ESTUDIO DE POTENCIA DE PRUEBAS DE NORMALIDAD USANDO DISTRIBUCIONES DESCONOCIDAS CON DISTINTOS NIVELES DE NO NORMALIDAD.

STUDY OF THE POWER OF TEST FOR NORMALITY USING UNKNOWN DISTRIBUTIONS WITH DIFFERENT LEVELS OF NON NORMALITY.

[1]Pablo Flores Muñoz*, [2]Laura Muñoz Escobar, [1]Tania Sánchez Acalo

[1]Escuela Superior Politécnica de Chimborazo, Facultad de Ciencias, Grupo de Investigación en Ciencia de Datos

[2]Universidad Nacional de Chimborazo, Facultad de Ciencias de la Educación, Humanas y Tecnologías

* p_flores@espoch.edu.ec

Resumen

La mayoría de pruebas de hipótesis paramétricas están sujetas al cumplimiento de normalidad. Debido a la gran variedad de opciones para contrastar este supuesto, se considera como mejor alternativa el test que presente una mayor potencia. Investigaciones preliminares estiman este valor a partir de muestras provenientes de distribuciones no normales conocidas pero cuyo alejamiento o contaminación respecto a la normalidad se desconoce. En el presente estudio seleccionamos siete pruebas que además de ser las más comunes parecen ser las mejores. Mediante un proceso de simulación, estimamos la potencia de cada una de ellas usando muestras provenientes de distribuciones desconocidas, pero con un alejamiento medible de la normalidad. Parece ser que el test de Shapiro – Wilk es la mejor opción, su potencia es muy elevada, pero solo para muestras no normales grandes y alejamientos fuertes. Para distribuciones con alejamientos débiles y muestras pequeñas parece ser que ninguna de las pruebas tradicionales en estudio es buena. Se discute un posible mal planteamiento de estos test y su incidencia en los resultados obtenidos. Finalmente se introduce la posibilidad de incluir pruebas basadas en el enfoque de equivalencia, las cuales quizás podrían resultar mejores que las pruebas estudiadas.

Palabras claves: Pruebas de normalidad, Potencia, Coeficientes de Fleishman, Equivalencia, Simulación.

Abstract

Most parametric tests are subject to normality. There are forty different tests to prove this assumption. Preliminary researches to determine the best tests are based on the estimation of their power using samples from known non-normal distributions but whose distance or contamination from normality is unknown. In the present study, we selected seven better and more known tests. Through a simulation process, we estimate the power of each one using samples from unknown distributions but with a measurable distance from normality. It seems that Shapiro - Wilk test is the best option, its power is very high, but only for large non-normal samples and strong distances. For distributions with weak distances and small samples it seems that none of the traditional tests are good. We discuss a possible poor approach to these tests and their impact on the results obtained. Finally, the possibility of including a hypothesis test based on the equivalence approach is analyzed; perhaps this option is better than the traditional tests introduced.

keywords: Normality test, Power, Fleishman Coefficients, Equivalence, Simulation.

Fecha de recepción: 27-12-2018 **Fecha de aceptación:** 25-01-2019

I. INTRODUCCIÓN

La mayoría de técnicas utilizadas para realizar inferencia estadística (pruebas de hipótesis, análisis de regresión, modelos de pronósticos, etc.) se basan en modelos paramétricos, los cuales a su vez están sujetos principalmente al cumplimiento del supuesto de normalidad. En la literatura estadística, varias pruebas de hipótesis se han propuesto con la finalidad de probar si un grupo de datos provienen o no de esta distribución teórica, de hecho, los trabajos de revisión reportan aproximadamente 40 test diferentes desarrollados con este fin (1,2,3).

Ante esta diversa gama de opciones, suponemos que la interrogante más frecuente con la que se encuentra un investigador que usa instrumentos estadísticos paramétricos tiene que ver con cuál es la mejor alternativa para probar el supuesto de normalidad en sus datos. En este sentido, múltiples investigaciones se han desarrollado con el fin de establecer un criterio de comparación que permita determinar objetivamente cuál opción es la más adecuada (4,5,6,7).

El criterio de comparación que usan estas investigaciones se basa en estimar mediante un proceso de simulación la potencia que tienen las distintas pruebas con el fin de recomendar aquellas cuya potencia resulte ser la más alta, es decir aquella prueba que maximice la probabilidad de rechazar la hipótesis nula de existencia de normalidad cuando realmente esta es falsa. Para asegurar esta condición de no normalidad, los trabajos mencionados usan muestras provenientes de distribuciones conocidas (gamma, exponencial, t – Student, uniforme, etc.)

En (6), se estimó la potencia de 33 pruebas de normalidad para distintos tamaños muestrales (n=25,50,100) y niveles de significancia (α=0.05,0.10), además con el fin de asegurar la condición de no normalidad en las muestras generadas utilizó distintas distribuciones no normales simétricas (Beta, Cauchy, Laplace, Logística, t – Student), asimétricas (chi - cuadrado, Gamma, Gumbelt, Log - Normal, Weibull) y normales modificadas (normal truncada, normal contaminada, normal mezclada y normal con atípicos).
La potencia estimada en todos estos casos no se altera significativamente cuando el nivel de significancia es diferente y no se pudo definir alguna prueba en específico como más potente, puesto que existen varias opciones diferentes que dependen de la naturaleza de la no normalidad. Otro estudio más detallado que el anterior, el cual usa las mismas distribuciones teóricas para simular muestras no normales, determinó que para muestras simétricas, Shapiro – Wilk junto con Jarque – Bera son los test más potentes, mientras que paras muestras asimétricas Shapiro – Wilk y Anderson – Darling parecen ser la mejor opción (7). En tra investigación, usando tamaños muestrales (n=20,30,40,50), se determinó que si se generan muestras a parir de una Distribución Beta, el test más potente es Anderson – Darling, si las muestras son generadas por una Gamma o por una Log – normal, el test más potente es Jarque – Bera (5).

Todos estos estudios, tienen en común que para estimar la potencia generan muestras de distribuciones de probabilidad que, aunque es verdad que no son normales, muchas de ellas tienden a serlo de manera asintótica, lo cual creemos puede estar afectando de alguna manera los resultados encontrados. Además, pensamos que en un análisis real no necesariamente los datos disponibles deben ajustarse a alguna distribución no normal conocida como las presentadas en estudios previos, sino que en algunos casos (o quizás en la mayoría) tendrán una distribución totalmente desconocida. Aunque la generación de muestras desconocidas, a nuestro parecer es tan importante como aquellas que provienen de distribuciones conocidas, estas no se han presentado en ninguna de las investigaciones antes mencionadas. Es así que proponemos el siguiente estudio, donde mediante un proceso de simulación estocástica estimamos la potencia de siete pruebas de normalidad, usando el sistema de Fleishman (8) para la creación de muestras de distribución desconocida (pero que se sabe se encuentran en datos reales) con distintos niveles de alejamiento o contaminación de la normal.

II. MATERIALES Y MÉTODOS

REVISIÓN DE PRUEBAS DE NORMALIDAD

Tomando en cuenta los resultados de investigaciones previas y considerando los test de normalidad más comunes en la mayoría de libros y software estadístico, proponemos el estudio de potencia de las siguientes pruebas para contrastar las hipótesis:

H_0 : Los datos provienen de una distribución normal

H_1 : Los datos no provienen de una distribución normal

Prueba de Kolmogorov – Smirnov (9; 10)

Sea $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ una muestra ordenada aleatoria, con función de distribución acumulada $F\left(x_{(i)}\right)$, con $1 \le (i) \le n$, y sea Z_i la distribución de probabilidad acumulada de una distribución normal estándar. El estadístico de prueba para este test viene dado por:

$$D = \max\left(D^+, D^-\right)$$

Donde, $D^+ = \max\left\{F\left(x_{(i)}\right) - Z_i\right\}$ y $D^- = \max\left\{Z_i - F'\left(x_{(i)}\right)\right\}$. Se rechaza H_0. cuando $D \ge d$ siendo d los puntos críticos que se encuentran en la tabla tabulada (9).

Test de Lilliefors (11)

Esta prueba es una modificación de Kolmogorov-Smirnov. Se sabe de antemano que KS es apropiada cuando se conoce los parámetros de la distribución hipotética (normal), sin embargo a veces o casi siempre es difícil conocer estos valores. En este sentido, el test Lilliefors usa las estimaciones de μ y σ en función de los datos de la muestra. El estadístico y los valores críticos siguen siendo los mismos que KS.

Test de Shapiro – Wilk (12).

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria, el estadístico de prueba para este test viene dado por:

$$W = \frac{b^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Con $b = \sum_{i=1}^{n} a_i \left[X_{(n-i-1)} - x_i\right]$ y $a_i = m'V^{-1}\left(m'V^{-1}m\right)^{-1/2}$. Donde, $\left[X_{(n-i-1)} - x_i\right]$ son las diferencias sucesivas que se obtiene al restar el primer valor al último valor, el segundo al antepenúltimo y así sucesivamente. a_i, son los coeficientes tabulados en la tabla de Shapiro, $m = \left(m_1, \ldots, m_n\right)$ son los valores medios del estadístico ordenado, de variables aleatorias independientes e idénticamente distribuidas, muestreadas de distribuciones normales. V es la matriz de covarianzas de ese estadístico de orden.

Se rechazará H_0 si $\mathrm{W} \le \mathrm{W}_{\propto,\mathrm{n}}$, donde $\mathrm{W}_{\propto,\mathrm{n}}$ son los puntos críticos tabulados en (12).

Test de Anderson – Darling (13)

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de tamaño mayor que 5, el estadístico de prueba para este test viene dado por:

$$A^2 = -\frac{\sum_{i=1}^{n}(2i-1)\left(\ln\left(Z_i\right) + \ln\left(Z_{n+1-i}\right)\right)}{n} - n$$

Donde Z_i es el cuantil de una distribución de probabilidad acumulada de una distribución normal estándar. De acuerdo al nivel de significancia y tamaño de muestra, rechazamos la hipótesis de normalidad cuando A^2 es mayor que $A^2_{c,\alpha}$,donde $A^2_{c,\alpha}$ son los puntos críticos tabulados en (14).

Test de Jarque – Bera (15)

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria, el estadístico de prueba para este test viene dado por:

$$JB = n\left(\frac{S^2}{6} + \frac{(k-3)^2}{24}\right)$$

Donde, S representa el coeficiente de asimetría y k el coeficiente de curtosis. La hipótesis nula de normalidad se rechaza cuando $JB > JB_{c,\alpha}$, donde $JB_{c,\alpha}$ es el valor crítico de una distribución chi – cuadrado, que deja a la derecha un área de α con 2 grados de libertad.

Test de Bondad de Ajuste χ^2 (16)

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño mayor que 5, el estadístico de prueba para este test viene dado por:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Donde, k es el número de clases existentes, O_i son las frecuencias observadas en cada clase y E_i son las frecuencias esperadas. La hipótesis nula de normalidad se rechaza cuando $\chi^2 > \chi_a^2$, donde χ_a^2 es el valor crítico de una distribución chi cuadrado con $k-1$ grados de libertad que deja a la derecha un área de α.

Test de Cramer – Von Mises (17).

Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de tamaño mayor que 5, el estadístico de prueba para este test viene dado por:

$$W = \frac{1}{12n} + \sum_{i=1}^{n} \left(p_i - \frac{2i-1}{2n} \right)^2$$

Donde $p_i = \phi \frac{x_i - \overline{x}}{s}$, siendo ϕ la función de distribución acumulada de una distribución normal estándar y además $\overline{x}$ y s representan respectivamente la media y desviación estándar muestral. Los valores críticos son calculados a partir del estadístico modificado $Z = W\left(1 + \frac{0.5}{n}\right)$, de acuerdo a la Tabla 4.9 presentada en (2).

OBTENCIÓN DE MUESTRAS PROVENIENTES DE DISTRIBUCIONES DESCONOCIDAS.

Allen Fleishman, en el año 1978, desarrolló un método mediante el cual se pueden generar muestras aleatorias no normales cuya distribución se desconoce (8). Partiendo de una normal estándar Z ($Z \sim N(\mu = 0, \sigma = 1)$), el autor demostró que una transformación lineal de esta variable, de la forma $Y = a + bZ + cZ^2 + dZ^3$, sigue una distribución totalmente desconocida de parámetros $(\mu = 0, \sigma = 1, \gamma_1, \gamma_2)$, donde γ_1 y γ_2 representan respectivamente los coeficientes poblacionales de simetría y curtosis, los cuales a su vez definen los grados de separación o contaminación de la normal.

En una posterior investigación (18), donde se recopiló 693 variables con distribución desconocida, provenientes de bases de datos reales (de las cuales 193 correspondieron a poblaciones diferentes), se estimó los coeficientes γ_1 de simetría y γ_2 de curtosis basados en el tercer y cuarto momento no central respectivamente. Los resultados mostraron un rango de valores para γ_1 comprendido entre -2.49 y 2.33, mientras que para γ_2 los valores encontrados fluctuaban en un intervalo de -1.92 a 7.41. Establecer puntos de corte para valores combinados de estos coeficientes permitió determinar diferentes niveles de alejamiento o contaminación de la normalidad.

Con la ayuda de funciones existentes en el software estadístico R (19), en el presente estudio utilizamos la transformación de Fleishman para generar muestras provenientes de distribuciones desconocidas. Los valores para los coeficientes (a,b,c,d) son determinados de acuerdo a los valores γ_1 y γ_2 propuestos a partir de las investigaciones desarrolladas en (18,20). Estos coeficientes se obtuvieron con la función "fleishman.coef" del paquete "BinNonNor" (21). Los valores de simetría, curtosis y coeficientes de Fleishman calculados con sus respectivos niveles de alejamiento de la normalidad, se muestran en la Tabla 1.

Nivel de Contaminación	Simetría	Curtosis	Coeficientes de Fleishman
Leve	0.25	0.7	(-0.037, 0.933, 0.037, 0.021)
Moderada	0.7	1	(-0.119, 0.956, 0.119, 0.0098)

Alta	1.3	2	(-0.249, 0.984, 0.249, -0.016)
Severa	2	6	(-0.314, 0.826, 0.314, 0.023)

Tabla 1. Simetría, Curtosis y coeficientes de Fleishman usados para generar muestras no normales.

MÉTODO DE ESTIMACIÓN DE LA POTENCIA.

La potencia fue estimada a partir de una función propia de R, la cual implementa un algoritmo de simulación de montecarlo para dicho fin. El algoritmo consiste en crear muestras no normales de tamaño **n=5,10,15,20,30,50,100,** a partir de la transformación de Fleishman (usando los coeficientes dados en la Tabla 1.), luego las siete técnicas descritas en la Sección II para probar la hipótesis de normalidad son aplicadas sobre estas muestras, finalmente nos interesará guardar en un vector, un valor lógico que indique "TRUE" si la hipótesis se rechaza o "FALSE" en caso contrario. Este proceso se repite **m=100000** veces y la potencia queda estimada como la proporción de rechazos de la hipótesis nula. Dado que en investigaciones previas se comprobó que los resultados de la potencia son independientes del nivel de significancia, el valor α usado en cada una de las pruebas fue de **0.05.**

III. RESULTADOS

La Figura 1, resume la potencia determinada para cada una de las pruebas de normalidad para diferentes alejamientos de la distribución y distinto tamaño muestral.

De manera general, en todos los escenarios analizados se puede observar un comportamiento casi invariante de las potencias calculadas. Esta potencia se incrementa conforme aumenta el grado de alejamiento de la normal y el tamaño muestral crece, de este modo, por ejemplo para una contaminación leve y un tamaño muestral de 5, la potencia de todas las siete pruebas es demasiado baja (tiende al nivel de significancia 0.05), mientras que en el otro extremo, es decir cuando existe una contaminación severa y el tamaño muestral es 100, la potencia de todas las pruebas tiende a 1 (nivel de potencia más grande que puede alcanzar una prueba).

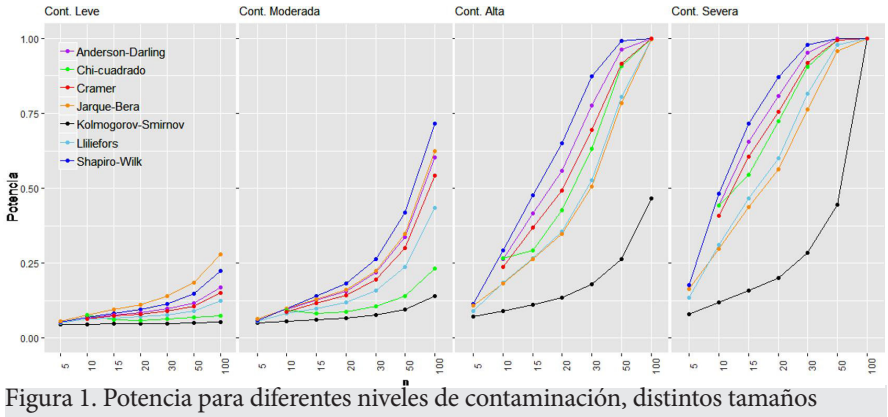

Figura 1. Potencia para diferentes niveles de contaminación, distintos tamaños muestrales y α=0.05

De manera específica, cuando la contaminación es leve, se puede observar que no muy distante de la Prueba de Shapiro – Wilk, el test de Jarque – Bera parece tener la potencia más alta independientemente del tamaño muestral usado, sin embargo, esta potencia decrece conforme el nivel de contaminación aumenta y a partir de un alejamiento moderado, la prueba de Shapiro Wilk parece ser la más potente de todas, por lo se recomienda su uso. Contrario a esto, independientemente del alejamiento y el tamaño muestral, la prueba de Kolmogorov – Smirnov resulta ser siempre la peor y se desaconseja su uso. Los demás test analizados (Anderson D, Chi – cuadrado, Cramer y LLilefors), siempre se encuentran entre los dos test más y menos potentes descritos.

IV. DISCUSIÓN

Los resultados obtenidos muestran evi-

dencia a favor del uso del test de Shapiro Wilk, el cual parece tener una mayor potencia respecto a los otros test analizados. Sin embargo, es muy importante hacer notar que, aunque este test es relativamente alto en todos los tamaños muestrales, en realidad en el intervalo [0, 1] donde toma valores, nunca supera un valor que podría ser aceptable (mínimo de 0.8) para muestras pequeñas. A partir de una contaminación alta, solo se observa una potencia elevada para el tamaño muestral n≥30. Esto parece ser razonable y coherente ya que solo estamos confirmando algo que en estadística suele ser muy común "todas las cosas mejoran cuando aumenta el tamaño muestral".

Llama la atención, lo poco (o casi nada) potentes que resultan ser todas las pruebas analizadas cuando existe una contaminación leve o moderada, en este sentido podemos concluir que los test de normalidad solo pueden ser buenos cuando el alejamiento de la distribución teórica es fuerte, mientras que, para muestras con alejamientos débiles, estas pruebas tradicionales están lejos de ser efectivas. A nuestro criterio, lo que podría estar ocurriendo es que las muestras generadas con alejamiento leve y moderado, presentan un grado de contaminación tan pequeño que podría considerarse despreciable o irrelevante como para ser consideradas no normales.

Si analizamos un poco más a fondo, la hipótesis nula de los test en estudio sugiere que los datos siguen una normalidad perfecta, al respecto George Box en 1979 dijo "En la vida real no existe una distribución perfectamente normal, sin embargo, con modelos que se sabe que son falsos, a menudo se puede derivar resultados que coinciden, con una aproximación útil a los que se encuentran en el mundo real" (22). Entonces, según este enunciado, no tiene sentido realizar test de hipótesis que prueben una perfecta normalidad (ya sabemos que esto no existe). A nuestro criterio y de acuerdo al enunciado de Box, tendría más sentido que en lugar de probar normalidad perfecta, se contraste una hipótesis que pruebe la existencia de una normalidad aproximada, es decir una normalidad que acepte desviaciones tan pequeñas que se puedan considerar despreciables o irrelevantes. Aunque es nuevo y no tan común, este enfoque para contrastar hipótesis, distinto al enfoque tradicional, estudiado en la mayoría de la literatura estadística existe y se denomina "pruebas de equivalencia" (23). Entendemos por equivalencia a una forma dilatada de una relación de identidad, la cual se induce al añadir en la hipótesis tradicional una zona de irrelevancia alrededor del correspondiente punto en el espacio paramétrico que denota perfecta normalidad, este enfoque conduce a diseñar un estudio que pretende demostrar ausencia de una diferencia relevante entre los efectos de dos distribuciones comparadas (una de ellas podría ser la normal), es decir equivalencia. En este sentido, resulta interesante la realización de investigaciones futuras sobre estimación de la potencia pero incorporando el criterio de equivalencia, de hecho existen ciertos estudios donde ya se demuestra la superioridad de este enfoque sobre el tradicional (24). Podría ser, que este criterio de equivalencia sea más potente, pero eso no lo sabremos hasta que se realicen las mediciones respectivas.

V. CONCLUSIONES

A diferencia de otras investigaciones, donde para asegurar la no existencia de normalidad teórica, se usa distribuciones no normales conocidas, en el presente trabajo se usó el sistema de Fleishman, el cual nos permitió crear muestras no normales con un grado medible del alejamiento o contaminación. Este nuevo aporte para estudiar la potencia de las pruebas de normalidad, permitió tener la seguridad de que las muestras provienen de una distribución desconocida, independientemente del tamaño muestral, lo cual podría influir en una normalidad asintótica. Además, de esta forma se pudo obtener grados de contaminación a partir de los cuales se presenta una potencia relativamente alta, lo cual permitió detectar grados de alejamientos que se pueden considerar irrelevantes.

Luego de comparar la potencia de siete pruebas de hipótesis para contrastar normalidad, hallamos que Shapiro – Wilk es el test más potente. Por tanto, esta prueba maximiza la probabilidad de cometer un acierto en el sentido de rechazar la hipótesis de normalidad perfecta dado que teóricamente esta es falsa. Sin embargo, la potencia de esta prueba es óptima siempre y cuando el tamaño muestral sea mayor que 30 (n≥30) y para contaminaciones de

normalidad alta y severa.

Para contaminaciones leves y moderadas, a pesar que la potencia del test de Shapiro – Wilk es la más alta, no se puede considerar que es óptima, puesto que en ningún caso supera al menos un valor razonable de 0.80. Parece ser que estas dos contaminaciones de la normalidad, son alejamientos que podrían considerarse tan irrelevantes que se podrían despreciar.

Un planteamiento hipotético de normalidad irrelevante en lugar de normalidad perfecta podría ser una mejor alternativa.

Referencias

1. Recent and classical tests for normality-a comparative study. Baringhau, L, Danschke, R y Henze, N. 1, 1989, Communications in Statistics-Simulation and Computation, Vol. 18.
2. D'Agostino, Ralph y Stephens, Michael. Goodness-of-fit techniques (Statistics, a series of textbooks and monographs. New York : Marcel Deker, 1986.
3. Tests of unvariate and multivariate normality. Mardia, Kanti. 1980, Handbook of statistics, Vol. 1.
4. Simulation based finite sample normality test in linear regressions. Dufour, Jean-Marie, y otros. 1998, Econometrics Journal, Vol. 1.
5. A comparison of various tests of normality. Yazici, Berna y Yolacan, Senay. 2, 2007, Journal of Statistical and Simulation, Vol. 77.
6. An empirical power comparison of univariate goodness of fit test for normality. Romao, Xavier, Delgado, Raimundo y Costa, Aníbal. 5, May de 2010, Journal of Statistical Computation and Simulation, Vol. 80.
7. Comparisons of various types of normality tests. Yap, B y Sim, C. 12, December de 2011, Journal of Statistical Computation and Simulation, Vol. 81.
8. A method for simulating non-normal distributions. Fleishman, Allen. 4, 1978, Psycometrika, Vol. 43.
9. Sulla determinazione empirica di una lgge di distribuzione. Kolmogorov, Andrey. 1933, Inst. Ital. Attuari, Giorn., Vol. 4.
10. Table for estimating the goodness of fit of empirical distributions. Smirnov, Nickolay. 2, 1948, The annals of mathematical statistics, Vol. 19.
11. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Lilliefors, Hubert. 318, 1967, Journal of the American statistical Association, Vol. 62.
12. An analysis of variance test for normality. Shapiro, SS y Wilk, MB. 3, 1965, Biometrika, Vol. 52.
13. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. Anderson, Theodore y Darling, Donald. 1952, The annals of mathematical statistics}, Vol. 23.
14. EDF statistics for goodness of fit and some comparisons. Stephens, Michael. 347, 1974, Journal of the American statistical Association, Vol. 69.
15. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Jarque, Carlos y Bera, Anil. 3, 1980, Economics Letters, Vol. 6.
16. Greenwood, Priscilla y Nikulin, Michael. A guide to chi-squared testing. 280. s.l. : John Wiley & Sons, 1996.
17. On the composition of elementary errors: First paper: Mathematical deductions. Cramér, Harald. 1928, Scandinavian Actuarial Journal, Vol. 1.
18. Skewness and kurtosis in real data samples. Blanca, María, y otros. 2013, Methodology.
19. R: A Language and Enviroment for Statistical Computing. Team, R Core. 2018, R Foundation For Statistical Computing. https://www.R-project.org.

20. Comparison of the procedures of Fleishman and Ramberg et al. for generating non-normal data in simulation studies. Bendayan, Rebecca, y otros. 1, 2014, Annals of Psychology, Vol. 30.
21. BinNonNor: Data Generation with Binary and Continuous Non-Normal Components. Inan, Gul y Demirtas, Hakan. 2018, R package version 1.4.
22. Robustness in the strategy of scientific model building. Box, George. 1979, Army res off Work Robustness, Vol. 1.
23. Wellek, Sthepan. Testing St atistical Hypothesis of equivalence and non Inferiority. 2010.
24. Heteroscedasticity Irrelevance when Testing Means Difference. Flores, Pablo y Ocaña, Jordi. 1, June de 2018, SORT, Vol. 42.
25. EDF statistics for goodness of fit and some comparisons. Stephens, Michael. 1974, Journal of the American statistical Association, págs. 730--737.