

Distribución de probabilidad de normal

TAREA 5

Alumno:

Joaquín Arturo Velarde Moreno

1. Introducción

El objetivo del siguiente reporte es describir el comportamiento de una distribución normal, ver su representación matemática y la forma de su curva en histogramas, para lo cual usaremos el programa R 4.0.2 [1] y de este modo, haremos cálculos con conjuntos con la finalidad de mostrarlos gráficamente. Además, intentaremos simular una distribución normal a partir de valores uniformes utilizando la transformación Box-Muller. Para cumplir con esta finalidad, usaremos como apoyo el material de la Dra. Elisa Schaefer [1].

2. Definición

La distribución normal o Gaussiana, nombrada así en honor a Carl Friederich Gauss, es una distribución con forma de campana usada para aproximarse al valor de una variable aleatoria continua a una situación ideal. Esta distribución contiene dos parámetros que la definen: μ y σ donde

- μ es la media de la distribución y
- σ es la desviación estándar

Por ejemplo, si elaboráramos una encuesta de la edad de cada uno de los 150 estudiantes de una escuela y obtuviéramos que su promedio es de 18 años, tendríamos un histograma donde se muestra que hay una mayor cantidad de alumnos cerca del promedio o la media Figura 4.1, es decir, hay la tendencia de querer aproximarse a una curva de distribución normal Figura 4.2. Pero si esto lo aproximamos a una situación ideal, no tomando la edad de solo 150 individuos sino de 5000 estudiantes, entonces veremos que la forma de nuestra distribución se acerca a la curva de la distribución normal Figura 4.3, esta curva nos permite ver dos características de distribución, la asimétrica y la asintótica:

a) **Asimétrica**, lo cual significa que la media es igual a los valores de la moda y la mediana, en otras palabras, los alumnos que estén más cerca del promedio, son más numerosos que los que están más alejados de este promedio, esto también implica que se cumple la siguiente ecuación:

$$f_X(\mu - x) = f_X(\mu + x) \forall x \in \mathbb{R}$$

b) **Asintótica**, significa que se acerca continuamente a la recta del eje x sin llegar nunca a encontrarla, es decir, puede extenderse hasta el infinito. La desviación estándar se representa con la letra griega (σ), y nos indica lo lejos que están distribuidos los resultados de la media, lo cual puede producir dos situaciones: una situación homogénea o una heterogénea.

Si la situación es homogénea la desviación estándar σ es muy reducida y los sujetos están muy cercanos a la media μ y la curva será muy afilada Figura 4.3. En cambio, si la situación es heterogénea la desviación σ es muy alta y los sujetos están muy alejados a la media μ y la curva será menos afilada Figura 4.4.

3. Formula

La fórmula para la función de probabilidad normal está definida por la siguiente expresión:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

donde:

- μ es la media de la distribución.
- σ es la desviación estándar.

Dado que esta ecuación es demasiado complicada, incluso haciendo uso de la calculadora, para este tipo de cálculos se suele hacer uso de la herramienta R [1], en la cual ya está predefinida con la siguiente función *dnorm()*, y si queremos generar un vector de una distribución normal, podemos utilizar el método *rmnorm()*.

```
Promedio <- 18
Edad     <- rmnorm(500000, Promedio, 1)
```

Si esto lo graficamos mediante un histograma, tendremos la misma distribución normal que se espera Figura 4.3

4. Método de Box-Muller

También podemos generar una distribución normal por otros métodos tales como el método de Box-Muller, el cual es un método de generación de pares de números aleatorios independientes con distribución normal estándar, es decir con esperanza cero y varianza unitaria, expresada en las siguientes fórmulas.

$$Z_0 = R \cos(\Theta) = \sqrt{-2 \log(U_1)} \cos(2\pi U_2)$$
$$Z_1 = R \sin(\Theta) = \sqrt{-2 \log(U_1)} \sin(2\pi U_2)$$

donde:

- U_1 y U_2 son variables aleatorias independientes con una distribución uniforme entre los valores 0 a 1.
- Z_1 y Z_2 son variables aleatorias independientes con una distribución normal con una desviación estándar de 1.

Esto expresado en R puede ser definido como:

```
u = runif(2);
z0 = sqrt(-2 * log(u[1])) * cos(2 * pi * u[2]);
z1 = sqrt(-2 * log(u[1])) * sin(2 * pi * u[2]);
```

Si queremos tener nuestra propia media y desviación, podemos afectar la variable aleatoria:

```
mu      = 18;
sigma   = 1;
u       = runif(2);
z0      = sqrt(-2 * log(u[1])) * cos(2 * pi * u[2]);
z1      = sqrt(-2 * log(u[1])) * sin(2 * pi * u[2]);
z0      = sigma * z0 + mu;
z1      = sigma * z1 + mu;
```

Si utilizamos este método para generar nuestras variables aleatorias, se obtendrían resultados similares a la distribución normal Figura 4.5.

Es necesario que ambas funciones se obtengan al calcular con dos variables aleatorias de distribución uniforme, en caso de cambiar el cálculo usando únicamente una sola variable uniforme, se rompería la distribución normal Figura 4.6 . Esto lo podemos verificar con el método *Shapiro.test()* el cual nos devuelve un valor *p-value*, si el valor obtenido es mayor a 0.05, entonces significa que nuestro vector muy probablemente viene de una distribución normal, sin embargo este método solo puede trabajar con un vector de 1 a 5000 valores por lo cual, es necesario utilizar el método *sample()* para tratar con vectores muy grandes.

```

test <- shapiro.test(runif(100))
print(test)
data:  runif(100)
      W = 0.95081, p-value = 0.0009379

# prueba negativa 0.0009379 < 0.05

test <- shapiro.test(sample(Edades, 5000, replace = TRUE))
print(test)
data:  sample(Edades, 5000, replace = TRUE)
      W = 0.99963, p-value = 0.5049

# prueba positiva 0.5049 > 0.05

```

Si cambiáramos una variable uniforme del método de Box-Muller a por ejemplo esta fuera dependiente de la primera elevándola al cuadrado, también afectaría nuestra distribución Figura 4.7 .

```

u      <- runif(1);
u_2    <- u * u;
z0     <- sqrt(-2 * log(u)) * cos(2 * pi * u_2 );
z1     <- sqrt(-2 * log(u)) * sin(2 * pi * u_2 );
datos  <- c(z0, z1);

```

Referencias

- [1] The R Foundation. *The R Project for Statistical Computing*. <https://www.r-project.org/>. 2019.

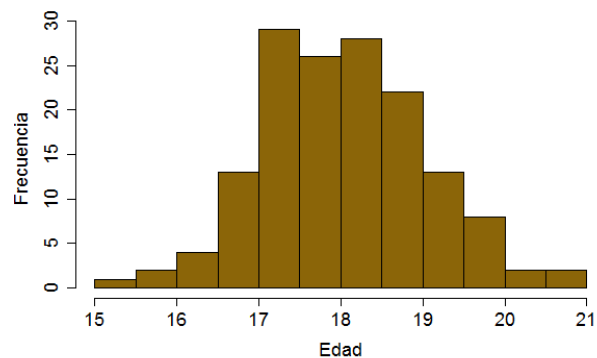


Figura 4.1: Distribución de frecuencia de la encuesta.

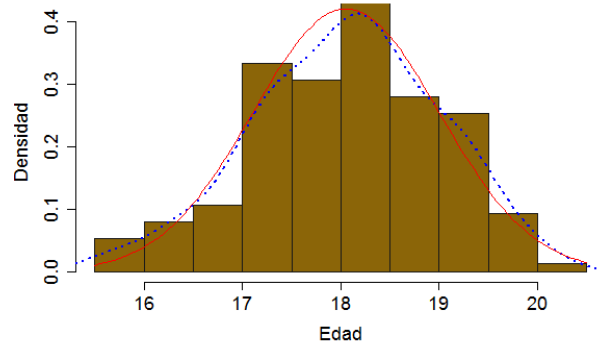


Figura 4.2: Densidad de resultados en la encuesta.

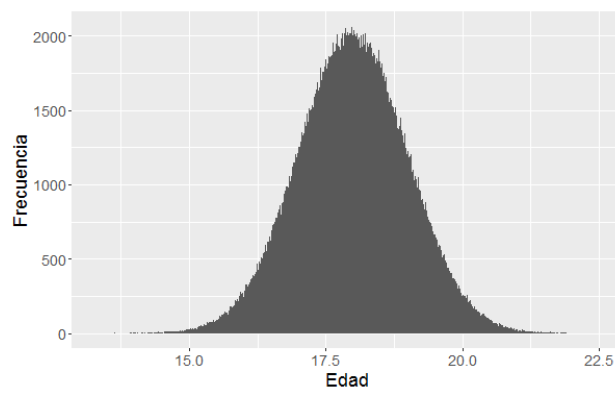


Figura 4.3: Distribución de frecuencia de la encuesta con 50000 estudiantes.

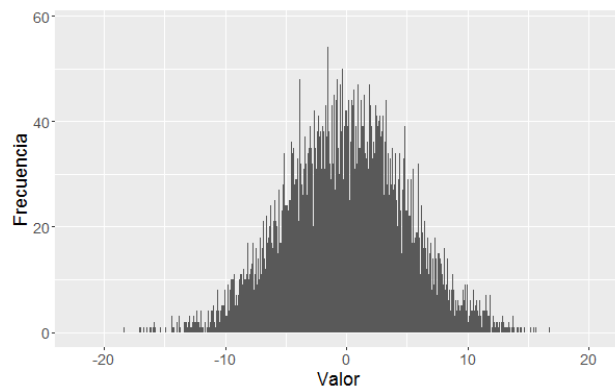
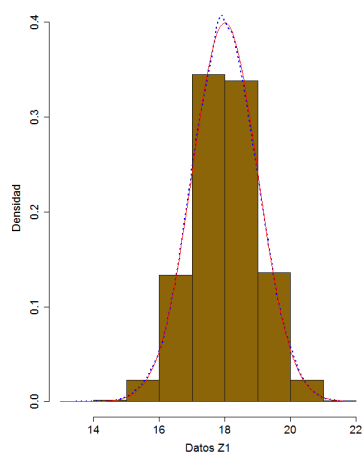
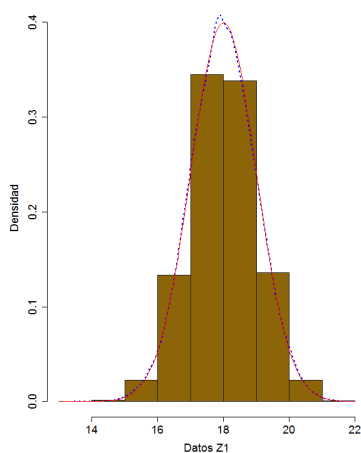


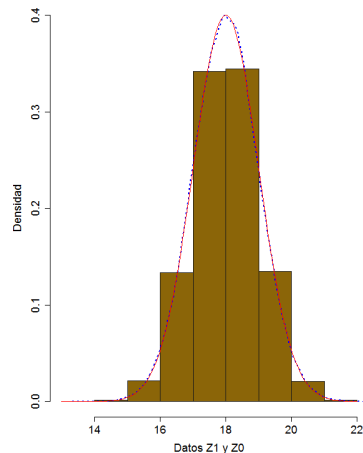
Figura 4.4: Distribución de frecuencia normal heterogénea.



(a) Distribución de frecuencia con una sola variable Z_1 .

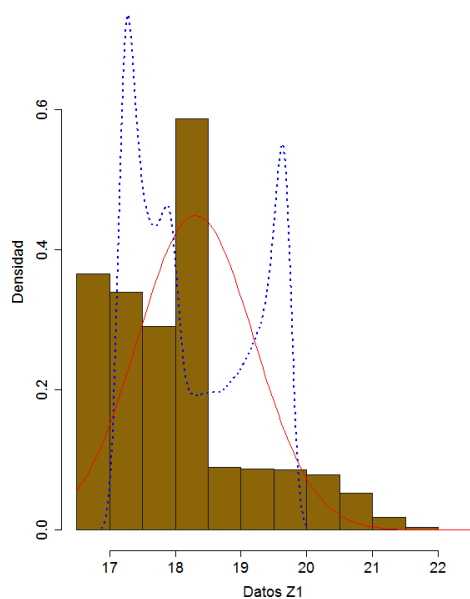


(b) Distribución de frecuencia con una sola variable Z_0 .

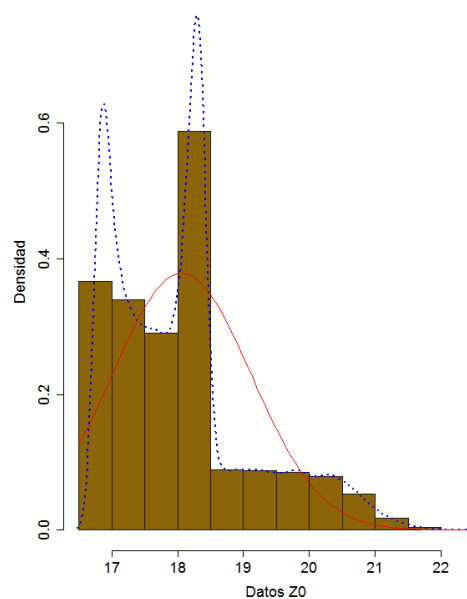


(c) Distribución de frecuencia con ambas variables.

Figura 4.5: Densidad de las distribuciones generadas con el método de Box-Muller.

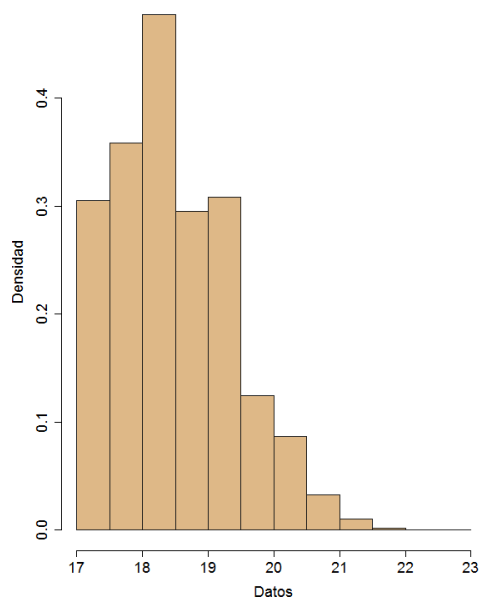


(a) Distribución de frecuencia de la variable Z_0 con solo una variable uniforme.

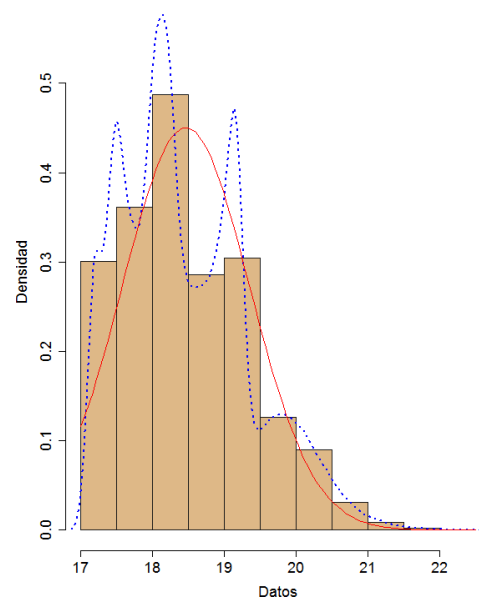


(b) Distribución de frecuencia de la variable Z_1 con solo una variable uniforme.

Figura 4.6: Distribución de frecuencia de las variables de Box-Muller con solo una variable uniforme.



(a) Histograma que muestra la frecuencia de un vector obtenido por el método de Box-Muller.



(b) Densidad de la distribución resultado de una variable dependiente en el método Box-Muller.

Figura 4.7: Distribución de frecuencia de la variables de Box-Muller con una variable uniforme dependiente en una población de 50000 estudiantes.