Sección 2

Probabilidad y Estadística ejercicios

Daniela Jijón, Juan Francisco Cisneros y Luciana Valdivieso
10 de julio de 2022

1 Espacio Muestral y Probabilidad

El espacio muestral con todas posibles combinaciones entre todas las variables que recolectamos se calculó utilizando la fórmula para combinaciones:

$$C_n^r = \frac{n!}{(n-r)!r!} \tag{1}$$

Cada observación de nuestra muestra consta de 6 variables, cuatro cuantitativas y dos categóricas. Se tiene 3 opciones para genero, 11 de edad, 10 de colegio académico, 6 año de estudio, 6 de hora de estudio y 18 opciones de GPA

$$\frac{3!}{2!1!} \cdot \frac{11!}{10!1!} \cdot \frac{10!}{9!1!} \cdot \frac{6!}{5!1!} \cdot \frac{6!}{5!1!} \cdot \frac{18!}{17!1!} = 213840$$

El espacio muestral contiene 213840 elementos

2 Variables Aleatorias

La primera variable aleatoria X muestra la diferencia entre los valores de la variable cuantitativa GPA y la media de la muestra que es 3.47.

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx \tag{2}$$

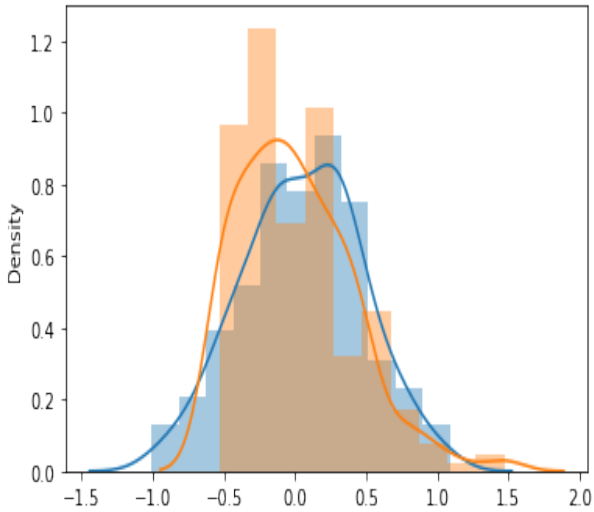
$$\begin{aligned} V(x) &= E(x - \mu x)^2 \\ &= E(x^2) - \mu x^2 \end{aligned} \tag{3}$$

Los resultados de los cálculos son:

$$E(x) = 0.026$$

$$V(x) = 0.21$$

La variable aleatoria sigue una distribución aproximadamente normal



La segunda variable aleatoria Y muestra la diferencia entre los valores de la variable cuantitativa Horas de Estudio Semanales y la media de su muestra que es 12.22
Se calculó la covarianza y el coeficiente de correlación entre estas dos variables aleatorias utilizando las fórmulas 4 y 5 respectivamente

$$\rho xy = \frac{Cov(x,y)}{\sqrt{V(x) \cdot V(y)}} \tag{4}$$

$$\begin{aligned} Cov(x,y) &= E[(x-\mu x)(y-\mu y)] \\ &= E(xy) - \mu x \cdot \mu \end{aligned} \tag{5}$$

Los resultados de los cálculos son:

$$Cov(x,y) = 2.69$$

$$\rho xy = 0.79$$

La covarianza es positiva, lo que significa que existe una relación lineal positiva entre ambas variables. De la misma manera el coeficiente de correlación es un valor positivo considerablemente cercano a uno, lo que indica que existe una fuerte correlación entre las variables, se puede decir que entre mas horas de estudio mayor es el promedio GPA en los estudiantes.

Se seleccionaron dos variables aleatorias adicionales que describen los posibles resultados de las variables cualitativas Sexo y Colegio académico. La función de masa de probabilidad conjunta se obtuvo al calcular las frecuencias de los pares de datos en las observaciones de la muestra.

La independencia de las variables aleatorias se puede comprobar multiplicando las funciones de probabilidad marginal:

$$P(x, y) = P_x(x) \cdot P_x(y) \tag{6}$$

Las variables son dependientes

3 Muestreo

Se han seleccionado dos variables aleatorias, entre ellas el GPA de los estudiantes y las horas de estudio semanales. De cada una de estas variables hemos obtenido 40 muestras aleatorias el 25% de los datos usando Python y la función ".sample(n=50)", para cada una de las 80 muestras se ha obtenido la media y la varianza de estas muestras graficando así las distribuciones para el GPA y las horas de estudio semanales, además de un gráfico QQ plot.

Los datos arrojados nos muestran que ambas distribuciones de medias y varianzas tienen una distribución normal. Esto se observa tanto visualmente debido a la forma de campana que muestra el gráfico de distribución y el gráfico QQ plot donde los datos se muestran en una línea de 45 grados.

Gráfico Distribución de Media para GPA y QQ Plot

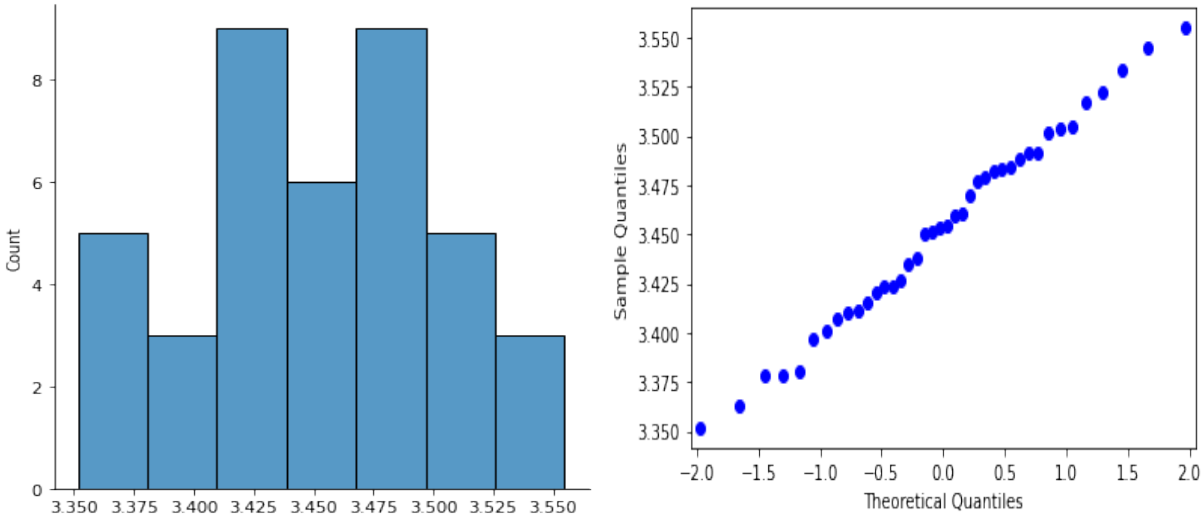
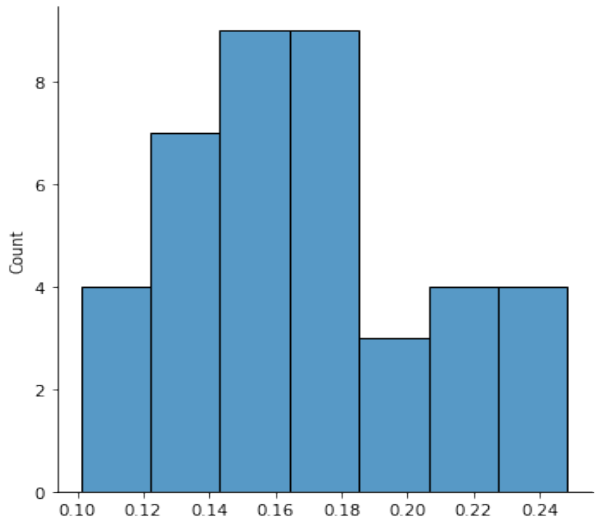


Gráfico Distribución de Varianza para GPA y QQ Plot



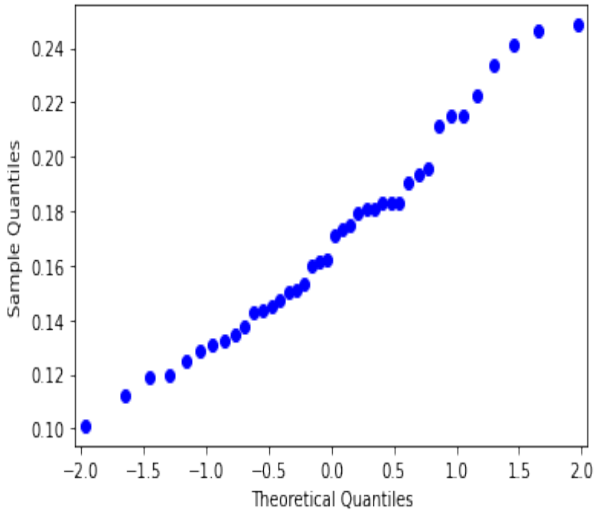
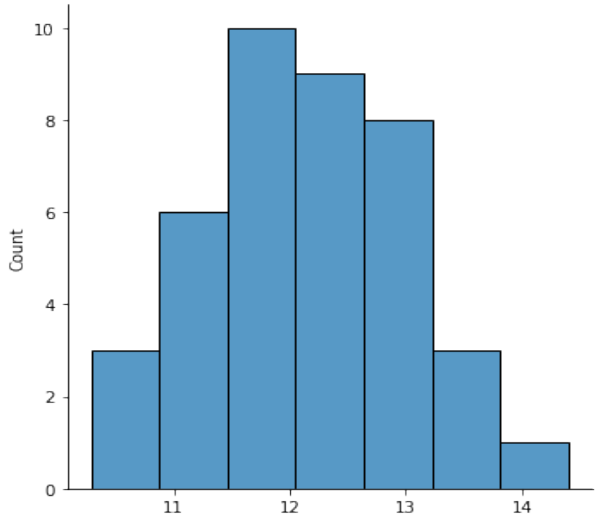


Gráfico Distribución de Media para Horas de Estudio Semanasy QQ Plot



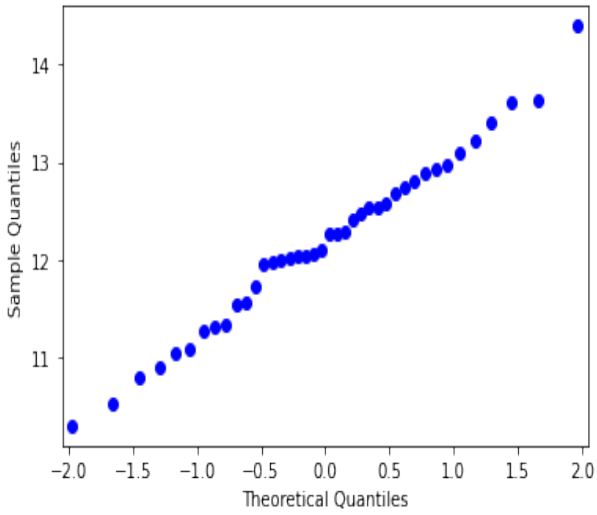
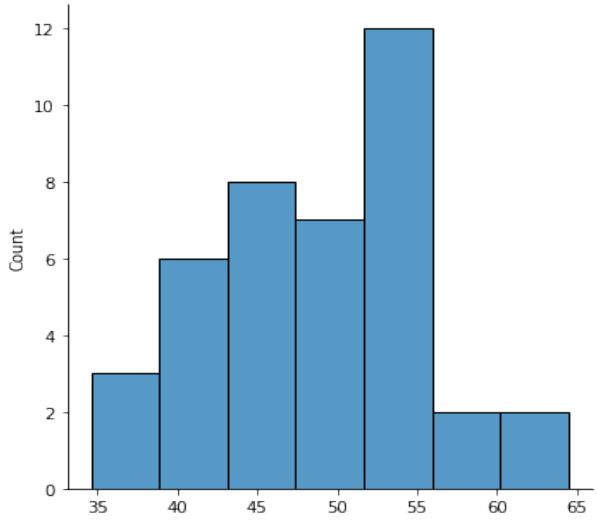
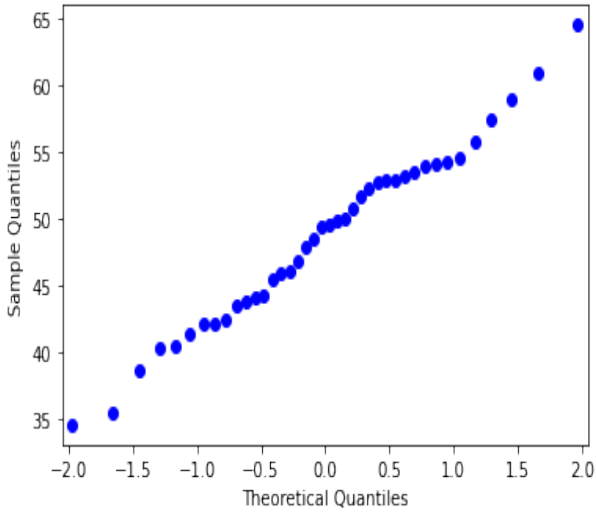


Gráfico Distribución de Varianza para Horas de Estudio Semanasy QQ Plot





Repitiendo el ejercicio anterior pero solo obteniendo 5 muestras del 25% de datos tanto para gpa como para horas de estudio, encontramos que los resultados ya no son distribuciones normales o por lo menos ya no se puede decir que se distribuyen de forma normal sus medias y varianzas. Esto puede ser debido a la cantidad de muestras que no es significativa para el estudio.

Gráfico Distribución de Media para GPA y QQ Plot (5 muestras)

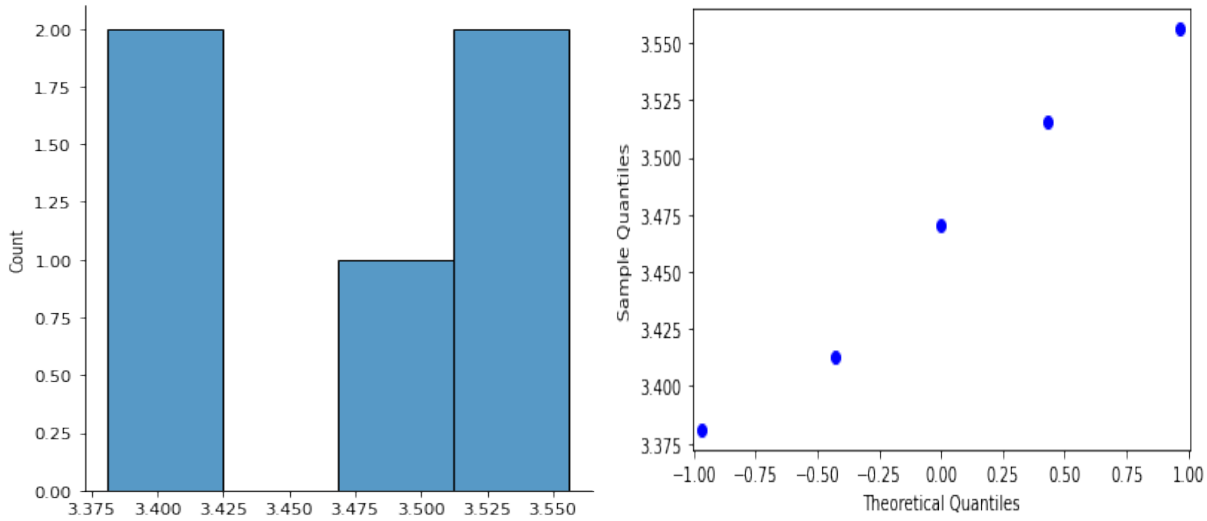


Gráfico Distribución de Varianza para GPA y QQ Plot (5 muestras)

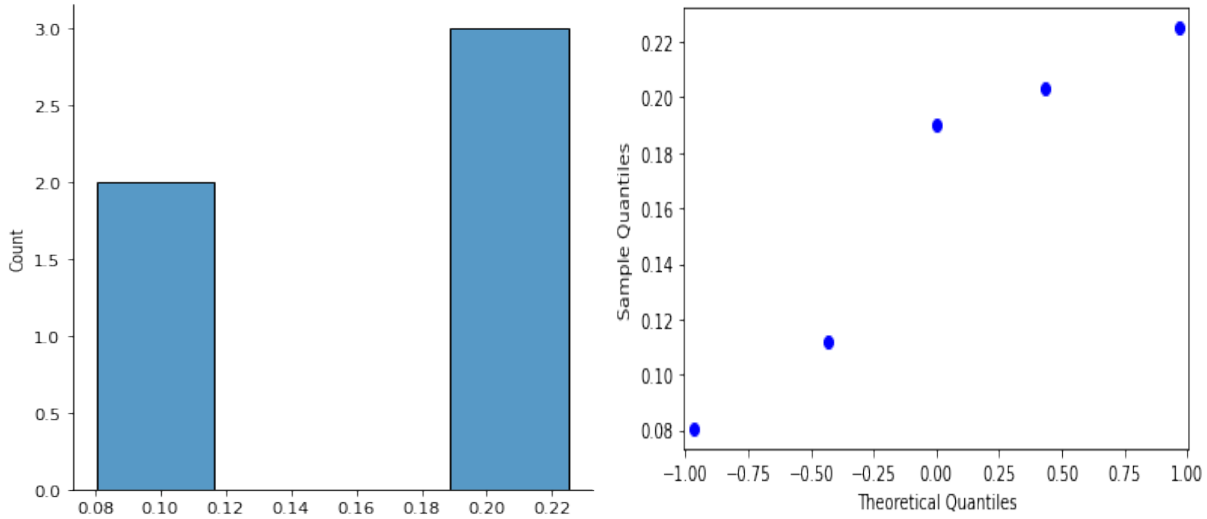


Gráfico Distribución de Media para Horas de Estudio Semanasy QQ Plot (5 muestras)

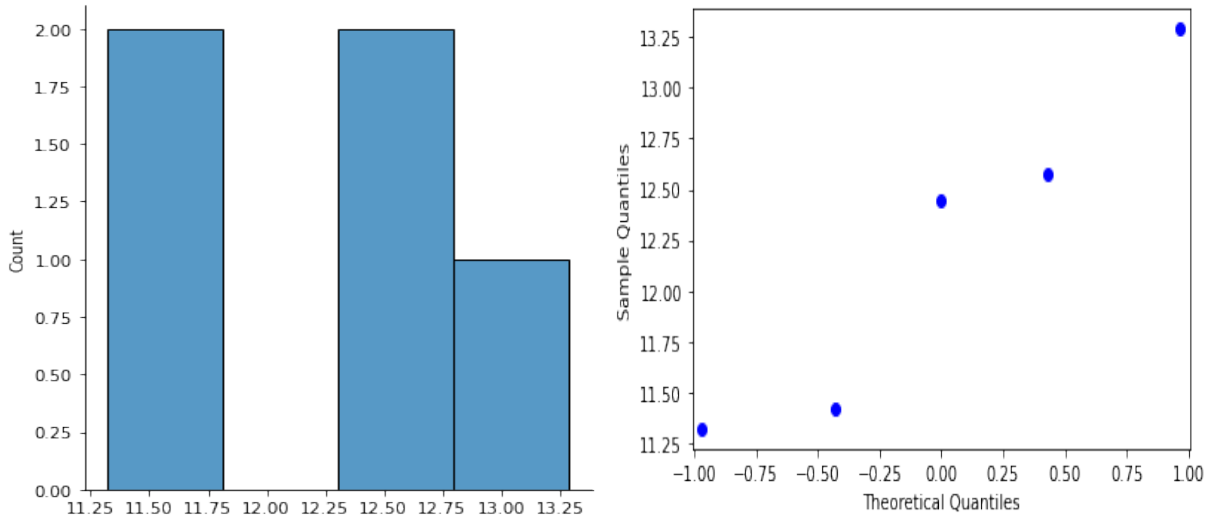
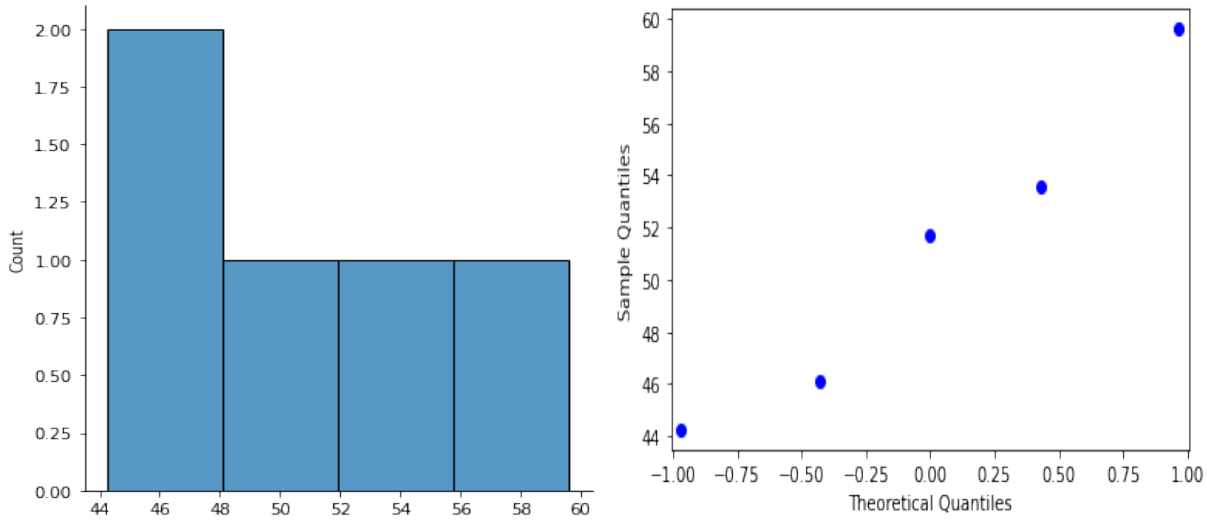


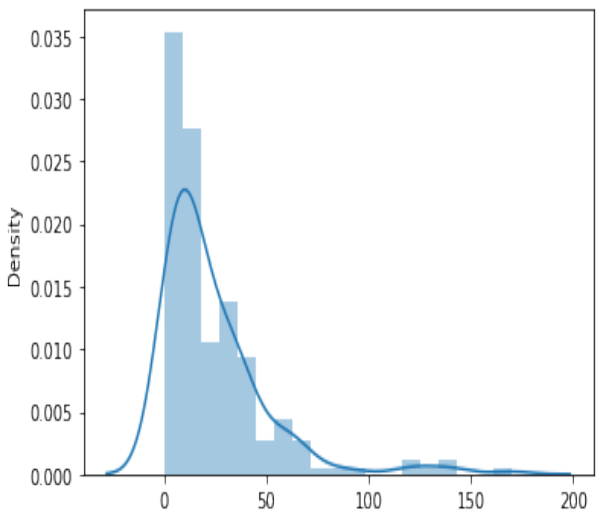
Gráfico Distribución de Varianza para Horas de Estudio Semanasy QQ Plot (5 muestras)

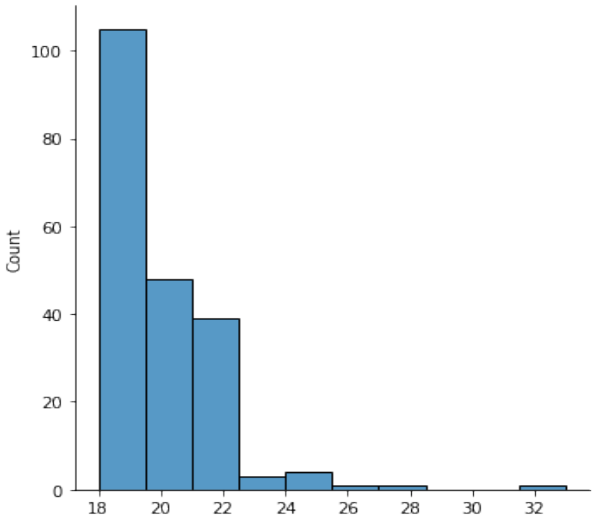


Continuando con el experimento hemos obtenido una tabla mostrando la media y desviación estándar de cada muestra, la diferencia de estas con la media y desviación estándar de la población que previamente habíamos calculado en la sección 1.

Asumiendo que la media de cada muestra es un estimador de la media poblacional se ha encontrado el error estándar del estimador de la media poblacional dando como resultado una media de error estándar de 0.05785757152145068 para el GPA y de 0.9842321350674442 para las horas de estudio semanales.

Finalmente hemos obtenido una nueva variable cuantitativa de nuestra encuesta la cual fue la edad de los estudiantes encuestados y hemos obtenido que las edades de se distribuyen de forma exponencial tal como se muestra en la gráfica a continuación. La primera gráfica muestra una simulación de distribución exponencial mientras que la segunda gráfica es la distribución de las edades de los estudiantes.





Asi mismo hemos obtenido que el valor del parámetro poblacional a partir de el método de máxima verosimilitud es de teta igual a 17.99.