

DOI: <https://doi.org/10.56712/latam.v4i4.1242>

Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS

Histogram and normal distribution: Shapiro-Wilk and Kolmogorov Smirnov applied in SPSS

Héctor Alberto Luzuriaga Jaramillo

ha.luzuriaga@uta.edu.ec

<https://orcid.org/0000-0002-7844-1805>

Universidad Técnica de Ambato

Ambato – Ecuador

Carlos Alberto Espinosa Pinos

carlospinos@indoamerica.edu.ec

<https://orcid.org/0000-0002-7841-8090>

Universidad Indoamérica

Ambato – Ecuador

Alexander Fernando Haro Sarango

alexander.haro@iste.edu.ec

<https://orcid.org/0000-0001-7398-2760>

Instituto Superior Tecnológico España

Ambato – Ecuador

Hermel David Ortiz Román

hd.ortiz@uta.edu.ec

<https://orcid.org/0000-0003-3529-3571>

Universidad Técnica de Ambato

Ambato – Ecuador

Artículo recibido: 28 de septiembre del 2023. Aceptado para publicación: 13 de octubre de 2023.
Conflictos de Interés: Ninguno que declarar.

Resumen

En el ámbito de los estudios relacionados con la administración, la economía y la sociedad, los histogramas se pueden emplear con el propósito de visualizar la distribución de los datos, por otra parte, las distribuciones normales son frecuentemente utilizadas para modelar fenómenos que ocurren en la vida real para efectuar predicciones basadas en análisis estadísticos viables y sólidos. Por la relevancia, el estudio tiene como objetivo generar una guía para el análisis de la normalidad de una variable específica utilizando métodos estadísticos y gráficos aplicados en SPSS. El método aplicado en este estudio posibilitó un examen minucioso de una variable específica piloteada empleando enfoques tanto estadísticos como visuales. Los histogramas revelaron la disposición de las frecuencias y los esquemas de los datos, en contraste, los gráficos Q-Q ofrecieron una confrontación directa con una distribución normal de referencia. Las pruebas de Shapiro-Wilk y Kolmogorov-Smirnov resultaron en la obtención de valores "p" que se cotejaron con niveles predefinidos de significancia, lo cual contribuyó a la identificación de la conformidad de los datos con una distribución normal.

Palabras clave: estadística, histograma, datos, normalidad

Abstract

In the field of studies related to management, economics and society, histograms can be employed for the purpose of visualizing the distribution of data, on the other hand, normal distributions are frequently used to model phenomena occurring in real life to make predictions based on viable and robust statistical analysis. Because of the relevance, the study aims to generate a guide for the analysis of the normality of a specific variable using statistical methods and graphics applied in SPSS. The method applied in this study enabled a thorough examination of a specific piloted variable using both statistical and visual approaches. Histograms revealed the arrangement of frequencies and data patterns, in contrast, Q-Q plots offered a direct confrontation with a reference normal distribution. Shapiro-Wilk and Kolmogorov-Smirnov tests resulted in p-values that were checked against predefined levels of significance, which contributed to the identification of the conformity of the data with a normal distribution.

Keywords: statistics, histogram, data, normality

Todo el contenido de LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, publicados en este sitio está disponibles bajo Licencia Creative Commons . 

Como citar: Luzuriaga Jaramillo, H. A., Espinosa Pinos, C. A., Haro Sarango, A. F. & Oriz Román, H. D. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades* 4(4), 596– 607. <https://doi.org/10.56712/latam.v4i4.1242>

INTRODUCCIÓN

El histograma y la distribución normal son conceptos importantes en estadística y análisis de datos. Permiten comprender la distribución de frecuencias y la forma de los datos (Tauber, 2001). Para evaluar la normalidad de una distribución, se suelen utilizar pruebas estadísticas como la prueba de Shapiro-Wilk y la prueba de Kolmogorov-Smirnov. Estas pruebas ayudan a determinar si un conjunto de datos sigue una distribución normal o se desvía significativamente de ella (Pedrosa et al., 2015).

Los histogramas y las distribuciones normales se utilizan habitualmente en estudios administrativos, económicos y sociales. Un histograma es una representación gráfica de la distribución de un conjunto de datos, en la que los datos se dividen en intervalos y la frecuencia de cada intervalo se representa en el eje Y (Hederich Martínez, 2023). Una distribución normal, también conocida como distribución de Gauss, es una distribución de probabilidad simétrica y en forma de campana, con la mayoría de los datos agrupados en torno a la media (Tomás-Sábado, 2010).

En estudios administrativos, económicos y sociales, los histogramas pueden utilizarse para visualizar la distribución de datos, como los ingresos o las notas de los exámenes, y para identificar patrones o valores atípicos (Haro Sarango et al., 2023; Roldán Molina, 2019). Las distribuciones normales se utilizan a menudo para modelizar fenómenos del mundo real, como la distribución de alturas o pesos en una población, y para hacer predicciones basadas en análisis estadísticos (Guachimboza-Villalva et al., 2021; Pinkovetskaia et al., 2021). Además de los histogramas y las distribuciones normales, en los estudios administrativos, económicos y sociales suelen utilizarse otras herramientas estadísticas, como parámetros como la media y la desviación típica, y técnicas como los gráficos de control y el diseño experimental, un ejemplo, es el estudio ejecutado en el sector cooperativo de Altamirano et al. (2023).

La prueba de Shapiro-Wilk es una prueba paramétrica que mide la correlación entre los datos y las puntuaciones normales correspondientes. Calcula un estadístico de prueba (W) y lo compara con los valores críticos para determinar si los datos se desvían significativamente de la normalidad (Juárez Manayay, 2021). Por otro lado, la prueba de Kolmogorov-Smirnov es una prueba no paramétrica que compara la función de distribución empírica de una muestra con una distribución de referencia, como una distribución normal. Cuantifica la diferencia máxima (D) entre las dos funciones de distribución acumulativa (Platas García, 2021).

Ambas pruebas desempeñan un papel crucial en la evaluación de la normalidad de los datos y se utilizan a menudo en diversos campos, como la investigación, las finanzas y el control de calidad. Sin embargo, es importante señalar que los resultados de estas pruebas no siempre coinciden con las evaluaciones visuales, como los histogramas. Las discrepancias pueden deberse a factores como el tamaño de la muestra, la forma de la distribución subyacente y la sensibilidad de las pruebas.

Entrando en contexto, el histograma y la distribución normal son herramientas fundamentales en el análisis de datos. La prueba de Shapiro-Wilk y la prueba de Kolmogorov-Smirnov son pruebas estadísticas utilizadas para evaluar la normalidad de una distribución. Combinando representaciones visuales como los histogramas con estas pruebas, los investigadores pueden obtener una comprensión exhaustiva de la distribución de sus datos y tomar decisiones fundamentadas basándose en los resultados. Por aquello, el estudio plantea como base generar una guía para el análisis de la normalidad de una variable específica utilizando métodos estadísticos y gráficos aplicados en SPSS.

MÉTODO

Diseño del Estudio

Para lograr el propósito del estudio, se aplicaron las pruebas de normalidad de Shapiro-Wilk y Kolmogorov-Smirnov, así como la creación de histogramas y gráficos Q-Q en el software estadístico SPSS (Statistical Package for the Social Sciences).

Muestra

La muestra utilizada en este estudio consiste en una selección de una base de datos piloto, resultando en un tamaño de muestra de 100 datos respectivamente. Estos datos fueron elaborados de forma proactiva para divisar el proceso de ejecución y como se da lectura a los resultados.

Procedimiento

Preparación de los datos: Los datos fueron importados en formato electrónico al software SPSS para su posterior análisis.

Creación de Histogramas: Se generaron histogramas para visualizar la distribución de la variable de interés. Los histogramas proporcionan una representación gráfica de la frecuencia de los valores en intervalos específicos, lo que ayuda a evaluar la forma general de la distribución.

Prueba de Shapiro-Wilk: Se realizó la prueba de normalidad de Shapiro-Wilk para verificar si los datos se ajustan a una distribución normal. El p-valor resultante de esta prueba se comparó con un nivel de significancia predefinido para determinar si se rechaza o no la hipótesis nula de normalidad.

Prueba de Kolmogorov-Smirnov: Se aplicó la prueba de normalidad de Kolmogorov-Smirnov para evaluar la normalidad de los datos. Al igual que en la prueba de Shapiro-Wilk, el p-valor obtenido se comparó con un nivel de significancia para tomar una decisión sobre la normalidad.

Gráficos Q-Q (Quantile-Quantile): Se generaron gráficos Q-Q para comparar los cuantiles de los datos con los cuantiles esperados de una distribución normal. Estos gráficos proporcionan una visualización adicional de la normalidad y pueden revelar desviaciones en la distribución de los datos.

Análisis de Datos

El análisis de los resultados se basó en la interpretación de los histogramas, los valores de p-valor obtenidos de las pruebas de normalidad y la observación de los gráficos Q-Q. Se compararon los resultados de ambas pruebas de normalidad y se consideró la validez de la hipótesis nula de normalidad en función de los p-valores obtenidos.

RESULTADOS

En primera instancia cargamos los datos en el sistema SPSS, recordemos que, para el cálculo de los test de normalidad Shapiro-Wilk y el test de Kolmogorov-Smirnov son pruebas estadísticas utilizadas para evaluar si una muestra de datos sigue una distribución normal (gaussiana) o no. Sin embargo, estas pruebas están diseñadas principalmente para datos continuos, no categóricos. Se detalla la visualización de la data en el software:

Figura 1

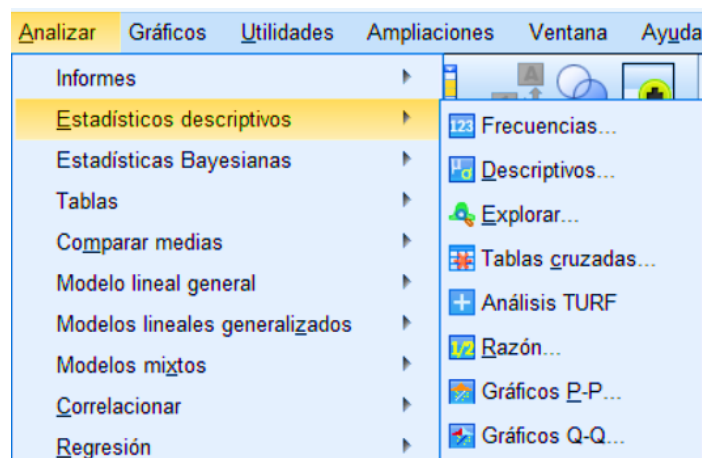
Carga de los datos

Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
Identificador	Númerico	8	1	ID	Ninguna	Ninguna	8	Centro	Escala	Entrada
Edad	Númerico	8	1	Edad (número de años)	Ninguna	Ninguna	8	Centro	Escala	Entrada
Peso	Númerico	8	1	Peso (Kilogramos)	Ninguna	Ninguna	8	Centro	Escala	Entrada

Posteriormente, en la barra de opciones nos dirigimos en >Analizar< luego >Estadísticos descriptivos< finalmente en la opción >Explorar<.

Figura 2

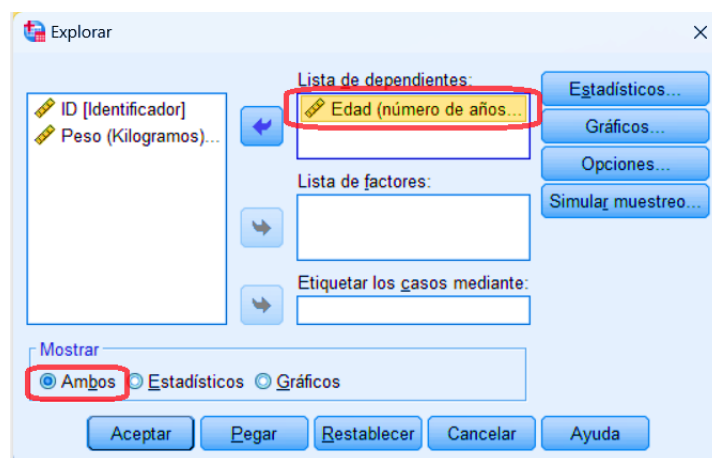
Direccionando a la opción



Consecutivamente, nos aparecerá la siguiente ventana en la opción >Explorar<, donde, en la >Lista de independientes< procedemos a insertar la variable cuantitativa continua y, manteniendo la opción >Ambos< en la parte inferior para un mejor detalle.

Figura 3

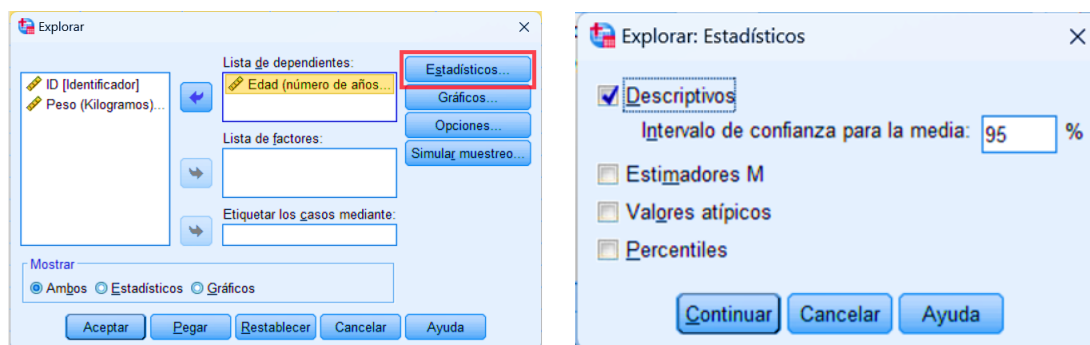
Ventana explorar



Consecutivamente, explicaremos los distintos botones que nos ofrece la opción >Explorar<, en el primero denominado >Estadísticos< donde, por preferencia, dejamos las opciones que están marcadas y >Continuar<.

Figura 4

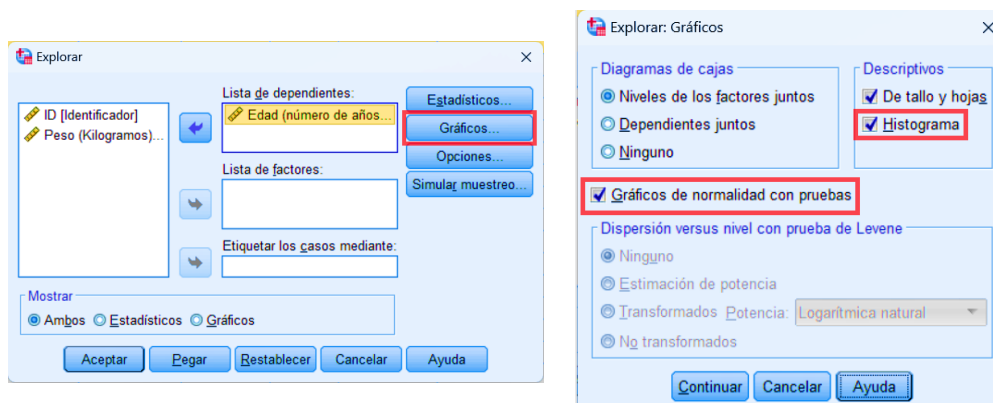
Ventana explorar: Estadísticos



Inmediatamente, explicaremos en el siguiente botón que nos ofrece la opción >Explorar<, en el denominado >Gráficos< donde, por preferencia, dejamos las opciones que están marcadas y, adicional, seleccionamos la casilla >Histograma< y >Gráficos de normalidad con pruebas<, >Continuar< y, en la primera ventana, >Aceptar<.

Figura 5

Ventana explorar: Gráficos



Luego de haber dado click en la opción >Aceptar< se despliega todo el modelo tal y con las especificaciones mencionadas. En primera instancia saldrá el >Resumen de procesamiento de casos< en donde podemos visualizar la cantidad de datos procesados en el análisis y la cantidad de datos omitidos por algún error encontrado, un ejemplo podría ser cuadros vacíos o datos fuera de las especificaciones adecuadas.

Tabla 1

Resumen de procesamiento de casos

Resumen de procesamiento de casos						
	Válido		Casos Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Edad (número de años)	100	100,0%	0	0,0%	100	100,0%

Consecutivamente, se genera una tabla de >Descriptivos< donde se genera un resumen del comportamiento de la variable en distintas especificidades como, medidas de tendencia central, dispersión o de forma. Se procede a dar lectura a los datos de la tabla y su forma de interpretación:

Media: La media aritmética de la variable "Edad", que es la suma de todas las edades dividida por el número de observaciones. En este caso, la media es de aproximadamente 41,39 años.

95% Intervalo de Confianza para la Media: Un intervalo estimado en el que se espera que esté la verdadera media poblacional con un 95% de confianza. En este caso, se estima que la verdadera media de la población de edades se encuentra entre aproximadamente 39,513 y 43,267 años.

Media Recortada al 5%: La media calculada después de eliminar el 5% de los valores extremos más altos y bajos. En este caso, tiene aproximadamente 41,522 años.

Mediana: El valor medio cuando los datos se ordenan en orden ascendente. En este caso, la mediana es 41 años, lo que significa que al menos la mitad de las edades son iguales o menores que 41, y al menos la mitad son iguales o mayores que 41.

Varianza: Una medida de cuánto se dispersan los datos alrededor de la media. En este caso, la varianza es aproximadamente 89,513.

Desviación Estándar: La raíz cuadrada de la varianza, que representa la dispersión promedio de los datos alrededor de la media. Aquí, tiene aproximadamente 9,4611 años.

Mínimo: El valor más bajo en la variable "Edad". En este caso, es 12 años.

Máximo: El valor más alto en la variable "Edad". Aquí, tiene 61 años.

Rango: La diferencia entre el valor máximo y el valor mínimo. En este caso, tiene 49 años.

Rango Intercuartil: La diferencia entre el tercer cuartil y el primer cuartil. Representa el rango que abarca el 50% central de los datos. En este escenario, es 11,8 años.

Asimetría: Mide la asimetría de la distribución. Un valor negativo indica una asimetría hacia la izquierda (cola larga a la izquierda). En este conjunto de datos, es aproximadamente -0,197, lo que sugiere una leve asimetría hacia la izquierda.

Nota: sí el valor de asimetría es (0) indica la existencia de una distribución simétrica, sí es negativa (-) indica una cola larga a la izquierda, sí resulta positivo (+) es una cola larga a la derecha.

Curtosis: Mide la forma de la distribución y cuánto se desvía de la distribución normal. Un valor positivo indica una distribución más puntiaguda que la normal. En este cálculo, es aproximadamente 0,179, indicando una curtosis relativamente baja y cercana a la normalidad.

Nota: si el valor de curtosis es (>0) es distribución leptocúrtica, colas más pesadas y puntas más agudas que la distribución normal (mesocúrtica). Las observaciones se agrupan más cerca de la media y las colas de la distribución son más pronunciadas. (0) distribución mesocúrtica (o normal), tiene una curtosis igual a cero, lo que indica que las colas y las puntas son similares a las de una distribución normal. (<0) distribución platocúrtica, las observaciones tienden a ser menos extremas y la distribución puede parecer más dispersa en comparación con una distribución normal.

Tabla 2

Descriptivos

Descriptivos			Estadístico	Error estándar
Edad (número de años)	Media		41,390	,9461
	95% de intervalo de confianza para la media	Límite inferior	39,513	
		Límite superior	43,267	
	Media recortada al 5%		41,522	
	Mediana		41,000	
	Varianza		89,513	
	Desviación estándar		9,4611	
	Mínimo		12,0	
	Máximo		61,0	
	Rango		49,0	
	Rango intercuartil		11,8	
	Asimetría		-,197	,241
	Curtosis		,179	,478

Por consiguiente, damos interpretación a las pruebas de normalidad de Shapiro-Wilk y Kolmogorov Smirnov de manera individual:

Prueba de Kolmogorov-Smirnov

El p-valor (0,200) es mayor que un nivel de significancia típico de 0,05. Esto sugiere que no tienes suficiente evidencia para rechazar la hipótesis nula de que los datos siguen una distribución normal. En otras palabras, los datos de edad en tu conjunto de datos no muestran desviaciones significativas de la normalidad según la prueba de Kolmogorov-Smirnov.

Prueba de Shapiro-Wilk

Al igual que en la prueba de Kolmogorov-Smirnov, el p-valor (0,598) es mayor que un nivel de significancia comúnmente utilizado, como 0,05. Esto también indica que no hay suficiente evidencia para rechazar la hipótesis nula de normalidad según la prueba de Shapiro-Wilk.

En sinopsis, según los resultados de ambas pruebas, no se tiene suficiente evidencia para afirmar que los datos de edad en tu conjunto de datos se desvían significativamente de una distribución normal. Esto sugiere que los datos podrían considerarse razonablemente aproximados a una distribución normal en términos de su forma.

Tabla 3

Pruebas de normalidad

Pruebas de normalidad						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad (número de años)	,064	100	,200 [*]	,989	100	,598

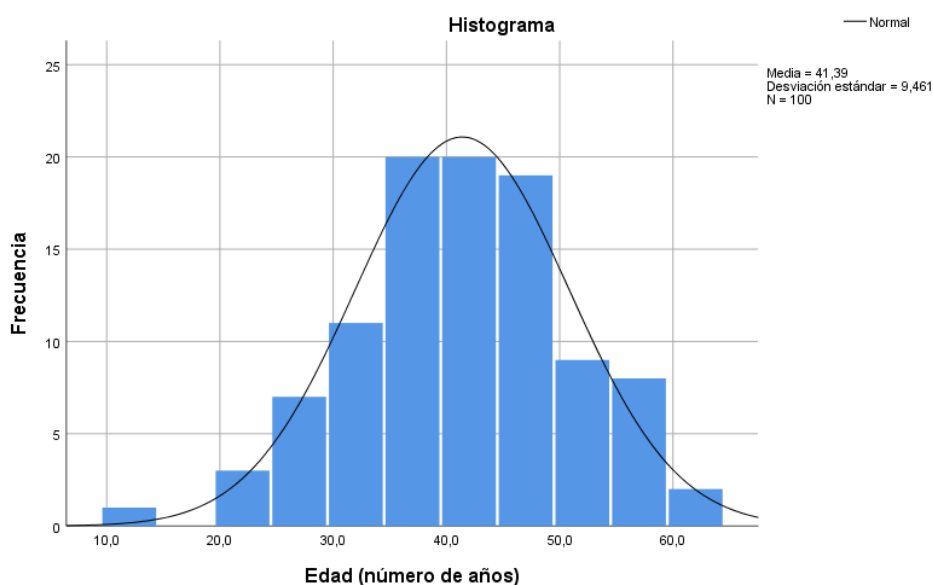
*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Finalmente, se generarán diversos gráficos para generar una visualización de la información, en este caso el histograma el eje horizontal representa los valores de la variable y se divide en intervalos (bins). El eje vertical muestra la frecuencia o el recuento de observaciones en cada intervalo. El histograma permite visualizar la forma de la distribución de tus datos, si son simétricos, sesgados, bimodales, etc. En este caso, podemos visualizar una línea negra que superpone al gráfico de barras, esto implica un ejemplo de cómo debería estar la distribución de las barras para que sea un gráfico normal.

Gráfico 1

Histograma



Adicional, se generan tres gráficos más:

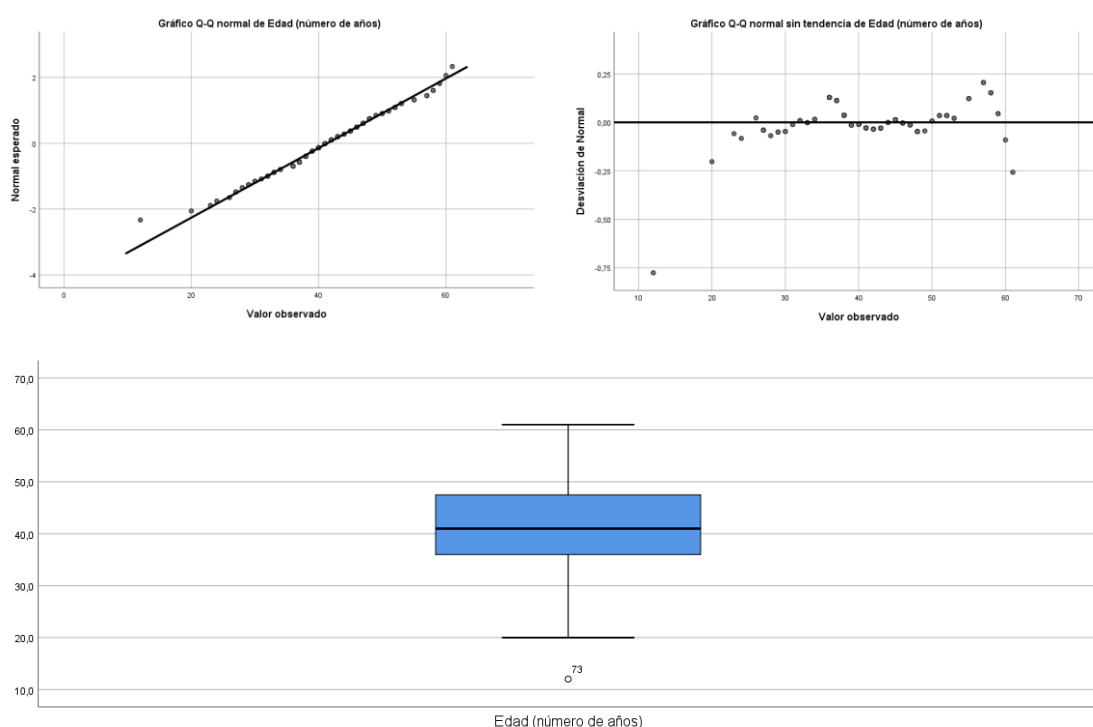
Gráfico Q-Q Normal (Normal Q-Q Plot): El gráfico Q-Q normal compara los cuantiles de tus datos con los cuantiles esperados de una distribución normal. Si los puntos en el gráfico caen aproximadamente en una línea recta diagonal, esto sugiere que tus datos se ajustan a una distribución normal. Si los puntos se desvían de esta línea, podrían indicar desviaciones de la normalidad.

Gráfico Q-Q Normal Sin Tendencia (Normal Q-Q Plot without Trend): A veces, los datos pueden tener tendencias o patrones que afectan la interpretación del gráfico Q-Q. El gráfico Q-Q normal sin tendencia elimina estos patrones, permitiéndote evaluar la normalidad más directamente. Puede ser útil cuando hay una fuerte tendencia en los datos que podría distorsionar la evaluación de la normalidad.

Diagrama de Caja (Box Plot): Un diagrama de caja es un gráfico que muestra la distribución de los datos a través de un rectángulo (la "caja") que representa el rango intercuartil (IQR), con líneas (los "bigotes") que se extienden hasta los valores mínimo y máximo. Los puntos fuera de los bigotes se consideran valores atípicos. Los diagramas de caja son útiles para visualizar la dispersión de los datos y detectar valores extremos.

Gráfico 2

Q-Q normal, Q-Q normal sin tendencia y diagrama de caja



CONCLUSIÓN

Este estudio ha explorado y analizado la importancia de los histogramas y la distribución normal en el análisis de datos, proporcionando herramientas cruciales para comprender la forma y la frecuencia de las observaciones. Las pruebas estadísticas de Shapiro-Wilk y Kolmogorov-Smirnov se utilizan para evaluar la normalidad de los datos y se complementaron con la creación de gráficos Q-Q y diagramas de caja para una comprensión más completa.

El método empleado en este estudio permitió un análisis exhaustivo de una variable específica utilizando tanto enfoques estadísticos como visuales. Los histogramas revelaron la distribución de frecuencias y los patrones de los datos, mientras que los gráficos Q-Q proporcionaron una comparación directa con una distribución normal ideal. La prueba de Shapiro-Wilk y la prueba de Kolmogorov-Smirnov dieron lugar a p-valores que se compararon con niveles de significancia predefinidos, lo que ayudó a determinar si los datos se ajustaban a una distribución normal.

Los resultados obtenidos en este caso piloteado de las pruebas sugieren que, para la variable analizada en este estudio, no existe suficiente evidencia para afirmar que los datos se desvían significativamente de una distribución normal. Tanto la prueba de Shapiro-Wilk como la prueba de Kolmogorov-Smirnov indicaron que los datos podrían considerarse razonablemente aproximados a una distribución normal en términos de su forma.

La combinación de enfoques estadísticos y visuales proporciona una evaluación más sólida de la normalidad de los datos, lo que es esencial para tomar decisiones informadas basadas en el análisis de datos. En el ámbito de los estudios administrativos, económicos y sociales, esta metodología puede ser valiosa para comprender la distribución de variables críticas y tomar decisiones estratégicas basadas en resultados confiables.

En última instancia, este estudio subraya la relevancia de comprender y aplicar adecuadamente las técnicas de evaluación de normalidad en el análisis de datos. A través de la combinación de enfoques estadísticos y visuales, los investigadores pueden obtener una perspectiva completa de la distribución de sus datos y fortalecer la base para la toma de decisiones informadas en diversas disciplinas.

REFERENCIAS

Altamirano, G. E. P., Sarango, A. F. H., Romo, H. E. M., Pérez, M. E. A., & Acuña, L. G. V. (2023). Cartera de crédito en las cooperativas de ahorro y crédito: Un estudio del efecto pandemia: Credit unions' loan portfolio: a study of the pandemic effect. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, 4(1), Article 1. <https://doi.org/10.56712/latam.v4i1.442>

Guachimboza-Villalva, M. V., Guachimbosa-Villalba, V. H., Luzuriaga Jaramillo, H. A., & Sánchez-Guerrero, J. (2021). Web Application for the Management of Reagents, Based on MEAN Stack Tools. En M. Botto-Tobar, M. Zambrano Vizuite, & A. Díaz Cadena (Eds.), *Innovation and Research* (pp. 319–333). Springer International Publishing. https://doi.org/10.1007/978-3-030-60467-7_27

Haro Sarango, A., López Solís, O., Córdova Pacheco, A., & Pérez Briceño, J. (2023). El teorema Modigliani-Miller: Un análisis desde la estructura de capital mediante modelos Data Mining en pymes del sector comercio. *Revista Finanzas y Política Económica*, 15(1), 45–66.

Hederich Martínez, C. (2023). Análisis cuantitativo de datos para la investigación educativa y social. Universidad Pedagógica Nacional. <http://repository.pedagogica.edu.co/handle/20.500.12209/18427>

Juárez Manayay, A. M. (2021). Sistema de información gerencial para el proceso de administración de ventas en el área gerencia de negocios telefónica de Teleatento del Perú S.A.C. Repositorio Institucional - UCV. <https://repository.ucv.edu.pe/handle/20.500.12692/70122>

Pedrosa, I., Juarros-Basterretxea, J., Robles-Fernández, A., Basteiro, J., & García-Cueto, E. (2015). Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar? *Universitas Psychologica*, 14(1), 245–254. <https://doi.org/10.11144/Javeriana.upsy13-5.pbad>

Pinkovetskaia, I. S., Nuretdinova, Y. V., Nuretdinov, I., & Lipatova, N. (2021). Mathematical modeling on the base of functions density of normal distribution. *Revista de La Universidad Del Zulia*, 12(33), 34–49.

Platas García, V. (2021). Contrastes de normalidad. <https://minerva.usc.es/xmlui/handle/10347/28978>

Roldán Molina, L. E. (2019). Aplicación de pruebas no paramétricas para k-muestras para identificar cambios en la distribución del Ingreso en Ecuador en el periodo 2006-2016 por provincia y actividad económica [bachelorThesis, Quito, 2019.]. <http://bibdigital.epn.edu.ec/handle/15000/20538>

Tauber, L. M. (2001). La construcción del significado de la distribución normal a partir de actividades de análisis de datos [Http://purl.org/dc/dcmitype/Text, Universidad de Sevilla]. <https://dialnet.unirioja.es/servlet/tesis?codigo=13551>

Tomás-Sábado, J. (2010). Fundamentos de bioestadística y análisis de datos para enfermería. Univ. Autònoma de Barcelona.