
16      return x
17
18  #INICIA PROGRAMA PRINCIPAL
19
20  #Defino variables a utilizar
21  n = 50
22
23  #Parámetros de la distribución hipergeométrica
24  tn = 100000 #Total de elementos de la población
25  ns = 10 #Cantidad de elementos de la muestra
26  p = 0.2 #Proporción de elementos de la clase I sobre la población total
27
28  for i in range(n):
29      x = hypergeometric(tn, ns, p)
30      #Imprimo valores
31      print(x)

```

Listing 7: Programa que genera e imprime 50 números con una distribución $X \sim HG(N = 100.000, M = 0.2, n = 10)$

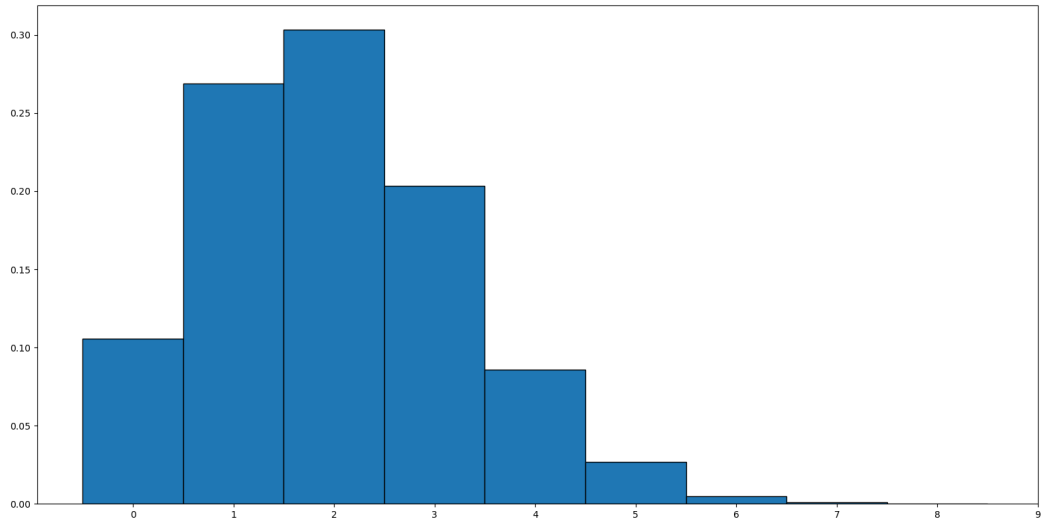


Fig. 23: Frecuencia relativa de 50.000 valores generados con distribución $X \sim HG(N = 100.000, M = 0.2, n = 10)$

Gráfica esperada de la Distribución Hipergeométrica:

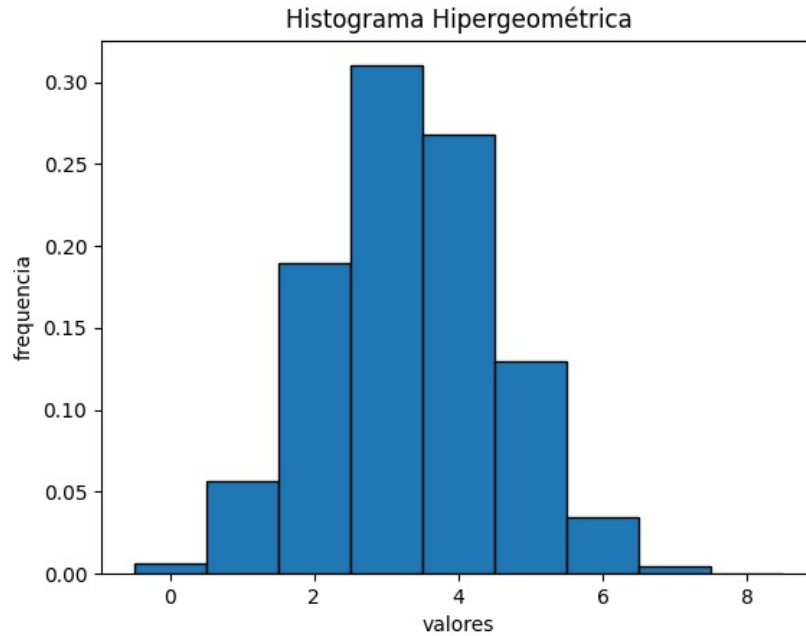


Fig. 24: Frecuencia relativa de 50.000 valores generados con distribución $X \sim HG(N = 100.000, M = 0.2, n = 10)$

4.8 Distribución de Poisson

```

1  # Importe de librerías
2  import random
3  import math
4  import matplotlib.pyplot as plt
5  import numpy as np
6
7  xAxis = []
8  yAxis = []
9  #Definición de funciones
10 def poisson(p):
11     x = 0
12     b = math.exp(-p)
13     tr = 1
14     r = random.random()
15     tr = r * tr
16     if (tr - b) >= 0:
17         x = x + 1
18         r = random.random()
19         tr = r * tr
20         tr *= r
21     return tr
22
23 #INICIA PROGRAMA PRINCIPAL
24
25 #Definición de variables
26
27 n = 10000 #cantidad de números a generar
28 lambda_param = 10 #parámetro Lambda
29 for i in range(n):
30     x = poisson(lambda_param)
31     print(x)

```

Listing 8: Programa que genera e imprime 10000 números con una distribución de $X \sim P(\lambda = 10)$

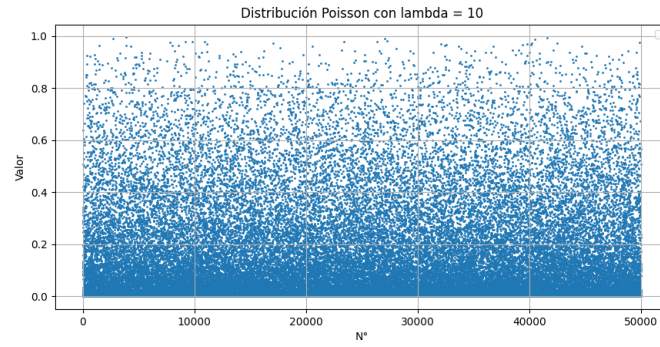


Fig. 25: Representación de 50.000 valores resultantes del generador poisson $X \sim P(\lambda = 10)$

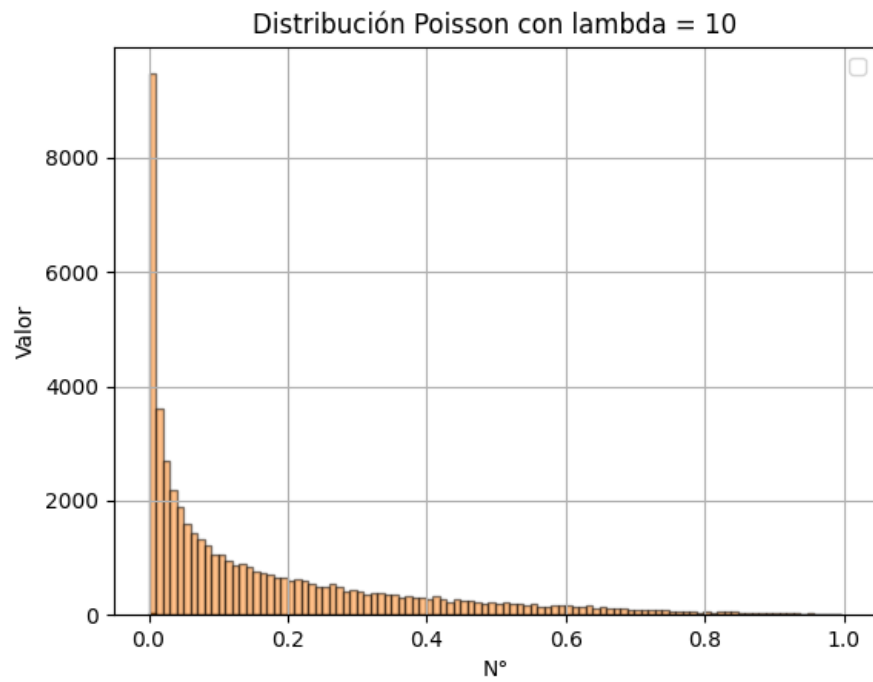


Fig. 26: Histograma de 50.000 valores resultantes del generador poisson $X \sim P(\lambda = 10)$

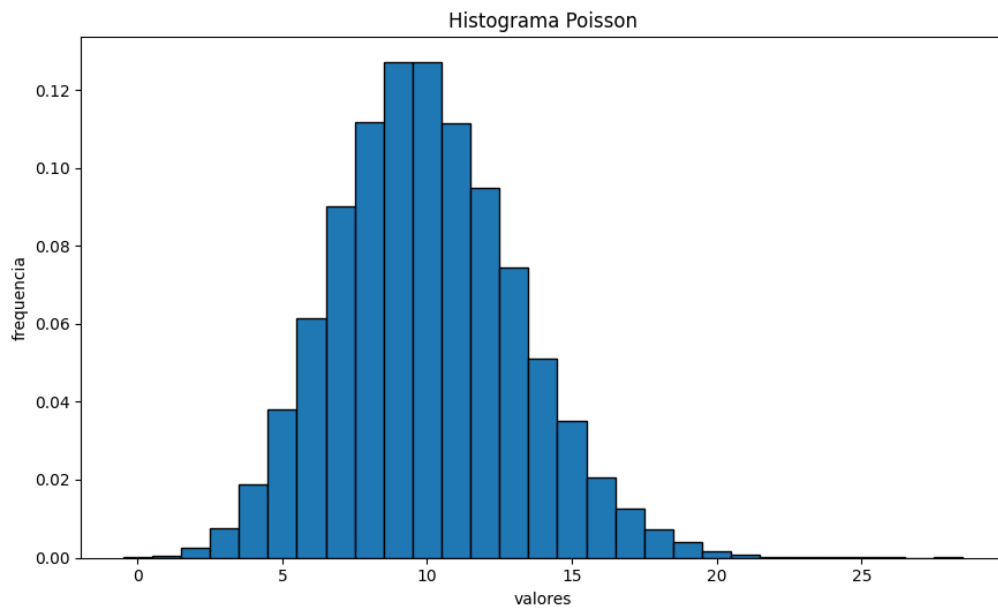


Fig. 27: Representación de 50.000 valores resultantes del generador exponencial con $\lambda = 10$

Gráfica esperada de la Distribución Poisson:

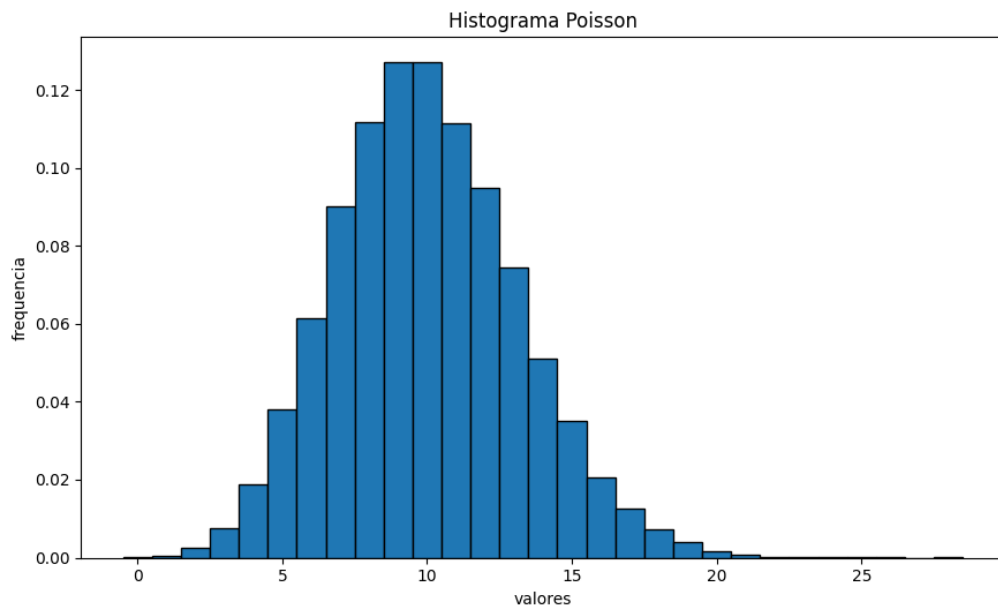


Fig. 28: Representación de 50.000 valores resultantes del generador exponencial con $\lambda = 10$

4.9 Distribución empírica discreta

```

1  # Importe de librerías
2  import random
3  import matplotlib.pyplot as plt
4  import numpy as np
5
6  #Cadenas discretas de Markov
7
8  pos2 = 0
9  pos = 0
10 x = []
11 def list_empirica(limit, M, N, I):
12     p=np.zeros((limit, limit), dtype=np.float) #creo la matriz p y la lleno de valores random
13     pos=0 #posicion del valor random dentro del listado
14     for i in range(limit):
15         for j in range(limit):
16             val=np.random.rand(pos)
17             p[i, j]=val
18             pos+=i+j+limit #incremento la posicion para evitar valores randoms repetidos
19     p=np.sort(p)
20     x=np.zeros((limit), dtype=np.float) #creo el array x y lo lleno de 0's
21
22     pos2=pos+1 #posicion del valor random dentro del listado generado por gcl
23     for l in range(N):
24         r=np.random.rand(pos2)
25         pos2+=1 #incremento la posicion para evitar valores randoms repetidos
26         J = -1
27         for j in range(M):
28             if j>=M-1:
29                 break
30             else:
31                 if p[I, j]<r<p[I, j+1]:
32                     J=j+1
33                     break
34             else:
35                 continue
36         if J!=-1:
37             I=J
38             x[I]+=1
39
40     lista=x.tolist() #para acomodar los resultados
41     if M<10:
42         while len(lista)>M:
43             lista.pop()
44     result=[]
45     for i in lista:
46         result.append(int(i))
47     return result
48
49 #Defino los valores
50 list_empirica(100,10,10,10)

```

Listing 9: Programa que genera la matriz P con distribución empírica

5 Conclusión

Después de realizar los experimentos destinados a verificar la aleatoriedad (o no) de los generadores con las distribuciones estudiadas, podemos concluir que, en la mayoría de las mismas, las gráficas obtenidas por medio del código no son exactamente iguales a las gráficas esperadas (hecho que esperábamos que sucediera), pero presentan importantes similitudes, lo que nos permite decir que pasaron el test. Siendo específicos, solamente dos distribuciones (Hipergeométrica y binomial negativa) presentaron significativas diferencias entre gráficas; por lo tanto, en estos casos se podría decir que se produjo una excepción, es decir que no pasaron el test.

6 Referencias

Función de probabilidad - Wikipedia
Función de densidad - Wikipedia
Función de distribución - Wikipedia
Función cuantil - Wikipedia
Método de la transformada inversa - Wikipedia
Distribuciones de probabilidad - Datanalytics
Simulación, algunos métodos y aplicaciones - fcea uruguay
Método de Aceptación y rechazo - columbia edu