Teorema del límite central

Miguel Angel Soto Hernandez

Centro de Investigación en Computación, Instituto Politécnico Nacional

Ciudad de México, México

msotoh2021@cic.ipn.mx

I. Introducción

El teorema del límite central es un concepto muy importante en el campo de la estadística, y la idea principal nos dice que:

"La media muestral se distribuirá aproximadamente normalmente para tamaños de muestra grandes independientemente de la distribución que estemos muestreando"

Es decir, el teorema del límite central estudia el comportamiento de la suma de las variables aleatorias, cuando crece el numero de sumandos asegura su convergencia hacia una distribución normal en condiciones generales. Este teorema tiene una gran aplicación en la inferencia estadística, ya que muchos parámetros de diferentes distribuciones de probabilidad pueden expresarse en función de la suma de variables, así como también nos permite aproximar distribuciones tales como: binomial, Poisson, chi cuadrado, t-student, entre otras. Esto lo podemos hacer ya que los parámetros tienden a crecer y el calculo se hace difícil, por lo tanto, el uso del teorema del limite central es de gran ayuda.

II. Desarrollo

Antes de conocer de lleno que es el teorema del límite central, primero se retomarán algunas de las características de la distribución muestral de la media muestral para entrar en contexto.

Supóngase que se toman muestras de una población con una media μ, una desviación estándar Ơ y sea $\bar{x}$ una variable aleatoria que represente la media muestral de n observaciones extraídas independientemente de esta distribución. Se sabe que la media de la distribución muestral de la media muestral es igual a la media poblacional, es decir, $\mu_{\bar{x}} = \mu$. Y la desviación estándar de la distribución muestral de $\bar{x}$ es igual a $Ơ_{\bar{x}} = \frac{Ơ}{\sqrt{n}}$. Por lo tanto, si la población se distribuye normalmente, entonces la $\bar{x}$ de la media muestral también se distribuye normalmente.

¿Qué pasa si la población no es normal?

El teorema del limite central aborda el cuestionamiento anterior. La distribución de la media muestral tiende hacia la distribución normal a medida que aumenta el tamaño de la muestra independientemente de la distribución que se este muestreando.

A continuación, se ejecuta una simulación en el lenguaje de programación Python, donde se tiene una población total de 10,000 muestras. Esto se realiza con el fin de ejemplificar la importancia del teorema del límite central.

Primero que nada, se importan las librerías que serán necesarias para poder graficar y ejecutar las muestras.

```
#la librería 'numpy' sirve para generar el espacio muestral
import numpy as np

#la librería 'seaborn' sirve para graficar
import seaborn as sns

#la librería 'statistics' sirve para obtener funciones de estadística
import statistics as stat
```

Una vez importadas las librerías correspondientes procedemos a definir una función la cual nos permitirá calcular la media muestral:

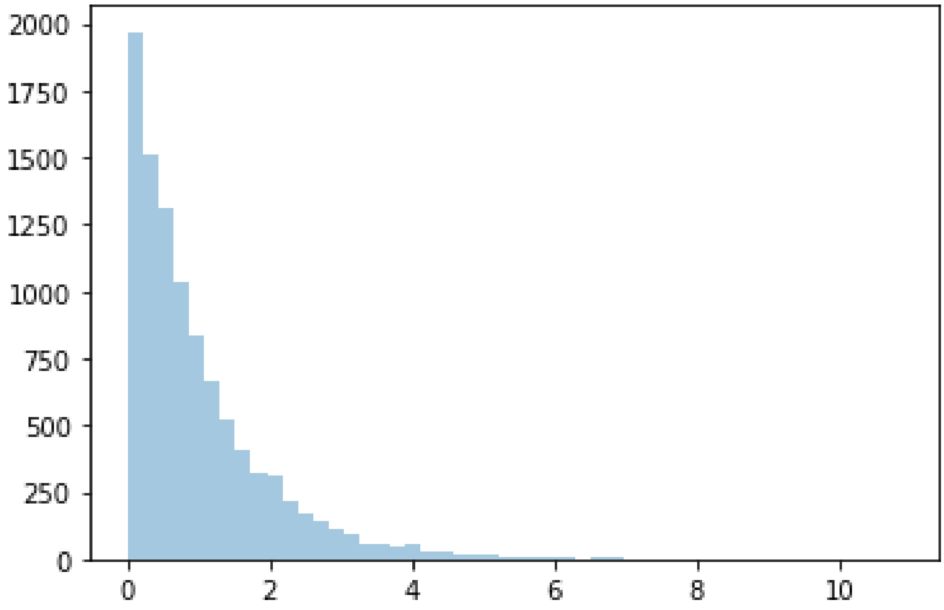
```
def calculadora_media_muestral(array_poblacion, tamano_muestra,
                               n_muestras):
  medias_muestrales = []

  for i in range(n_muestras):
    muestra = np.random.choice(array_poblacion,
                               size= tamano_muestra,
                               replace= False)
    media_muestra = stat.mean(muestra)
    medias_muestrales.append(media_muestra)

  return medias_muestrales
```

Ya que tenemos nuestra función para calcular la media muestral, procedemos a crear un espacio muestral, en este ejemplo será un espacio muestral exponencial con 10,000 muestras. Esto se define en código de la siguiente manera:

```
espacio_muestral = np.random.exponential(size=10000)
```

Hasta el momento se tiene guardado el espacio muestral en una variable llamada '*espacio_muestral*'. Para verificar que nuestro espacio muestral esté tal cual lo queremos, lo graficamos con la siguiente línea de código:

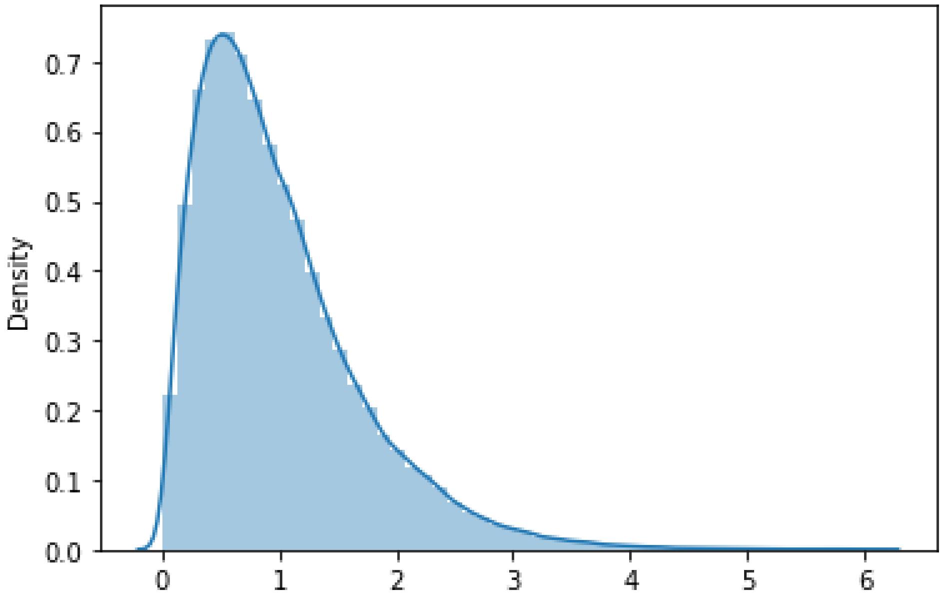
```
sns.distplot(espacio_muestral, kde=False)
```



Gráfica 1: Distribución exponencial

Tomando en cuenta la gráfica 1, si se sacan muestras de 2 observaciones, entonces tendríamos que n=2, si esto se realiza repetidas ocasiones, por ejemplo, en cien mil veces, tendríamos como resultado lo siguiente:

```
_ = sns.distplot(calculadora_media_muestral(espacio_muestral, 2, 100000))
```

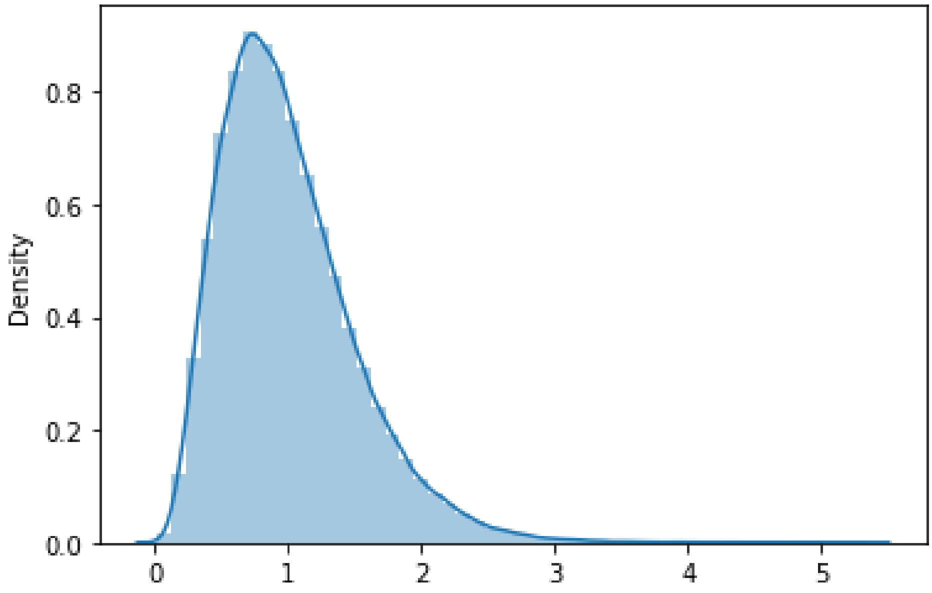


Gráfica 2 Muestras de 2 observaciones ó n=2

El histograma de barras es en la gráfica 2 es un histograma de las medias muestrales donde n=2, y esta será aproximadamente la distribución muestral de $\bar{x}$ en este escenario. En este caso particular se puede calcular matemáticamente la distribución muestral exacta, sin embargo, esto es solo una simulación. De la misma manera, se puede observar que la distribución conserva parte de la distribución original ya que se tiene cierto sesgo y esto no es normal. Por lo tanto, cuando n=2 la distribución muestral de la media muestral no es normal.

Si aumentamos el tamaño de las muestras a 4 observaciones, tal que n=4 y de igual manera se realiza en cien mil ocasiones como se observa en la gráfica 3 se puede observar nuevamente un poco de asimetría, sin embargo, esta va disminuyendo.

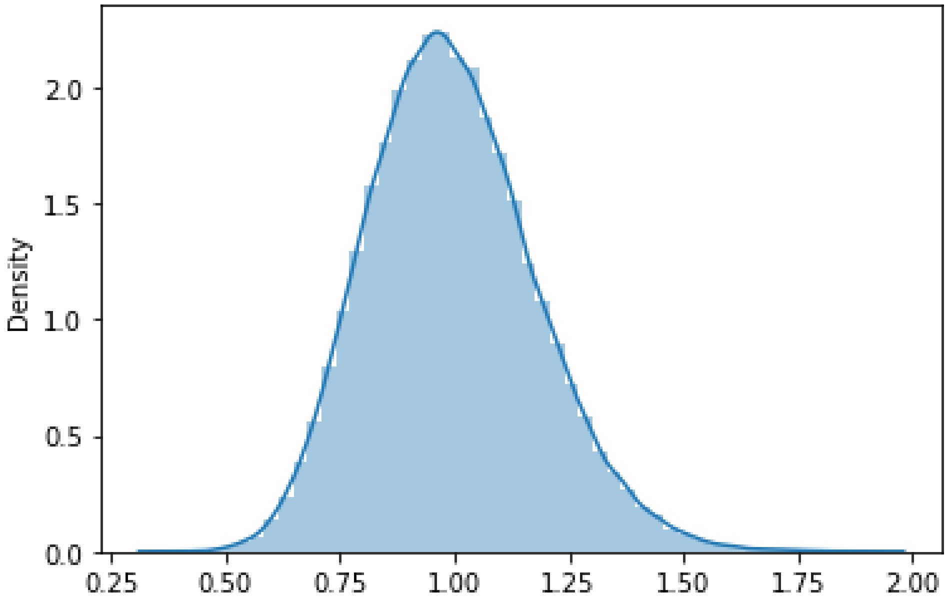
```
_ = sns.distplot(calculadora_media_muestral(espacio_muestral, 4, 100000))
```



Gráfica 3: Muestras de 4 observaciones ó n=4

Cuando tenemos que n=30 como se puede observar en el gráfico 4, se acerca un poco más, a pesar de esto se puede seguir viendo asimetría.

```
_=sns.distplot(calculadora_media_muestral(espacio_muestral, 30, 100000))
```



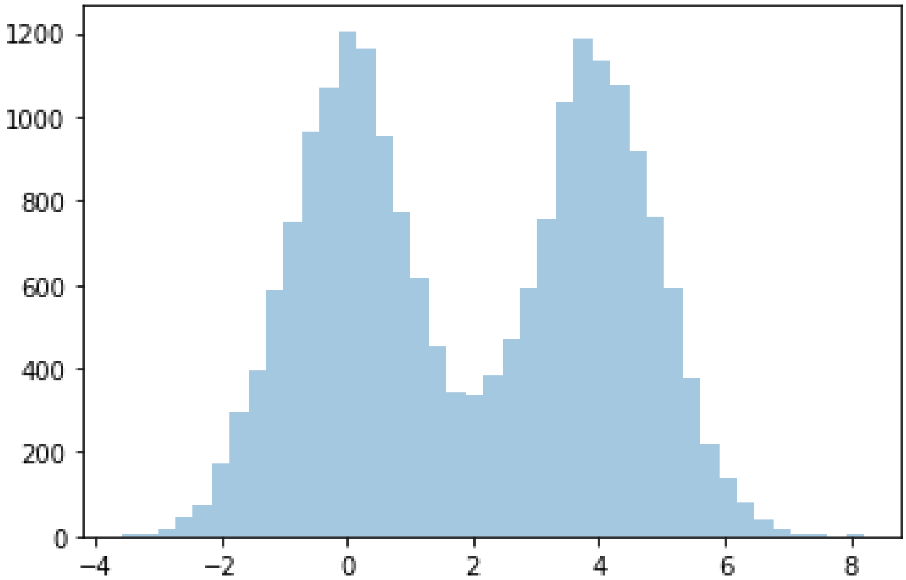
Gráfica 4: Muestras de 30 observaciones ó n=30

Sin embargo, el teorema del limite central no solamente funciona con un solo tipo de muestreo estadístico. En el ejemplo anterior se puede observar que el espacio muestral es exponencial.

Ahora, con el fin de comprobar que este teorema no solo aplica a este tipo de espacios muestrales, se ejemplificará con otro tipo de espacio muestral al cual es un espacio muestral multimodal. El primer paso para crear este espacio muestral en Python, tenemos que definirlo, esto se hace de la siguiente manera:

```
multimodal = np.concatenate((np.random.normal(size= 10000),
                             np.random.normal(loc= 4.0, size= 10000)))
```

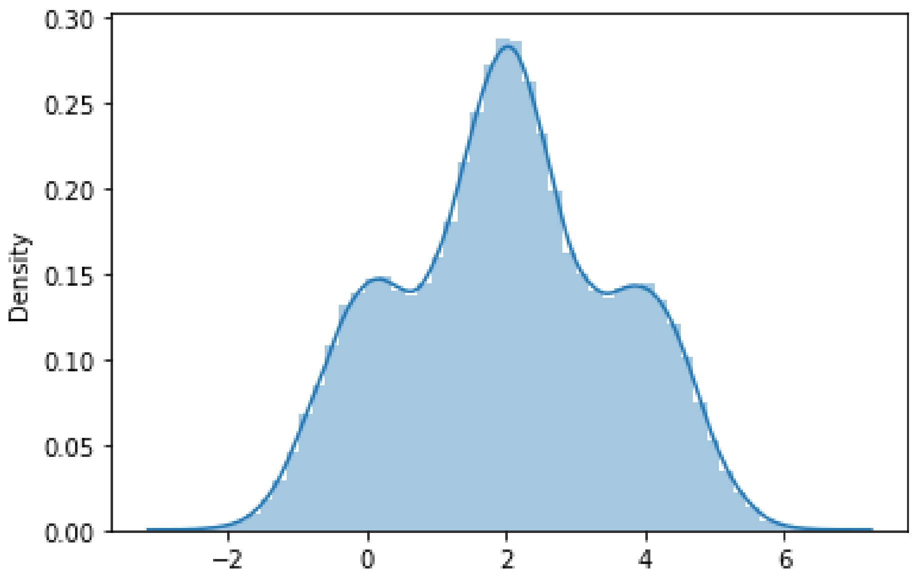
A continuación, se verifica el espacio muestral, como se muestra en la gráfica 5:



Gráfica 5: Distribución multimodal

Una vez verificada la información, se procede a obtener muestras de 2 observaciones, entonces tendríamos que n=2, si esto se realiza repetidas ocasiones, por ejemplo, en cien mil veces, tendríamos como resultado la gráfica 6:

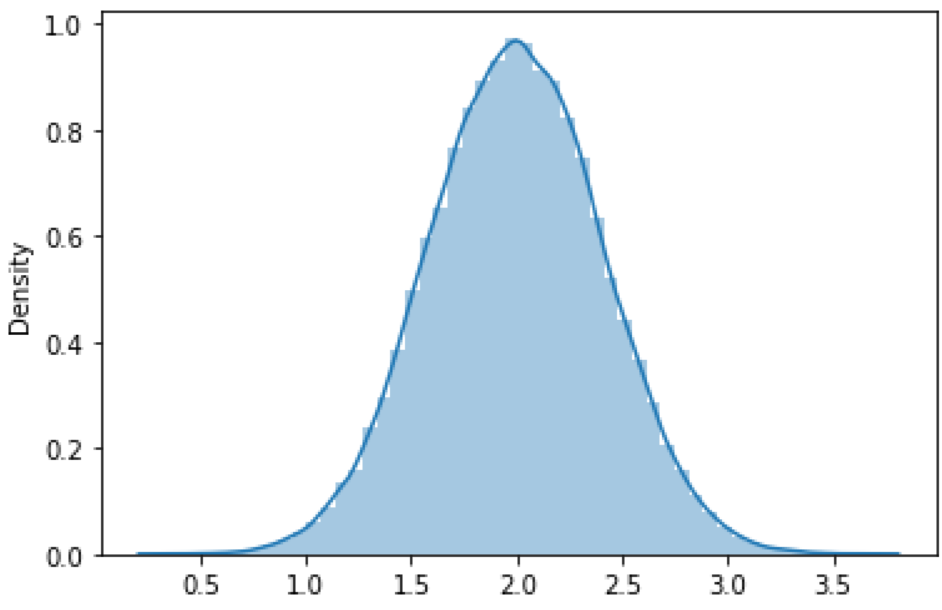
```
_ = sns.distplot(calculadora_media_muestral(multimodal, 2, 100000))
```



Gráfica 6: Muestras de 2 observaciones ó n=2

Por último, se realiza este mismo procedimiento, pero con un mayor número de observaciones, de tal manera que se pueda observar de una mejor manera la distribución, y con el fin de verificar el teorema del limite central, esto se muestra en la gráfica numero 7:

```
_ = sns.distplot(calculadora_media_muestral(multimodal, 30, 100000))
```



Lo que se ilustra con los ejemplos anteriores es que cuando se toman muestras de poblaciones no normales, la distribución de la media muestral tiende hacia la distribución normal a medida que aumenta el tamaño de la muestra.

Como una pauta muy aproximada, se puede considerar que la media de la muestra tiene una distribución aproximadamente normal si el tamaño de la muestra al menos es de 30 si la muestra n es al menos 30. No obstante, esto es solo una pauta, ya que se pueden construir o dar escenarios en los que a pesar de tener un tamaño de muestra de millones no seria suficiente para otorgarnos una normalidad aproximada. Esta pauta esta dada, ya que en la mayoría de los casos prácticos cuando el tamaño de muestra comienza a superar los 30 la distribución de la media muestral será aproximadamente normal.

¿Por qué esto es importante?

El teorema del limite central nos dice que muchas estadísticas tienen distribuciones que son aproximadamente normales para tamaños de muestra grandes incluso cuando se toman muestras de una distribución que no es normal.

Esto significa que a menudo se pueden utilizar procedimientos de inferencia estadística y cálculos de probabilidad bien desarrollados que se basan en una distribución normal incluso si se están tomando muestras de una población que no es normal siempre que se tenga un tamaño de muestra grande.

Visto de una manera mas formal, el teorema del limite central nos dice que el valor de puntuación z habitual involucra la media muestral que tiende en distribución a la distribución normal estándar cuando el tamaño de la muestra tiende al infinito.

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} N(0, 1)\ como\ n \to \infty$$

(dado que μ y σ^2 $son\ finitas$)

III. *Conclusiones*

El teorema del límite central es uno de los conceptos más importantes de la estadística, ya que nos permite converger hacia una distribución normal a partir de una distribución estadística cualquiera. Esto gracias a que, en condiciones muy generales, la distribución de la suma de las variables aleatorias tiende a una distribución normal cuando la cantidad de variables es muy grande. Por lo tanto, se puede garantizar una distribución normal cuando n es suficientemente grande.

IV. *Referencias*

García Cruz, J.A. (2002). Inferencia y significación estadística. En E. Palacián y J. Sancho (Eds.), Actas de las X JAEM (Vol. II, 457-466). ICE Universidad de Zaragoza, Sociedad Aragonesa de Profesores de Matemáticas.

Glencross, M. (1988). A practical approach to the central limit theorem. En R. Davidson y J. Swift (Eds.), Proceedings of the Second International Conference on Teaching Statistics (pp. 287-291). Victoria BC: University of Victoria.

StatQuest with Josh Starmer. (2018, 3 septiembre). The Central Limit Theorem [Vídeo]. YouTube. https://www.youtube.com/watch?v=YAlJCEDH2uY

jbstatistics. (2012, 29 diciembre). Introduction to the Central Limit Theorem [Vídeo]. YouTube. https://www.youtube.com/watch?v=Pujol1yC1_A

Jon Krohn. (2020, 16 marzo). The Central Limit Theorem – With Examples in Python [Vídeo]. YouTube. https://www.youtube.com/watch?v=1p8pBje5SOE