

Tarea 5

Generación de números aleatorios

Erick Cervantes Mendieta

Matrícula: 2032430

Modelos Probabilistas Aplicados

Octubre 2020

1. Números aleatorios Uniformes

Existen diferentes procedimientos para generar números aleatorios uniformes, algunos de ellos se discutirán más adelante, sin embargo, se debe asegurar ciertas características deseables que aseguren la confiabilidad de los resultados obtenidos, como lo son:

- a. Uniformemente distribuidos.
- b. Estadísticamente independientes.
- c. Reproducibles.
- d. Período largo.
- e. Generados a través de un método rápido.
- f. Generados a través de un método que no requiera mucha capacidad de almacenamiento de la computadora.

Los números generados por una computadora no se deben llamar *números aleatorios* porque son predecibles y se pueden reproducir, por ello, en ocasiones se les llama **números pseudoaleatorios**. Varios esquemas han sido propuestos [2] para la generación de dichos números a través de relaciones matemáticas de recurrencia, de los cuales los más populares son los *métodos Congruenciales*: aditivo, multiplicativo (Lehmer, 1951) y mixto (Rotenburg, 1960). El método congruencial mixto incluye características de los otros dos, por lo que se detalla a continuación.

Números uniformes		
n	$(a = 5 \text{ y } c = 7)$	$(a = 3 \text{ y } c = 11)$
0	0.375	0.875
1	0.750	0
2	0.625	0.375
3	0	0.500
4	0.875	0.875
5	0.250	0
6	0.125	0.375
7	0.500	0.500
8	0.375	0.875
9	0.750	0
10	0.625	0.375

Cuadro 1: Números pseudoaleatorios generados con $m = 8$ y $x_0 = 4$.

1.1. Generador congruencial Mixto

Los generadores congruenciales lineales generan una secuencia de números pseudoaleatorios en la cual el próximo número pseudoaleatorio es determinado a partir del último número generado. Para este caso, la relación de recurrencia es la siguiente:

$$x_{n+1} = (ax_n + c) \text{ mód } m$$

donde: x_0 = la semilla ($x_0 > 0$); a , c y m son enteros positivos ($a < m$, $c < m$). Esta notación matemática significa que x_{n+1} es el residuo cuando $ax_n + c$ se divide entre m . En consecuencia, los valores posibles de x_{n+1} son $0, 1, \dots, m - 1$, de manera que m representa el número deseado de valores diferentes que se puede generar como números aleatorios. La cantidad de números consecutivos de una sucesión antes de que se repita se conoce como *longitud de ciclo*.

Suponga que se tiene un generador (implementado en lenguaje R [1]), con los siguientes parámetros: $a = 5$, $c = 7$, $x_0 = 4$ y $m = 8$, la secuencia de **números uniformes** (x_{n+1}/m) se muestran en el cuadro 1, en dónde, se puede apreciar que la longitud de ciclo es 8. En este primer ejemplo se puede pensar que la longitud de ciclo es igual a m , es por esto que se decidió cambiar los valores de $a = 3$, $c = 11$ para observar el comportamiento de los números generados (cuadro 1), y se puede observar que, aunque se volvió a considerar un valor de $m = 8$, la longitud de ciclo en este caso es de 4.

De los ejemplos antes mencionados, se puede inferir que la asignación de los valores que toman los parámetros en el generador son importantes, ya que se desearía al menos tener una longitud de ciclo completo (igual a m). Coss (1995) propone algunas de estas reglas, las cuales se mencionan a continuación:

- a. Selección de m : seleccionar de modo que sea el número primo más grande posible y que a su vez sea menor que p^d , donde p es la base del sistema (binario, decimal,

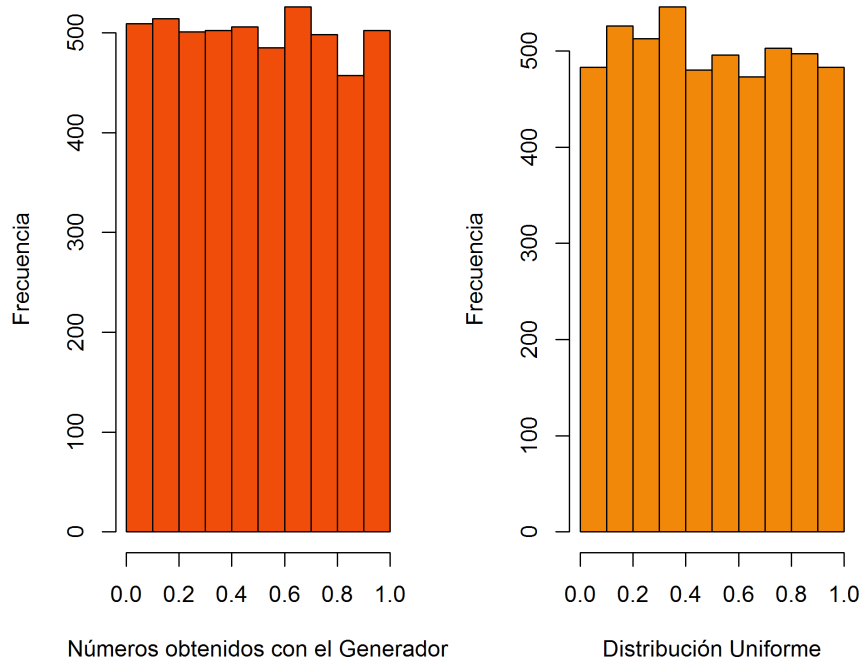


Figura 1: Comparación de gráficas de los números pseudoaleatorios generados *vs* datos provenientes de una distribución Uniforme.

hexadecimal, etc.) que se está utilizando y d es el número de bits que tiene una palabra de computadora en ese sistema.

- Selección de a : el valor seleccionado de a debe ser entero impar, y además no debe ser divisible por 3 o 5.
- Selección de c : el valor seleccionado para este parámetro puede ser cualquier constante, sin embargo se recomienda un valor entero impar y relativamente primo a m .
- Selección de x_0 : el valor de este parámetro resulta tener poca o ninguna influencia ya que solo afecta a la sucesión en el punto en el que comienza y no en la progresión de los números.

Luego entonces, con base a las recomendaciones anteriores, procedemos a generar $n = 5000$ números uniformes con los siguientes parámetros: $a = 7^5$, $c = 19813$, $x_0 = 10169$ y $m = 2^{31} - 1$, en la figura 1 se puede apreciar visualmente que al parecer nuestros números obtenidos con el generador parecieran provenir de una distribución Uniforme, sin embargo, se necesita de una herramienta más especializada para poder llegar a esa conclusión.

En el lenguaje R esta implementada una prueba estadística la cuál comprueba si un histograma es significativamente diferente de una distribución uniforme, esto nos permitió verificar si nuestros números uniformes generados sustentan la hipótesis de que vienen de una distribución Uniforme, la instrucción es la siguiente: `uniform.test(hist(uniforme(5000,10169)))`. El resultado nos arroja un $p\text{-value} = 0.7246$, lo cual nos hace pensar que al parecer nuestros datos si provienen de una distribución Uniforme.

Otra prueba que se puede hacer para probar esta hipótesis es la **prueba de los promedios**, la cual parte del hecho que el valor esperado y la varianza de una variable aleatoria uniformemente distribuida son iguales a $1/2$ y $1/12$, respectivamente. En esta prueba se plantean las siguientes hipótesis: $H_0 : \mu = 1/2$ vs $H_1 : \mu \neq 1/2$. Luego, para realizar esta prueba se requiere obtener el promedio aritmético de n números pseudoaleatorios:

$$\bar{x} = \frac{U_1 + U_2 + \dots + U_n}{n}$$

En seguida se determina el valor del estadístico Z_0 , con la siguiente expresión:

$$Z_0 = \frac{(\bar{x} - 1/2)\sqrt{n}}{\sqrt{1/12}}$$

Si $|Z_0| < Z_{\alpha/2}$, entonces no se puede rechazar la hipótesis de que los números pseudoaleatorios generados provienen de una distribución Uniforme con media igual a 0.5. Para nuestro caso, se obtuvo que de los 5000 números generados su media es igual a 0.4952859, por lo que el estadístico Z_0 resulta ser:

$$Z_0 = \frac{(0.4952859 - 1/2)\sqrt{5000}}{\sqrt{1/12}} = -1.154722$$

Así, si suponemos un nivel de significancia del 5 %, entonces $Z_{\alpha/2} = 1.96$, luego es fácil ver que $|Z_0| < 1.96$, por lo que no hay evidencia para rechazar H_0 , es decir, los números generados tienen una media igual a 0.5, por lo que es posible que si sigan una distribución Uniforme.

Ahora, probemos generar otros $n = 5000$ números uniformes, pero ahora con parámetros que no cumplan las reglas citadas anteriormente: $a = 80$, $c = 1000$, $x_0 = 10169$ y $m = 115$, en la figura 2 podemos observar que al parecer nuestros números obtenidos con el generador no parecen provenir de una distribución Uniforme. Al realizar la prueba con instrucción: `uniform.test(hist(uniforme(5000,10169)))`, el resultado nos arroja un $p\text{-value} < 2.2e - 16$, lo cual apoya que al parecer nuestros datos no provienen de una distribución Uniforme.

1.2. Generador de Wichmann-Hill

Uno de los primeros generadores combinados de números pseudoaleatorios fue el propuesto por Wichmann y Hill en 1982. Este generador emplea los siguientes tres generadores lineales congruenciales para obtener tres números enteros (x_i, y_i, z_i) :

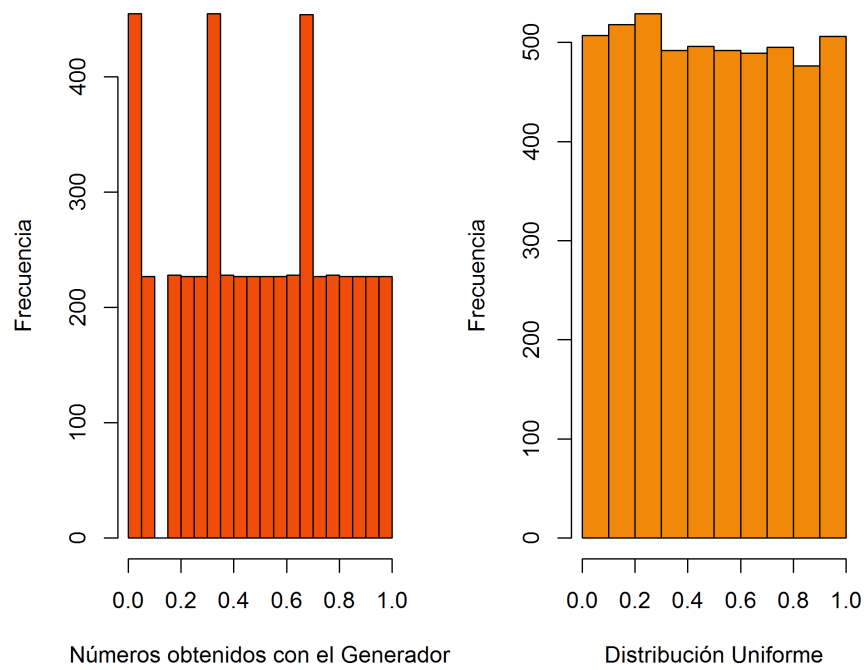


Figura 2: Comparación de gráficas de los números pseudoaleatorios generados *vs* datos provenientes de una distribución Uniforme.

$$\begin{aligned}x_i &= (171x_{i-1}) \text{ mód } 30269 \\y_i &= (172y_{i-1}) \text{ mód } 30307 \\z_i &= (170z_{i-1}) \text{ mód } 30323\end{aligned}$$

de los cuales se puede obtener un número pseudoaleatorio u_i en el intervalo $(0, 1)$, de la siguiente forma:

$$u_i = \left(\frac{x_i}{30269} + \frac{y_i}{30307} + \frac{z_i}{30323} \right) \text{ mód } 1$$

La semilla de este generador son tres números enteros (x_0, y_0, z_0) , los cuales deben tomar valores entre 1 y 30000.

Suponga que se tiene un generador (implementado en lenguaje R) con los siguientes parámetros: $x_0 = 3$, $y_0 = 11$, $z_0 = 10865$, en la figura 3 podemos observar que al parecer nuestros números obtenidos con el generador parecen provenir de una distribución Uniforme. Al realizar la prueba con instrucción: `uniform.test(hist(uniforme2(n,3,11,10865)))`, el resultado nos arroja un $p\text{-value} = 0.6865$, lo cual apoya la idea de que al parecer nuestros datos si provienen de una distribución Uniforme.

2. Números aleatorios Normales

Puesto que no es posible expresar la distribución acumulada de la distribución Normal en forma explícita, se utiliza el teorema del límite central para generar números al azar, dicho teorema establece que la suma de n variables aleatorias independientes se aproxima a una distribución Normal a medida que n se aproxima a infinito. Si las variables que se están sumando son uniformes en el intervalo $(0, 1)$, entonces

$$Z = \frac{\sum_{i=1}^n R_i - n/2}{\sqrt{n/12}}$$

tiene una distribución normal estándar. Se sabe que la normal estándar de una variable aleatoria x distribuida normalmente se obtiene de la siguiente manera:

$$Z = \frac{x - \mu}{\sigma}$$

Una manera simplificada para obtener valores simulados es mediante la siguiente expresión:

$$x = \mu + \sigma \left(\sum_{i=1}^{12} R_i - 6 \right)$$

Así, un método para generar números aleatorios que sigan una distribución Normal es el proponer dos números aleatorios (u_1, u_2) independientes, obtenidos de la misma función de densidad en el intervalo $(0, 1)$, luego considerar las siguientes variables:

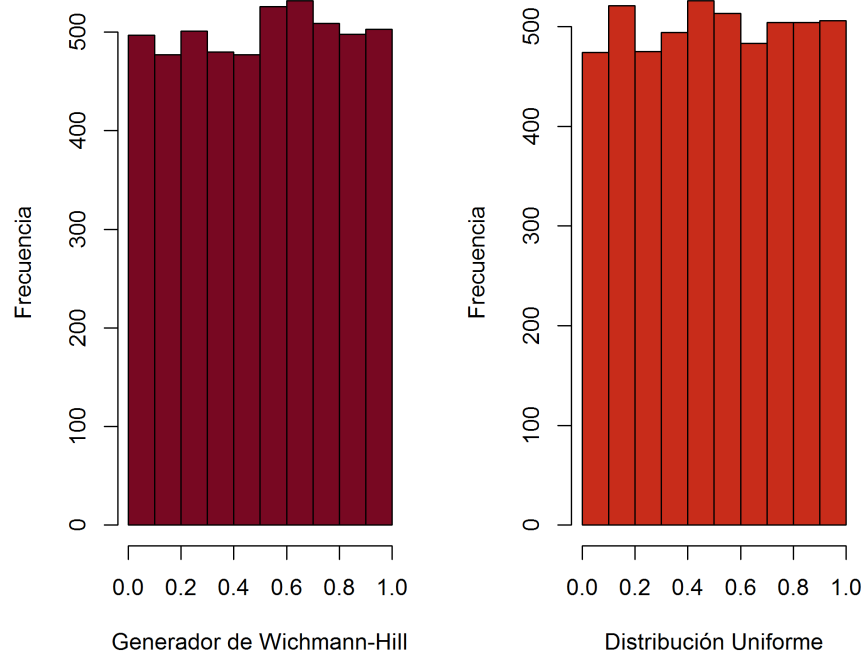


Figura 3: Comparación de gráficas de los números pseudoaleatorios generados *vs* datos provenientes de una distribución Uniforme.

$$\begin{aligned}x_1 &= \sqrt{-2 \ln u_1} \cos(2\pi u_2) \\x_2 &= \sqrt{-2 \ln u_1} \sin(2\pi u_2)\end{aligned}$$

Con las dos expresiones anteriores se pueden obtener dos números pseudoaleatorios que siguen una distribución Normal:

$$\begin{aligned}z_0 &= [\sqrt{-2 \ln u_1} \cos(2\pi u_2)]\sigma + \mu \\z_1 &= [\sqrt{-2 \ln u_1} \sin(2\pi u_2)]\sigma + \mu\end{aligned}$$

En la figura 4 se muestra el comportamiento de números pseudoaleatorios obtenidos con el método descrito anteriormente, para esta ocasión se consideraron los valores de z_0 y de z_1 en la misma base de datos generados ($n = 5000$), y podemos observar visualmente que al parecer si siguen una distribución Normal. Para apoyar este supuesto, se obtuvo una gráfica QQ Normal (ver figura 5), y se realizó una prueba de normalidad de Shapiro-Wilk, la cual muestra un $p\text{-value} = 0.4027$, por lo que supongo que los datos si presentan un comportamiento Normal.

En la figura 6 se muestra el comportamiento de números pseudoaleatorios, en el cual se consideraron los valores de z_0 y de z_1 por separado, y podemos observar visualmente que al parecer si siguen una distribución Normal los dos conjuntos de datos generados. Para apoyar este supuesto, se obtuvo una gráfica QQ Normal (ver figura 7), y se realizó una prueba de normalidad de Shapiro-Wilk, la cual muestra un $p\text{-value} = 0.4027$ para los valores de z_0 , y un $p\text{-value} = 0.8381$ para los valores de z_1 , por lo que hay evidencia para decir que los datos obtenidos si presentan un comportamiento Normal.

Referencias

- [1] R Core Team. R: A Language and Environment for Statistical Computing. <https://www.R-project.org/>, 2020.
- [2] Coss Bu Rauil. *Simulación: un enfoque práctico*. Limusa, México, 1995.

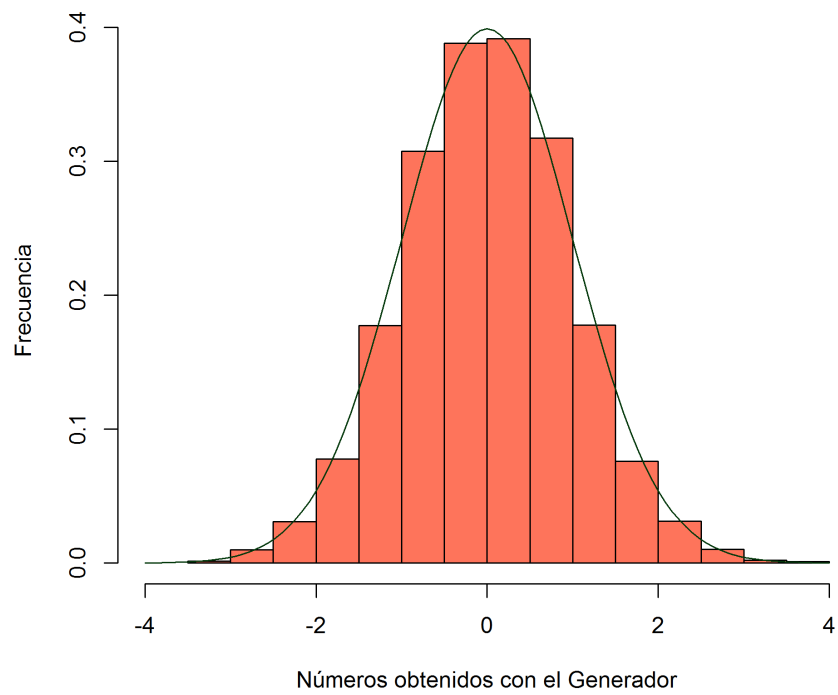


Figura 4: Distribución teórica Normal (z_0 y z_1) sobre el histograma obtenido con los números generados.

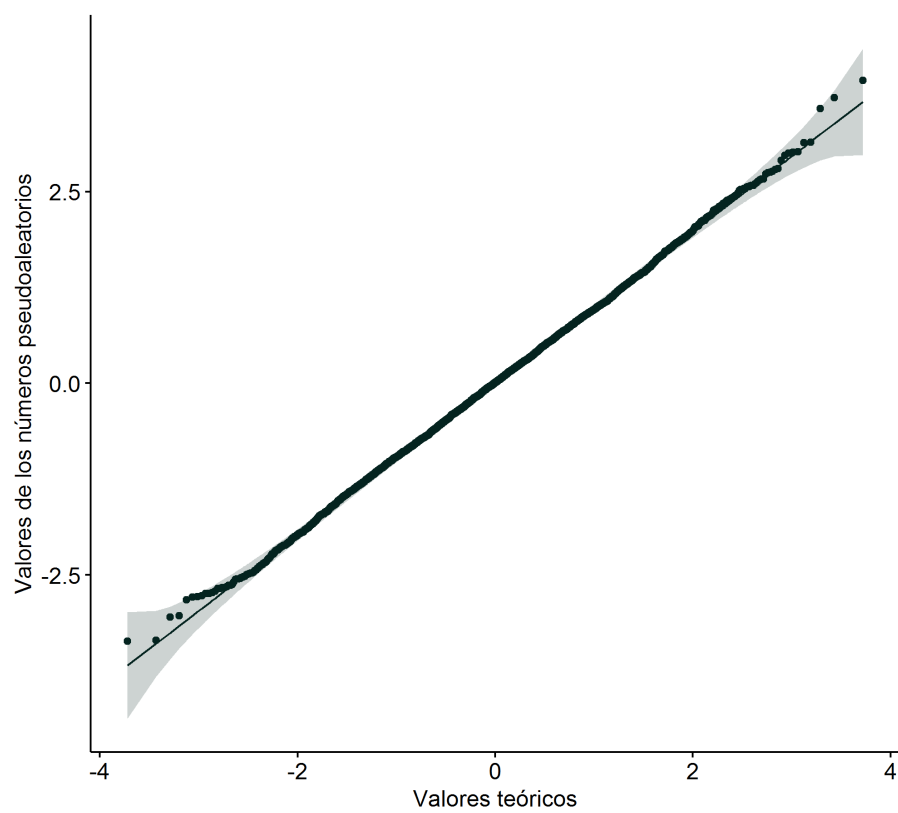


Figura 5: Gráfica QQ Normal para los números pseudoaleatorios generados (z_0 y z_1).

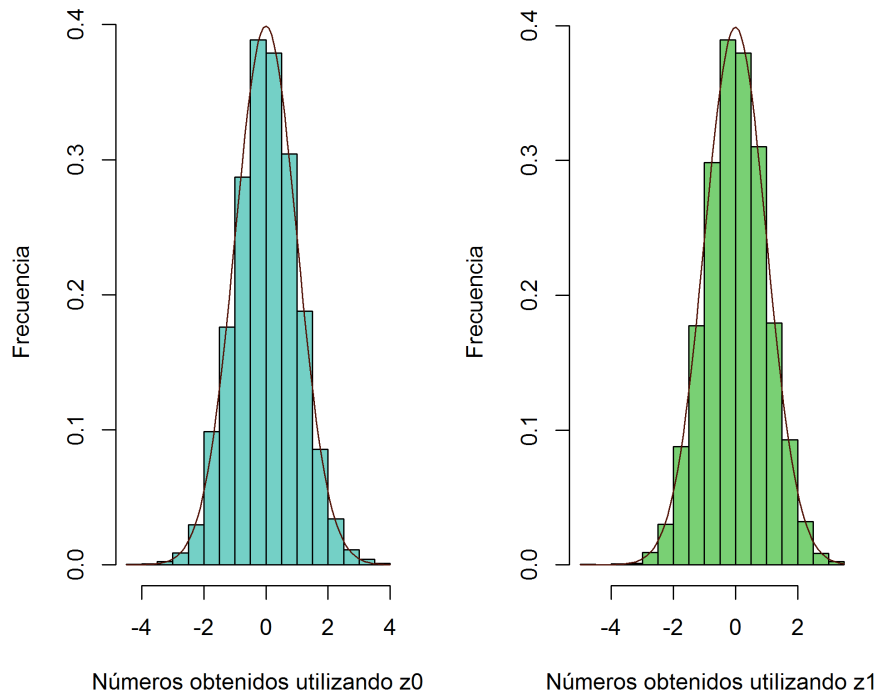


Figura 6: Distribución teórica Normal *vs* histograma de los números generados.

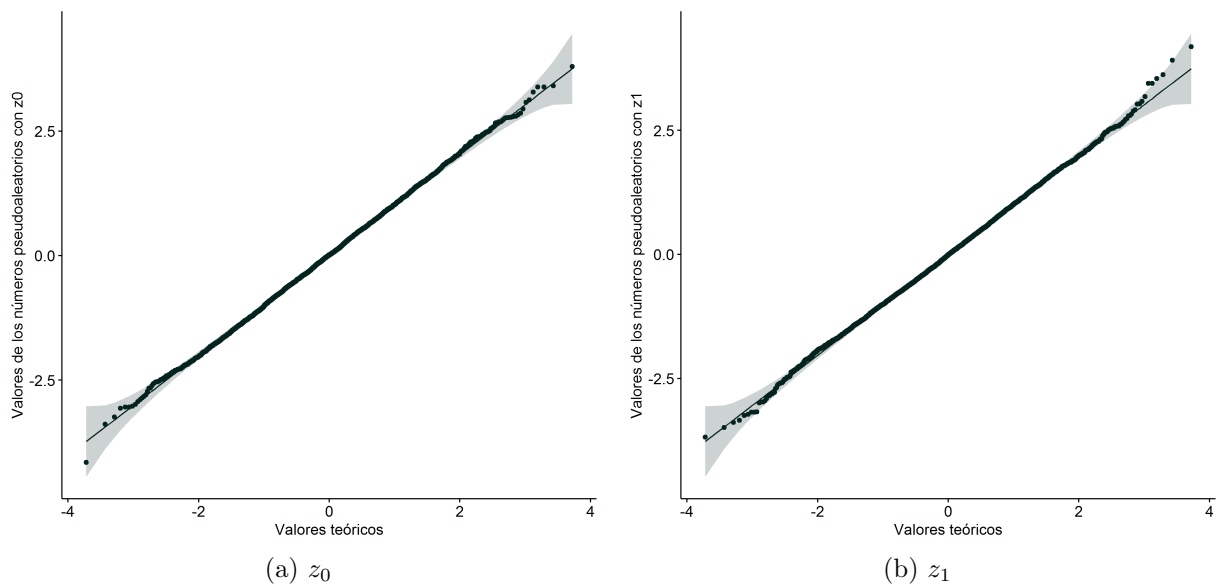


Figura 7: Gráfica QQ Normal para los números pseudoaleatorios generados.