UNIVERSIDAD NACIONAL DE CHIMBORAZO

VICERRECTORADO DE INVESTIGACIÓN, VINCULACIÓN Y POSGRADO

DIRECCIÓN DE POSGRADO

Desarrollo de un modelo matemático para establecer la relación entre los estadísticos de prueba de normalidad y el p-valor: Una primera aproximación.

Trabajo de Titulación para optar al título de Magíster en Matemática Aplicada con mención en Matemática Computacional

AUTOR:

Mendoza Rodríguez, José Fernando

TUTOR:

Meneses Freire Manuel Antonio, PhD.

Riobamba, Ecuador. 2025

Declaración y Cesión de Derechos de Autoría

Yo, **José Fernando Mendoza Rodríguez** con número único de identificación **180375585-7**, declaro y acepto ser responsable de las ideas, doctrinas, resultados y lineamientos alternativos realizados en el presente trabajo de titulación denominado: "Desarrollo de un modelo matemático para establecer la relación entre los estadísticos de prueba de normalidad y el p-valor: Una Primera Aproximación." previo a la obtención del grado de Magíster en Matemática Aplicada con mención en Matemática Computacional.

- Declaro que mi trabajo investigativo pertenece al patrimonio de la Universidad Nacional de Chimborazo de conformidad con lo establecido en el artículo 20 literal j) de la Ley Orgánica de Educación Superior LOES.
- Autorizo a la Universidad Nacional de Chimborazo que pueda hacer uso del referido trabajo de titulación y a difundirlo como estime conveniente por cualquier medio conocido, y para que sea integrado en formato digital al Sistema de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor, dando cumplimiento de esta manera a lo estipulado en el artículo 144 de la Ley Orgánica de Educación Superior LOES.

Riobamba, 21 de julio de 2025

Ing. José Fernando Mendoza Rodríguez

N.U.I. 180375585-7

Dirección de
Posgrado
VICERRECTORADO DE INVESTIGACIÓN,
VINCULACIÓN Y POSGRADO

SGC

ACTA DE CULMINACIÓN DE TRABAJO DE TITULACIÓN

En la ciudad de Riobamba, a los 18 días del mes de julio del año 2025, los miembros del Tribunal designado por la Comisión de Posgrado de la Universidad Nacional de Chimborazo, reunidos con el propósito de analizar y evaluar el Trabajo de Titulación bajo la modalidad Proyecto de titulación con componente investigación aplicada y/o desarrollo, CERTIFICAMOS lo siguiente:

Que, una vez revisado el trabajo titulado: "**Desarrollo de un modelo matemático para establecer la relación entre los estadísticos de prueba de normalidad y el p-valor: Una Primera Aproximación.**", perteneciente a la línea de investigación: Ingeniería Informática, presentado por el maestrante **Mendoza Rodríguez José Fernando**, portador de la cédula de ciudadanía No. 1803755857, estudiante del programa de **Maestría en Matemática Aplicada con mención en Matemática Computacional**, se ha verificado que dicho trabajo cumple al 100% con los parámetros establecidos por la Dirección de Posgrado de la Universidad Nacional de Chimborazo.

Es todo cuanto podemos certificar, en honor a la verdad y para los fines pertinentes.

Atentamente,

PhD. Manuel Antonio Meneses Freire	Mgs. Lidia del Rocío Castro Cepeda	Mgs. Lorena Paulina Molina Valdiviezo
TUTOR	**MIEMBRO DEL TRIBUNAL 1**	**MIEMBRO DEL TRIBUNAL 2**

Dirección de
Posgrado
VICERRECTORADO DE INVESTIGACIÓN,
VINCULACIÓN Y POSGRADO

SGC
SISTEMA DE GESTIÓN DE LA CALIDAD
UNIVERSIDAD NACIONAL DE CHIMBORAZO

Riobamba, 24 de julio de 2025

CERTIFICADO

De mi consideración:

Yo, Manuel Antonio Meneses Freire, certifico que José Fernando Mendoza Rodríguez con cédula de identidad No. 1803755857 estudiante del programa de Maestría en Matemática Aplicada con mención en Matemática Computacional, tercera cohorte (2023-2024), presentó su trabajo de titulación bajo la modalidad de Proyecto de titulación con componente de investigación aplicada/desarrollo denominado: Desarrollo de un modelo matemático para establecer la relación entre los estadísticos de prueba de normalidad y el p-valor: Una Primera Aproximación, el mismo que fue sometido al sistema de verificación de similitud de contenido COMPILATION identificando el **2%** de similitud en el texto y el **6%** de similitud en inteligencia artificial.

Es todo en cuanto puedo certificar en honor a la verdad.

Atentamente,

PhD. Manuel Antonio Meneses Freire
CI: 1802515849

Adj.-

- Resultado del análisis de similitud (Compilation)

Dedicatoria

Con mucho cariño dedico este trabajo a la memoria de mi padre (+) y a mi madre, por educarme con valores firmes que han guiado mi formación personal.

A mi esposa, por su apoyo y motivación constante.

Con profundo amor a mis hijos, quienes han sido mi fuerza e inspiración en todo momento. Espero que en su momento y con madurez sean mejores que su padre en todo sentido.

Agradecimiento

Agradezco a Dios por ser mi guía y brindarme la sabiduría necesaria para completar este proceso de estudio. Mi agradecimiento especial a todos los docentes que compartieron su conocimiento en cada asignatura y utilizaron metodologías óptimas para que la enseñanza sea más significativa. A mi tutor, el PhD. Manuel Antonio Meneses Freire, por su aporte y colaboración en todo el proceso de este trabajo de investigación.

Este trabajo también lo dedico a todos quienes hacemos investigación y tratamos de dejar un legado, que sea una pequeña contribución para que las futuras generaciones, con la esperanza de que puedan descubrir o formular teorías que ayuden a resolver problemas de nuestra sociedad.

Índice General

Declaración y Cesión de Derechos de Autoría ..... **ii**

Acta de Culminación de Trabajo de Titulación ..... **iii**

Certificado de Contenido de Similitud ..... **iv**

Dedicatoria ..... **v**

Agradecimiento ..... **vi**

Índice General ..... **vii**

Índice de Tablas ..... **ix**

Índice de Figuras ..... **x**

Resumen ..... **xii**

Abstract ..... **xiii**

Introducción **1**

Capítulo 1 Generalidades **2**

1.1 Planteamiento del Problema 2
1.2 Justificación 3
1.3 Objetivos 5
1.3.1 Objetivo general 5
1.3.2 Objetivos específicos 5

Capítulo 2 Estado del Arte y la Práctica **7**

2.1 Antecedentes Investigativos 7
2.2 Fundamentación Teórica 10
2.2.1 Tipos de Variables y Naturaleza de los Datos 10
2.2.2 Distribución Normal 14
2.2.3 Pruebas de Normalidad Univariadas 18

2.2.4 Modelos matemáticos para primera aproximación....24
Capítulo 3 Diseño Metodológico....27
3.1 Enfoque de la Investigación....27
3.2 Diseño de la Investigación....27
3.3 Tipo de Investigación....27
3.4 Nivel de Investigación....27
3.5 Técnicas e Instrumentos de Recolección de Datos....28
3.6 Técnicas para el procesamiento e Interpretación de los Datos....29
3.7 Población y Muestra....30
3.7.1 Población....30
3.7.2 Tamaño de la Muestra....31
3.8 Tratamiento Estadístico....31
Capítulo 4 Análisis y Discusión de los Resultados....36
4.1 Análisis Descriptivo de los Resultados....36
4.2 Discusión de los Resultados....65
Capítulo 5 Marco Propositivo....69
5.1 Planificación de la Actividad Preventiva....69
Conclusiones....71
Recomendaciones....72
Referencias....74

Índice de Tablas

Tabla 1. Paquetes y librerías para el análisis 29

Tabla 2. Funciones en R para cada prueba de normalidad univariada 29

Tabla 3. Descripción de variables..... 30

Tabla 4. Muestra de cada variable 31

Tabla 5. Valores mínimo y máximo para las iteraciones de cada submuestra 32

Tabla 6. Comparación de modelos - valor de p y estadístico D 53

Tabla 7. Comparación de modelos - valor de p y estadístico A^2 54

Tabla 8. Comparación de modelos - valor de p y estadístico JB 56

Tabla 9. Comparación de modelos - valor de p y estadístico W^2 58

Tabla 10. Comparación de modelos - valor de p y estadístico D' 60

Tabla 11. Comparación de modelos - valor de p y estadístico P 61

Tabla 12. Comparación de modelos - valor de p y estadístico W' 63

Tabla 13. Comparación de modelos - valor de p y estadístico W 64

Tabla 14. Umbrales referenciales de R^2 67

Índice de Figuras

Figura 1. Clasificación tipos de variable 10

Figura 2. Características de los datos 11

Figura 3. Esquema general para la generación de modelos predictivos 12

Figura 4. Esquema - análisis exploratorio de los datos 13

Figura 5. Esquema – distribución de los datos 15

Figura 6. Esquema – relación R^2 y SCE 30

Figura 7. Función de densidad normalizado para las variables 36

Figura 8. Función de densidad para la prueba *KS* 37

Figura 9. Función de densidad para la prueba *AD* 38

Figura 10. Función de densidad para la prueba *JB* 38

Figura 11. Función de densidad para la prueba *CvM* 39

Figura 12. Función de densidad para la prueba *M.KS* 39

Figura 13. Función de densidad para la prueba *P* χ^2 40

Figura 14. Función de densidad para la prueba *SF* 40

Figura 15. Función de densidad para la prueba *SW* 41

Figura 16. Comportamiento de la prueba *KS* 43

Figura 17. Comportamiento de la prueba *AD* 44

Figura 18. Comportamiento de la prueba *JB* 44

Figura 19. Comportamiento de la prueba *CvM* 45

Figura 20. Comportamiento de la prueba *M.KS* 45

Figura 21. Comportamiento de la prueba *P* χ^2 46

Figura 22. Comportamiento de la prueba *SF* 46

Figura 23. Comportamiento de la prueba *SW* 47

Figura 24. Matriz de correlación para la variable precipitación 49

Figura 25. Matriz de correlación para la variable crudo.....50

Figura 26. Matriz de correlación para la variable temperatura.....51

Figura 27. Ajuste de curva – valor de p y el estadístico D.....52

Figura 28. Ajuste de curva – valor de p y estadístico A^2.....54

Figura 29. Ajuste de curva – valor de p y estadístico JB.....56

Figura 30. Ajuste de curva – valor de p y estadístico W^2.....58

Figura 31. Ajuste de curva – valor de p y estadístico D'.....60

Figura 32. Ajuste de curva – valor de p y estadístico P.....61

Figura 33. Ajuste de curva – valor de p y estadístico W'.....63

Figura 34. Ajuste de curva – valor de p y estadístico W.....64

Resumen

Las pruebas de normalidad son consideradas como el punto de partida para la estadística paramétrica; sin embargo, es importante comprender el comportamiento de los estadísticos y el valor de p para tomar decisiones significativas. El objetivo de la investigación fue diseñar un modelo matemático que permita profundizar en los patrones o tendencias existentes entre el estadístico y el valor de p. La metodología utilizada consistió en aplicar las pruebas de normalidad univariadas más comunes a variables como la precipitación, crudo y temperatura, mediante un análisis de submuestras con incrementos o decrementos de una observación, según las características de cada prueba. Cada submuestra proporcionó el valor del estadístico y el valor de p, los cuales fueron procesados mediante modelos de regresión clásicos como primera aproximación: lineal, polinomial de grado 2 y 3, exponencial y logarítmica. La validación de los modelos se realizó mediante el coeficiente de determinación, obteniendo como resultados que los modelos polinomiales de grado 3 y exponenciales presentaron un mejor ajuste a las variables estudiadas. La aplicación de diferentes tamaños muestrales (submuestras) evidenció una alta variabilidad o dispersión en el comportamiento del estadístico de las diferentes pruebas de normalidad univariadas, lo cual provocó que las variables analizadas no se distribuyan normalmente. Realizar este tipo de análisis preliminar permite a los investigadores tomar decisiones más efectivas y óptimas al momento de comprobar el supuesto de la normalidad en un conjunto de datos.

Palabras clave: Estadístico de prueba, modelo matemático, normalidad, regresión, valor de p.

ABSTRACT

Normality tests are considered the starting point for parametric statistics; however, it is important to understand the behavior of the test statistics and the p-value to make meaningful decisions. The objective of this research was to design a mathematical model that allows for a deeper understanding of the patterns or trends existing between the test statistic and the p-value. The methodology consisted of applying the most common univariate normality tests to variables such as precipitation, crude oil, and temperature, through a subsampling analysis with increments or decrements of one observation, according to the characteristics of each test. Each subsample provided the value of the test statistic and the p-value, which were then processed using classical regression models as a first approximation: linear, polynomial of degree 2 and 3, exponential, and logarithmic. Model validation was performed using the coefficient of determination, with results indicating that third-degree polynomial and exponential models provided a better fit for the variables analyzed. The use of different sample sizes (subsamples) revealed high variability or dispersion in the behavior of the test statistic across the various univariate normality tests, which led to the analyzed variables not following a normal distribution. Conducting this type of preliminary analysis enables researchers to make more effective and optimal decisions when assessing the normality assumption in a dataset.

Keywords: Test statistics, Mathematical model, Normality, Regression, p-value.

Reviewed by:
MsC. Edison Damian Escudero
ENGLISH PROFESSOR
C.C.0601890593

Introducción

En un mundo donde la tecnología está en constante desarrollo en los últimos años y el crecimiento de la población global ha generado que las personas se conviertan en una fuente principal de información para la captación de datos. Las personas cada vez se vuelven en consumidores potenciales, cada día el factor de compra está latente; recalcar que no solo la acción de comprar algún objeto tangible puede generar información, al solicitar una atención por motivos de salud, educación, entre otras, proporcionan datos valiosos donde no existe un intermediario económico para tal fin.

Los datos constituyen una fuente invaluable de información para la toma de decisiones estratégicas. No obstante, es fundamental realizar un análisis exhaustivo previo que permita identificar patrones, tendencias y comportamientos en función de las variables en estudio.

Al definir las variables de estudio, es fundamental garantizar su contextualización y contraste para obtener datos significativos. Más allá de las características que puedan ofrecer las variables, resulta esencial establecer una estrategia que determine la relevancia de los datos y asegure su adecuada vinculación con la forma en que se miden, todo en función de los objetivos planteados y el propósito de la investigación.

En muchas ocasiones, medir una variable para obtener un dato parece sencillo, pero esto depende en gran medida de las variables previamente definidas y si el instrumento de medición está diseñado acorde con sus características. Es importante señalar que toda variable es medible; sin embargo, al ser medida, siempre existe un margen de error. Esto se debe a que no hay un instrumento de medición perfecto, y dicha imprecisión puede aumentar cuando el instrumento es manipulado por el ser humano.

Capítulo 1

Generalidades

1.1 Planteamiento del Problema

En el ámbito de la investigación, los datos se han considerado como un papel importante para mostrar hallazgos representativos en un tema en particular y sobre todo para la toma de decisiones. En toda área del conocimiento donde los datos estén presentes y con énfasis cuando se trate de datos cuantitativos, utilizar estadística descriptiva e inferencial se ha convertido en un proceso crucial para contrastar hipótesis. El punto de partida una vez obtenidos datos, indistintamente del cómo fueron recolectados o medidos, sin dejar de lado su importancia, es aplicar los supuestos de la estadística paramétrica. Uno de los supuestos es evaluar la normalidad de los datos, y existen diferentes pruebas de normalidad que pueden analizar de forma univariada, bivariada o multivariada y que depende de las variables de estudio y del objetivo de la investigación para seleccionar la más adecuada.

Las pruebas de normalidad están sujetas en comprobar si la hipótesis nula cumple con la normalidad de los datos en base a una muestra predefinida. Cada prueba de normalidad proporciona o genera un valor de p, que representa la probabilidad de obtener un resultado tan extremo basado en la suposición de que la hipótesis nula es verdadera (Pawitan, 2020). El valor de p toma valores distintos cuando los tamaños muestrales crecen o decrecen, es decir, que están ligados por el tamaño de la muestra inicial. En muchos de los casos, la interpretación del valor de p debe ser un punto de inflexión para seguir profundizando en el análisis, el valor de p no debe ser la única vía para tomar decisiones en una investigación (Diaz-Ballve, 2020). El valor de p debe ser respaldado por otros parámetros estadísticos que evidencien el comportamiento de una variable. Forzar en rechazar la hipótesis nula basado en un valor de p, puede ocasionar un error en seleccionar una prueba estadística para realizar inferencia sobre ciertas variables definidas en un estudio.

A modo de ejemplo, cuando se requiere comprobar la normalidad de los datos, los valores de p pueden presentar una alta variabilidad al momento de utilizar diferentes tamaños muestrales. Para profundizar, si se define una muestra inicial n y otra con $n+1$, los valores de p obtenidos probablemente puedan estar: ambos por debajo, ambas por encima, o una por debajo y otro por encima del nivel de significancia. En otras palabras, el valor de p es sensible al tamaño de la muestra, por tal razón, es recomendable complementar su análisis con representaciones gráficas o estadísticos descriptivos, a fin de fortalecer la interpretación.

Al tener un tamaño de muestra, es probable que algunos de los datos presenten un comportamiento distinto a lo esperado. Estos datos, denominados como valores atípicos, influyen estrictamente en el comportamiento del estadístico de la prueba de normalidad y en los valores de p. La naturaleza de los datos juega un papel muy importante en el funcionamiento de la prueba estadística. No obstante, eliminar los valores atípicos no siempre es la estrategia correcta para garantizar que la prueba sea consistente. Considerar los valores atípicos puede ser útil para utilizar otros de tipo de pruebas de normalidad que sean más robustas o estables ante este tipo de alteraciones en los datos.

1.2 Justificación

La relación entre las pruebas de normalidad y los valores de p es cada vez evidente, y existen factores que influyen en su variabilidad al momento de analizar la posible causalidad entre ambas. El uso de un modelo matemático permite identificar la relación que existe entre las variables, sin embargo, seleccionar el modelo adecuado, no debe basarse únicamente en qué modelo se ajusta mejor a los datos, sino también en cómo se comportan las variables a lo largo tiempo y qué factores internos o externos afectan dicho comportamiento.

La finalidad de la investigación es proporcionar, en una primera instancia o aproximación, un modelo matemático que relacione los estadísticos de las pruebas de normalidad y los valores de p, modificando el tamaño de la muestra desde una muestra inicial hasta una muestra total. Es importante indicar que, los valores de p dependen del estadístico de prueba, puede afirmarse, a priori, que el estadístico de prueba actúa como variable independiente, mientras que el valor de p corresponde a la variable dependiente.

Independientemente de las variables que se analice, ya sean variables reales, aleatorias, o generadas mediante números aleatorios simulados por computadora, es importante modelar el comportamiento de las variables con la finalidad de ofrecer una solución óptima para interpretar adecuadamente los resultados. En el caso de la presente investigación, contar con un modelo matemático permite a los investigadores ser más objetivos en el análisis, esto contribuye a optimizar los procesos cuando se realice inferencia estadística en una investigación, especialmente al verificar el supuesto de normalidad en los datos.

Diseñar un modelo matemático para relacionar el estadístico de prueba de normalidad y el valor de p puede requerir un conjunto de datos donde se incorpore la variabilidad y la presencia de valores extremos o atípico, ya que algunas pruebas de normalidad son sensibles a este tipo de condiciones. En un modelo matemático se puede considerar diversos criterios, parámetros iniciales, así como factores internos y externos que convierte al proceso de análisis más eficaz para su interpretación. No obstante, es fundamental evaluar constantemente cada información que se incorpore al modelo, con el fin de identificar posibles hallazgos que puedan representar aspectos positivos o negativos en el proceso de diseño.

Al diseñar un modelo matemático en una primera aproximación, se pueden emplear modelos o expresiones matemáticas comunes o de mayor utilidad, como los modelos

lineales, cuadráticas (polinomio de grado 2), cúbicas (polinomio de grado 3), exponenciales, logarítmicas, entre otros. Este tipo de modelos se utilizan para realizar ajustes de curva a los datos que provienen de una muestra. Además, son considerados como modelos iniciales o introductorios que orientan al investigador en la exploración de patrones o comportamientos de las variables analizadas.

La validación de un modelo es fundamental debido a que cada expresión matemática tiene un comportamiento distinto que depende de los coeficientes definidos en el modelo. El uso del coeficiente de determinación, coeficiente de determinación ajustado y la suma de los cuadrados de los errores son estadísticos que permiten estimar la calidad del ajuste del modelo a los datos.

1.3 Objetivos

1.3.1 Objetivo general

Diseñar un modelo matemático que establezca la relación entre los estadísticos de prueba de normalidad y el valor de p, con el fin de mejorar la interpretación y aplicación de estas pruebas en análisis de datos.

1.3.2 Objetivos específicos

- Analizar las pruebas de normalidad más utilizadas para identificar sus respectivos estadísticos de prueba y el valor de p.
- Realizar un análisis exploratorio de los datos para identificar patrones y comportamientos relevantes.
- Desarrollar un modelo matemático preliminar que describa la relación entre los estadísticos de prueba de normalidad y el valor de p.
- Validar el modelo propuesto mediante análisis estadístico apropiados.

- Proponer mejoras en la interpretación de los resultados de las pruebas de normalidad, con base en el modelo matemático desarrollado.

Capítulo 2

Estado del Arte y la Práctica

2.1 Antecedentes Investigativos

Es importante indicar que las pruebas de normalidad univariadas adoptan distintos enfoques. En muchos de los casos, analizan distribuciones empíricas y teóricas para conocer si un conjunto de datos cumple con el supuesto de normalidad. Por otro lado, existen pruebas que se centrar en medir momentos estadísticos, como el grado de asimetría de una distribución (tercer momento) y curtosis (cuarto momento), evaluando si los momentos muestrales se ajustan a los esperados en una distribución normal.

En la investigación titulada: *NormalityAssessment: una herramienta interactiva para el aula que permite evaluar la normalidad de forma visual,* realizada por Casement y McSweeney, donde el objetivo principal de la investigación fue describir una aplicación interactiva gratuita desarrollada con Shiny, que se puede descargar como un paquete de R, la cual implementa dos procedimientos desarrollados recientemente para la inferencia gráfica, con un énfasis específico en la evaluación de la normalidad, la metodología utilizada fue orientada a la didáctica aplicada debido a la creación de la aplicación y evaluación realizada a un grupo de estudiantes. Concluyen que la aplicación NormalityAssessment diseñada específicamente para abordar esta dificultad sobre la normalidad de lo datos, basándose en avances recientes en inferencia gráfica, ayudó a comprender de mejor manera mediante los gráficos de tipo Rorschach como los line-up fueron útiles al momento de evaluar el criterio de normalidad (Casement & McSweeney, 2022). Este tipo de investigaciones permite utilizar otro tipo de herramientas para realizar un análisis exploratorio de los datos. El uso de aplicaciones interactivas incorporadas en paquetes y librerías en R para el análisis de la normalidad, beneficia la comprensión y formas de aplicación a partir de una representación gráfica.

En el estudio titulada: *Aplicaciones de las pruebas de normalidad en el análisis estadístico,* realizada por Khatun, el objetivo principal fue comparar diferentes procedimientos de las pruebas de normalidad univariada con la aplicación de un nuevo algoritmo con tamaños muestrales con distribución uniforme. La metodología utilizada en la investigación tuvo un enfoque cuantitativo, debido al análisis estadísticos de un conjunto de datos donde la bondad de ajuste permitió probar la normalidad de los residuos, evaluar si dos muestras provienen de distribuciones idénticas o si las frecuencias observadas siguen una distribución específica. Concluye que, con el aumento del tamaño de la muestra, la potencia estadística aumenta; sin embargo, las pruebas de Shapiro-Wilk (SW), Shapiro-Francia (SF) y Anderson-Darling (AD) son las más potentes entre todas las pruebas analizadas (Khatun, 2021). Este tipo de investigaciones permiten tener un panorama más amplio del comportamiento de las pruebas de normalidad univariadas con respecto al incremento del tamaño de la muestra. Es esencial, observar el comportamiento de los estadísticos de cada prueba de normalidad con variables reales y aleatorias.

En la investigación titulada: *Prueba robusta de normalidad en presencia de valores atípicos,* realizada por Rana et al., donde el objetivo del trabajo fue evaluar la potencia estadística de algunas pruebas de normalidad univariada con la presencia de valores atípicos o distribuciones asimétricas de un conjunto de datos. La metodología utilizada tuvo un análisis cuantitativo-estadístico mediante un experimento de simulación para comparar el rendimiento de algunas pruebas de normalidad. Se concluye que, los resultados de un ejemplo de la vida real y un estudio de simulación muestran que la potencia del MJB (modificación de la prueba Jarque-Bera) es superior para detectar la normalidad de los datos en comparación con las pruebas Jarque-Bera y Jarque-Bera robusta (Rana et al., 2021). Este análisis permite, en base a sustento teórico y simulaciones experimentales, proponer mejoras en la formulación, estructura y objetividad de las pruebas de normalidad clásicas,

especialmente cuando existen distribuciones asimétricas donde los datos atípicos influyen en el comportamiento de las pruebas estadísticas.

Una de las investigaciones titulada: *El papel de los valores p en la evaluación de la fuerza de la evidencia y las expectativas realistas de replicación,* realizada por Gibson, donde el objetivo de la investigación fue ofrecer una guía sobre el papel de los valores *p* al evaluar si la fuerza de la evidencia reflejada en un conjunto de datos es convincente. Utiliza una metodología teórica con enfoque explicativo, la finalidad fue esclarecer el rol de los valores de p para evaluar la evidencia científica. Concluye que la evidencia aparente se inflama por factores como la multiplicidad, la inferencia selectiva, el sesgo y el bajo poder estadístico, la esperanza de replicación puede convertirse en una expectativa poco realista (Gibson, 2020). Este tipo de estudio permite un análisis más profundo sobre el significado del valor de p, especialmente cuando se realiza inferencia estadística. El valor de p no debe ser un punto determinante para tomar decisiones, más bien, debe entenderse como un punto de partida para en lo posterior examinar los datos con mayor profundidad.

En el estudio titulado: *Significación estadística, valores de p y la comunicación de la incertidumbre,* realizada por Imbens, el objetivo de la investigación fue analizar a profundidad sobre los criterios relacionados con el nivel de significancia y la interpretación de los valores de p. La metodología se basó en realizar un análisis crítico sobre las controversias de tomar decisiones basados a los valores de p. Tanto la significancia estadística como los valores de p han sido sobredimensionados. Esto implica que, los resultados derivados de los valores de p no deben ser considerados concluyentes por más pequeño que sea dicho valor. El autor concluye que, los valores de p pueden ser útiles como punto de partida, pero que deben ir acompañados por los intervalos de confianza, errores estándar o aún mejor los intervalos posteriores bayesianos (Imbens, 2021).

2.2 Fundamentación Teórica

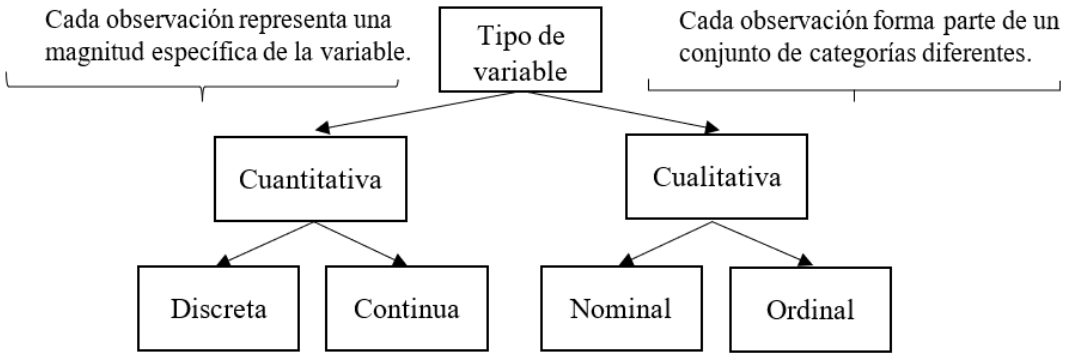
2.2.1 Tipos de Variables y Naturaleza de los Datos

Las variables cualitativas (categóricas) se dividen en nominales y ordinales. Son nominales cuando los datos que proceden no tienen un orden o secuencia específica, es decir, no existe una jerarquía principal entre ellos. Por lo contrario, se denominan ordinales debido a que el orden de los datos genera un posicionamiento jerárquico o patrón de nivel.

Por otro lado, las variables cuantitativas (numéricas) se dividen en discretas o continuas. Son discretas cuando los datos toman un valor entero positivo (natural), negativo o cero (ausencia de valor). Mientras que los datos de tipo continuo pueden asumir valores con cifras decimales, lo que implica una mayor variabilidad entre valores. Esto significa que los datos pertenecen al conjunto de los números racionales.

Figura 1

Clasificación tipos de variable

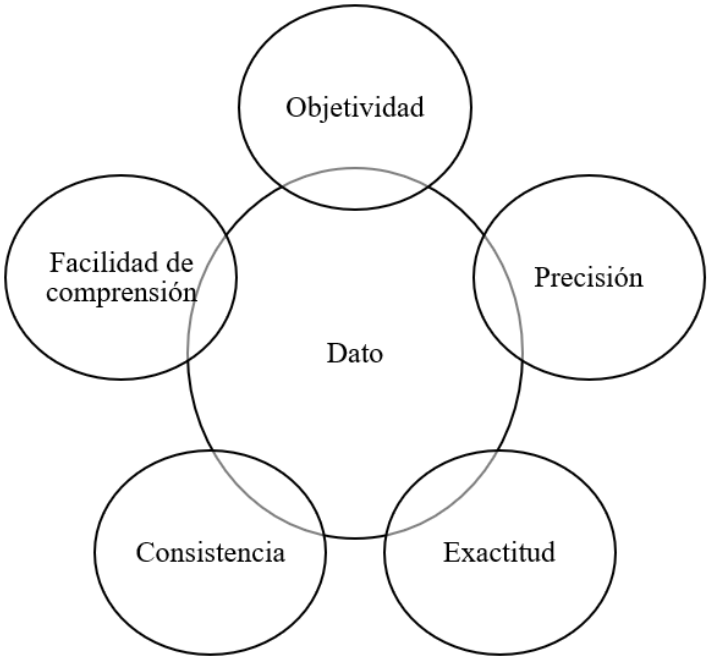


Nota. Adaptado de Shukla (2023).

La obtención de un dato debe caracterizarse por la objetividad, precisión, exactitud, consistencia y facilidad de comprensión. Si bien en condiciones ideales estas características se pueden cumplir, en contextos reales la interacción dentro de pares o grupos pueden generar datos significativos que sean de mucha utilidad para un procesamiento óptimo.

Figura 2

Características de los datos



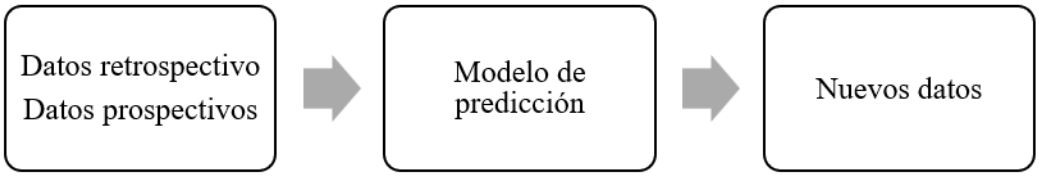
Los datos se derivan de las variables de estudio que por lo general están clasificadas como variables cuantitativas o cualitativas según la naturaleza de las mismas.

Los datos pueden estar almacenados en una base de datos, es decir, datos históricos recolectados para recabar información sobre variables. Estos también son denominados como datos retrospectivos y que son de gran importancia para realizar predicciones y toma de decisiones que mejoren el impacto de las variables en contextos sociales, científicos, educativos, culturales, políticos, entre otras.

Por otro lado, cuando las variables se miden por primera vez, los datos reciben el nombre prospectivo, que de igual manera que los datos retrospectivos, estos son fundamentales para realizar un análisis estadístico descriptivo e inferencial siempre que los datos presenten el menor error o variabilidad al momento de la medición. En ambos casos, la predicción es posible, aclarando que los datos masivos (*big data*) suelen ser más confiables y óptimos según el contexto en el que se utilice y las posibles aplicaciones que se diseñen.

Figura 3

Esquema general para la generación de modelos predictivos



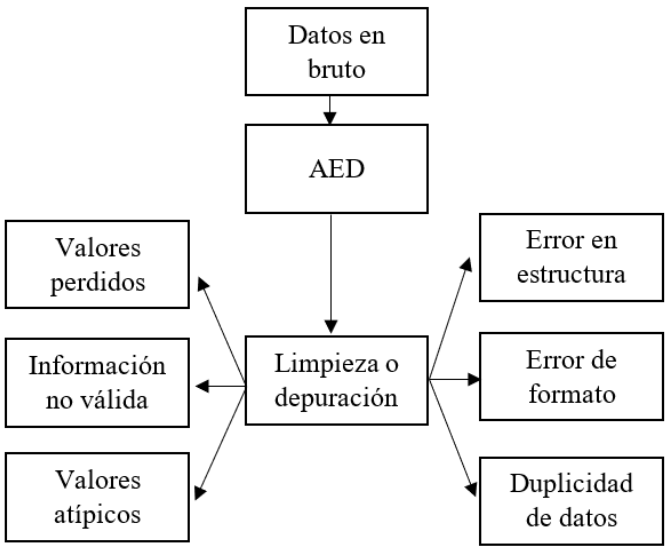
Nota. Adaptado de Wang et al. (2020).

Los datos pueden generar resultados incorrectos si no se realiza un análisis previo de los mismos. En muchos casos, es importante aplicar un filtrado y limpieza de los datos con la finalidad de que cumplan un formato específico y acorde a lo que se desea analizar, corrigiendo o eliminando errores e inconsistencias en la base de datos, el objetivo es obtener datos de calidad (Ferencek & Kljajić, 2020). Es importante indicar que, si las técnicas de recolección de información son eficientes, el tiempo de la preparación de los datos disminuye considerablemente y resulta tomar decisiones más creíbles. En cambio, el uso de técnicas deficientes provocaría incertidumbre y resultados ambiguos (Taplin, 2024).

Es fundamental realizar un Análisis Exploratorio de los Datos (AED) con la finalidad de entender y comprender las características principales de un conjunto de datos. No es necesario comenzar a explorar los datos con técnicas visuales, más bien, es recomendable considerar una limpieza o depuración de los datos en bruto. Esto implica observar posibles errores de formato, duplicidad de datos, errores en la estructura, valores atípicos, información no válida, valores perdidos o en blanco. Esta etapa permite tener una mejor comprensión y refinamiento de los datos, lo que resulta esencial para cumplir con los objetivos planteados (Kross et al., 2020).

Figura 4

Esquema - análisis exploratorio de los datos



Nota. Adaptado de Guo et al. (2023).

Luego de analizar las características de los datos en base a su calidad o representatividad, es necesario realizar un análisis estadístico descriptivo con la finalidad de indagar o explorar el comportamiento de los datos. Una opción es mediante una representación gráfica, donde el diagrama de dispersión (*scatter plot*), el diagrama de cajas (*box plot*) y los gráficos Q-Q (Q-Q plot) son las herramientas fundamentales y representativas. Además, complementar el estudio con medidas de tendencia central y dispersión juega un papel importante para una correcta interpretación de los datos (Ag-Yi & Aidoo, 2022).

La elección de un gráfico estadístico depende en gran medida de las variables de estudio, además, dicha representación debe adaptarse a las condiciones específicas y a la naturaleza de cada una. Es importante comprender el comportamiento del tipo de variable para que los datos a recolectar tengan mayor validez y puedan encajar en un gráfico cuya interpretación sea significativa, enriquecedora y de fácil comprensión (Inglis et al., 2022). La visualización se convierte en una herramienta indispensable para representar un conjunto de datos (*data set*), a través de los gráficos estadísticos de pueden generar patrones,

relaciones entre variables, tendencias y posibles anomalías. La adaptabilidad es una de las características importantes de los gráficos, se ajustan a las necesidades de las variables, sus datos y el tipo de análisis a realizar, ya sea univariado, bivariado o multivariado (Hudiburgh & Garbinsky, 2020).

2.2.2 Distribución Normal

La distribución normal también definida como una distribución gaussiana, tiene la finalidad de verificar que los datos se ajusten a una línea de tendencia basado en la media poblacional (μ) o muestral ($\bar{x}$) y a la desviación estándar poblacional (σ) o muestral (s) como medida de tendencia central y dispersión respectivamente. Tiene como característica principal la simetría de los datos donde existe una concentración en forma de campana de Gauss, lo que indica que la mayoría de los valores se agrupan cerca del centro de la distribución, reduciendo gradualmente hacia los extremos (Avram & Mărușteri, 2022). Si se refiere a una muestra, el criterio de simetría se convierte en una condición ideal, conociendo que los datos son recolectados con un error muy probable desde cómo se define variable de estudio hasta su respectiva medición. En otras palabras, la existencia de una asimetría (*skewness*) en los datos tiene referencia a los sesgos o ligeros desvíos que se producen por los valores atípicos (*outliers*) o por la propia naturaleza de los datos (Kumagai et al., 2022).

En contexto real basado en una muestra, se puede mencionar que las medidas de tendencia central media ($\bar{x}$), mediana ($\tilde{x}$) y la moda (Mo) son distintas y que se aproximan o tienden a una distribución normal bajo ciertos criterios expuestos anteriormente.

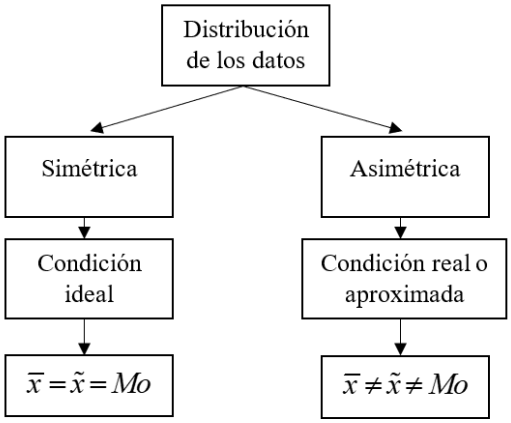
$$\bar{x} \neq \tilde{x} \neq Mo \tag{1}$$

La importancia de la distribución normal interfiere en la capacidad de modelar fenómenos en distintas áreas del conocimiento que son de base para la aplicación de técnicas de inferencia estadísticas que permiten predecir o estimar sobre los resultados basados en la

normalidad. El tamaño de la muestra juega un papel crucial en la disminución del error muestral, mientras mayor sea el tamaño de la muestra, mayor precisión de las estimaciones y mejor representatividad con respecto a la población (Pawel & Held, 2025). En muchos de los casos, los grandes volúmenes de datos al ser procesados representan un costo computacional significativo, por eso la importancia de trabajar con una muestra adecuada. En este contexto, seleccionar un muestreo probabilístico es la mejor alternativa para que los datos sean considerados en base a un criterio estadístico según las características propias de la población (Javed & Irfan, 2020).

Figura 5

Esquema – distribución de los datos



Mediante la distribución normal se permite tomar decisiones en lo que corresponde a la inferencia estadística. Existen pruebas estadísticas que consideran como un criterio o supuesto la normalidad de los datos, que al ser aplicados se puede obtener resultados óptimos y confiables (Savalei & Rosseel, 2021). Las pruebas que se basan a este criterio tienen mayor precisión en los resultados y minimiza el riesgo de error. Además, el teorema central del límite sirve como refuerzo para la aplicabilidad de este tipo de pruebas que, independientemente de la distribución normal de los datos originales, la media muestral puede obtener la normalidad a medida que se aumenta el tamaño de la muestra, sin dejar de

lado el nivel de significancia (α) como valor umbral o criterio de decisión (Uhm & Yi, 2021).

Es importante explicar que los datos bajo la condición real o aproximada tienen una inclinación hacia uno de los extremos en referencia a la campana de Gauss, la tendencia depende de los sesgos que se originen independientemente de los valores atípicos existentes en la base de datos. Cuando la campana de Gauss es sesgada a la derecha $\bar{x} > \tilde{x}$ recibe el nombre de asimetría positiva. En cambio, si el sesgo es a la izquierda $\bar{x} < \tilde{x}$ su asimetría es negativa. Para el cálculo de la asimetría se utiliza el coeficiente de asimetría muestral de Fisher (g_1) simple (ver ecuación 2) y ajustado (ver ecuación 3), que permite medir que tan simétrica o asimétrica es una distribución, y que está sujeto a que, si $g_1 > 0$ la asimetría es positiva, si $g_1 < 0$ la asimetría es negativa y bajo la condición ideal $g_1 = 0$ la distribución es simétrica (Demir, 2022).

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^{3/2}} \tag{2}$$

Donde n es el tamaño de la muestra, x_i cada valor de la muestra, $\bar{x}$ es la media muestral y s es la desviación estándar muestral. Donde el numerador representa el tercer momento central de la distribución (medida de asimetría) y el denominador se interpreta como la desviación estándar muestral elevada al cubo (esto permite estandarizar, medida adimensional).

$$g_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3 \tag{3}$$

Donde n es el tamaño de la muestra, x_i cada valor de la muestra, $\bar{x}$ es la media muestral y s es la desviación estándar muestral.

Un análisis de asimetría permite identificar según la forma de la distribución de los datos posibles distorsiones o desviaciones que influyen directamente en la elección del método estadístico, especialmente cuando se considera el supuesto de normalidad. La existencia de desviaciones pronunciadas puede afectar directamente dicho supuesto, lo que ocasiona dificultad en la interpretación de los resultados y equivocaciones tanto en las conclusiones como en la toma de decisiones (Lian et al., 2024).

Otra de las medidas estadísticas que intervienen en los cálculos para algunas pruebas de normalidad direccionadas a momentos es la curtosis (*Kurtosis*), también denominado como cuarto momento centrado, esto hace referencia a que los momentos son calculados respecto a la media muestral (centro) para describir su forma y estructura. Por otra parte, la curtosis (k) permite describir el grado de altura o apuntamiento de una distribución de datos, además, su interpretación está basado a tres categorías: mesocúrtica, leptocúrtica y platicúrtica (Korkmaz & Demir, 2023).

Cada categoría tiene una interpretación en base a un valor referencial de 3 bajo una distribución normal estándar $N(\mu,\sigma)$ o mediante su convección estandarizada $N(0,1)$. Si $k>3$ es leptocúrtica, mayor grado de apuntamiento; si $k=3$ es mesocúrtica, cumple con una distribución normal y si $k<3$ es platicúrtica, menor grado de apuntamiento o más plana.

$$k=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 \tag{4}$$

Donde n es el tamaño de la muestra, x_i cada valor de la muestra, $\bar{x}$ es la media muestral y s es la desviación estándar muestral.

2.2.3 Pruebas de Normalidad Univariadas

2.2.3.1 Kolmogorov Smirnov (KS).

La prueba KS (denotado como D, estadístico de prueba) conocida como la prueba de bondad de ajuste permite verificar la normalidad de los datos en cuestión. Por lo general, es empleada para variables cuantitativas de intervalo o continuas y para tamaños muestrales mayores de 50 (Paliz et al., 2021). La prueba KS tiene la funcionalidad de comparar una función de distribución acumulativa empírica de los datos observados de una función de distribución acumulativa teórica (Jaramillo et al., 2023). La expresión matemática para determinar el estadístico KS está definida en la ecuación 1 y 2.

$$KS = máx\left(D^{+}, D^{-} \right) \tag{5}$$

$$D^{+} = máx\left[\frac{i}{n} - F\left(x_{(i)} \right) \right] \tag{6}$$

$$D^{-} = máx\left[F\left(x_{(i)} \right) - \frac{i-1}{n} \right] \tag{7}$$

Donde n es el tamaño de la muestra; $F(x)$ es la función de distribución acumulativa con parámetros de media muestral y desviación estándar muestral; x_i son los valores ordenados de la muestra; $\frac{i}{n}$ y $\frac{i-1}{n}$ son los valores de la función de distribución empírica. La prueba KS tiene la particularidad que permite estudiar si el tamaño de la muestra procede de una población con una distribución de probabilidad en base a la media y desviación estándar establecida (Molina, 2022). Los datos de la muestra deben estar ordenados de manera creciente.

2.2.3.2 Shapiro Wilk (SW).

La prueba SW (denotado como W, estadístico de prueba) es muy utilizada para conocer si los datos recolectados provienen de una población con una distribución normal, por lo general, se emplea para un tamaño de muestra menor o igual a 50 (Roco-Videla et al., 2023). El estadístico W tiene la funcionalidad de verificar las diferencias que existe entre los valores observados y esperados. Además, examina las posibles discrepancias cuando los valores de W son altos o bajos (Goss-Sampson y Menese, 2019). La expresión matemática para determinar el estadístico SW está definida en la ecuación 3.

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)} \right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} \tag{8}$$

Donde n es el tamaño de la muestra; $x_{(i)}$ son los datos ordenados de menor a mayor; $\bar{x}$ es la media muestral y a_i son los coeficientes de la matriz de covarianza que está relacionado con el tamaño de la muestra y se determina en función de los momentos de la distribución normal estándar. Los datos de la muestra deben estar ordenados de manera creciente.

Uno de los criterios a considerar para los valores obtenidos de W es cuando se acerca a 1, esto se puede interpretar que las discrepancias entre los valores observados y esperados son bajas o pequeñas, esto sugiere que los datos provienen de una distribución normal, pero no implica que existan diferencias considerables cuando existen valores atípicos (Garibaldi et al., 2023).

2.2.3.3 Anderson Darling (AD).

La prueba AD (denotado como A^2, estadístico de prueba) es considerado como un estadístico que permite evaluar si un conjunto de datos proviene de una distribución normal.

Este tipo de prueba tiene como objetivo realizar la comparación de la función de distribución acumulada empírica basada en los datos observados con la función de distribución acumulada teórica asumida bajo una distribución específica (Tapia y Cevallos, 2021). Es necesario mencionar que la prueba AD requiere que los datos sean independientes e idénticamente distribuidos (Argüelles-Arellano, 2023). La expresión matemática para determinar el estadístico AD está definida en la ecuación 4 y 5.

$$A^2 = -n - s \tag{9}$$

$$s = \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F(Y_i) + \ln(1-F(Y_{n+1-i}))\right] \tag{10}$$

Donde n es el tamaño de la muestra; s es la desviación estándar muestral; Y_i son los datos ordenados de manera ascendente; $F(Y_i)$ es la función de distribución acumulada teórica y Y_{n+1-i} representa las observaciones en orden inverso. Los datos de la muestra deben estar ordenados de manera creciente.

Es necesario tomar en consideración que la prueba AD requiere que los datos que ingresan al análisis sean independientes e idénticamente distribuidos, además, este tipo de prueba puede generar sensibilidad al momento de la elección de los parámetros de la distribución acumulada teórica, tamaño de la muestra y posibles datos atípicos que infieran en los resultados y originen desviaciones en la distribución normal.

2.2.3.4 Jarque – Bera (JB).

La prueba (denotado como JB, estadístico de prueba) permite evaluar si una muestra de datos está distribuida normalmente, basándose en dos características principales de la distribución en medir la asimetría y curtosis. Sin embargo, puede generar inconsistencia en la interpretación al no detectar desviaciones para muestras pequeñas ($n < 30$), además,

puede la prueba no puede ser confiable para valores atípicos, delimitando su aplicación en ciertos contextos (Muñoz et al., 2019).

$$JB = n\left[\frac{{g_1}^2}{6} + \frac{(k-3)^2}{24}\right] \tag{11}$$

Donde n es el tamaño de la muestra; g_1 es el coeficiente de asimetría de Fisher simple de la muestra y k es coeficiente de curtosis de la muestra.

Es importante mencionar que el estadístico JB se puede comparar con una distribución chi-cuadrado específicamente con dos grados de libertad debido a que se utilizan dos momentos: asimetría y curtosis. Esta prueba es más confiable y potente para muestras grandes (Glinskiy et al., 2024).

2.2.3.5 Cramér-von Mises (CvM).

La prueba CvM (denotado como W^2, estadístico de prueba) es utilizada para evaluar si un conjunto de datos de una muestra proviene de una distribución normal. Su objetivo es medir la integral de las diferencias al cuadrado entre las funciones de distribución teórica y empírica, es decir que no se centra en el máximo valor o discrepancia como la prueba KS (Gandica de Roa, 2020).

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(x_{(i)}\right) - \frac{2i-1}{2n}\right]^2 \tag{12}$$

Donde n es el tamaño de la muestra y $F\left(x_{(i)}\right)$ es la función acumulada teórica evaluada en un dato específico $x_{(i)}$. Los datos de la muestra deben estar ordenados de manera creciente.

Para interpretar el estadístico de prueba W^2 se debe considerar que, si W^2 resulta un valor pequeño indica que las funciones de distribución teórica y empírica están cercanas, esto implica que los datos podrían ser normales, caso contrario si W^2 resulta un valor grande, se determina que existen diferencias significativas. Para validar si los datos son normales o existen diferencias es importante calcular el p-valor. Esta prueba mejora su potencia y es más eficiente para detectar desviaciones de la normalidad cuando se aumenta el tamaño de la muestra (Jiménez-Gamero, 2024).

2.2.3.6 Lilliefors (M.KS).

La prueba M.KS (denotado como D', estadístico de prueba) consiste en una versión modificada de la prueba KS, la diferencia hace énfasis en que no asume que media y desviación estándar poblacional sean conocidas, más bien son estimados a partir de un conjunto de datos, es decir de una muestra. Es aspectos de sensibilidad la prueba M.KS es más preciso para la normalidad de los datos con parámetros desconocidos (Hagag, 2022).

$$M.KS = máx\left(D'^{+}, D'^{-} \right) \tag{13}$$

$$D'^{+} = máx\left| F_n(x) - \Phi\left[\left(\frac{x-\mu}{s} \right) \right] \right| \tag{14}$$

$$D'^{-} = máx\left| \Phi\left[\left(\frac{x-\mu}{s} \right) \right] - F_n(x) \right| \tag{15}$$

Donde $F_n(x)$ es la función de distribución empírica; Φ es la función de distribución acumulativa de la distribución normal estándar; $\mu = \bar{x}$ es la media muestral y s es la desviación estándar muestral corregida con $n-1$. Además, $s = \sqrt{\sum\left(x_i - \bar{x}\right)^2 / \left(n-1\right)}$ siendo n el tamaño de la muestra. Los datos de la muestra deben estar ordenados de manera

creciente, además, esta prueba genera mayor potencia estadística al utilizar muestras de gran tamaño (Sulewski, 2021).

2.2.3.7 Pearson Chi-Cuadrado (P χ^2).

La prueba P χ^2 (denotado como P, estadístico de prueba) es utilizada para comparar la distribución observada de los datos con una distribución esperada, bajo la condición que se distribuyen normalmente con la misma media y desviación estándar que la muestra (Islam, 2021). El estadístico P se considera asintótico distribuido con $n-1$ grados de libertad. Además, es considerada como una prueba de bondad de ajuste basada en la agrupación de frecuencias, esto significa que mide la diferencia entre frecuencias observadas y esperadas (Arnastauskaitė et al., 2021).

$$\mathrm{P} = \sum_{i} \frac{\left(C_i - E_i\right)^2}{E_i} \tag{16}$$

Donde C_i indica el número de observaciones contadas de la clase i y E_i representa el número de observaciones esperadas de la clase i. Es importante considerar que las clases deben tener la misma probabilidad teórica, garantizando una comparación más eficaz.

2.2.3.8 Shapiro-Francia (SF).

La prueba SF (denotado como W', estadístico de prueba) consiste en una versión simplificada a la prueba de Shapiro Wilk, consiste en la correlación cuadrática entre los valores ordenados de la muestra y los cuantiles esperados de la distribución normal (Domański & Szczepocki, 2020). Tiene mayor potencia estadística y diseñada para muestras grandes (Uyanto, 2022).

$$W' = \frac{\left(\sum_{i=1}^{n} m_i x_i\right)^2}{\left(\sum_{i=1}^{n} x_i - \bar{x}\right)^2 \sum_{i=1}^{n} m_i^2} \tag{17}$$

Donde x_i es el dato específico, $\bar{x}$ es la media, n es el tamaño de la muestra y m_i representa los valores esperados de los datos ordenados de menor a mayor bajo el criterio de normalidad.

2.2.4 Modelos matemáticos para primera aproximación

2.2.4.1 Modelo Lineal.

Es un ajuste estadístico o una forma básica de regresión estadística, que permite describir la relación lineal aproximada entre una variable independiente (*x*) y una variable dependiente (*y*). La ecuación 18 representa la expresión matemática de un modelo lineal.

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{18}$$

Donde *x* es la variable de entrada o predictora, *y* es la variable de respuesta, β_0 es el corte o intercepto en el eje de las ordenadas, β_1 representa la pendiente de la recta y ε indica el error residual.

2.2.4.2 Modelo Polinomial de Grado 2.

Es un ajuste estadístico o una forma básica de regresión estadística, que permite describir la relación cuadrática aproximada entre una variable independiente (*x*) y una variable dependiente (*y*). La ecuación 19 representa la expresión matemática de un modelo polinomial de grado 2.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{19}$$

Donde *x* es la variable de entrada o predictora, *y* es la variable de respuesta, β_0 es el corte o intercepto en el eje de las ordenadas, β_1 representa el coeficiente lineal, β_2 es el coeficiente cuadrático (determina la curvatura), y ε indica el error residual.

2.2.4.3 Modelo Polinomial de Grado 3.

Es un ajuste estadístico o una forma básica de regresión estadística, que permite describir la relación cúbica aproximada entre una variable independiente (*x*) y una variable dependiente (*y*). La ecuación 20 representa la expresión matemática de un modelo polinomial de grado 3.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \tag{20}$$

Donde *x* es la variable de entrada o predictora, *y* es la variable de respuesta, β_0 es el corte o intercepto en el eje de las ordenadas, β_1 representa el coeficiente lineal, β_2 es el coeficiente cuadrático, β_3 es el coeficiente cúbico (determina inflexiones o cambios de curvatura), y ε indica el error residual.

2.2.4.4 Modelo Exponencial.

Es un ajuste estadístico o una forma básica de regresión estadística, que permite describir la relación exponencial aproximada entre una variable independiente (*x*) y una variable dependiente (*y*). La ecuación 21 representa la expresión matemática de un modelo exponencial.

$$y = \beta_0 e^{\beta_1 x} + \varepsilon \tag{21}$$

Donde *x* es la variable de entrada o predictora, *y* es la variable de respuesta, β_0 es el valor inicial, β_1 representa la tasa de crecimiento ($\beta_1 > 0$) o decrecimiento ($\beta_1 < 0$), y ε indica el error residual.

2.2.4.5 Modelo Logarítmico.

Es un ajuste estadístico o una forma básica de regresión estadística, que permite describir la relación logarítmica aproximada entre una variable independiente (*x*) y una variable dependiente (*y*). La ecuación 22 representa la expresión matemática de un modelo logarítmico.

$$y = \beta_0 + \beta_1 \log(x) + \varepsilon \tag{22}$$

Donde *x* es la variable de entrada o predictora, *y* es la variable de respuesta, β_0 es el valor inicial (cuando $\log x = 0$), β_1 representa el coeficiente logarítmico (corresponde la sensibilidad de *y* frente a cambios relativos en *x*), y ε indica el error residual.

Capítulo 3

Diseño Metodológico

3.1 Enfoque de la Investigación

El enfoque de la investigación se orientó a un análisis cuantitativo ya que se utilizaron datos retrospectivos previamente recolectados, asociados a distintas variables de estudio (Turnbull et al., 2020). Esto permitió la aplicación de un análisis estadístico para el diseño del modelo matemático que relacione los estadísticos de prueba de normalidad univariada y los p- valores. Por ende, el uso de datos numéricos justificó al enfoque cuantitativo como la principal opción de la investigación.

3.2 Diseño de la Investigación

El diseño de la investigación fue no experimental debido a que no se manipularon las variables de estudio. Se realizó un análisis estadístico descriptivo e inferencial con el propósito de indagar, analizar y validar la relación que existe entre los test de normalidad univariada aplicados a un conjunto de datos muestrales y sus respectivos valores de p.

3.3 Tipo de Investigación

El tipo de investigación se direccionó al modelado matemático, ya que el objetivo principal fue diseñar una función o expresión matemática que relacione los estadísticos de prueba de normalidad univariada con sus respectivos valores de p (Austin et al., 2024). Además, se consideró de tipo explicativo, dado que cada conjunto de datos seleccionados según las pruebas de normalidad, genera ciertos valores de p en función al test aplicado.

3.4 Nivel de Investigación

El estudio en lo referente al nivel de investigación fue exploratorio debido a que la intención era identificar patrones o relaciones en los datos como una primera aproximación

(Ebbelind, 2023). Además, la finalidad de diseñar el modelo era encontrar la función o curva más adecuada que se ajuste a su comportamiento, según los criterios de cada prueba de normalidad univariada, ya sean estos direccionados en medir distancias o momentos, tomando como referencia una distribución normal. Por tal razón, la exploración de los datos fue indispensable para lograr un mejor ajuste y llevar el proceso de validación correspondiente.

3.5 Técnicas e Instrumentos de Recolección de Datos

La técnica utilizada fue la revisión documental mediante un análisis a profundidad de las instituciones gubernamentales del Ecuador que recopilan y centralizan información estadística. En particular, los datos se concentran en un Catálogo de Datos Abiertos donde se permitió acceder a la información sin necesidad de seguir un proceso de solicitud de acceso a la información pública. La finalidad de este catálogo es proporcionar criterios técnicos y metodológicos a los usuarios para que puedan promover la investigación, el emprendimiento y la innovación de la sociedad. De esta forma, se obtuvo los datos para el desarrollo del estudio.

Los instrumentos de recolección de datos para este caso fueron los insumos o archivos digitales con diferentes tipos de extensiones abiertas, las mismas que fueron descargadas desde la plataforma oficial de datos abiertos del Ecuador. En cada uno de los archivos descargados se encontraba toda la información sobre las variables medidas con sus respectivos metadatos, la finalidad fue conocer a profundidad las características de las variables. Por conveniencia, se utilizó un conjunto de datos continuos debido a la facilidad en el procesamiento estadístico mediante técnicas avanzadas y facilidad en trasformaciones y normalización.

3.6 Técnicas para el procesamiento e Interpretación de los Datos

En lo que concierne al análisis estadístico, se empleó el software R para la programación de las pruebas de normalidad univariadas, con el propósito de obtener los estadísticos de prueba y sus respectivos valores de p.

Tabla 1

Paquetes y librerías para el análisis

Paquetes	Librería	Descripción
install.packages("readxl")	library(readxl)	Carga el archivo con extensión .xlsx
install.packages("ggplot2")	library(ggplot2)	Crea gráficos estadísticos.
install.packages("systemfonts")	library(systemfonts)	Utiliza un tipo de letra específico.
install.packages("nortest")	library(nortest)	Incorpora las pruebas AD, CvM, M.KS, P χ^2 y SF que no vienen por defecto.
install.packages("tseries")	library(tseries)	Incorpora la prueba JB que no viene por defecto.
install.packages("GGally")	library(GGally)	Realiza gráficos de correlación matricial.
install.packages("dplyr")	library(tidyr)	Combina múltiples data frames.
install.packages("tidyr")	library(tidyr)	Transforma y reorganiza data frames.

Tabla 2

Funciones en R para cada prueba de normalidad univariada.

Prueba	Función en R
KS	ks.test()
AD	ad.test()
JB	jarque.bera.test()
CvM	cvm.test()
M.KS	lillie.test()
P χ^2	pearson.test()
SF	sf.test()

Para el diseño del modelo matemático en su primera aproximación se utilizaron diferentes ajustes de curva según el comportamiento de los datos. Para definir el mejor ajuste se tomó como referencia los valores del coeficiente de determinación R^2 y la suma de los cuadrados de los errores SCE. R^2 oscila entre 0 a 1, su interpretación indica que, mientras

más cercano sea el valor de R^2 a 1, existe un mejor ajuste del modelo a la distribución de los datos. En cambio, SCE indica que, si su valor es bajo, mejor es el ajuste o se considera más preciso el modelo a un conjunto de datos.

Figura 6

Esquema – relación R^2 y SCE

3.7 Población y Muestra

3.7.1 Población

Los datos considerados para el análisis estadístico fueron: la temperatura mínima absoluta, la producción de petróleo mensual operativo y la precipitación, los mismos que fueron descargados de la plataforma de datos abiertos del Ecuador y proporcionados por el Instituto Nacional de Meteorología e Hidrología – INAMHI y la Empresa Pública de Hidrocarburos del Ecuador - EP PETROECUADOR.

Tabla 3

Descripción de variables

Variable	Descripción	Unidad de medida
Temperatura mínima absoluta	Datos sobre la temperatura máxima absoluta de las principales estaciones meteorológicas del INAMHI a nivel nacional.	Grado Celsius (°C)
Producción de petróleo mensual operativo	Datos sobre los volúmenes de barriles Promedio por mes (BPPM) por cada activo.	Barriles
Precipitación	Datos sobre la precipitación total mensual de las principales estaciones meteorológicas del INAMHI a nivel nacional.	Milímetros (mm)

3.7.2 Tamaño de la Muestra

En lo que concierne a la muestra, se utilizó un muestro no probabilístico por conveniencia dado que los datos provienen de fuentes oficiales y confiables. Esta elección permite realizar un análisis estadístico válido y transparente, considerando que los datos son reales y cuya disponibilidad es de acceso libre. Las variables detalladas en la tabla 3 fueron seleccionadas porque corresponden a datos de tipo continuo.

Para la temperatura mínima absoluta, se consideró como muestra el primer semestre del año 2019 de las diferentes estaciones meteorológicas ubicadas en puntos estratégicos del territorio, el objetivo fue obtener mediciones distintas y representativas. En lo que corresponde a la producción de petróleo mensual operativo, se emplearon datos comprendidos entre los años 2022 y 2025. Para la variable precipitación, se analizaron datos registrados en los años 2018 y 2019. Las muestras específicas de cada variable se detallan en la tabla 4.

Tabla 4

Muestra de cada variable

Variable	Muestra
Temperatura mínima absoluta (Temperatura)	418
Producción de petróleo mensual operativo (Crudo)	535
Precipitación	669

3.8 Tratamiento Estadístico

Sobre la metodología aplicada, se estructuró en base al enfoque, diseño y nivel de investigación. Primero, los estudios retrospectivos con base a la recolección de datos numéricos relacionados con variables cuantitativas como el rendimiento académico, permitieron extraer los datos de una base datos que almacena la unidad educativa, por ende, se estableció un enfoque cuantitativo por la naturaleza que provienen los datos. Segundo, se utilizó un diseño no experimental debido a que no se manipuló la variable de estudio

(Aragón et al., 2022)., mencionar que el objetivo fue analizar el comportamiento de la variable enfocado en los valores de las pruebas estadísticas preestablecidas con respecto a los valores de p. Finalmente, el nivel de investigación utilizado fue descriptivo debido a que el conjunto de datos analizados (muestra total n) se estableció en subconjuntos de datos (submuestras A_k) donde k es el valor incremental para cada iteración en base a un valor muestral inicial, n es el tamaño de la muestra para las pruebas KS, SW, AD, JB, CvM, M.KS, P χ^2 y SF.

Para las pruebas KS, AD, JB, CvM, M.KS, P χ^2 y SF, se utilizó $k > 50$ observaciones con incrementos de una observación adicional para cada submuestra y para la prueba SW $k \in \{50,49,48,...,3\}$ con un proceso inverso donde existe decrementos en una observación para cada submuestra. Se fijó $k = 50$ para la primera submuestra según las características propias de la prueba SW y las adaptaciones específicas en cuanto al tamaño muestral. Para las pruebas KS, SW, AD, JB, CvM, M.KS, P χ^2 y SF se definieron las muestras $n_1 = 418; n_2 = 535; n_3 = 669,$ según la variable de análisis.

Tabla 5

Valores mínimo y máximo para las iteraciones de cada submuestra

Prueba	Valor muestral inicial	Submuestra	Mínimo y máximo
KS AD JB CvM M.KS P χ^2 SF	$k = 51$	Incrementos de una observación hasta alcanzar el valor de n.	$\text{mín}(k) = 51,\ \text{máx}(k) = n$
SW	$k = 50$	Decrementos de una observación hasta alcanzar el valor mínimo de 3.	$\text{mín}(k) = 3,\ \text{máx}(k) = 50$

En las pruebas KS, AD, JB, CvM, M.KS, P χ^2 y SF con n designada $\{x_1, x_2, x_3, ..., x_n\}$, se definió las submuestras $A_k = \{x_1, x_2, x_3, ..., x_k\}$, con $k \in \{51, 52, 53, ..., n\}$. En el caso de la prueba SW se establecieron bloques sistemáticos (B_m) como se indica en la ecuación 24 con $k \in \{50, 49, 48, ..., 3\}$ La finalidad de utilizar bloques fue cubrir la mayor cantidad de datos hasta acercarse al valor de n. La forma ampliada para las pruebas las pruebas KS, AD, JB, CvM, M.KS, P χ^2 y SF se detallan a continuación:

$$M_k = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_{51} & & & & \\ x_1 & x_2 & x_3 & \cdots & x_{51} & x_{52} & & & \\ x_1 & x_2 & x_3 & \cdots & x_{51} & x_{52} & x_{53} & & \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \\ x_1 & x_2 & x_3 & \cdots & x_{51} & x_{52} & x_{53} & \cdots & x_n \end{pmatrix} \tag{23}$$

Donde M_k es una matriz de orden $n - 50$ fila y n columnas, las filas representa la cantidad de submuestras y las columnas indica el número de observaciones. A continuación, se detalla el conjunto de observaciones para las tres primeras submuestras:

Para la primera submuestra $A_{51} = \{x_1, x_2, x_3, ..., x_{49}, x_{50}, x_{51}\}$.

Para la segunda submuestra $A_{52} = \{x_1, x_2, x_3, ..., x_{50}, x_{51}, x_{52}\}$.

Para la tercera submuestra $A_{53} = \{x_1, x_2, x_3, ..., x_{51}, x_{52}, x_{53}\}$.

En la prueba SW, en lo referente a los bloques, cada bloque empieza con un total de 50 observaciones, luego se inicia el proceso de iteración hasta alcanzar el valor mínimo establecido.

$$B_m = \left\{ x_{(m-1)\bullet 50+1}, x_{(m-1)\bullet 50+2}, ..., x_{\min(m\bullet 50, n)} \right\} \tag{24}$$

Donde $m = 1,2,3,...,\left\lceil \frac{n}{50} \right\rceil$ y $\min(m \bullet 50, n)$ garantiza que no se exceda el tamaño muestral de cada variable de estudio.

Para cada bloque la ecuación 25 define su conjunto de observaciones que depende del número de bloque y sus iteraciones de manera decremental. La forma ampliada se representa en la ecuación 26.

$$A_{m,k} = \left\{ x_{(m-1)\bullet 50+1}, x_{(m-1)\bullet 50+2}, x_{(m-1)\bullet 50+3}, ..., x_{(m-1)\bullet 50+k} \right\} \tag{25}$$

$$M_{m,k} = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_{50} \\ x_1 & x_2 & x_3 & \cdots & x_{49} \\ x_1 & x_2 & x_3 & \cdots & x_{48} \\ \vdots & \vdots & \vdots & & \vdots \\ x_1 & x_2 & x_3 & \cdots & x_3 \end{pmatrix} \tag{26}$$

Donde cada fila de la matriz $M_{m,k}$ representa cada submuestra en base al bloque definido y las columnas indica el número de observaciones. A continuación, se detalla el conjunto de observaciones para las tres primeras submuestras con respecto al primer bloque:

Para la primera submuestra $A_{1,50} = \left\{ x_1, x_2, x_3, ..., x_{48}, x_{49}, x_{50} \right\}$.

Para la segunda submuestra $A_{1,49} = \left\{ x_1, x_2, x_3, ..., x_{47}, x_{48}, x_{49} \right\}$.

Para la tercera submuestra $A_{1,48} = \left\{ x_1, x_2, x_3, ..., x_{46}, x_{47}, x_{48} \right\}$.

Cada una de las pruebas según la cantidad de submuestras y bloques analizados anteriormente, generan los estadísticos de prueba y los valores de p respectivos en cada iteración, por lo tanto, cada valor obtenido se representó en forma de conjunto de pares ordenados (λ).

$$\lambda_{KS} = \left\{ \left(D_k, p_k \right) \mid k \in \{51, 52, ..., n\} \right\} \tag{27}$$

$$\lambda_{AD} = \left\{ \left({A^2}_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{28}$$

$$\lambda_{JB} = \left\{ \left(JB_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{29}$$

$$\lambda_{CvM} = \left\{ \left({W^2}_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{30}$$

$$\lambda_{M.KS} = \left\{ \left(D'_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{31}$$

$$\lambda_{P\chi 2} = \left\{ \left(P_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{32}$$

$$\lambda_{SF} = \left\{ \left(W'_k, p_k \right) \mid k \in \{51,52,...,n\} \right\} \tag{33}$$

$$\lambda_{SW} = \left\{ \left(W_k, p_k \right) \mid k \in \{50,49,...,3\} \right\} \tag{34}$$

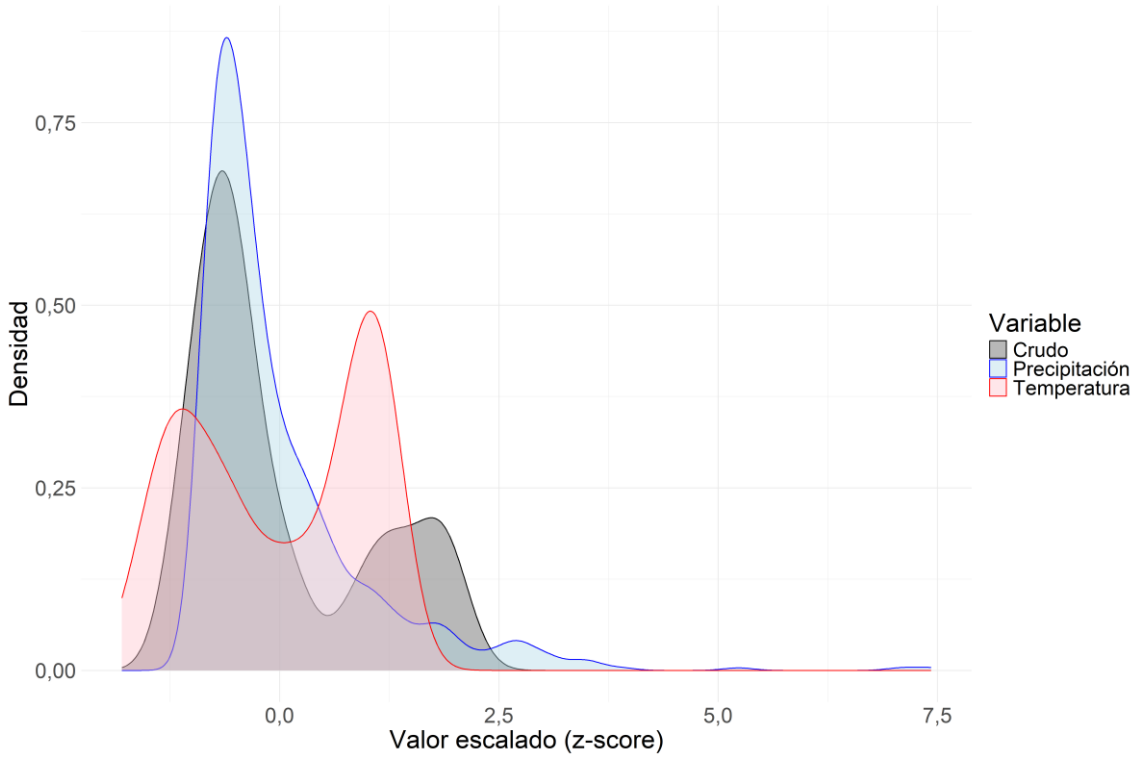
Capítulo 4

Análisis y Discusión de los resultados

4.1 Análisis Descriptivo de los Resultados

Figura 7

Función de densidad normalizado para las variables



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

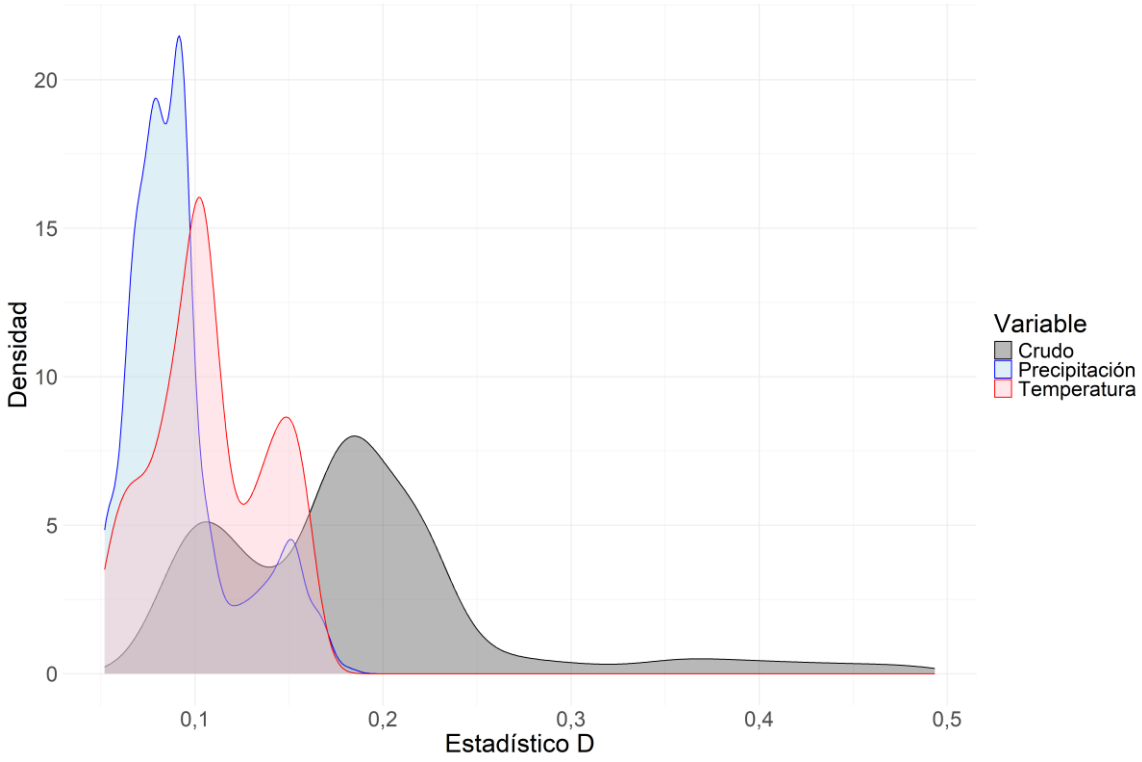
En la figura 7, se observa la función de densidad normalizado o escalado. Se utilizó el proceso de normalización debido a que las variables temperatura mínima absoluta, producción de petróleo mensual operativo y precipitación tienen unidades de medida distintas y se manejan escalas completamente diferentes. Por ende, la normalización de los datos permitió unificar las escalas de medición de las variables y mejorar la comparación visual de las curvas de densidad.

Se observa que las curvas de la temperatura mínima absoluta y producción de petróleo mensual operativo presentan dos picos, esto indica que ambas distribuciones son

bimodales y asimétricas. Por otro lado, la curva de precipitaciones tiene un sesgo positivo y esto se debe a que puede existir valores extremos o atípicos que visualmente no representa la forma típica de una campana de Gauss, lo que indica que no se trata de una distribución normal del conjunto de datos analizados.

Figura 8

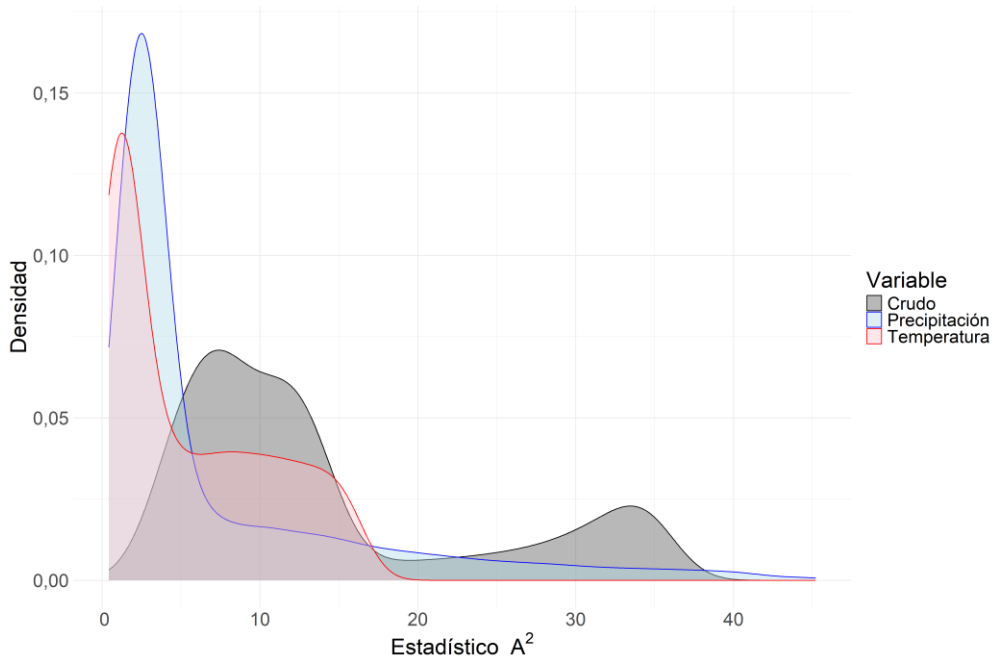
Función de densidad para la prueba KS



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 9

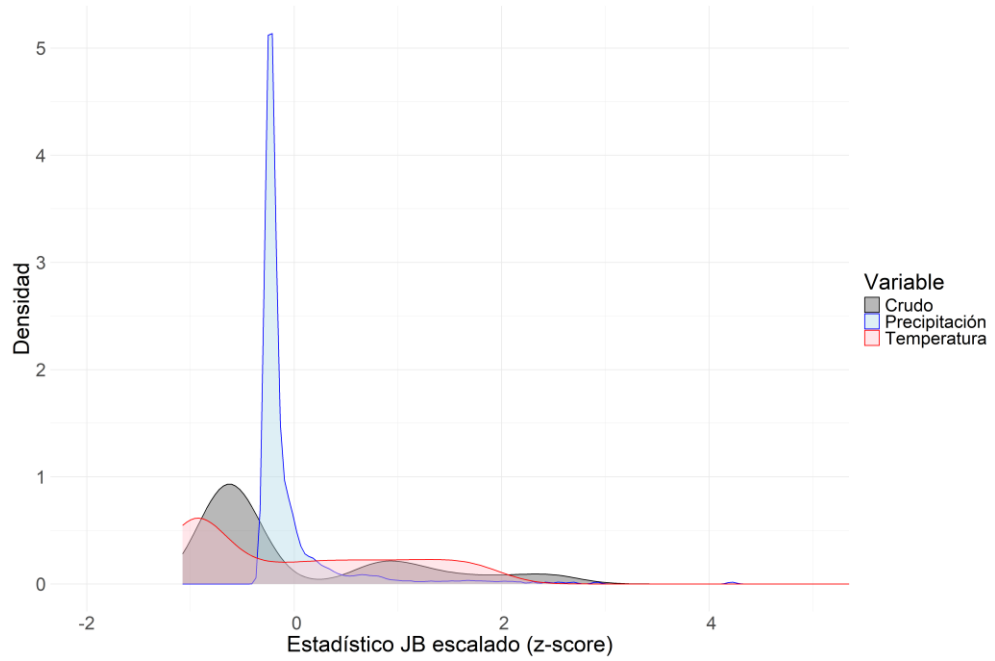
Función de densidad para la prueba AD



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 10

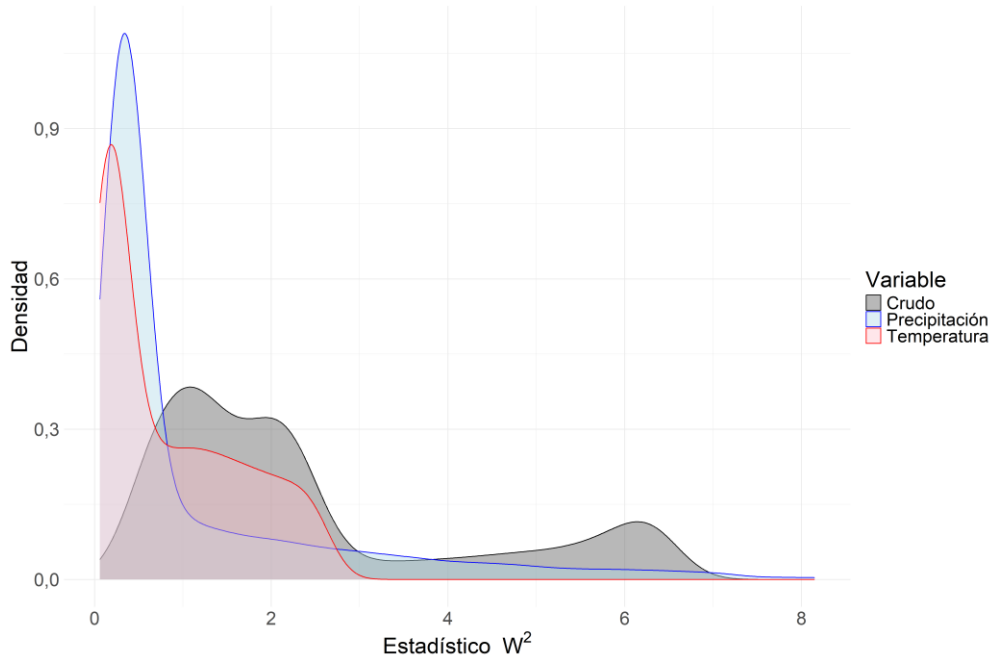
Función de densidad para la prueba JB



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 11

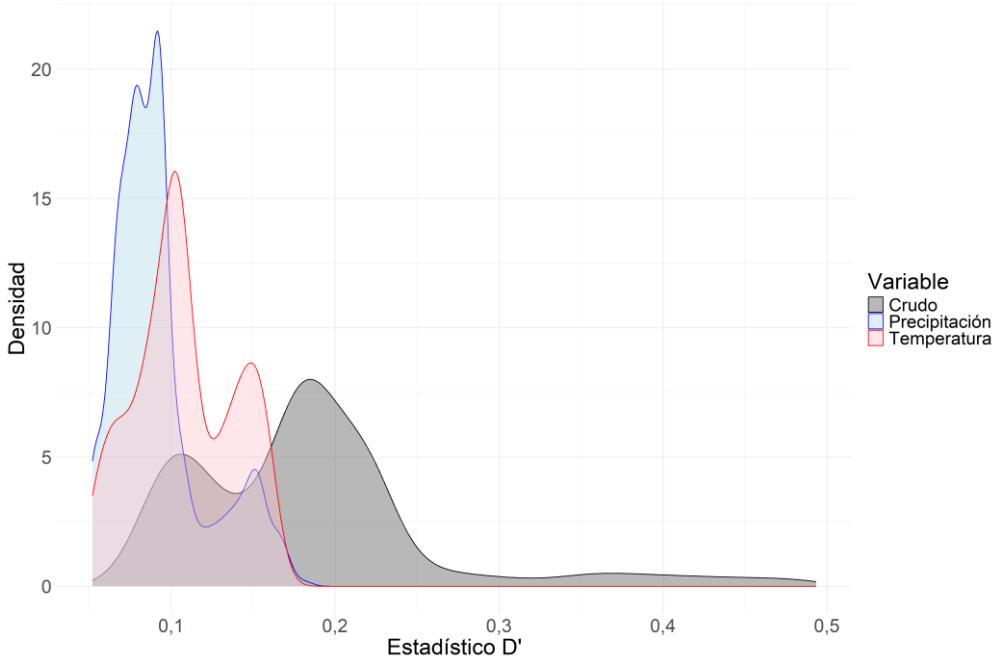
Función de densidad para la prueba CvM



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 12

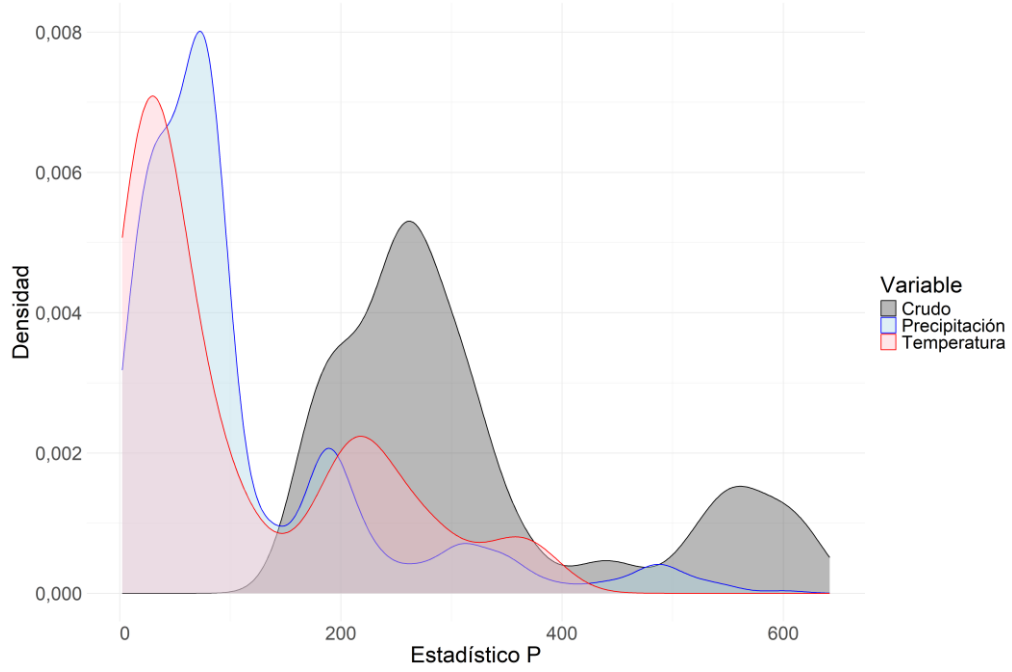
Función de densidad para la prueba M.KS



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 13

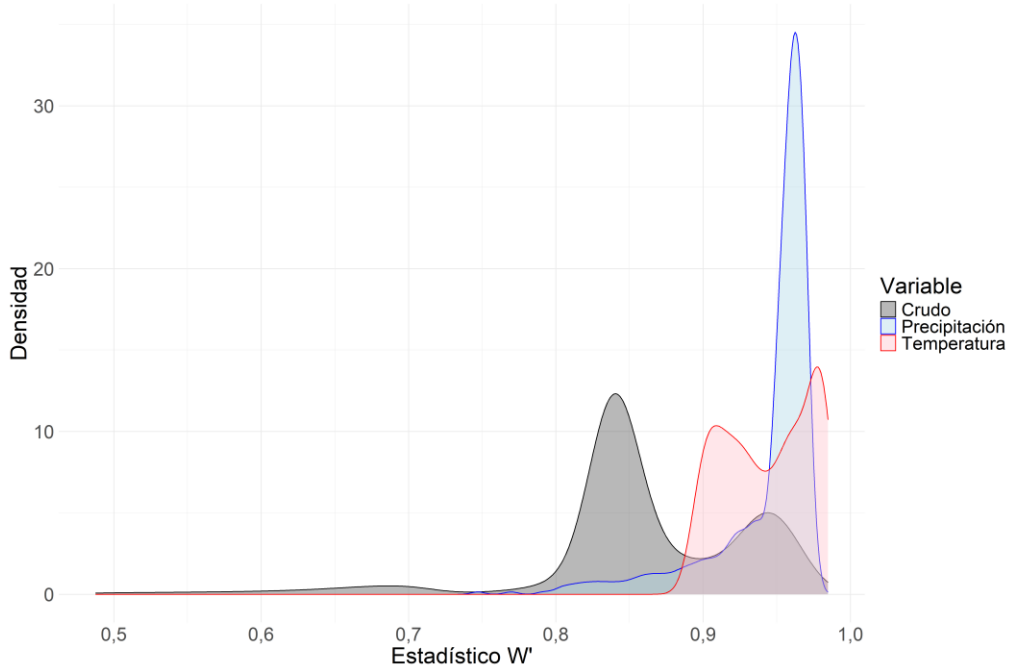
Función de densidad para la prueba P χ^2



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 14

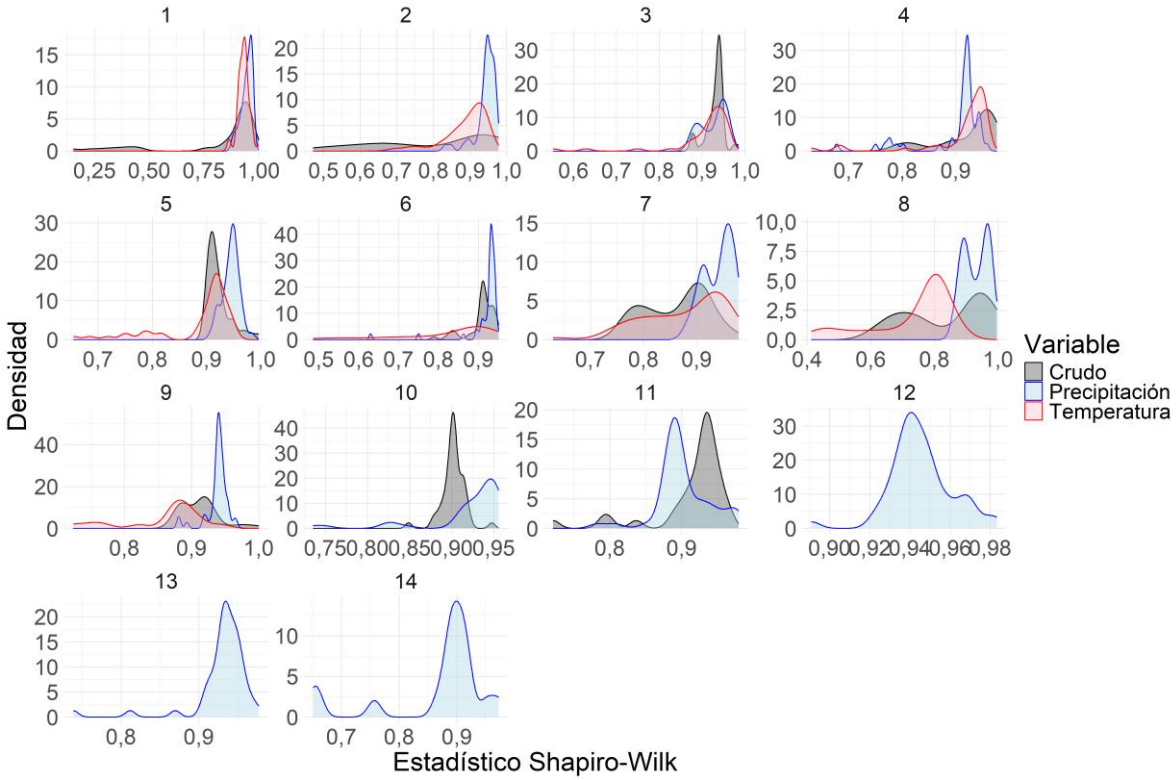
Función de densidad para la prueba SF



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 15

Función de densidad para la prueba SW



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Al analizar los datos de las variables temperatura mínima absoluta, producción de petróleo mensual operativo y precipitación en relación con la función de densidad, se observa que las pruebas KS, M.KS y P χ^2 (ver Figuras 8, 12 y 13) tienen distribuciones asimétricas con varios picos (bimodal y multimodal), dicha variabilidad de los datos depende de la naturaleza de los datos. La temperatura mínima absoluta tiene dispersión en su comportamiento porque la toma de datos abarca las diferentes regiones del Ecuador con sus respectivas estaciones meteorológicas, en otras palabras, la temperatura mínima varía en la región sierra y costa.

En lo que concierne a la producción de petróleo mensual operativo o crudo de igual forma existen diferentes niveles de producción. Esto se debe a los factores técnicos y geológicos. La capacidad limitada de extracción de los yacimientos en el trascurso del

tiempo, o a su vez por el mantenimiento preventivo, predictivo o correctivo de la maquinaria empleada. En lo referente a las precipitaciones, influyen factores atmosféricos y climáticos para que los datos originen variabilidad con respecto a la media, entre los principales se encuentran la humedad, fenómenos provocados por las corrientes oceánicas que afectan la temperatura y modifican los patrones de lluvia, con mayor influencia en la región costa.

Las pruebas AD, JB, CvM y SF (ver Figuras 9, 10, 11 y 14) con respecto a la variable precipitaciones tienen un comportamiento diferente. La curva tiene mejor distribución debido a que solo existe un pico (unimodal), a pesar de que existe asimetría positiva (cola de distribución hacia la derecha) en las pruebas AD, JB normalizado y CvM. La normalización de la prueba JB permitió reajustar las escalas de las variables y tener mejor visualización de las curvas de densidad y se basa en la trasformación de cada dato ($\bar{x} = 0$ y $s = 1$) mediante $z = \frac{x - \bar{x}}{s}$, donde z representan el valor normalizado o estandarizado, de esta manera se logró la trasformación sin modificar la validez de la prueba, manteniendo la asimetría y la curtosis. Con respecto a la prueba SF de igual forma tiene una distribución asimétrica, pero con un sesgo negativo (cola de distribución hacia la izquierda). Para las pruebas mencionadas en este apartado, la variable precipitación tiende a generar una campana de Gauss.

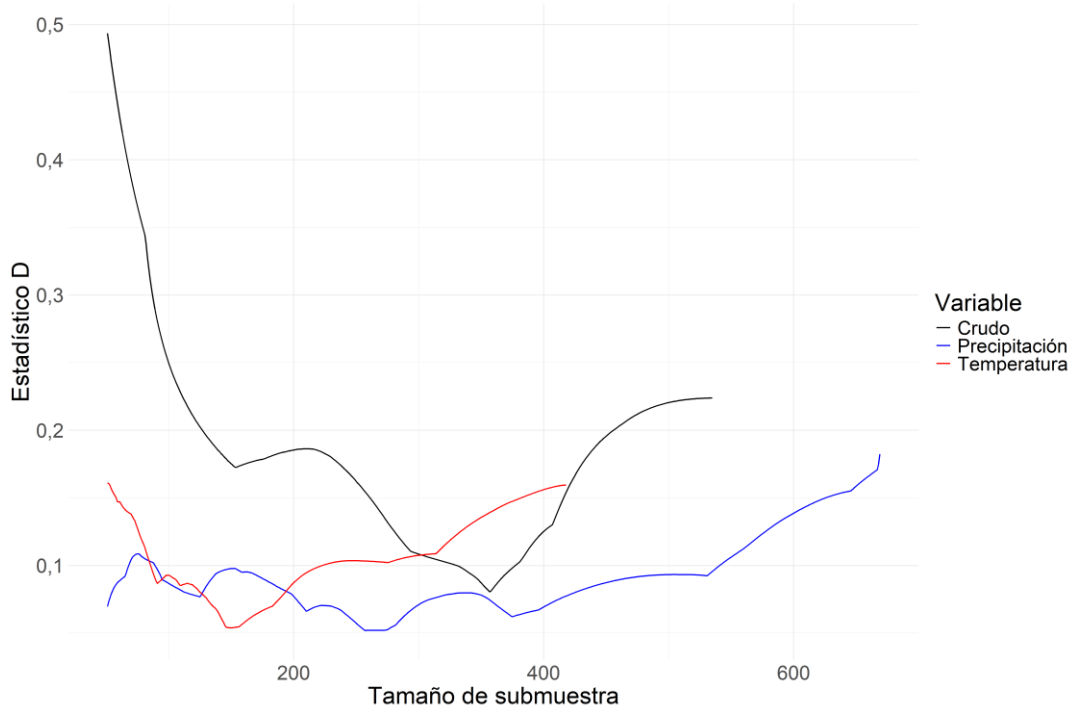
La prueba SW, las variables temperatura mínima absoluta, producción de petróleo mensual operativo y precipitación tienen las curvas de densidad concentradas con valores del estadístico entre 0,9 a 1. En todos los bloques, las variables tienen variabilidad en la forma de distribución, esta alta dispersión puede originarse por la presencia de valores atípicos o menor homogeneidad.

Las variables precipitación y temperatura mínima absoluta tienden aproximarse a una distribución normal especialmente en los bloques 1, 2, 4 y 5. Por otro lado, la variable producción de petróleo mensual operativo tiene un comportamiento similar en el bloque 5.

El uso de submuestras pequeñas permitió que las variables en base al estadístico se estabilicen de mejor manera a una distribución normal. En este contexto, aunque la variabilidad existente ocasiona efectos en la distribución de los datos, la adaptabilidad de las variables a la prueba SW fue significativa con respecto a las otras pruebas aplicadas a submuestras superiores a 50 observaciones.

Figura 16

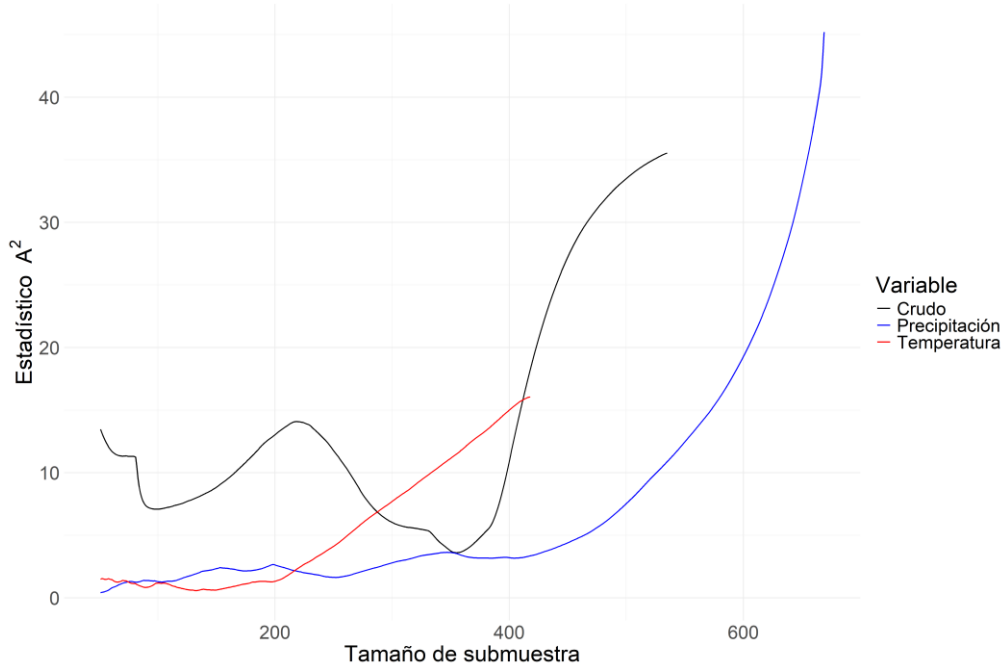
Comportamiento de la prueba KS



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 17

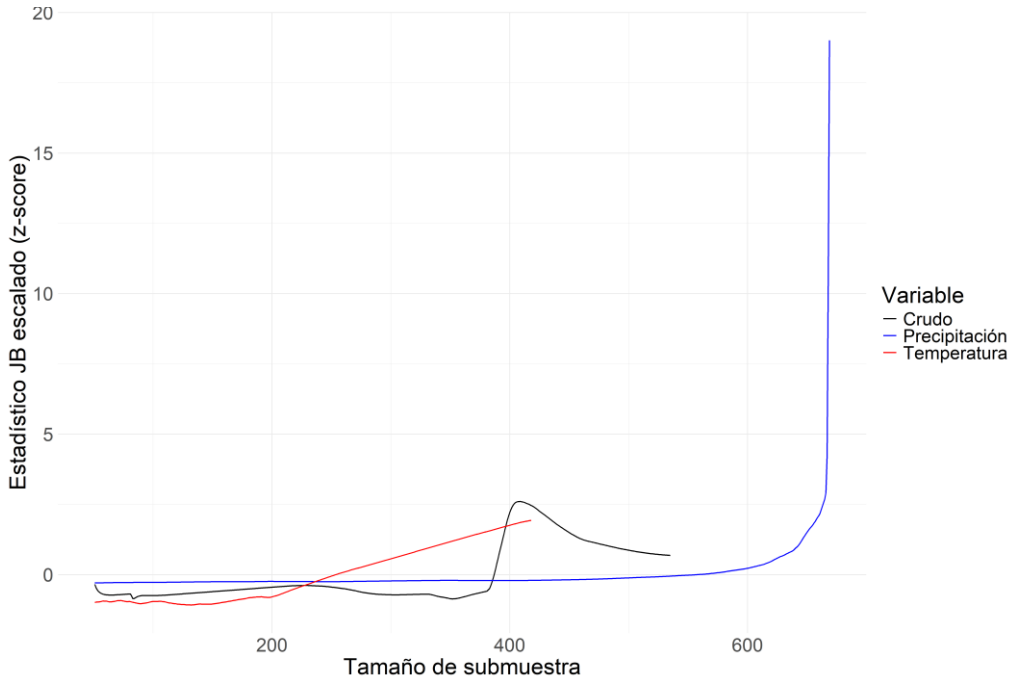
Comportamiento de la prueba AD



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 18

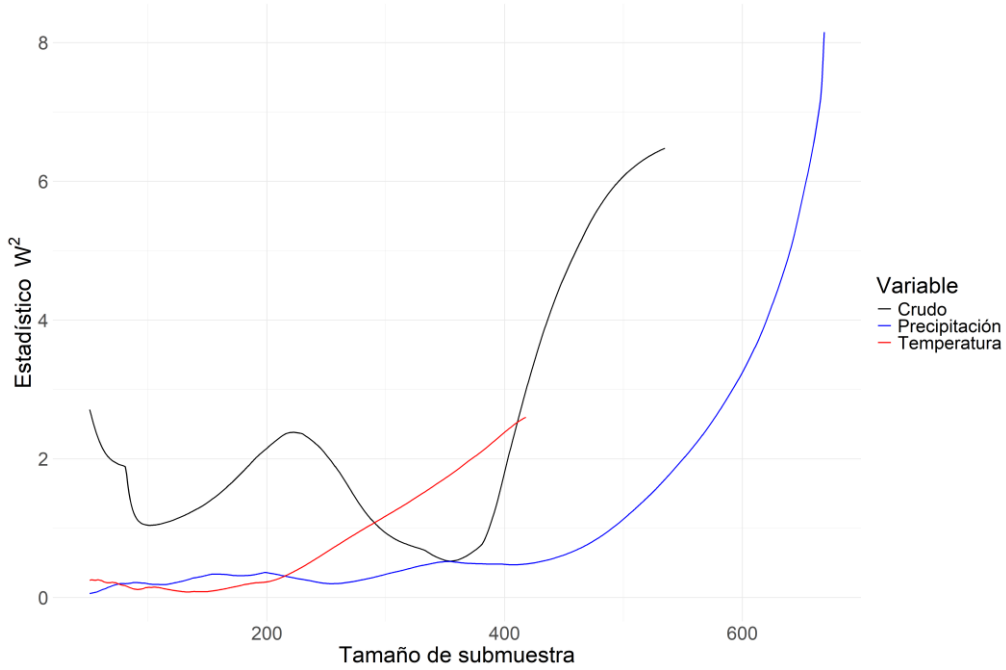
Comportamiento de la prueba JB



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 19

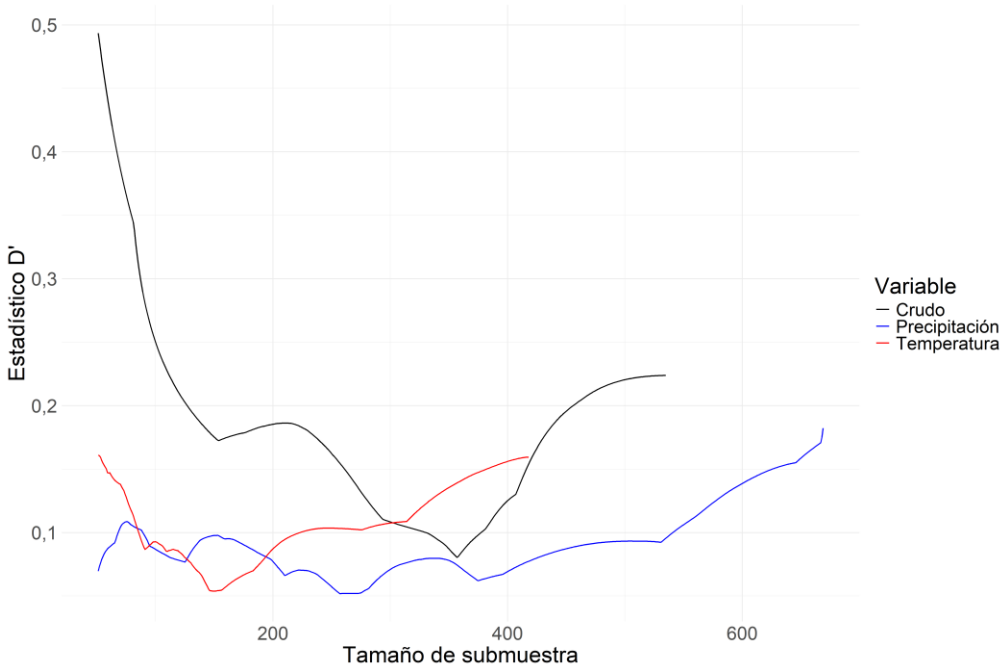
Comportamiento de la prueba CvM



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 20

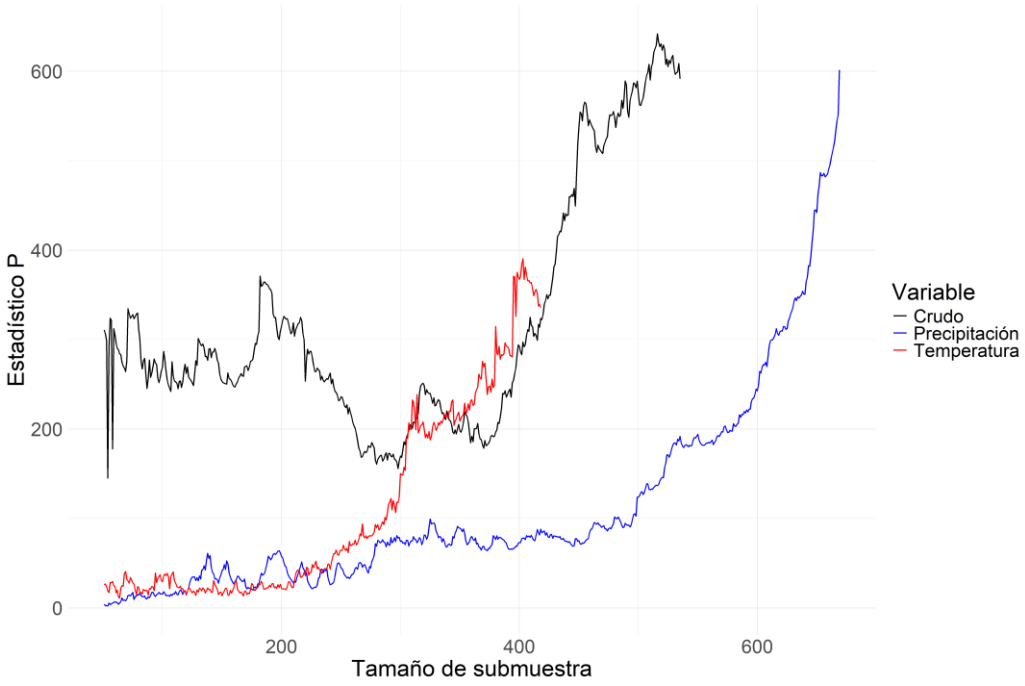
Comportamiento de la prueba M.KS



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 21

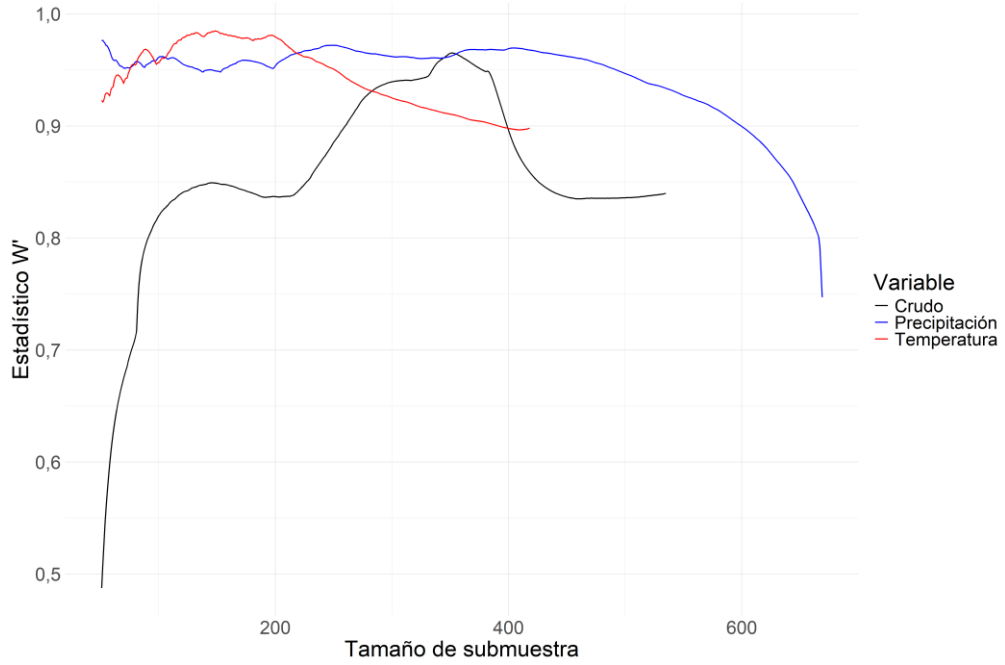
Comportamiento de la prueba P χ^2



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 22

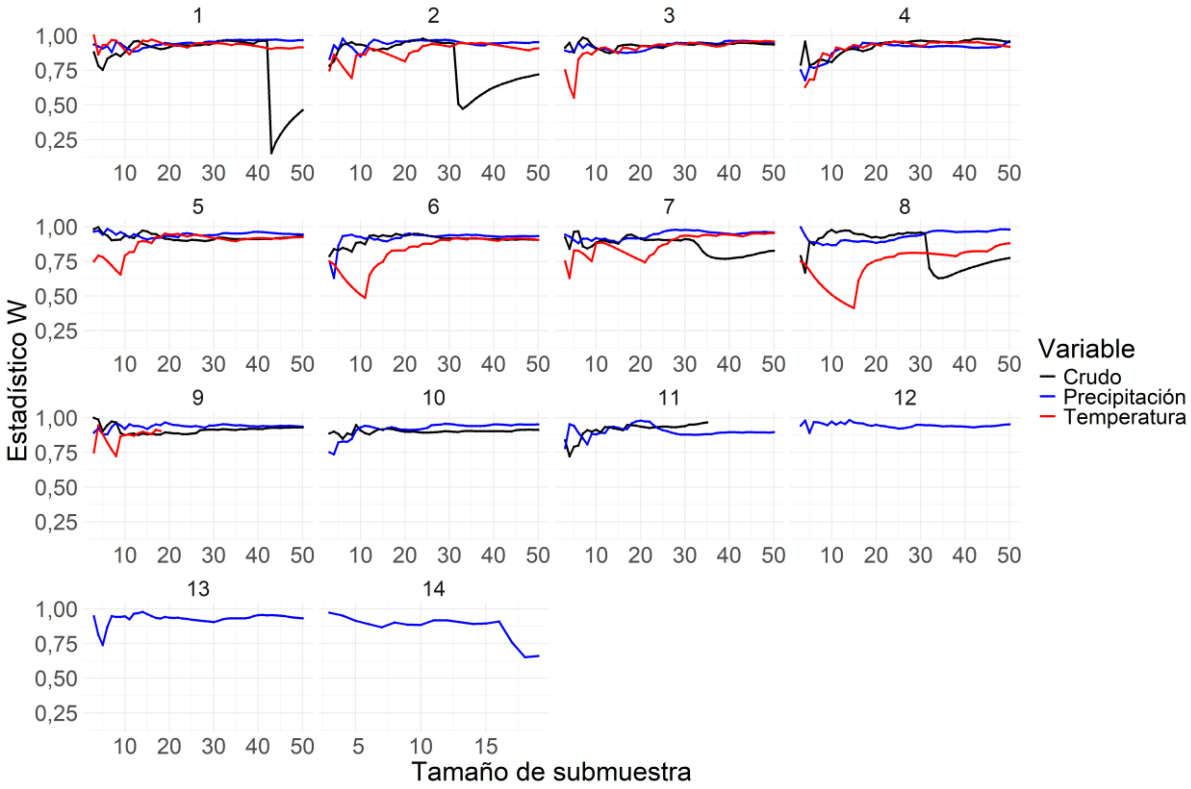
Comportamiento de la prueba SF



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Figura 23

Comportamiento de la prueba SW



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Las pruebas de normalidad univariadas utilizadas en la investigación tienen un comportamiento fuertemente influenciado por el tamaño de la muestra. Los estadísticos de las pruebas KS, AD, JB, CvM, M.KS, P χ^2 y SF se calcularon a partir de una muestra inicial de 51 datos u observaciones, incrementando una observación para cada submuestra, hasta completar los tamaños muestrales definidas en la tabla 4 según las variables.

Las primeras submuestras generan un crecimiento o decrecimiento inicial, dependiendo de la variable analizada. A medida que el tamaño de las submuestras va incrementando, los estadísticos tienen a estabilizarse. Sin embargo, al seguir incrementando los tamaños de las submuestras en las pruebas KS, AD, JB, CvM, M.KS y P χ^2 (ver Figuras 16 a 21) los valores de los estadísticos tienden a elevarse considerablemente conforme se aproxima al tamaño de muestra total. En contraste, la prueba SF tiene un comportamiento

inverso, mientras se aproxima progresivamente el tamaño de la submuestra a la muestra total, el estadístico tiende a decrecer o disminuir (ver Figura 22). Los estadísticos de las pruebas KS y M.KS son equivalentes, su diferencia está en el proceso de calcular el p-valor, KS clásica se basa en una aproximación teórica, mientras M.KS modificada utiliza un método ajustado para estimar la incertidumbre.

En lo que corresponde a la prueba SW, el análisis por bloques, utilizando submuestras decrecientes hasta completar el total de observaciones de cada variable, muestra que el estadístico tiene una mejor estabilidad con respecto a las submuestras cuando existen decrementos de una observación de manera progresiva. En la mayoría de los bloques, el estadístico tiende a acercarse a 1 (ver Figura 23), esto indica que la distribución de cada submuestra se aproxima a la normal.

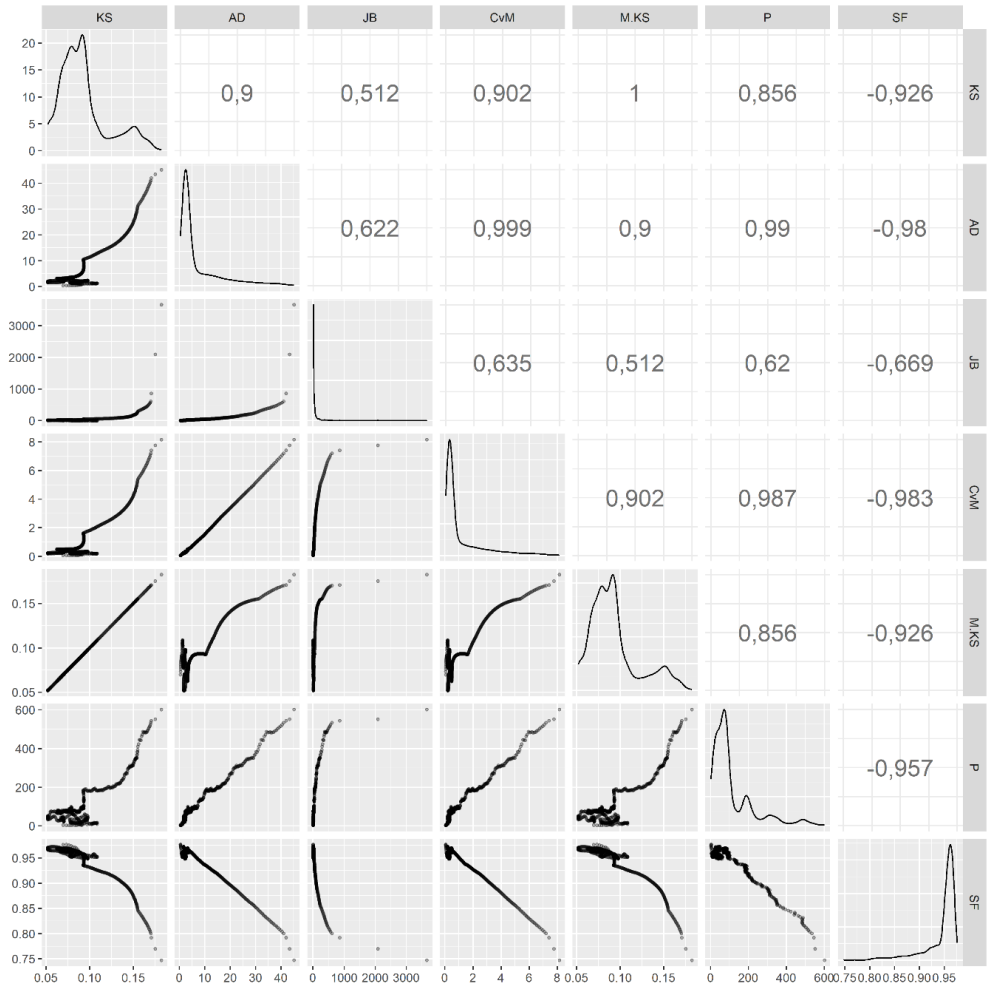
Sin embargo, en los bloques 1 y 2, se observa que la variable producción de petróleo mensual operativo tiene desviaciones notables en las primeras submuestras, que luego se estabilizan al reducir el tamaño muestral. De manera similar, la variable temperatura mínima absoluta muestra inestabilidad relevante en los bloques 3, 5, 6 y 8 especialmente en las últimas submuestras. Por otro lado, la variable precipitación tiene una alta estabilidad en el estadístico a lo largo de todos los bloques, lo que sugiere un comportamiento aceptable a la normalidad para submuestras pequeñas. La inestabilidad detectada en las variables podría estar asociadas a valores atípicos que pueden afectar levemente la aproximación a la distribución normal.

Las pruebas de normalidad con sus respectivos estadísticos, están orientadas a medir distancias o momentos estadísticos, dependiendo del enfoque con el que se analice el conjunto de datos. Es importante explorar cada estadístico de forma individual para identificar patrones o tendencia en su comportamiento según las variables de estudio. Sin

embargo, es adecuado, realizar una exploración por pares con el objetivo de evaluar la correlación y afinidad entre estadísticos.

Figura 24

Matriz de correlación para la variable precipitación



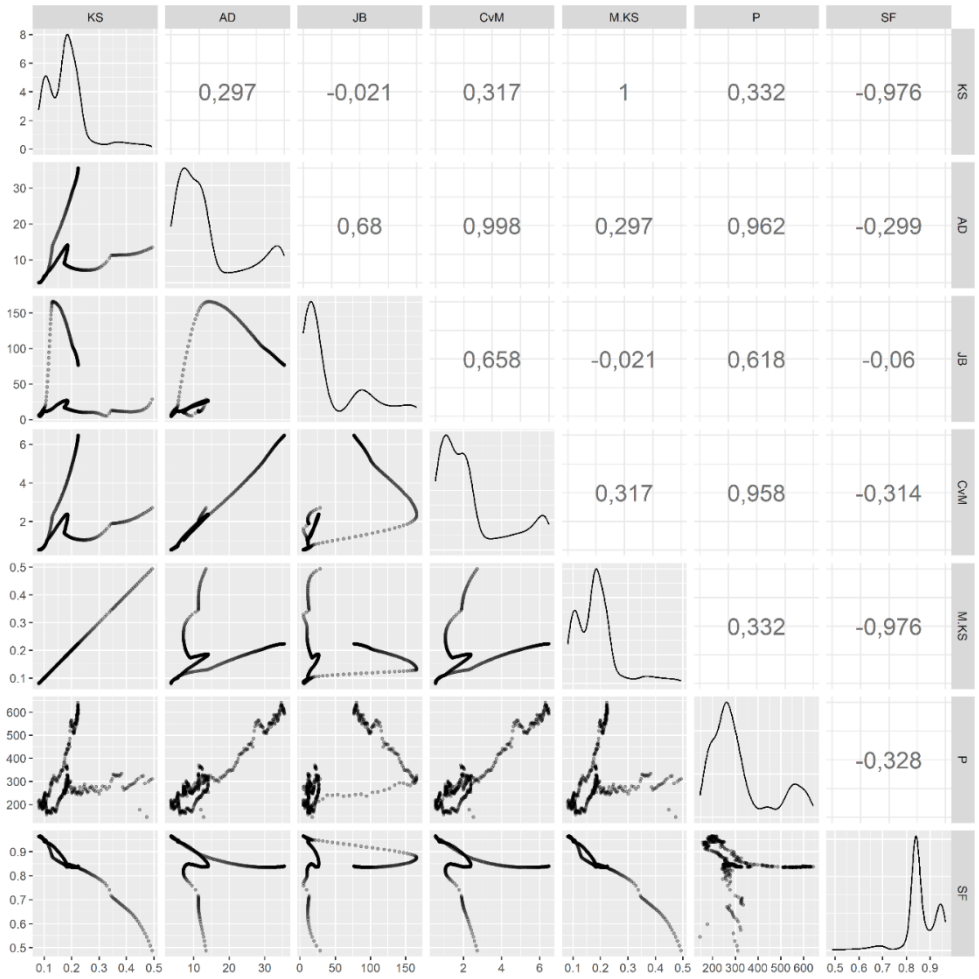
Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Si se habla de correlación se debe tener en cuenta que su estadístico está acotado entre $[-1,1]$ donde -1 indica una correlación perfecta negativa y 1 una correlación perfecta positiva. Para la variable precipitación (ver Figura 24) mediante una matriz de correlación entre estadísticos, se obtuvo que las pruebas KS, AD, CvM, M.KS, P χ^2 y SF tienen una alta correlación al compararse entre sí. El estadístico de la prueba JB tiene una correlación moderada en comparación con los demás estadísticos, esto sucede, debido a que la prueba

JB está direccionado en medir el tercer y cuarto momentos estadístico (asimetría y curtosis) y las otras pruebas se centran en medir distancias entre la distribución empírica y la teórica. Por otro lado, las pruebas KS y M.KS se obtuvo una correlación perfecta, esto se debe a que la prueba M.KS es una modificación o adaptación de la prueba KS, la diferencia está en el cálculo del valor de p.

Figura 25

Matriz de correlación para la variable crudo



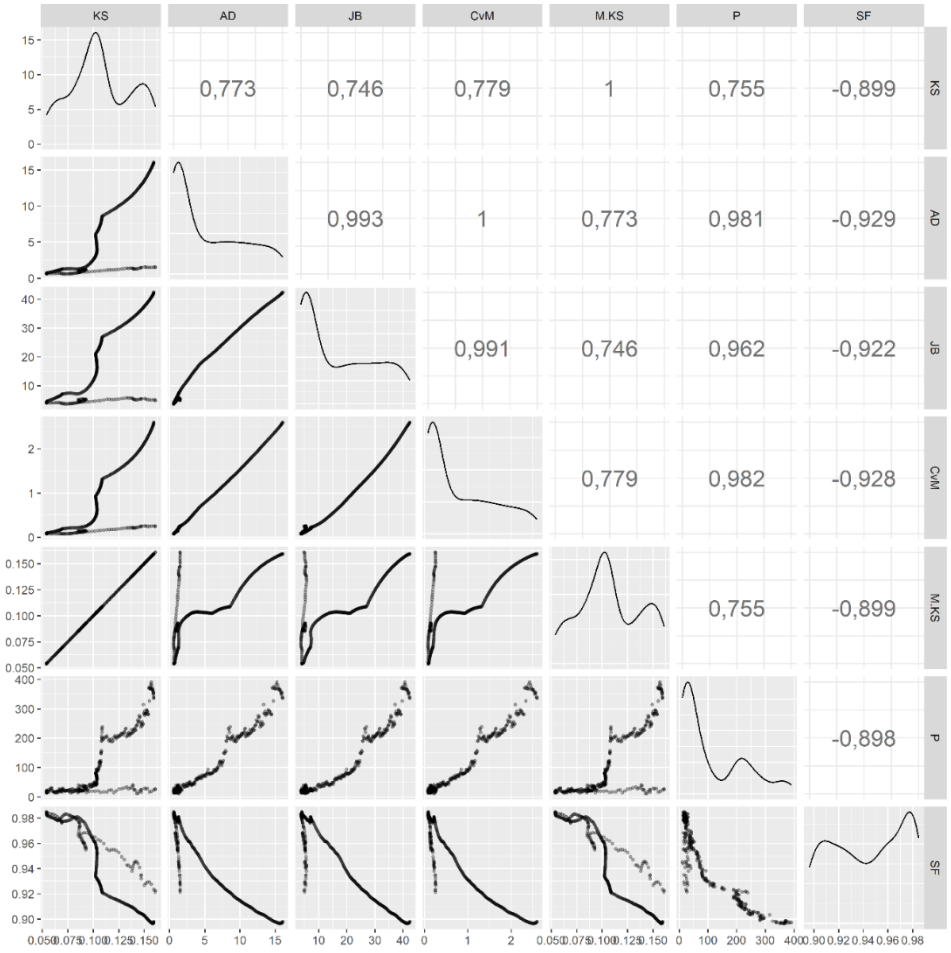
Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Para la variable producción de petróleo mensual operativo (ver Figura 25) el comportamiento de los estadísticos difiere considerablemente. La prueba KS tiene una alta correlación con las pruebas M.KS (correlación perfecta, como se explicó previamente) y SF, sin embargo, la correlación es baja en los demás estadísticos. La prueba AD mantiene una

alta correlación con las pruebas CvM y P χ^2, mientras que con el resto la correlación es baja y moderada. La prueba JB presenta una correlación baja y moderada con respecto a los otros estadísticos (evidente porque mide momentos). Por último, se observa una fuerte correlación entre las pruebas CvM y P χ^2.

Figura 26

Matriz de correlación para la variable temperatura



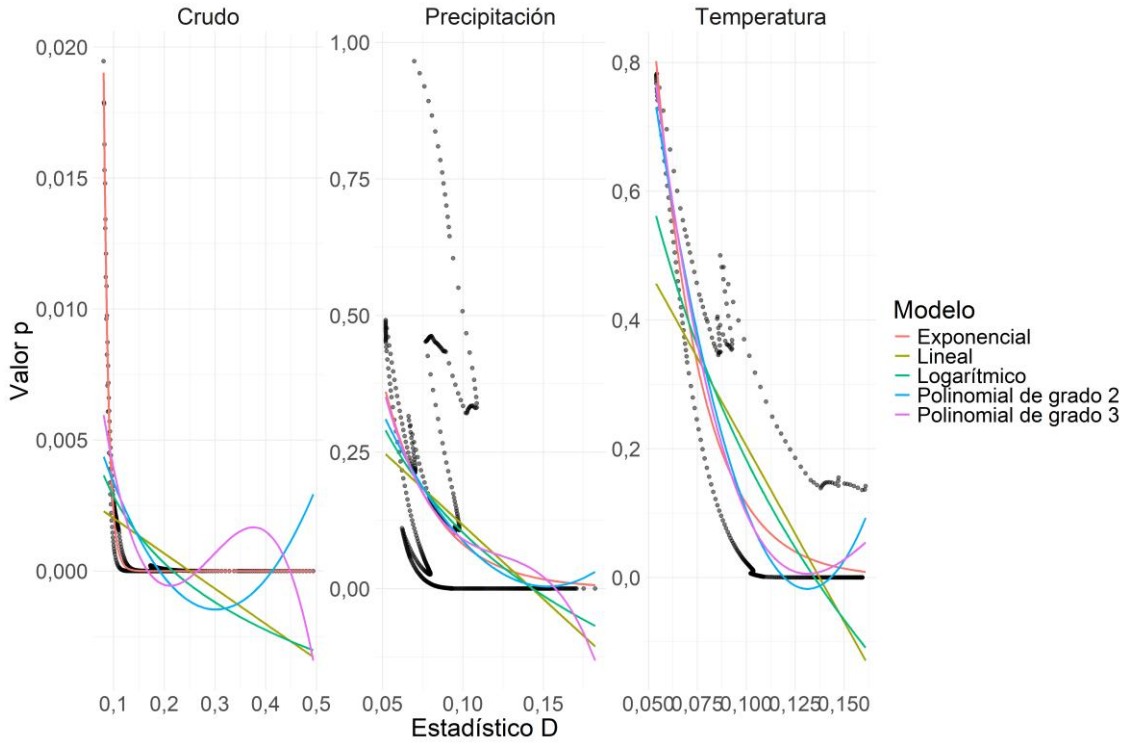
Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

La variable temperatura mínima absoluta (ver Figura 26) la correlación entre estadísticos varía de moderada a alta, tanto positivas como negativas. Esto sugiere una consistencia o mayor concordancia en las pruebas de normalidad, evidenciando un mejor comportamiento general de los estadísticos al evaluar la distribución de esta variable.

Es importante señalar que, en las figuras 24, 25 y 26, la matriz de correlación presenta, en la diagonal principal, las funciones de densidad de cada estadístico según la variable analizada. Debajo de la diagonal principal se encuentran los gráficos de dispersión entre pares, mientras que en la parte superior se muestran los coeficientes de correlación de Pearson correspondientes a cada par, indicando el grado de asociación entre estadísticos. Esta representación gráfica permite obtener una visión más clara sobre la relación de los estadísticos de las diferentes pruebas de normalidad univariada y así facilitar la interpretación de los resultados, identificación de patrones que se ajustan a un modelo determinado, corroboración de similitudes y diferencias entre pruebas que han sido modificadas o reajustadas y observación del comportamiento de los datos cuando se utilizan pruebas en medidas de distancia, en comparación con aquellas centradas en momentos estadísticos.

Figura 27

Ajuste de curva – valor de p y el estadístico D



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 6

Comparación de modelos - valor de p y estadístico D

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,1634	0,1621	17,9472
Polinomial de grado 2	0,1856	0,183	17,4715
Polinomial de grado 3	0,1953	0,1914	17,2635
Exponencial	0,1966	0,194	17,2356
Logarítmico	0,1834	0,1821	17,5191
Crudo			
Lineal	0,1632	0,1615	0,0027
Polinomial de grado 2	0,3798	0,3773	0,002
Polinomial de grado 3	0,5081	0,505	0,0016
Exponencial	0,9784*	0,9784	$1e^{-4}$
Logarítmico	0,3018	0,3003	0,0022
Temperatura			
Lineal	0,5326	0,5313	8,5229
Polinomial de grado 2	0,7972	0,7961	3,6974
Polinomial de grado 3	0,8019*	0,8003	3,612
Exponencial	0,7857	0,7845	3,908
Logarítmico	0,6438	0,6428	6,4957

Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

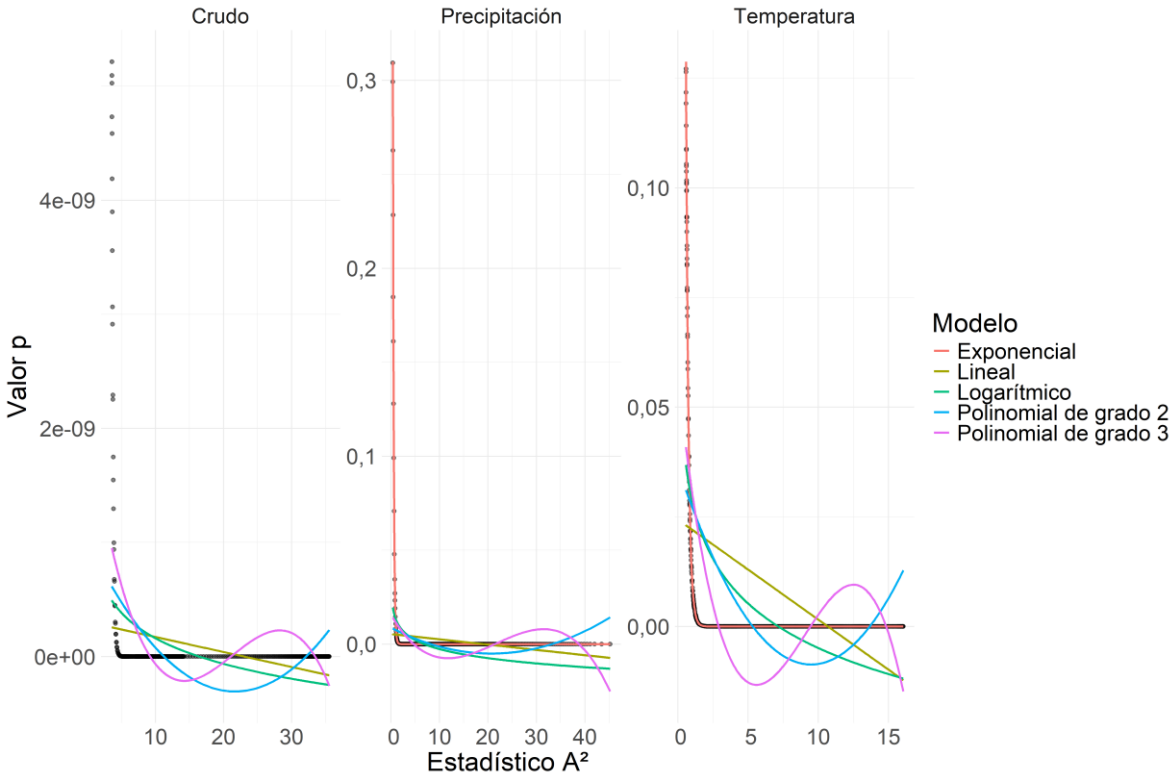
En lo que corresponde a la prueba KS, las variables producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante modelos exponencial (ver ecuación 35) y polinomial de grado 3 (ver ecuación 36) respectivamente. En el caso al modelo exponencial, se produce un decrecimiento debido al exponente negativo, esto significa que el valor de p disminuye exponencialmente a medida que el estadístico D aumenta, y viceversa. Por otro lado, en el modelo polinomial de grado 3, la interpretación de los coeficientes influye a medida que el estadístico D varía, el valor de p puede disminuir o incrementar dependiendo del equilibrio de los términos lineales, cuadrático y cúbico. Los modelos exponencial y polinomial de grado 3 están acotados con $D \in [0,0805; 0,4935]$ y $D \in [0,0539; 0,1610]$ respectivamente.

$$\hat{p} = 135,0617\, e^{-110,0911 D} \tag{35}$$

$$\hat{p} = 0{,}1621 - 3{,}1162\,D + 2{,}1967\,D^2 - 0{,}2921\,D^3 \qquad (36)$$

Figura 28

Ajuste de curva – valor de p y estadístico A^2



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 7

Comparación de modelos - valor de p y estadístico A^2

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,0098	0,0082	0,3926
Polinomial de grado 2	0,0245	0,0213	0,3868
Polinomial de grado 3	0,0523	0,0477	0,3758
Exponencial	1*	1	0
Logarítmico	0,074	0,0725	0,3672
Crudo			
Lineal	0,0394	0,0374	0
Polinomial de grado 2	0,1394	0,1358	0
Polinomial de grado 3	0,2316	0,2268	0
Exponencial	NA	NA	NA
Logarítmico	0,1045	0,1027	0
Temperatura			
Lineal	0,1697	0,1656	0,2354
Polinomial de grado 2	0,2793	0,2753	0,2039
Polinomial de grado 3	0,4002	0,3952	0,1697

Exponencial	0,9999*	0,9999	0
Logarítmico	0,3442	0,3424	0,1855

Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

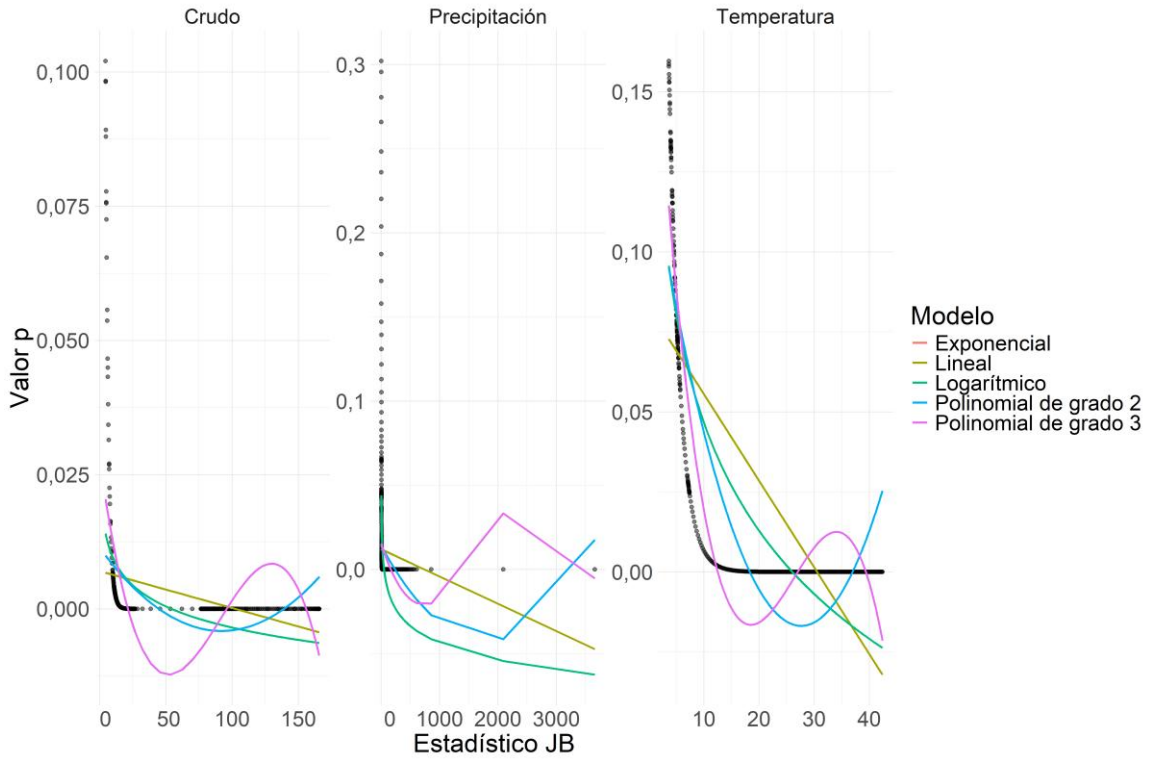
En lo que respecta a la prueba AD, las variables precipitación y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante modelos exponenciales (ver ecuaciones 37 y 38) en ambos casos. En los modelos exponenciales, se produce un decrecimiento debido al exponente negativo, esto significa que el valor de p disminuye exponencialmente a medida que el estadístico A^2 aumenta, y viceversa. Los modelos exponenciales están acotados con $A^2 \in [0,4227; 45,2029]$ y $A^2 \in [0,5825; 16,0537]$ respectivamente. Es importante indicar que, la notación matemática del estadístico es A^2, el superíndice no indica que al estadístico se debe elevar al cuadrado.

$$\hat{p} = 3,4294\, e^{-5,6865\left(A^2\right)} \tag{37}$$

$$\hat{p} = 3,5297\, e^{-5,6845\left(A^2\right)} \tag{38}$$

Figura 29

Ajuste de curva – valor de p y estadístico JB



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 8

Comparación de modelos - valor de p y estadístico JB

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,0072	0,0056	0,8001
Polinomial grado 2	0,0224	0,0192	0,7879
Polinomial grado 3	0,0322	0,0275	0,78
Exponencial	NA	NA	NA
Logarítmico	0,2003*	0,199	0,6445
Crudo			
Lineal	0,0568	0,0548	0,0824
Polinomial grado 2	0,113	0,1093	0,0775
Polinomial grado 3	0,2797*	0,2753	0,0629
Exponencial	NA	NA	NA
Logarítmico	0,1732	0,1715	0,0722
Temperatura			
Lineal	0,5315	0,5302	0,3931
Polinomial grado 2	0,765	0,7637	0,1972
Polinomial grado 3	0,8902*	0,8893	0,0921
Exponencial	NA	NA	NA
Logarítmico	0,7436	0,7429	0,2151

Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

Con respecto a la prueba JB, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante modelos logarítmico (ver ecuación 39) y polinomial de grado 3 (ver ecuaciones 40 y 41) respectivamente. En el caso al modelo logarítmico, el coeficiente negativo que acompaña a la expresión logarítmica indica que, el valor de p disminuye a medida que el estadístico JB aumenta, y viceversa. Por otro lado, en el modelo polinomial de grado 3, la interpretación de los coeficientes influye a medida que el estadístico JB varía, el valor de p puede disminuir o incrementar dependiendo del equilibrio de los términos lineales, cuadrático y cúbico. Los modelos logarítmico y polinomial de grado 3 están acotados con $JB \in [2,394; 3661,102]$, $JB \in [4,565; 165,741]$ y $JB \in [3,670; 42,403]$ respectivamente.

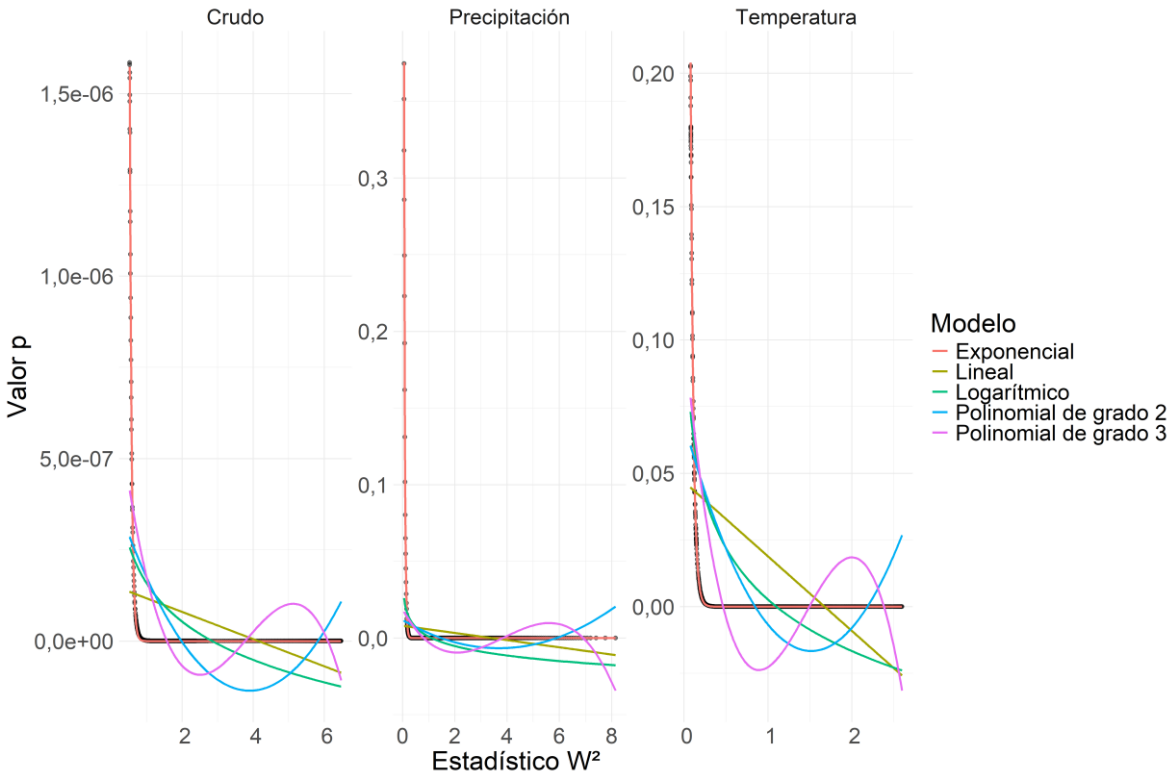
$$\hat{p} = 0,0567 - 0,0145 \log(JB) \tag{39}$$

$$\hat{p} = 0,004 - 0,0704\,JB + 0,0701\,JB^2 - 0,1207\,JB^3 \tag{40}$$

$$\hat{p} = 0,0352 - 0,6677\,JB + 0,4426\,JB^2 - 0,3241\,JB^3 \tag{41}$$

Figura 30

Ajuste de curva – valor de p y estadístico W^2



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 9

Comparación de modelos - valor de p y estadístico W^2

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,0126	0,011	0,6455
Polinomial de grado 2	0,028	0,0249	0,6354
Polinomial de grado 3	0,0535	0,0489	0,6188
Exponencial	1*	1	0
Logarítmico	0,0858	0,0843	0,5977
Crudo			
Lineal	0,0701	0,0681	0
Polinomial de grado 2	0,2161	0,2128	0
Polinomial de grado 3	0,3361	0,3319	0
Exponencial	0,9998*	0,9998	0
Logarítmico	0,1867	0,185	0
Temperatura			
Lineal	0,1975	0,1953	0,722
Polinomial de grado 2	0,3303	0,3266	0,6025
Polinomial de grado 3	0,4664	0,462	0,4801
Exponencial	1*	1	0
Logarítmico	0,4172	0,4156	0,5243

Nota. Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

En lo que corresponde a la prueba CvM, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante modelos exponenciales (ver ecuaciones 42 a 44) en todos casos. En los modelos exponenciales, se produce un decrecimiento debido al exponente negativo, esto significa que el valor de p disminuye exponencialmente a medida que el estadístico W^2 aumenta, y viceversa. Los modelos exponenciales están acotados con $W^2 \in [0,0597;8,1505]$, $W^2 \in [0,5255;6,4766]$ y $W^2 \in [0,0804;2,5972]$ respectivamente. Es importante indicar que, la notación matemática del estadístico es W^2, el superíndice no indica que al estadístico se debe elevar al cuadrado.

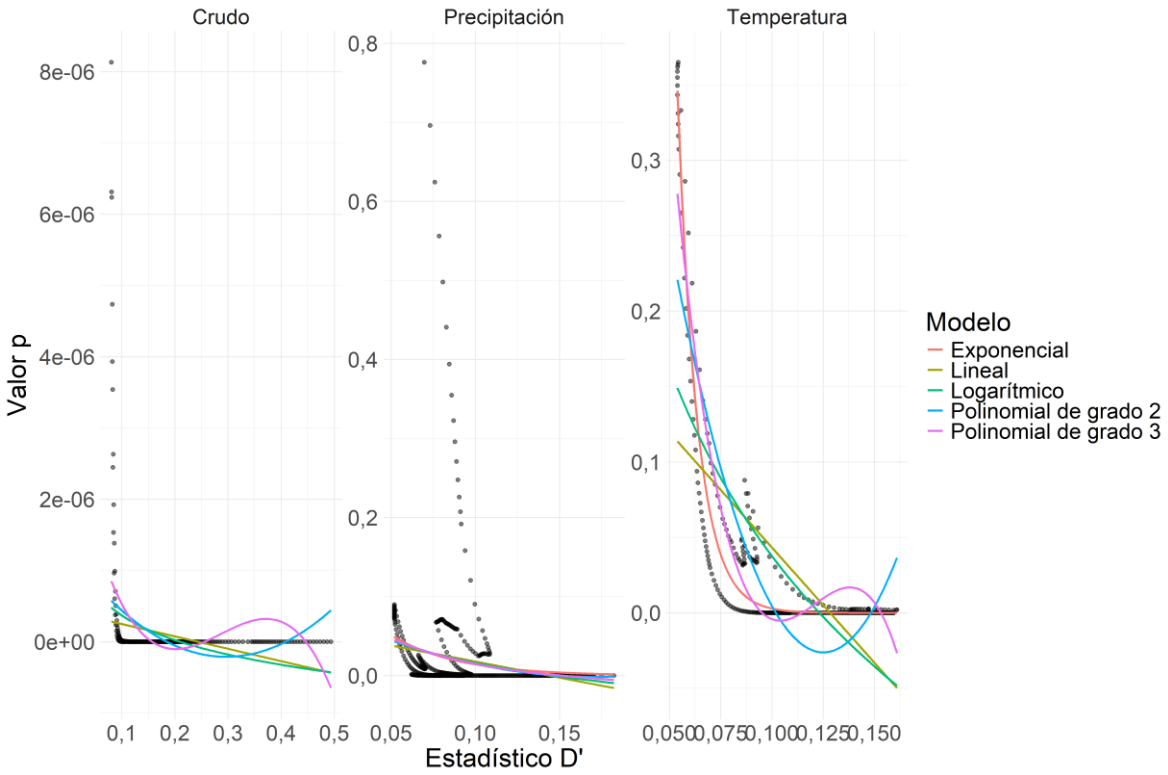
$$\hat{p} = 2,3317\, e^{-30,5783\left(W^2\right)} \tag{42}$$

$$\hat{p} = 0,0485\, e^{-19,6693\left(W^2\right)} \tag{43}$$

$$\hat{p} = 2,4986\, e^{-31,1508\left(W^2\right)} \tag{44}$$

Figura 31

Ajuste de curva – valor de p y estadístico D'



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 10

Comparación de modelos - valor de p y estadístico D'

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,0257	0,0241	3,0186
Polinomial de grado 2	0,0274	0,0243	3,0133
Polinomial de grado 3	0,0275	0,0227	3,0132
Exponencial	0,0272	0,0241	3,0139
Logarítmico	0,0275	0,0259	3,0131
Crudo			
Lineal	0,0408	0,0388	0
Polinomial de grado 2	0,1053	0,1015	0
Polinomial de grado 3	0,162	0,1567	0
Exponencial	NA	NA	NA
Logarítmico	0,0834	0,0815	0
Temperatura			
Lineal	0,3773	0,3756	1,2508
Polinomial de grado 2	0,7419	0,7405	0,5184
Polinomial de grado 3	0,8585	0,8573	0,2842
Exponencial	0,936*	0,9356	0,1286
Logarítmico	0,5055	0,5041	0,9933

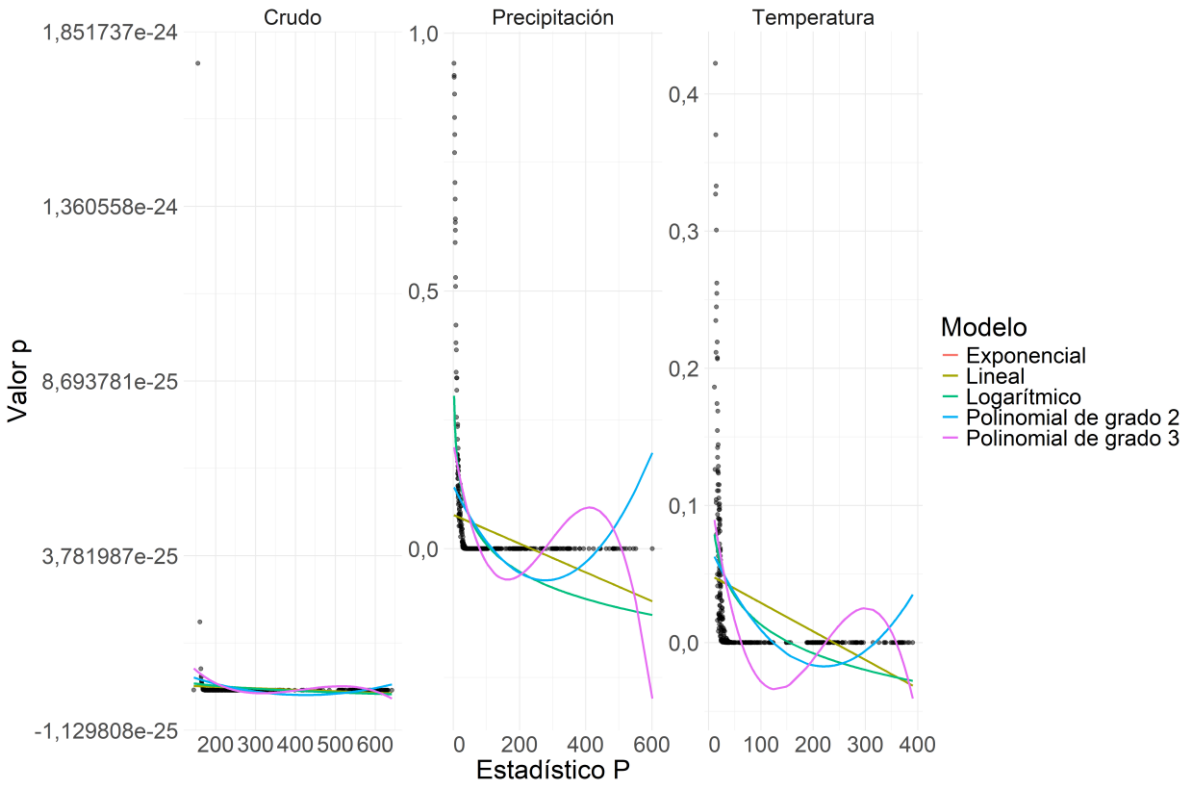
Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

En lo que respecta a la prueba M.KS, la variable temperatura mínima absoluta presentó un mejor ajuste de curva en su primera aproximación mediante un modelo exponencial (ver ecuación 45). El modelo produce un decrecimiento debido al exponente negativo, esto significa que el valor de p disminuye exponencialmente a medida que el estadístico D' aumenta, y viceversa. El modelo exponencial está acotado con $D' \in [0,0539; 0,1610]$.

$$\hat{p} = 93,6107\, e^{-103,8542\left(D'\right)} \tag{45}$$

Figura 32

Ajuste de curva – valor de p y estadístico P



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 11

Comparación de modelos - valor de p y estadístico P

Modelo	R^2	R^2 ajustado	SCE
	Precipitación		
Lineal	0,0637	0,0621	8,9564

Polinomial de grado 2	0,1686	0,1659	7,953
Polinomial de grado 3	0,3068	0,3034	6,6308
Exponencial	NA	NA	NA
Logarítmico	0,3798*	0,3788	5,9327
	Crudo		
Lineal	0,0049	0,0028	0
Polinomial de grado 2	0,0191	0,015	0
Polinomial de grado 3	0,0324	0,0264	0
Exponencial	NA	NA	NA
Logarítmico	0,0092	0,0071	0
	Temperatura		
Lineal	0,1395	0,1372	1,1687
Polinomial de grado 2	0,2254	0,2212	1,052
Polinomial de grado 3	0,3467*	0,3414	0,8873
Exponencial	NA	NA	NA
Logarítmico	0,2829	0,281	0,9739

Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

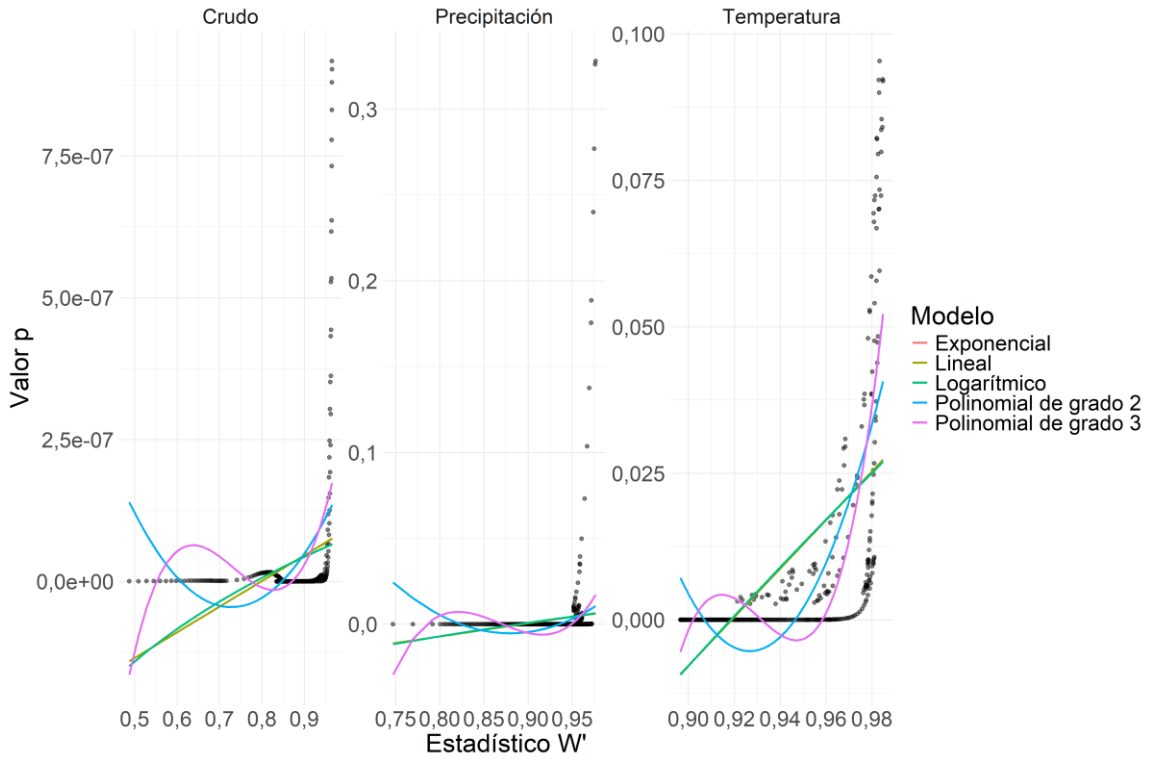
Con respecto a la prueba P χ^2, las variables precipitación y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante modelos logarítmico (ver ecuación 46) y polinomial de grado 3 (ver ecuación 47) respectivamente. En el caso al modelo logarítmico, el coeficiente negativo que acompaña a la expresión logarítmica indica que, el valor de p disminuye a medida que el estadístico P aumenta, y viceversa. Por otro lado, en el modelo polinomial de grado 3, la interpretación de los coeficientes influye a medida que el estadístico P varía, el valor de p puede disminuir o incrementar dependiendo del equilibrio de los términos lineales, cuadrático y cúbico. Los modelos logarítmico y polinomial de grado 3 están acotados con $P \in [2,296; 601,453]$ y $P \in [11,28; 390,59]$ respectivamente.

$$\hat{p} = 0,3598 - 0,0763 \log(P) \tag{46}$$

$$\hat{p} = 0,0271 - 0,4353\,P + 0,3416\,P^2 - 0,4059\,P^3 \tag{47}$$

Figura 33

Ajuste de curva – valor de p y estadístico W'



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 12

Comparación de modelos - valor de p y estadístico W'

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,0106	0,009	0,4465
Polinomial de grado 2	0,0246	0,0214	0,4402
Polinomial de grado 3	0,0435	0,0389	0,4316
Exponencial	NA	NA	NA
Logarítmico	0,01	0,0083	0,4468
Crudo			
Lineal	0,0885	0,0866	0
Polinomial de grado 2	0,193	0,1896	0
Polinomial de grado 3	0,2595	0,2549	0
Exponencial	NA	NA	NA
Logarítmico	0,0702	0,0683	0
Temperatura			
Lineal	0,3336	0,3318	0,108
Polinomial de grado 2	0,5072	0,5045	0,0799
Polinomial de grado 3	0,6019*	0,5986	0,0645
Exponencial	NA	NA	NA
Logarítmico	0,328	0,3262	0,1089

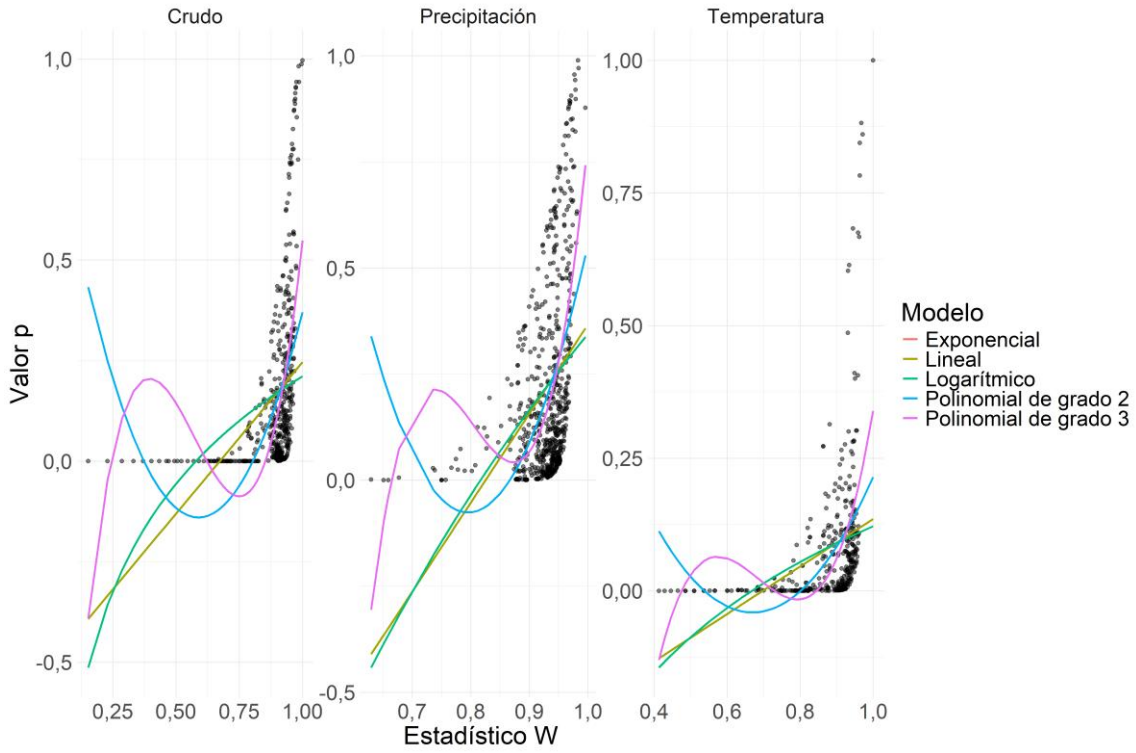
Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

En lo referente a la prueba SF, la variable temperatura mínima absoluta presentó un mejor ajuste de curva en su primera aproximación mediante un modelo polinomial de grado 3 (ver ecuación 48). En el modelo polinomial, la interpretación de los coeficientes influye a medida que el estadístico W' varía, el valor de p puede disminuir o incrementar dependiendo del equilibrio de los términos lineales, cuadrático y cúbico. El modelo polinomial de grado 3 está acotado con $W' \in [0,8965; 0,9848]$.

$$\hat{p} = -362,8148 + 1170,145(W') - 1257,5917(W')^2 + 450,3867(W')^3 \qquad (48)$$

Figura 34

Ajuste de curva – valor de p y estadístico W



Nota. Los datos fueron procesados utilizando el software R Core Team (2024).

Tabla 13

Comparación de modelos - valor de p y estadístico W

Modelo	R^2	R^2 ajustado	SCE
Precipitación			
Lineal	0,15	0,1487	28,6284
Polinomial de grado 2	0,2425	0,2401	25,514
Polinomial de grado 3	0,3081*	0,3048	23,3045

Exponencial	NA	NA	NA
Logarítmico	0,1348	0,1334	29,1406
Crudo			
Lineal	0,1431	0,1414	20,3126
Polinomial de grado 2	0,2623	0,2594	17,4858
Polinomial de grado 3	0,3712*	0,3675	14,9038
Exponencial	NA	NA	NA
Logarítmico	0,0945	0,0927	21,4644
Temperatura			
Lineal	0,1133	0,1111	6,5641
Polinomial de grado 2	0,1857	0,1815	6,0284
Polinomial de grado 3	0,2442*	0,2384	5,5952
Exponencial	NA	NA	NA
Logarítmico	0,0944	0,0921	6,704

Nota. Los datos fueron procesados utilizando el software R Core Team (2024). El símbolo * indica el modelo con mejor ajuste.

En lo que respecta a la prueba SW, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste de curva en su primera aproximación mediante un modelo polinomial de grado 3 (ver ecuaciones 49 a 51). En el modelo polinomial, la interpretación de los coeficientes influye a medida que el estadístico W varía, el valor de p puede disminuir o incrementar dependiendo del equilibrio de los términos lineales, cuadrático y cúbico. Las modelos polinomiales de grado 3 están acotados con $W \in [0,6298;0,9959]$, $W \in [0,1530;1,0000]$ y $W \in [0,4128;1,0000]$ respectivamente.

$$\hat{p} = -72,4629 + 274,1903(W) - 342,7814(W)^2 + 141,8534(W)^3 \quad (49)$$

$$\hat{p} = -1,6939 + 11,6496(W) - 22,4849(W)^2 + 13,0778(W)^3 \quad (50)$$

$$\hat{p} = -4,5067 + 21,0047(W) - 31,5604(W)^2 + 15,4017(W)^3 \quad (51)$$

4.2 Discusión de los Resultados

Un análisis preliminar de los datos es fundamental para identificar patrones o comportamientos en las variables de estudio. Las pruebas de normalidad univariadas se consideran como un paso previo de la estadística inferencial o supuestos de la estadística

paramétrica. Cada prueba de normalidad emplea distintos estadísticos, considerando momentos como la media aritmética, varianza, asimetría y curtosis. Uno factor importante a tomar en cuenta es el tamaño de muestra, ya que las pruebas de normalidad generan mayor efectividad cuando se utiliza muestras pequeñas, medianas o grandes. Asimismo, es relevante analizar los valores atípicos dado que pueden generan sensibilidad en los resultados de estas pruebas.

Diseñar un modelo matemático que relacione el estadístico de las pruebas de normalidad univariadas y el valor de p puede ser el inicio para explorar en profundidad el comportamiento de estas variables y explicar con mayor detalle las posibles hipótesis que se puedan plantear en un trabajo de investigación. Independientemente de las variables que se definen en un estudio, explorar los datos mediante modelos matemáticos aporta beneficios significativos al momento de interpretar los resultados y tomar decisiones más acertadas.

Para las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta consideradas en el análisis, se aplicaron las pruebas de normalidad univariadas KS, AD, JB CvM, M.KS, P χ^2 y SF en base a incrementos de una observación con muestra inicial de 50 hasta completar la muestra total correspondiente a cada variable. Cada submuestra generó un par ordenado compuesto por el estadístico de cada prueba y el valor de p.

Como primera aproximación, los modelos evaluados fueron: lineal, polinomio de grado 2 y 3, exponencial y logarítmica. La validación de cada modelo se realizó considerando el coeficiente de determinación (R^2), estableciendo un valor mínimo referencial (umbral mínimo) de 0,2 para aceptar la expresión matemática obtenida. Los modelos con un R^2 inferior a 0,2 no fueron considerados. El valor de R^2 está acotado entre $[0,1]$, donde 0 indica ausencia de relación entre las variables y 1 representa una relación perfecta o ajuste de curva ideal (ver umbrales de R^2 en la tabla 13).

Tabla 14

Umbrales referenciales de R^2

R^2	Interpretación
$R^2 < 0,25$	Relación muy débil.
$0,25 \leq R^2 < 0,50$	Relación débil.
$0,50 \leq R^2 < 0,75$	Relación moderada.
$R^2 \geq 0,75$	Relación sustancial.

Nota. Adaptado de Hair et al. (2011).

El coeficiente de determinación ajustado (R^2 ajustado) y la suma de los cuadrados de los errores (SCE) no fueron determinantes para validar el modelo. El R^2 ajustado resulta de mucha utilidad cuando se incorpora variables predictoras (variables de entrada o independientes), las cuales no formaron parte en la presente investigación. Además, en el análisis de cada variable según las pruebas de normalidad aplicadas, los valores de R^2 ajustado fueron inferiores a los de R^2. Con respecto a la SCE, se utilizó únicamente para verificar el comportamiento siguiendo el esquema relacional entre R^2 y SCE, como se muestra en la figura 6.

Para la prueba KS, las variables producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste a los modelos exponencial ($R^2 = 0,9784$) y polinomial de grado 3 ($R^2 = 0,8019$) respectivamente, al analizar la relación entre el estadístico *D* y el valor de p, dichos resultados son considerados como una relación sustancial. Por otro lado, en la prueba AD, las variables precipitación y temperatura mínima absoluta se ajustaron de mejor manera al modelo exponencial ($R^2 = 1$ y $R^2 = 0,9999$ respectivamente) en la relación entre el estadístico A^2 y el valor de p, siendo los resultados una relación sustancial.

En el caso de la prueba JB, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta presentaron un mejor ajuste a los modelos

logarítmicos ($R^2 = 0,2003$), polinomial de grado 3 en las dos últimas variables ($R^2 = 0,2797$ y $R^2 = 0,8902$), al relacionar el estadístico JB y el valor de p. Estos resultados se interpretan como relaciones muy débil y sustancial según los valores obtenidos. En lo referente a la prueba CvM, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta se ajustaron adecuadamente al modelo exponencial en todos los casos ($R^2 = 1$, $R^2 = 0,9998$ y $R^2 = 1$) respectivamente. Esto significa que el estadístico W^2 y el valor de p tienen una relación sustancial.

Con respecto a la prueba M.KS, la variable temperatura mínima absoluta tuvo un mejor ajuste al modelo exponencial ($R^2 = 0,936$), lo que indica que la relación entre el estadístico D' y el valor de p es sustancial. Por otro lado, para la prueba P χ^2, las variables precipitación y temperatura mínima absoluta mostraron un mejor ajuste a los modelos logarítmico ($R^2 = 0,3798$) y polinomial de grado 3 ($R^2 = 0,3467$) respectivamente. Estos resultados indican que la relación entre el estadístico P y el valor de p es débil. Finalmente, para la prueba SF, la variable temperatura mínima absoluta presentó un ajuste adecuado al modelo polinomial de grado 3 ($R^2 = 0,6019$), lo que sugiere que la relación entre el estadístico W' y el valor de p es moderada.

Finalmente, la prueba SW, las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta (analizado por bloques) mostraron un mejor ajuste al modelo polinomial de grado 3 en todos los casos con $R^2 = 0,3081$, $R^2 = 0,3712$ y $R^2 = 0,2442$. Estos valores obtenidos de R^2 indican que la relación entre el estadístico W y el valor de p es débil y muy débil.

Capítulo 5

Marco Propositivo

5.1 Planificación de la Actividad Preventiva

Diseñar un modelo matemático que relacione las pruebas de normalidad univariadas con sus respectivos estadísticos y los valores de p constituye el primer paso para profundizar en el criterio del nivel de significancia. El valor de p puede generar controversias en la toma de decisiones, por ende, es importante tomar una posición metodológica rigurosa para que el proceso de validar una hipótesis tenga sustento estadístico. Tener un valor de p menor que 0,05 no garantiza que las variables de estudio estén relacionadas, más bien, depende en gran medida del análisis previo de los datos que el investigador realice, a fin de que el estudio sea más significativo.

Verificar el supuesto de normalidad de los datos constituye en el proceso introductorio para el análisis estadístico. La generación previa de un modelo matemático que relacione la prueba de normalidad adaptada a las características del estudio y el valor de p permitiría obtener resultados más eficientes para la toma de decisiones, además, facilitaría la selección de una prueba estadística adecuada para el contraste de las hipótesis.

Al momento de utilizar las pruebas de normalidad univariadas es importante conocer la forma de aplicación. Como se mencionó en apartados anteriores, algunas pruebas se basan en medir distancias de cada observación con respecto a la media aritmética, donde el orden de los datos es importante. En cambio, pruebas como la de Jarque-Bera está direccionada en evaluar momentos estadísticos (asimetría y curtosis) sin importan el ordenamiento de los datos.

La prueba de Jarque-Bera está disponible en el paquete "tseries", en la librería library(tseries) de R donde viene incorporado por defecto el coeficiente de asimetría de Fisher clásico, por esta razón, el estadístico JB genera valores excesivamente altos,

afectando el comportamiento de las variables sobre todo cuando existen valores atípicos. Por esta razón, se propone reajustar el cálculo del estadístico JB mediante el uso del coeficiente de asimetría de Fisher ajustado, lo cual permitiría corregir los sesgos existentes en un conjunto de datos y podría mejorar la aproximación a la distribución normal.

Conclusiones

Se utilizaron distintas pruebas de normalidad univariadas con el objetivo de verificar el comportamiento de cada uno de los estadísticos y los respectivos valores de p. Este análisis permitió encontrar patrones consistentes que se ajustaron a los modelos de regresión clásicos para determinar la relación entre las variables.

La aplicación de diferentes tamaños muestrales (submuestras) evidenció una alta variabilidad o dispersión en el comportamiento del estadístico de las diferentes pruebas de normalidad univariadas, lo cual provocó que las variables analizadas no se distribuyan normalmente. Los gráficos utilizados fueron fundamentales para visualizar y comprender el comportamiento de los datos.

Con la aplicación de los modelos clásicos como primera aproximación se logró un avance significativo en el comportamiento de los estadísticos de las pruebas de normalidad univariadas (KS, AD, JB CvM, M.KS, P χ^2, SF y SW), así como los valores de p obtenidos en cada iteración al incrementar o decrementar el tamaño de la muestra.

Los modelos que presentaron un mejor ajuste de curva a las variables precipitación, producción de petróleo mensual operativo y temperatura mínima absoluta fueron los exponenciales y polinomial de grado 3. El coeficiente de determinación R^2 fue el indicador para validar cada uno de los modelos, obteniendo valores de R^2 aceptables como primera aproximación.

Recomendaciones

Para reforzar el análisis de los estadísticos de las pruebas de normalidad univariada, se recomienda incorporar otro tipo de pruebas actualizadas, con el objetivo de comparar la utilidad de los diferentes momentos estadísticos y evaluar el comportamiento del supuesto de normalidad, así como su aplicabilidad en diferentes tipos de variables.

Se recomienda ampliar los tamaños muestrales mediante el incremento progresivo de submuestras para verificar el comportamiento de los estadísticos de las pruebas de normalidad y los valores de p, esto permitiría observar patrones o tendencias que contribuyan en obtener resultados más eficientes, consistentes y robustos.

En estadística paramétrica, la verificación del supuesto de normalidad se ha convertido en el punto de partida en los trabajos de investigación, especialmente cuando se emplean variables continuas. Por ende, se recomienda diseñar un modelo matemático preliminar que se ajuste adecuadamente a un conjunto amplio de datos, con la finalidad de obtener valores de p más consistente y significativo.

Es recomendable para validar el modelo, utilizar de manera complementaria el uso de otros métodos de análisis con la finalidad de corroborar la calidad del ajuste y la viabilidad del modelo. Entre los principales métodos se encuentran: análisis de los residuos, tanto de forma analítica como gráfica, y la aplicación de medidas de error de ajuste como la raíz del error cuadrático medio que penaliza los errores grandes, y el error absoluto medio que proporciona el error sin penalización cuadrática.

En lo que concierne a los modelos más sofisticados y flexibles que se pueden aplicar para aumentar R^2 y tener un mejor ajuste de curva, se recomienda aplicar los modelos de regresión avanzada como: regresión spline (splines cúbicos, B-splines o P-splines) para dividir el conjunto de datos en tramos y realizar un ajuste flexible con continuidad y suavidad; regresión local (LOESS o LOWESS) cuando el comportamiento de los datos son

no paramétricos y necesita realizar un ajuste local para cada dato, permitiendo encontrar patrones complejos; regresión polinomial con regularización (ridge o lasso) para evitar sobreajuste en polinomios de grados superiores; además se propone aplicar modelos aditivos generalizados (GAM, por sus siglas en inglés) para modelar relaciones no lineales entre las variables de manera flexible.

Referencias

Ag-Yi, D., & Aidoo, E. N. (2022). A comparison of normality tests towards convoluted probability distributions. *Research in Mathematics*, *9*(1). https://doi.org/10.1080/27684830.2022.2098568

Aragón, E. E. P., Navas, A. R. V. y Aragón, H. A. P. (2022). Estrés académico en estudiantes de la Universidad de La Guajira, Colombia. *Revista de ciencias sociales*, *28*(5), 87-99. https://dialnet.unirioja.es/servlet/articulo?codigo=8471675

Arnastauskaitė, J., Ruzgas, T., & Bražėnas, M. (2021). An exhaustive power comparison of normality tests. *Mathematics*, *9*(7), 788. https://doi.org/10.3390/math9070788

Austin, P. C., Eekhout, I., & van Buuren, S. (2024). Evaluating the median *p*-value method for assessing the statistical significance of tests when using multiple imputation. *Journal of Applied Statistics*, *52*(6), 1161–1176. https://doi.org/10.1080/02664763.2024.2418473

Avram, C., & Mărușteri, M. (2022). Normality assessment, few paradigms and use cases. *Revista Romana de Medicina de Laborator*, *30*(3), 251-260. https://doi.org/10.2478/rrlm-2022-0030

Casement, C. J., & McSweeney, L. A. (2022). NormalityAssessment: An Interactive Classroom Tool for Testing Normality Visually. *Technology Innovations in Statistics Education*, *14*(1). https://doi.org/10.5070/T514156556

Demir, S. (2022). Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients. *International Journal of Assessment Tools in Education*, *9*(2), 397-409. https://doi.org/10.21449/ijate.1101295

Diaz-Ballve, L. P. M. (2020). El valor p: un concepto estadístico-metodológico omnipresente en la investigación biomédica. ¿Lo interpretamos correctamente?. *Argentinian Journal of Respiratory & Physical Therapy*, *2*(1). https://doi.org/10.58172/ajrpt.v2i1.91

Domański, C., & Szczepocki, P. (2020). Comparison of selected tests for univariate normality based on measures of moments. *Statistics in Transition new series*, *21*(5), 151-178. https://doi.org/10.21307/stattrans-2020-060

Ebbelind, A. (2023). A functional view on language: a methodology for mathematics education to study shifts in prospective teachers' discursive patterns. *International Journal of Mathematical Education in Science and Technology*, *54*(8), 1731–1745. https://doi.org/10.1080/0020739X.2023.2204506

Ferencek, A., & Kljajić Borštnar, M. (2020). Data quality assessment in product failure prediction models. *Journal of Decision Systems*, *29*(sup1), 79-86. https://doi.org/10.1080/12460125.2020.1776927

Gandica de Roa, E. M. G. (2020). Potencia y Robustez en pruebas de Normalidad con Simulación Montecarlo. *Revista Scientific*, *5*(18), 108-119. https://doi.org/10.29394/Scientific.issn.2542-2987.2020.5.18.5.108-119

Gibson, E. W. (2020). The Role of *p*-Values in Judging the Strength of Evidence and Realistic Replication Expectations. *Statistics in Biopharmaceutical Research*, *13*(1), 6–18. https://doi.org/10.1080/19466315.2020.1724560

Glinskiy, V., Ismayilova, Y., Khrushchev, S., Logachov, A., Logachova, O., Serga, L., Yambartsev, A., & Zaykov, K. (2024). Modifications to the Jarque–Bera Test. *Mathematics*, *12*(16), 2523. https://doi.org/10.3390/math12162523

Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., ... & Gao, R. (2023). Normal workflow and key strategies for data cleaning toward real-world data. *Interactive Journal of Medical Research*, *12*(1), e44310. https://doi.org/10.3390/sym10040099

Hagag, A. E. (2022). Normality tests Procedure with power comparison. المجلة العلمية للإقتصاد و التجارة, *52*(2), 499-556. https://doi.org/10.21608/jsec.2022.243208

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice*, *19*(2), 139-152. https://doi.org/10.2753/MTP1069-6679190202

Hudiburgh, L. M., & Garbinsky, D. (2020). Data Visualization: Bringing Data to Life in an Introductory Statistics Course. *Journal of Statistics Education*, *28*(3), 262–279. https://doi.org/10.1080/10691898.2020.1796399

Imbens, G. W. (2021). Statistical significance, p-values, and the reporting of uncertainty. *Journal of Economic Perspectives*, *35*(3), 157-174. https://doi.org/10.1257/jep.35.3.157

Inglis, A., Parnell, A., & Hurley, C. B. (2022). Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models. *Journal of Computational and Graphical Statistics*, *31*(3), 766–778. https://doi.org/10.1080/10618600.2021.2007935

Islam, T. U. (2021). Min-max approach for comparison of univariate normality tests. *Plos one*, *16*(8), e0255024. https://doi.org/10.1371/journal.pone.0255024

Jaramillo, H. A. L., Pinos, C. A. E., Sarango, A. F. H. y Román, H. D. O. (2023). Histograma y distribución normal: Shapiro-Wilk y Kolmogorov Smirnov aplicado en SPSS: Histogram and normal distribution: Shapiro-Wilk and Kolmogorov Smirnov applied in SPSS. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades*, *4*(4), 596-607. https://doi.org/10.56712/latam.v4i4.1242

Javed, M., & Irfan, M. (2020). A simulation study: new optimal estimators for population mean by using dual auxiliary information in stratified random sampling. *Journal of Taibah University for Science*, *14*(1), 557–568. https://doi.org/10.1080/16583655.2020.1752004

Jiménez-Gamero, M. D. (2024). Testing normality of a large number of populations. *Statistical Papers*, *65*(1), 435-465. https://doi.org/10.1007/s00362-022-01384-y

Khatun, N. (2021) Applications of Normality Test in Statistic al Analysis. Open Journal of Statistics, *11*(01), 113-122. https://doi.org/10.4236/ojs.2021.111006

Korkmaz, S., & Demir, Y. (2023). Investigation of some univariate normality tests in terms of type-i errors and test power. *Journal of Scientific Reports-A*, (052), 376-395. https://doi.org/10.59313/jsr-a.1222979

Kross, S., Peng, R. D., Caffo, B. S., Gooding, I., & Leek, J. T. (2020). The democratization of data science education. *The American Statistician*, *74*(1), 1-7. https://doi.org/10.1080/00031305.2019.1668849

Kumagai, M., Ando, Y., Tanaka, A., Tsuda, K., Katsura, Y., & Kurosaki, K. (2022). Effects of data bias on machine-learning–based material discovery using experimental property data. *Science and Technology of Advanced Materials: Methods*, *2*(1), 302–309. https://doi.org/10.1080/27660400.2022.2109447

Lian, M., Chen, L., Hui, C., Zhu, F., & Shi, P. (2024). On the Relationship Between the Gini Coefficient and Skewness. *Ecology and Evolution*, 14. https://doi.org/10.1002/ece3.70637.

Muñoz, P. F., Escobar, L. M., & Acalo, T. S. (2019). Estudio de potencia de pruebas de normalidad usando distribuciones desconocidas con distintos niveles de no normalidad. *Perfiles*, *1*(21), 4-11. https://doi.org/10.47187/perf.v1i21.42

Paliz, C., Perugachi, N., Martínez, J., Moreno, M., Yaucán, C. y Palaguachi, R. (2021). Análisis estadístico de datos de las precipitaciones usando métodos robustos y bootstrap. *FIGEMPA: Investigación y Desarrollo*, *12*(2), 52-61. https://doi.org/10.29166/revfig.v12i2.3515

Pawel, S., & Held, L. (2025). Closed-Form Power and Sample Size Calculations for Bayes Factors. *The American Statistician*, 1–15. https://doi.org/10.1080/00031305.2025.2467919

Pawitan, Y. (2020). Defending the P-value. *arXiv preprint.* https://doi.org/10.48550/arXiv.2009.02099

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rana, S., Eshita, N. N., & Al Mamun, A. S. M. (2021). Robust normality test in the presence of outliers. In *Journal of Physics: Conference Series* (Vol. 1863, No. 1, p. 012009). IOP Publishing. https://doi.org/10.1088/1742-6596/1863/1/012009

Savalei, V., & Rosseel, Y. (2021). Computational Options for Standard Errors and Test Statistics with Incomplete Normal and Nonnormal Data in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 163–181. https://doi.org/10.1080/10705511.2021.1877548

Shukla, D. (2023). A narrative review on types of data and scales of measurement: An initial step in the statistical analysis of medical data. *Cancer Research, Statistics, and Treatment*, *6*(2), 279-283. https://doi.org/10.4103/crst.crst_1_23

Sulewski, P. (2021). Two component modified Lilliefors test for normality. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *16*(2), 429-455. https://doi.org/10.24136/eq.2021.016

Taplin, R. H. (2024). Quantifying data quality after removing respondents who fail data quality checks. *Current Issues in Tourism*, 1–12. https://doi.org/10.1080/13683500.2024.2378611

Thornberg, R., Forsberg, C., Hammar Chiriac, E., & Bjereld, Y. (2020). Teacher–Student Relationship Quality and Student Engagement: A Sequential Explanatory Mixed-Methods Study. *Research Papers in Education*, *37*(6), 840–859. https://doi.org/10.1080/02671522.2020.1864772

Turnbull, D., Chugh, R., & Luck, J. (2020). Learning management systems: a review of the research methodology literature in Australia and China. *International Journal of Research & Method in Education*, *44*(2), 164–178. https://doi.org/10.1080/1743727X.2020.1737002

Uhm, T., & Yi, S. (2021). A comparison of normality testing methods by empirical power and distribution of *P*-values. *Communications in Statistics - Simulation and Computation*, *52*(9), 4445–4458. https://doi.org/10.1080/03610918.2021.1963450

Uyanto, S. S. (2022). An Extensive Comparisons of 50 Univariate Goodness-of-fit Tests for Normality. Austrian Journal of Statistics, 51(3), 45–97. https://doi.org/10.17713/ajs.v51i3.1279

Wang, Z., Wang, Y., & Sun, D. (2022). A retrospective and prospective study to establish a preoperative difficulty predicting model for video-assisted thoracoscopic lobectomy and mediastinal lymph node dissection. *BMC surgery*, *22*(1), 135. https://doi.org/10.1186/s12893-022-01566-3