Influencia de la asimetría en el tamaño de la muestra para el cumplimiento del teorema central del límite en distribuciones continúas

Influence of asymmetry in the sample size for the fulfillment of the central limit theorem.

Carlos Andrés Tobón Orozco, José Rubiel Bedoya Sánchez
Departamento de Matemáticas, Universidad Tecnología de Pereira, Pereira, Colombia
joserubiel@utp.edu.co
caantoor@gmail.com

***Resumen*—. Este artículo muestra la influencia de la asimetría de una población para la escogencia de los tamaños de muestra que garantizan el cumplimiento del Teorema Central del Límite. Se realizó un proceso de simulación con el software estadístico R 3.1.0 con diferentes tipos de poblaciones continuas. Se utilizaron las pruebas de bondad de ajuste de normalidad Shapiro-Wilk, Anderson-Darling y Lilliefors para evaluar la normalidad de las diferentes distribuciones muestrales. Por último se muestran los tamaños de muestra que garantizan de acuerdo a la asimetría el cumplimiento del Teorema.**

***Palabras clave*— Teorema Central del Límite, distribuciones muestrales, asimetría, pruebas de bondad de ajuste.**

***Abstract*— This paper shows the influence of the asymmetry of a population for the selection of sample sizes to ensure compliance of the Central Limit Theorem. A process simulation was performed using the statistical software R 3.1.0 with different types of continuous populations. The goodness of fit tests of normality Shapiro-Wilk, Lilliefors and Anderson-Darling normality to evaluate different sampling distributions were used. Finally the sample sizes that guarantee according to the asymmetry of Theorem compliance is.**

***Key Word*- Central Limit Theorem, sampling distributions, asymmetry, goodness of fit tests.**

I. INTRODUCCIÓN

El conocimiento de la distribución muestral de un estadístico como por ejemplo la media muestral, juega un papel relevante en el proceso inferencial. Estas herramientas [1] teóricas y prácticas son las que permiten asociar a una variable aleatoria una distribución de probabilidad que modele su comportamiento, y de esta forma dar solución a problemas relacionados con estimación, decisión y predicción, de eventos o valores en contextos condicionados por incertidumbre y error. La mayoría de los métodos estadísticos paramétricos se basan en la media muestral y se han construido bajo el supuesto de normalidad. El Teorema Central del Límite (TCL) es el argumento más fuerte que sustenta todas estas teorías.

La aplicación del TCL a la distribución de la media muestral permite sacar diferentes conclusiones sobre alguna población. Dicha aplicación lleva a la utilización de la distribución normal, quizás, la más importante o una de las más relevantes en el campo de la estadística. Existen distribuciones como la normal que son simétricas y permiten el cumplimiento del TCL hasta en tamaños de muestras muy pequeños. Pero también hay otras distribuciones como la Gamma que son muy sesgadas y por ende tienen unos coeficientes de asimetría grandes. En la información existente sobre este teorema se encuentra una la regla empírica [2], que generaliza y establece que un tamaño de muestra mínimo de 30 permite el uso del TCL y además, la escogencia de este valor es independiente de la distribución sobre la cual se hizo el muestreo. Con este trabajo se pretende verificar si esta regla empírica es o no general, e indagar si existe alguna dependencia de la forma que tenga la distribución de donde se realiza el muestreo con el tamaño de muestra. Además, sugiere los tamaños mínimos de muestras de acuerdo a los coeficientes de asimetría, que permitan el uso del teorema sobre poblaciones que tengan distribuciones continuas como la Exponencial, Gamma y Ji cuadrada. De esta manera se tendrá una regla empírica, objetivo general de esta investigación.

Fecha de Recepción: 14 de Noviembre de 2014
Fecha de Aceptación: 30 de septiembre de 2017

II. METODOLOGÍA

Para la realización de este trabajo se utilizaron diferentes distribuciones de probabilidades continuas y de cada distribución se simularon 10000 datos con sus respectivos parámetros que permitieran obtener un valor específico de asimetría. Para la simulación de estas poblaciones se utilizó el software estadístico Infostat versión 2011 I.

La tabla 1 muestra las poblaciones simuladas, los parámetros y sus correspondientes valores de asimetría. Además se simularon siete poblaciones Gamma e igual cantidad de poblaciones Beta. El propósito de la simulación de estas últimas poblaciones fue determinar si existe alguna influencia, además de la asimetría, del tipo de población de la cual se hace el muestreo para que el TCL se cumpla, es decir, comparar valores de asimetría similares de poblaciones diferentes.

Distribución	Parámetros	Asimetría
Beta	α=3,5 ; β =5	0,20
F	u=200 ; v=200	0,40
Weibull	a=0,5 ; b=2	0,61
χ^2	v=7; λ=0	1,09
Beta	α =0,8; β =5	1,39
Gamma	α =2; β =0,3	1,54
Exponencial	λ=2	2,10

Tabla 1. Poblaciones con distribución de probabilidad continúa.

Posteriormente se seleccionó por medio de un muestreo aleatorio simple 250 muestras de cada población, de tamaños 10, 20, 30, 40, 50,…,480, 490, 500 y para cada una de ellas se calculó la media muestral, con esta información se obtuvo la distribución de la media muestral para cada uno de los tamaños de muestra. A cada una de las distribuciones de la media muestral se les aplicó la prueba de bondad de ajuste para distribución normal Shapiro-Wilk, por ser la prueba más potente [3][4]. Además se utilizó otra prueba potente como es la de Anderson-Darling y una de las menos potentes, la prueba de Lilliefors, con el fin de realizar un análisis comparativo entre las tres pruebas. El proceso de aplicar las tres pruebas a un tamaño específico de muestra se realizó 1000 veces y se calculó el valor p en cada una de las repeticiones del proceso. Con esta información se obtuvo la frecuencia relativa de las veces que se cumplió el Teorema del Límite Central, tomando como criterio de aceptación de la hipótesis nula valores p≥ 0,05. Todo este proceso de simulación se hizo por medio de un algoritmo en el software estadístico R versión 3.1.0 y sus librerías nortest y moments.

Se tomó como criterio para el análisis del cumplimiento del TCL un porcentaje del 90%, valor que se estimó de los gráficos de dispersión de los tamaños de muestra contra los porcentajes de aceptación según las tres pruebas de bondad de ajuste.

III. ANÁLISIS DE RESULTADOS

A continuación se realiza el análisis de los resultados obtenidos en la investigación de acuerdo a la secuencia planteada en la metodología. Las figuras 1, 2, 3 y 4 corresponden a los gráficos de dispersión de los tamaños de muestra contra los porcentajes de aceptación para un valor de asimetría y de acuerdo a la aplicación de las tres pruebas de normalidad Shapiro-Wilk (S-W) , Anderson-Darling (A-D) y Lilliefors (LIL).

La figura 1 corresponde a una población relativamente simétrica y se observa de acuerdo a las tres pruebas de bondad de ajuste de normalidad, que con un tamaño mínimo de 10 en la muestra se logra más del 90 % de las veces, el cumplimiento del TCL. Inclusive se puede ver que hasta con tamaños de muestra inferiores a 10 se alcanza el mismo porcentaje. La figura 2 muestra un igual desempeño de las pruebas de Lilliefors y Anderson-Darling y sugieren un tamaño mínimo de muestra de 15. En esta última población la prueba de Shapiro-Wilk sugiere un tamaño mínimo de 20 en la muestra para alcanzar el criterio establecido en la investigación.

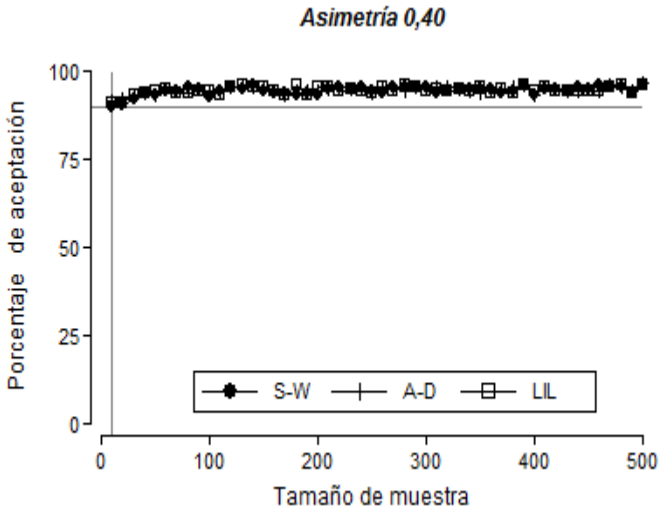

Figura 1. Distribución F (u=200; v=200).

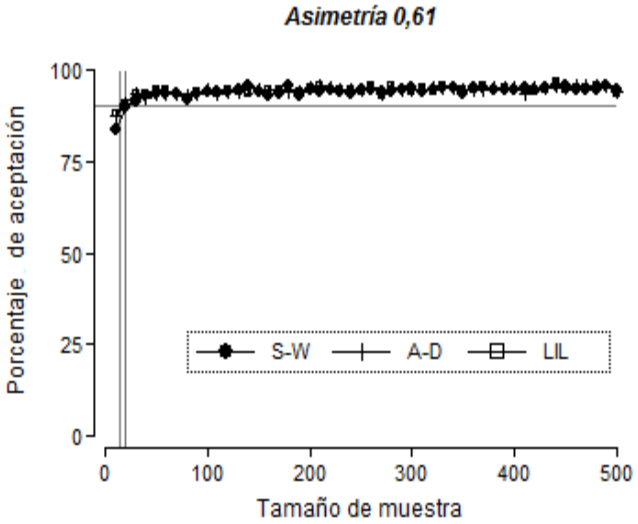

Figura 2. . Distribución Weibull (a=0,5; b=2).

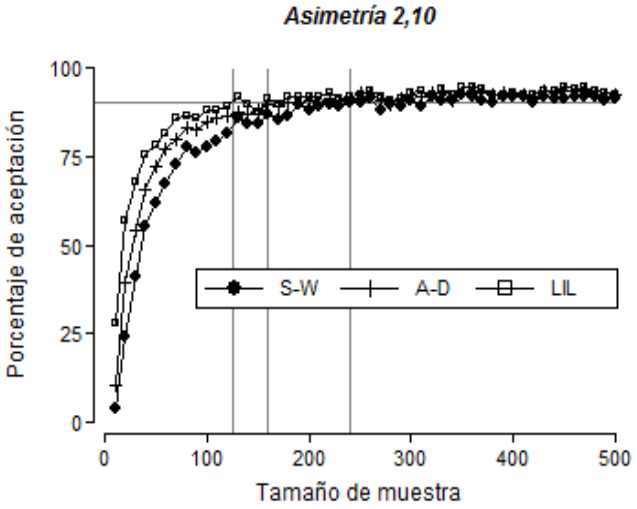

Figura 4. Distribución exponencial (λ=2)

Las figuras 3 y 4 corresponden a poblaciones con una alta asimetría. Los tamaños de muestra mínimos se van incrementando a medida que la asimetría también lo hace de acuerdo a los resultados de la aplicación de las tres pruebas de normalidad. Otro aspecto relevante es que la prueba Shapiro-Wilk al ser la prueba más potente [3][4], es decir, la que tiene más capacidad de detectar no normalidad en los datos, requiere tamaños de muestra mayores en comparación con las otras dos pruebas. Además la prueba de Lilliefors es la prueba menos potente y es la que evidencia tamaños de muestra menores, es decir, es la prueba más eficiente [3], por utilizar menos cantidad de unidades muestrales en comparación con las otras dos pruebas.

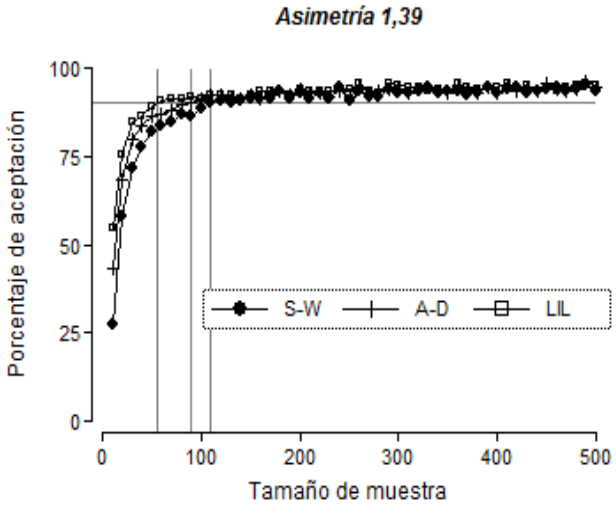

Figura 3. Distribución Beta (α =0,8; β =5).

La tabla 2 muestra las siete poblaciones continuas simuladas, los tamaños mínimos de muestra de acuerdo a las tres pruebas de bondad de ajuste de normalidad para lograr un 90% del cumplimiento del TCL y la asimetría de cada una.

Asimetría	Shapiro-Wilk	Anderson-Darling	Lilliefors
0,20	10	10	10
0,40	10	10	10
0,61	20	15	15
1,09	80	47	30
1,39	110	90	55
1,54	140	90	80
2,10	240	160	125

Tabla 2. Tamaños mínimos de muestra de acuerdo a cada prueba de normalidad.

Las tablas 3 y 4 muestran los resultados de las simulaciones de las siete poblaciones con distribución Gamma y las siete poblaciones con distribución Beta respectivamente.

Asimetría	Shapiro-Wilk	Anderson-Darling	Lilliefors
0,03	10	10	20
0,32	10	10	20
0,83	40	40	40
1,08	80	50	40
1,41	110	85	65
1,54	120	120	70
1,99	220	160	120

Tabla 3. Tamaños mínimos de muestra de acuerdo a cada prueba de normalidad simulados de una población *Gamma*.

Asimetría	Shapiro-Wilk	Anderson-Darling	Lilliefors
0,20	10	10	10
0,33	10	10	10
0,60	20	20	20
1,18	100	60	50
1,39	110	90	60
1,65	170	120	100
1,93	180	150	110

Tabla 4. Tamaños mínimos de muestra de acuerdo a cada prueba de normalidad simulados de una población *Beta*.

Los resultados de estas dos últimas tablas muestran al igual que la tabla 2 la relación existente entre la asimetría y los tamaños de muestra mínimos para lograr en un 90% el cumplimiento del TCL. A mayor asimetría en la población mayor deben ser los valores mínimos de la muestra para obtener el criterio establecido. Además si se comparan coeficientes de asimetría similares en las tres tablas y sus correspondientes tamaños de muestra se observa en la mayoría de los casos una similaridad entre estos valores, lo que conlleva a pensar que no importa el tipo de población continua de la cual se realice el respectivo muestreo.

IV. CONCLUSIONES

La asimetría de una población influye en el tamaño de la muestra que garantiza el cumplimiento del TCL. A medida que el coeficiente de asimetría aumenta, el tamaño mínimo de la muestra que garantiza el cumplimiento del Teorema también lo hace. Esto desmitifica el hecho de que un valor de 30 en el tamaño muestral es suficiente para el cumplimiento del teorema.

Si se realiza un muestreo de poblaciones continuas con diferentes distribuciones de probabilidades y coeficientes de asimetría similares, los tamaños mínimos de muestra para el cumplimiento del TCL también son similares, es decir, no importa el tipo de población.

La prueba de bondad de ajuste de normalidad de Lilliefors se comportó como la más eficiente en cada una de las poblaciones donde se aplicó, al necesitar menos unidades muestrales para alcanzar el criterio establecido que garantiza el cumplimiento del TCL.

Se pueden sugerir tamaños mínimos de muestra para ciertos intervalos de coeficientes de asimetría, por su similaridad en los resultados. Sin embargo los valores no muy similares y mayores que los demás se tomaron como los valores mínimos en cada intervalo. Además estos valores se tomaron de acuerdo a la prueba de Shapiro-Wilk por ser la más potente de las tres pruebas utilizadas en esta investigación y así disminuir la probabilidad de cometer un error tipo II. Para poblaciones con coeficientes de asimetría no superiores a 0,5 el cumplimiento del TCL con un criterio del 90% se logra con un tamaño mínimo de muestra de 10. Poblaciones con coeficientes de asimetría superiores a 0,5 e inferiores a 1,0 requieren aproximadamente un tamaño mínimo de muestra de 80. Para poblaciones con coeficientes de asimetría entre 1,0 y 1,5 requieren un tamaño mínimo de muestra de 110. Poblaciones con coeficientes de asimetría entre 1,5 y 2,0 requieren un tamaño mínimo de muestra de 240. Estos resultados se muestran en la siguiente tabla.

Intervalo coeficientes de asimetría	Tamaño mínimo de muestra de acuerdo a la prueba Shapiro-Wilk
[0,0; 0,5)	10
[0,5; 1,0)	80
[1,0; 1,5)	110
[1,5; 2,0)	240

Tabla 5. Tamaños mínimos de muestra para intervalos de asimetría.

RECOMENDACIONES

Las investigaciones que sirvieron como fundamento teórico para la realización de este artículo muestran cuales son las pruebas de bondad de ajuste de normalidad más potentes y eficientes, pero las simulaciones realizadas sólo fueron con poblaciones continuas. Se sugiere realizar una investigación donde se analice la potencia y eficiencia de estas pruebas utilizando poblaciones discretas.

REFERENCIAS

[1] D. A. Lenis, J. A. Ospina, and A. Barrientos, “Estudio exploratorio de las propiedades asintóticas del teorema central del límite asociadas a la distribución del estimador media muestral,” Revista Heurística, 14,27, pp. 27-34, 2007. Recuperado de http://hdl.handle.net/10893/6113.

[2] J. L. Devore, *Probabilidad y estadística para ingeniería y ciencias*, sexta edición, Thomson, 2006, p. 240.

[3] E. Zuluaga, J. Millán y J. Mosquera, “Análisis comparativo del desempeño de algunas pruebas de normalidad bajo diferentes escenarios de simulación,” Revista Heurística, 15, 13, pp. 13-21, 2008. Recuperado de http://hdl.handle.net/10893/6123.

[4] N. Razali and Y. Wah, "Power Comparison of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors
and Anderson Darling tests," Journal of Statistical Modeling and Analytics, Selangor, Malaysia, (2011).

[5] J. H. Mayorga, *Inferencia estadística*, Universidad Nacional de Colombia, Bogotá, D.C., Unibiblos, 2003, p. 15.

[6] E. González, "Pruebas de bondad de ajuste para distribuciones estables," Tesis doctoral. Colegio de Posgraduados Campus Montecillo, Montecillo, Texcoco, México, 2007.

[7] W. Mendenhall, D. Wackerly and R. Scheaffer, *Estadística matemática con aplicaciones*, segunda edición. Grupo editorial iberoamericano, U.S.A., 1994, p. 298.