Distribución de probabilidad de normal

TAREA 5

Alumno:
Joaquín Arturo Velarde Moreno

1. Introducción

En el siguiente reporte vamos a definir el comportamiento de una distribución normal, ver sus formulas y sus representaciones en histogramas, haremos uso del programa R 4.0.2 [1] para hacer cálculos con conjuntos de datos y mostrarlos gráficamente usando como apoyo el material de la Doctora Elisa Schaefer [1]. También intentaremos simular una distribución normal a partir de valores uniformes utilizando la transformada de box-Muller.

2. Definición

la distribución normal o Gaussiana, nombrada así por Gauss, es una distribución usada para aproximarse al valor de una variable aleatoria continua a una situación ideal. Esta consiste de 2 parámetros μ y σ. donde:

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

si sacáramos 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 empieza a ver que hay una mayor cantidad de alumnos cerca del promedio a la media Figura 4.1, esto da la tendencia de querer aproximarse a una curva de distribución normal Figura 4.2 .

Si a esto lo aproximamos a una situación ideal no tomando solo 150 si no tomando la de 5000 estudiantes entonces veremos mas fácilmente la forma de nuestra distribución acercándose a la curva de la distribución normal Figura 4.3, esta curva nos indica 2 cosas sobre esta distribución.

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

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

Asintotica Que se acerca continuamente a la recta de los eje x sin llegar nunca a encontrarla, es decir puede extenderse hasta el infinito.

la desviación estándar se representa con la letra griega σ (sigma), y nos indica lo lejos que están distribuidos los resultados de la media, esto puede producir 2 situaciones una situación homogénea o heterogénea.

Si la situación es homogénea la desviación σ es muy reducida y los sujetos están todos 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 todos muy alejados a la media μ y la curva sera menos afilada Figura 4.4.

3. Formula

La formula para la función de probabilidad normal esta definida por la siguiente expresión

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

donde:

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

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

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

si lo graficamos con 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 formulas.

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

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

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 un 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 obtendríamos que son similares a la distribución normal Figura 4.5.

Es necesario que ambas funciones se obtengan al calcular con 2 variables aleatorias de distribución uniforme, en el caso de cambiar el calculo usando solo una sola variable uniforme, 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, debido a esto es bueno 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
```

Referencias

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

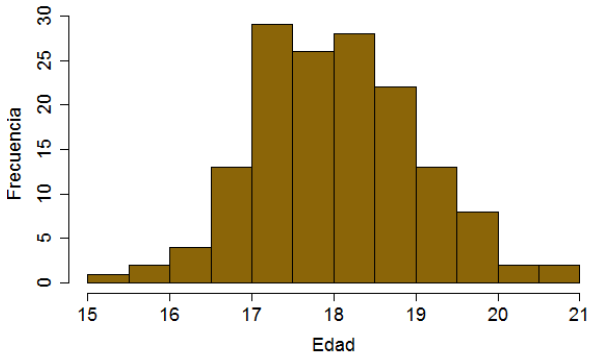


Figura 4.1: Distribución de frecuencia de la encuesta.

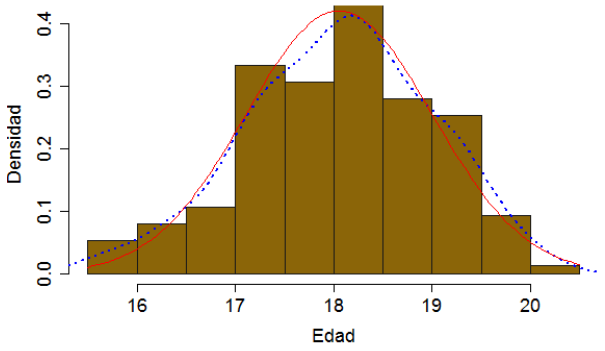


Figura 4.2: Densidad de resultados en la encuesta.

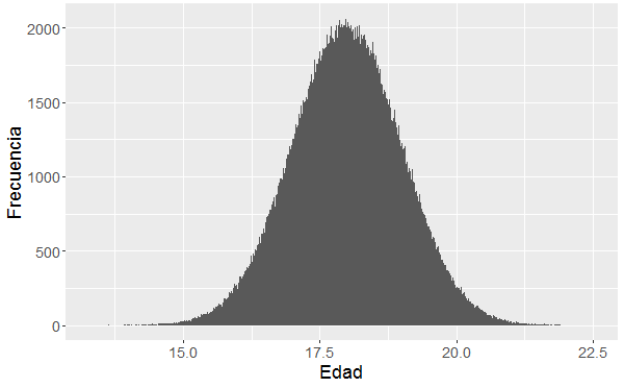


Figura 4.3: Distribución de frecuencia de la encuesta con 50000 estudiantes.

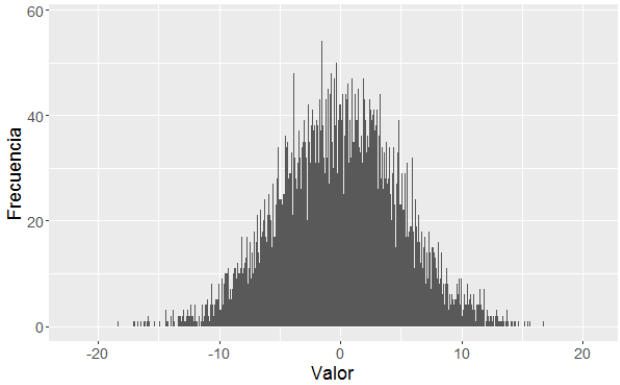
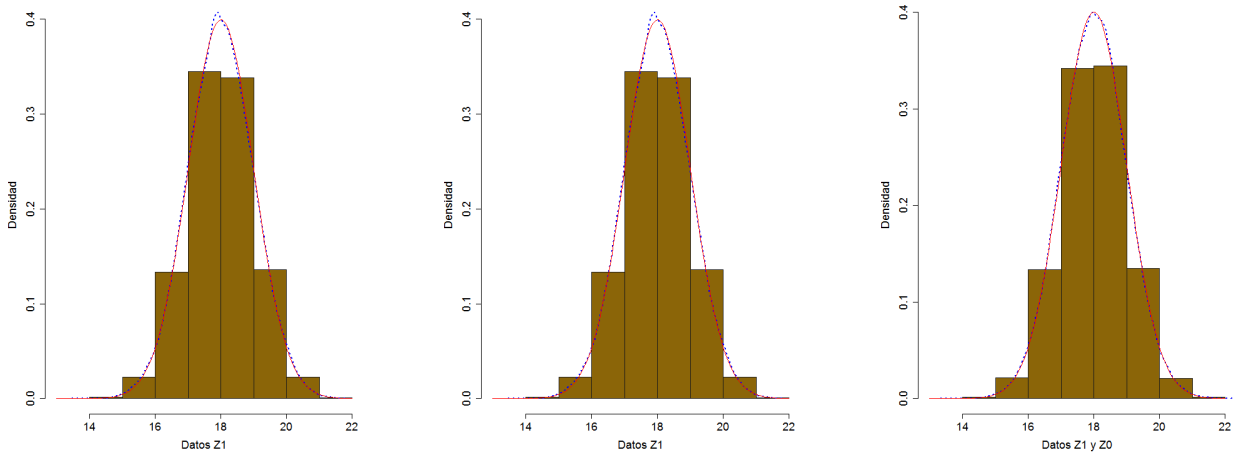


Figura 4.4: Distribución de frecuencia normal heterogénea.

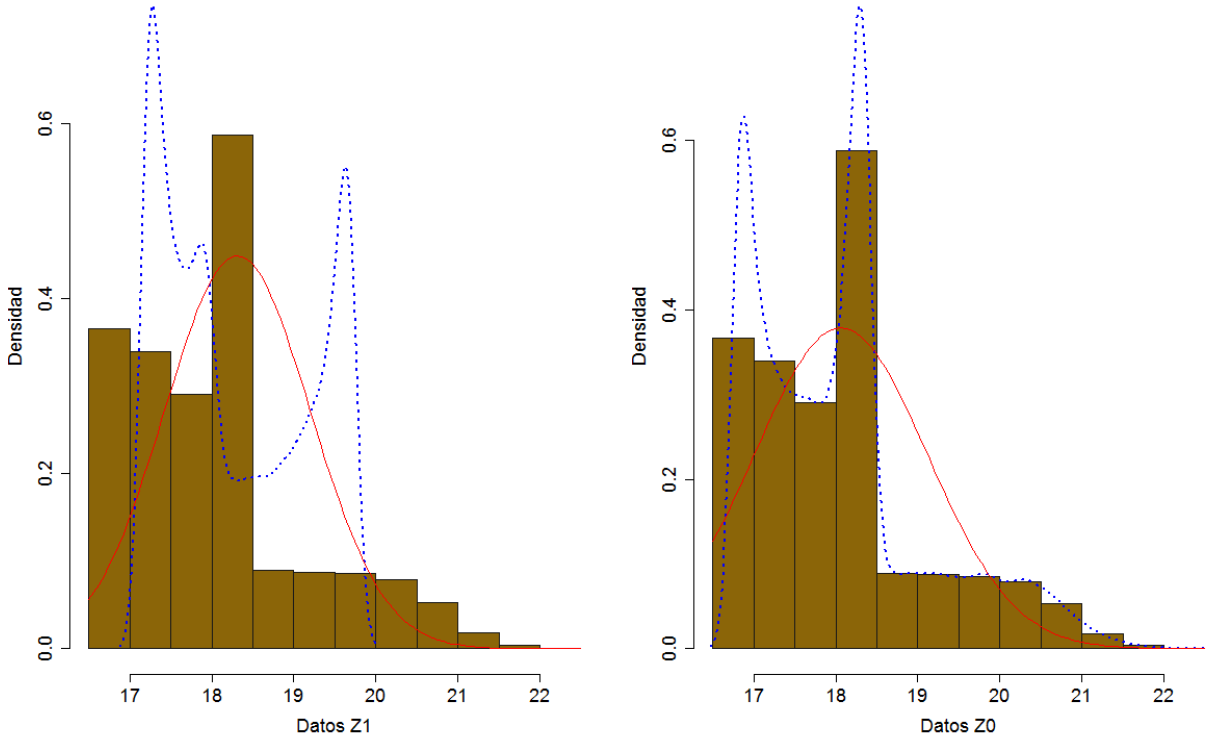


(a) Distribución de frecuencia con una sola variable Z_1.

(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_0con solo una variable uniforme.

(b) Distribución de frecuencia de la variable Z_1con solo una variable uniforme.

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