



Universidad de Valladolid

Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Estadística

**Un nuevo Test de Normalidad basado en el
Estimador Núcleo de la Densidad**

Autor: Jorge Arranz Simón

Tutor: Dr. D. Bonifacio Salvador González

Junio de 2023

Gracias a los que me habéis visto llegar hasta aquí, apoyándome y animándome siempre a ser una mejor persona. Me hacéis muy feliz. Gracias.

Agradecimientos

Me gustaría agradecer todo el apoyo recibido no sólo para realizar este Trabajo de Fin de Grado, sino a todos aquellos que habéis estado conmigo durante este maravilloso camino.

Gracias a mi familia, por animarme, por aguantarme y sobre todo, por quererme.

Gracias a mis amigos, por escucharme aunque no les importe de lo que hable, por hacerme mejor persona y por seguir creciendo juntos.

Gracias a Boni, por haber hecho tan fácil trabajar juntos, y por haberme enseñado tanto en tan poco tiempo.

Y por último, me gustaría agradecer a toda la Facultad de Ciencias, desde profesores hasta el personal de la biblioteca y cafeterías y, en particular, a todas aquellas aquellas que habéis formado parte de mi evolución durante todos estos años, por habernos brindado a todos los estudiantes un magnífico lugar dónde poder desarrollarnos como personas.

Resumen

La importancia de la distribución normal es innegable en el mundo de la estadística, ya que muchos de los procedimientos de análisis de datos ampliamente utilizados hoy en día asumen la normalidad en los datos como requisito fundamental. Es por esto, que el esfuerzo en el avance de este campo no tiene fin, y la mejora en los métodos que ayudan a detectar la normalidad de los datos amplía el abanico de procedimientos que pueden ser aplicados en el análisis de los mismos.

En el presente trabajo, se propone un nuevo de test de normalidad basado en el Estimador Núcleo de la Función de Densidad y se compara a través de la potencia en una simulación de Monte Carlo, en la que se incluyen distribuciones alternativas de diferente índole. Para cada una de las distribuciones, se consideran diferentes tamaños de muestra, y se comenta el desempeño obtenido por este nuevo test frente a algunos de los más utilizados en la actualidad.

Palabras clave

Test de normalidad, Simulación de Monte Carlo, Estimador Núcleo de la Función de Densidad.

Abstract

The importance of the normal distribution is undeniable in the world of statistics, as many widely used data analysis procedures nowadays assume normality as a fundamental requirement. Therefore, the effort in studying this field is endless, and the improvement in methods that enhance in detecting data normality expands the range of procedures that can be applied in their analysis.

In this paper, a new test of normality, based on the Kernel Density Estimator, is proposed and compared through power analysis in a Monte Carlo simulation, which includes alternative distributions with different behaviour. For each distribution, different sample sizes are considered, and the performance of this new test is discussed in comparison to some of the most commonly used tests today.

Key words

Normality test, Monte-Carlo simulation, Kernel Density Estimator.

Índice general

| | | |
|----------|--|-----------|
| 1 | Introducción | 9 |
| 2 | Tests de normalidad | 11 |
| 2.1 | Descripción de algunos tests de normalidad conocidos | 11 |
| 2.1.1 | Tests basados en la función de distribución empírica | 12 |
| | Test de Lilliefors | 12 |
| | Test de Anderson-Darling | 13 |
| | Test de Cramer von Mises | 13 |
| 2.1.2 | Tests basados en regresión y correlación | 14 |
| | Test de Saphiro-Willks | 14 |
| | Test de Shapiro-Francia | 14 |
| 2.1.3 | Tests basados en el estimador núcleo de la función de densidad | 15 |
| | Descripción del Estimador Núcleo de la Función de Densidad | 15 |
| | El parámetro de suavizado | 17 |
| | La función núcleo | 18 |
| | Métodos de elección del parámetro de suavizado | 20 |
| | Test de normalidad basado en el estimador núcleo | 21 |
| 2.2 | Test propuesto | 23 |
| 2.2.1 | Definición formal | 23 |
| 2.2.2 | Región crítica | 24 |
| 3 | Comparación de los tests de normalidad | 27 |
| 3.1 | Tests incluidos en la comparación | 28 |
| 3.2 | Obtención de las Regiones Críticas | 28 |
| 3.3 | Simulación de la potencia | 29 |
| 3.4 | Comparación de la potencia de los tests de normalidad | 29 |
| 3.4.1 | Distribuciones simétricas | 30 |
| | T-student con 3 G.L. | 30 |
| | Cauchy | 31 |
| | Laplace | 32 |
| | Uniforme(0, 1) | 33 |
| 3.4.2 | Distribuciones Asimétricas | 34 |
| | Distribución exponencial | 34 |
| | Distribución Weibull(4, 1) | 35 |
| | Distribución Weibull(1, 1.5) | 37 |
| | Distribución Chi-Cuadrado | 38 |
| | Lognormal | 39 |

| | | |
|----------|--|-----------|
| 3.4.3 | Mixturas | 39 |
| | Mixtura normal | 40 |
| | Mixtura bimodal | 42 |
| | Mixtura contaminada por media | 44 |
| | Mixtura contaminada por varianza | 45 |
| 4 | Conclusiones y trabajos futuros | 47 |
| A | Funciones desarrolladas en R | 51 |
| A.1 | Cálculo del estadístico para el test propuesto | 51 |
| A.2 | Simulacion de Valores Críticos | 51 |
| A.3 | Cáculo de la Potencia | 53 |
| A.4 | Generacion de gráficos | 56 |

Índice de figuras

| | | |
|------|--|----|
| 2.1 | Efecto del del parámetro de suavizado en el ENFD | 17 |
| 2.2 | Comparación de la estimación a través de diferentes funciones núcleo | 19 |
| 2.3 | Ejemplo de QQ plot y de comparación de dos densidades | 22 |
| 2.4 | Comparación Q-Q Plot - Distribución Normal | 25 |
| 2.5 | Comparación Q-Q Plot - Distribución Exponencial | 25 |
| 2.6 | Comparación Q-Q Plot - Distribución Tstudent 3GL | 26 |
| | | |
| 3.1 | Distribución Tstudent(3) | 30 |
| 3.2 | Distribución Cauchy(0,1) | 31 |
| 3.3 | Distribución Laplace(0, 1) | 32 |
| 3.4 | Distribución Uniforme(0, 1) | 33 |
| 3.5 | Distribución Exponencial | 34 |
| 3.6 | Distribución Weibull(4, 1) | 35 |
| 3.7 | Distribución Weibull(4, 1) | 36 |
| 3.8 | Distribución Weibull(4, 1) | 36 |
| 3.9 | Distribución Weibull(1, 1.5) | 37 |
| 3.10 | Distribución Chi-Cuadrado con 4 GL | 38 |
| 3.11 | Distribución LogNormal(0, 1) | 39 |
| 3.12 | Mixtura con aspecto Normal | 40 |
| 3.13 | Aumento de tamaño en la mixtura $0.7*N(0, 0.8) + 0.3*N(0.7, 1.1)$ | 41 |
| 3.14 | Mixtura de Normales con aspecto Bimodal | 42 |
| 3.15 | ENFD en mixtura bimodal | 43 |
| 3.16 | Estudio en el entorno del parámetro de suavizado | 43 |
| 3.17 | Mixtura contaminada por la media | 44 |
| 3.18 | Mixtura contaminada por la varianza | 45 |

Índice de cuadros

| | | |
|------|---|----|
| 2.1 | Valores críticos para los distintos tamaños muestrales | 25 |
| 3.1 | Regiones críticas con 100.000 simulaciones para $\alpha = 0.05$ | 29 |
| 3.2 | Potencia obtenida con $\alpha = 0.05$ | 31 |
| 3.3 | Potencia obtenida con $\alpha = 0.05$ | 31 |
| 3.4 | Potencia obtenida con $\alpha = 0.05$ | 32 |
| 3.5 | Potencia obtenida con $\alpha = 0.05$ | 34 |
| 3.6 | Potencia obtenida con $\alpha = 0.05$ | 35 |
| 3.7 | Potencia obtenida con $\alpha = 0.05$ | 38 |
| 3.8 | Potencia obtenida con $\alpha = 0.05$ | 39 |
| 3.9 | Potencia obtenida con $\alpha = 0.05$ | 40 |
| 3.10 | Potencia obtenida con $\alpha = 0.05$ | 42 |
| 3.11 | Potencia obtenida con $\alpha = 0.05$ | 44 |
| 3.12 | Potencia obtenida con $\alpha = 0.05$ | 45 |

Siglas

AD Anderson-Darling. 13, 28

CVM Cramer Von Mises. 28

ENFD Estimador Núcleo de la Función de Densidad. 4, 9, 15, 17, 18, 22–25, 28, 30, 36, 42, 43, 45, 47

Lillie Lilliefors. 12, 28

SF Shapiro-Francia. 23, 28, 30, 43, 47

SW Shapiro-Wilks. 28

Capítulo 1

Introducción

Asegurar la hipótesis de normalidad es fundamental en el campo de la investigación científica ya que, para aplicar muchos de los procedimientos estadísticos que se utilizan hoy en día para el análisis de datos, se debe asegurar que estos siguen una distribución Normal. Para ello, se pueden emplear distintas técnicas, desde las comparaciones visuales hasta procedimientos formales como los tests de hipótesis, también conocidos como tests de normalidad cuando se emplean para la hipótesis nula de Normalidad. Son herramientas estadísticas utilizadas para evaluar si una muestra de datos desconocida sigue una distribución Normal.

En este estudio se introducirá un nuevo test de normalidad univariante, que aprovecha el buen rendimiento del Estimador Núcleo de la Función de Densidad (ENFD). Se realizará una comparación de diferentes tests de normalidad en términos de la potencia, medida empleada para cuantificar la efectividad de un test de hipótesis para detectar una hipótesis alternativa, frente a algunos de los tests más empleados hoy en día, como son el Test de Shapiro-Francia, el Test de Lilliefors o el Test de Anderson-Darling. Para ello se simulará el comportamiento de todos los tests sobre diferentes distribuciones alternativas y se medirá el rendimiento de cada uno de ellos.

Para el desarrollo de la simulación, se requiere realizar un esfuerzo de cómputo para obtener, a través de la simulación de Monte-Carlo, unos valores críticos que se puedan considerarse consistentes. Para la obtención de la Región Crítica de cada uno de los tests se realiza una simulación de 100.000 muestras normales de cada uno de los tamaños estudiados, que varían desde muestras con 10 observaciones hasta muestras con 200 observaciones, sobre las que se extraerán estos valores críticos para cada uno de los tests. Una vez obtenidas las regiones críticas necesarias para el estudio, se obtendrán 5000 muestras para cada una de las distribuciones alternativas y tamaños seleccionados para el estudio, para estimar con una alta fiabilidad el valor de la potencia para cada una de las alternativas.

Se tratará de explicar con claridad cada uno de los nuevos conceptos introducidos, haciendo énfasis en aquellos detalles del ENFD que aportan un valor añadido al conocimiento recogido en los documentos que componen la bibliografía y se realizará una amplia comparación de todos los tests sobre cada una de las hipótesis alternativas, incluyendo los procedimientos *ad hoc* para aquellas distribuciones que requieran de una mayor explicación o presenten unos resultados sobre los que se pueda extraer más información que ayude a comprender mejor el comportamiento de este nuevo procedimiento.

Capítulo 2

Tests de normalidad

En este capítulo se expondrán de forma ordenada todos los tests que se van a incluir en el estudio, haciendo especial énfasis en los tests basados en el estimador núcleo de la función de densidad. Finalmente se introducirá de forma teórica el nuevo test propuesto.

2.1. Descripción de algunos tests de normalidad conocidos

Los tests de normalidad son una serie de técnicas estadísticas utilizadas para determinar si una muestra de datos sigue una distribución normal. Estos tests son útiles para evaluar si una muestra de datos puede ser analizada mediante técnicas estadísticas paramétricas que asumen la normalidad, como la prueba t de Student [3] o el análisis de varianza (ANOVA) [4].

Los tests de normalidad generalmente se dividen en dos categorías: tests gráficos y tests de hipótesis formales. Los tests gráficos se basan en la visualización de los datos a través de un histograma, gráficos de probabilidad normal, Q-Q plots o gráficos de cajas y bigotes. Por otro lado, los tests de hipótesis se basan en estadísticos que miden la discrepancia de los datos observados y la hipótesis de normalidad.

Algunos de los tests formales más comunes son: el de Kolmogorov-Smirnov, el de Lilliefors, el de Shapiro-Wilk, el de Shapiro-Francia y el de Anderson-Darling, que serán brevemente descritos a continuación. Para una mejor clasificación de los tests se van a agrupar en función de la característica principal que utilizan para la comparación con la distribución normal [5]. Cabe destacar que existen otros tests de normalidad, como los basados en el apuntamiento y la kurtosis, pero no se estudiarán en este documento. Los tests que se incluyen en el estudio se han clasificado en:

- Tests basados en la función de distribución empírica
- Tests basados en regresión y correlación
- Tests basados en la función núcleo

2.1.1. Tests basados en la función de distribución empírica

Este tipo de tests utilizan las características de la función de distribución para tratar de detectar las diferencias, de forma que la construcción del estadístico se realiza a partir de la comparación entre la función de distribución bajo la hipótesis nula de normalidad y la función de distribución de la muestra. Destacan los tests de Lilliefors, Anderson-Darling y Cramer von Mises.

Test de Lilliefors

El test de Lilliefors es un test de normalidad basado en el test de Kolmogorov-Smirnov.

El test de Kolmogorov-Smirnov es la versión más sencilla, para hipótesis nulas simples. Es un test no paramétrico que determina la bondad de ajuste de dos distribuciones de probabilidad cuales sean entre sí. El estadístico que se construirá tiene como objetivo detectar si la muestra proviene o no de una distribución normal estandarizada, por lo que esta comparación se basa en la diferencia entre la función de distribución empírica de la muestra con la función de distribución Normal teórica a la que queremos ver si se ajusta. La comparación se hace por medio del estadístico:

$$D_n = \sup |F_n(x) - F_0(x)| \quad (2.1)$$

Donde $F_n(x)$ es la función de distribución empírica de la muestra y F_0 es la función de distribución de referencia, en nuestro caso de estudio, la distribución $N(0, 1)$. Este estadístico tiene la propiedad de ser de libre distribución, lo que significa que su distribución no se ve afectada por la función de referencia considerada bajo la hipótesis nula.

Para aquellos casos en los que tanto la media como desviación típica son desconocidas y deben ser estimadas para su comparación, se emplea el test de Lilliefors. El estadístico de este test sigue la misma lógica que el de Kolmogorov-Smirnov, pero para el caso de hipótesis compuestas. El estadístico de Lilliefors tiene esta forma:

$$\hat{D}_n = \sup |F_n(x) - \hat{F}_{N(\hat{x}, \hat{\sigma})}(x)| \quad (2.2)$$

Donde se ha sustituido la función de distribución de referencia por la distribución teórica normal con los parámetros estimados: $\hat{F}_{N(\hat{x}, \hat{\sigma})}$. Cabe resaltar que las distribuciones de D_n y \hat{D}_n no son iguales, por lo que no tendrán los mismos umbrales de aceptación. Al igual que en el anterior caso, se rechazará la hipótesis nula en el caso de que las diferencias obtenidas sean suficientemente grandes.

$$RC = \{\hat{D}_n > C.Lillie_{n,\alpha}\} \quad (2.3)$$

Típicamente se realiza la comparación con los valores obtenidos por simulación de Monte-Carlo para comprobar si esta diferencia entre las funciones de distribución acumuladas es suficiente para considerarlas distintas y, consecuentemente, rechazar la hipótesis de Normalidad, o si por el contrario no consideramos significativa esta diferencia podemos asumir que los datos sobre los que hemos realizado el test provienen de una distribución normal.

La implementación que se utilizará en este estudio es *lillie.test*, desarrollada en el paquete de R *nortest* [6]

Test de Anderson-Darling

El test de Anderson-Darling (AD) es un test no paramétrico que se utiliza también para determinar si un conjunto de datos proviene de una distribución en particular. La prueba se basa en la comparación de la distribución empírica de los datos con la distribución teórica deseada.

El estadístico de AD mide también la diferencia entre la distribución acumulada empírica y la distribución acumulada teórica. El cálculo del estadístico tiene como objetivo determinar si los datos, que necesariamente deben estar ordenados, $Y_1 < \dots < Y_n$, provienen de una distribución acumulativa F .

Rechaza H_0 para valores grandes del estadístico

$$W_n^2 = n \int_{-\infty}^{\infty} \frac{[\hat{F}_n(x) - F_0(x)]^2}{F_0(x)[1 - F_0(x)]} dF_0(x) \quad (2.4)$$

$$RC = \{W_n^2 > C.Anderson_{n,\alpha}\} \quad (2.5)$$

La implementación que se utilizará en este estudio es *ad.test*, desarrollada en el paquete de R *nortest* [6].

Test de Cramer von Mises

El test propuesto por Harald Cramér y Richard Elder von Mises entre los años 1928 y 1930 se considera una alternativa al test de Kolmogorov-Smirnov. El estadístico se construye a través de la suma de las diferencias de dos distribuciones continuas.

$$\omega^2 = n \int_{-\infty}^{\infty} [\hat{F}_n(x) - F_0(x)]^2 dF_0(x) \quad (2.6)$$

En nuestro caso de estudio F_0 representará la distribución de una normal con los parámetros estimados por la media y la desviación típica muestrales. Debido a que el estadístico mide las diferencias entre las distribuciones, la Región Crítica queda definida por el rechazo de valores altos del estadístico, que harían notar que las diferencias entre las dos distribuciones también son grandes.

$$RC = \{\omega^2 > C.CVM_{n,\alpha}\} \quad (2.7)$$

La implementación que se utilizará en este estudio es *cvm.test*, desarrollada en el paquete de R *nortest* [6].

2.1.2. Tests basados en regresión y correlación

Los tests que se presentan en esta sección tratan de aprovechar que la variable $y \sim N(\mu, \sigma^2)$ puede ser expresada como $y = \mu + \sigma x$ donde $x \sim N(0, 1)$, y están asociados de forma natural a los gráficos de probabilidad. Los de regresión se enfocan en la pendiente de la línea cuando los estadísticos de orden de la muestra se confrontan con su valor esperado bajo normalidad. Los de correlación se enfocan en la fuerza de esta relación.

Test de Saphiro-Willks

Como todos los anteriores, este test se utiliza para comprobar si un conjunto de datos proviene de una distribución normal o no. Samuel Shapiro y Martin Wilk publicaron en 1965 un nuevo estadístico con este mismo objetivo. El estadístico utilizado para el test es:

$$W = \frac{(\sum_{i=1}^n a_i x_i)^2}{\sum_{i=1}^n (x_i - (\bar{x}))^2} \quad (2.8)$$

donde las variables a_i se calculan como

$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \quad (2.9)$$

siendo m_1, \dots, m_n los valores medios del estadístico ordenado correspondiente a una muestra aleatoria de tamaño n de una población normal y V su matriz de covarianzas.

Es uno de los tests más potentes para contrastar la normalidad, sobre todo con tamaños reducidos como muestras de tamaño $n \leq 30$. La distribución bajo la hipótesis nula solo está tabulada hasta $n \leq 50$. Los p-valores para tamaños de $n \geq 50$ se obtienen mediante el algoritmo de Royston. Cabe destacar sobre este test que se confrontan dos estimadores alternativos de la varianza

Se rechazará la hipótesis nula de normalidad si el estadístico W , acotado entre 0 y 1, obtiene un valor demasiado pequeño.

$$RC = \{W < C.S ShapiroWilk_{n,\alpha}\} \quad (2.10)$$

La implementación que se utilizará en este estudio es *shapiro.test*, desarrollada en el paquete de R *nortest* [6]

Test de Shapiro-Francia

Es una aproximación al estadístico test de Shapiro-Wilk, basada en la idea de que en muestras grandes los componentes del estadístico ordenado pueden tratarse como si fueran independientes. Lo que supone considerar la matriz $V = I$.

Se rechaza H_0 para valores pequeños del estadístico W_n^2 , cuadrado del coeficiente de correlación lineal para los puntos que representan en Q-Q plot normal, es decir

$$W_n^2 = \text{Cor}^2((X_{(i)}, G^{-1}(\frac{i - \frac{3}{8}}{n + \frac{1}{4}})), i = 1, \dots, n) \quad (2.11)$$

Donde G es la función de distribución de una $N(0, 1)$. En el Q-Q plot se representa la muestra ordenada en el eje de abscisas, y en el eje de ordenadas los valores corregidos de los cuantiles de una distribución normal. De esta manera, si los puntos se alinean sobre una recta, podemos concluir que los datos sustentan la hipótesis nula. Para no tomar la decisión de una forma visual, que carece completamente de una base científica, se utiliza como estadístico del test la correlación lineal al cuadrado entre los puntos.

La Región crítica o de rechazo queda definida por

$$RC = \{W_n^2 < C.S ShapiroFrancia_{n,\alpha}\} \quad (2.12)$$

ya que una correlación baja indica falta de linealidad en los puntos y por lo tanto falta de normalidad en la distribución.

La implementación que se utilizará en este estudio es *sf.test*, desarrollada en el paquete de R *nortest* [6].

2.1.3. Tests basados en el estimador núcleo de la función de densidad

En primer lugar presentaremos el concepto de Estimador Núcleo de la Función de Densidad (ENFD) y, en la siguiente sección, comentaremos el test basado en el error cuadrático integrado estudiado en Bowman (1992) [7].

Descripción del Estimador Núcleo de la Función de Densidad

El Estimador Núcleo de la Función de Densidad es un método utilizado para estimar una función de densidad de probabilidad subyacente a una muestra de datos. Fue introducido por Rosenblatt en 1956 con una idea principal: distribuir la masa $1/n$ de cada una de las observaciones sobre el total, de forma que la estimación de la densidad en cada punto no se vea influida solo por sí misma, sino que el resultado sea una combinación del entorno de cada X_i . Para el caso de una muestra aleatoria simple X_1, \dots, X_n procedente de una variable aleatoria continua queda definida por

$$\hat{f}_n(y, X_1, \dots, X_n, h) = \frac{1}{n} \sum_{i=1}^n w(y - X_i; h) \quad (2.13)$$

donde w representa la función núcleo y h el parámetro positivo conocido como el ancho de banda o tamaño de ventana. Lo nombraremos a partir de ahora como parámetro de suavizado y será utilizado para controlar la varianza de la función núcleo, que regula el grado de suavidad de la estimación. Por simplicidad y claridad a partir de ahora la función se referirá solo por $\hat{f}_n(y)$, pero se incide en el hecho de que la estimación es dependiente de la muestra de datos utilizada y del parámetro de suavizado.

Como características básicas y con el objetivo de conocer un poco mejor el comportamiento del estimador, se comentará el comportamiento del sesgo y la varianza, comentados en Bowmann y Azzalini [7].

La esperanza del estimador de la densidad se puede escribir como

$$E\{\hat{f}(y)\} = \int w(y-z; h)f(z)dz \quad (2.14)$$

Esto es una convolución de la verdadera función de densidad con w como función núcleo. El suavizado ha producido un estimador sesgado, cuya media es una versión suavizada de la verdadera densidad. A través de un desarrollo de Taylor y con el cambio de notación en la función núcleo que quedará parametrizada por $(1/h)w(z/h)$ se llega a la aproximación

$$E\{\hat{f}(y)\} \approx f(y) + \frac{h^2}{2}\sigma_w^2 f''(y) \quad (2.15)$$

donde σ_w^2 es la varianza de la función núcleo. Debido a que $f''(y)$ mide el ratio de curvatura de la función de densidad, se aprecia como \hat{f} subestima f en los picos de la verdadera densidad y sobreestima en los valles. El tamaño del sesgo se ve afectado por el parámetro de suavizado h . El componente σ_w^2 será 1 si la función núcleo escogida tiene varianza unitaria.

A través de otro desarrollo de Taylor se estima la varianza a través de

$$var\{\hat{f}(y)\} \approx \frac{1}{nh}f(y)\alpha(w) \quad (2.16)$$

donde $\alpha(w) = \int w^2(z)dz$. Como siempre, la varianza es inversamente proporcional al tamaño de la muestra. En resumen, se puede ver como el término nh actúa como controlador del tamaño de la muestra local, ya que h controla el número de observaciones cuyo peso en la función núcleo afectan a la estimación en y .

Estas aproximaciones de la media y la varianza del estimador de la densidad encierran el efecto del parámetro de suavizado. En la medida que h decrece, el sesgo disminuye y la varianza aumenta, mientras que si h aumenta, el sesgo aumenta y la varianza disminuye. El efecto combinado de esta propiedad es que, para poder producir un estimador que converja a la función de densidad verdadera, es necesario que tanto h como $1/nh$ decrezcan con el aumento del tamaño muestral.

Es útil definir una medida combinada de sesgo cuadrático y varianza para medir cuán efectivo es \hat{f} a la hora de estimar f . Una elección sencilla es el *error cuadrático medio integrado* (MISE) que, en el caso unidimensional y gracias a las aproximaciones de Taylor mostradas en 2.14 y 2.15, se aproxima a

$$MISE(\hat{f}) \approx \frac{1}{4}h^4\sigma_w^4 \int f''(y)^2 dy + \frac{1}{nh}\alpha(w) \quad (2.17)$$

El parámetro de suavizado

La elección del parámetro de suavizado es de vital importancia a la hora de emplear el Estimador Núcleo de la Función de Densidad. En términos generales el parámetro de suavizado controla la suavidad del ENFD.

Valores "pequeños" de este parámetro provocan que el número de sumandos significativamente no nulos en la estimación de la densidad sea "pequeño" para cada uno de los puntos, lo que nos lleva a una estimación irregular y con mucho ruido. Esto conduce a una estimación con un menor sesgo pero hace que las estimaciones dependan demasiado de la variabilidad muestral. Esto se conoce como *infrasuavización*.

En cambio, valores grandes para este parámetro producirían un aumento del sesgo, ya que utilizaría un gran número de observaciones en el promedio del punto de interés. Esto se conoce como *sobresuavización*.

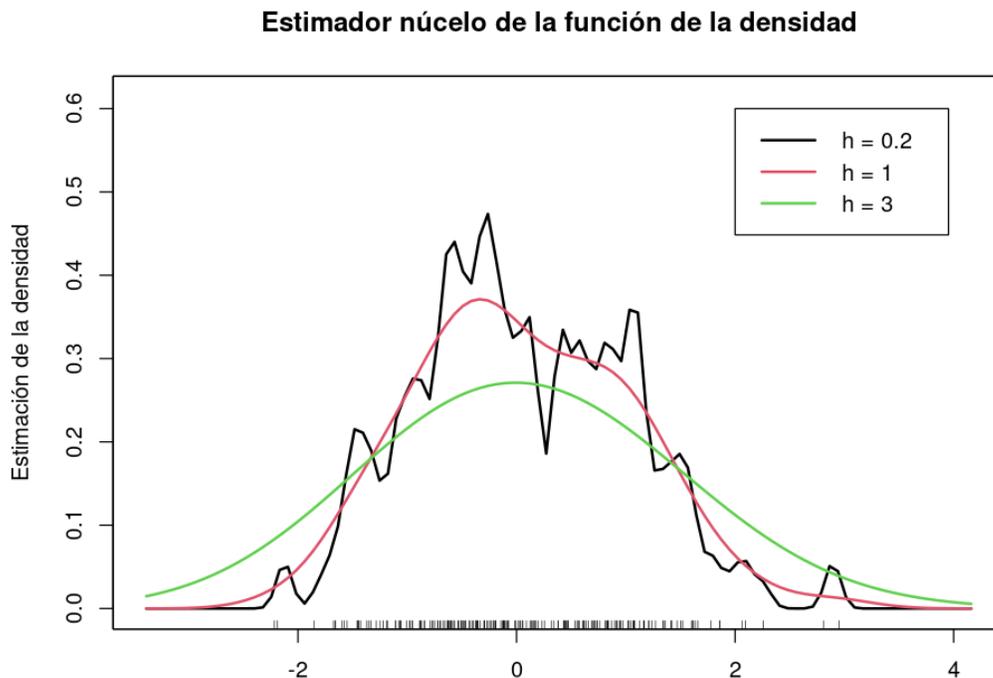


Figura 2.1: Efecto del del parámetro de suavizado en el ENFD

En la Fig. 2.1 podemos ver como la elección de valores menores de h conllevan un mayor ajuste a la distribución subyacente, sobreajustando a los datos, mientras que los valores mayores crean una distribución con demasiada influencia del entorno de y , produciendo una gran pérdida de información.

En resumen, la elección del parámetro de suavizado es crucial para obtener una estimación precisa y útil. Cabe destacar que no hay un valor óptimo para todos los casos, sino que depende de las características de la distribución subyacente, que es desconocida, y los obje-

tivos de la investigación. Más adelante se presentarán los métodos de elección del parámetro de suavizado más comunes para esta estimación.

- Un valor de h "pequeño" conduce a que el estimador utilice pocas observaciones para la ponderación, disminuyendo el sesgo, pero haciendo que las estimaciones dependan demasiado de la variabilidad muestral. Esta situación recibe el nombre de *infrasuavización*.
- Un valor de h "grande" produce un aumento del sesgo, ya que utiliza un gran número de observaciones en el promedio en el punto x de interés, lo que se conoce como *sobresuavización*.

La función núcleo

La otra decisión que se debe tomar para utilizar el ENFD es qué tipo de función núcleo vamos a emplear. Esta decisión no es tan relevante como la del parámetro de suavizado, ya que las diferencias que encontramos al usar diferentes funciones no son tan significativas. Si bien se observa el efecto de cada una de las funciones, este no es para nada diferencial en el comportamiento del estimador. Esta función debe cumplir unas características fundamentales como son ser suave, no negativa y tener una integral finita.

Algunos ejemplos de funciones núcleo que cumplen estas condiciones son la normal, la uniforme o la triangular. Cada una de ellas tiene sus propiedades y características, pero como se aprecia en la fig 2.2, con un parámetro de suavizado fijo no se aprecian grandes diferencias al variar la función núcleo.

La importancia de la elección de esta función radica en que, al no tener un efecto diferencial en el estimador, podemos escoger una función conocida con un comportamiento que favorezca el objetivo para el que se plantea el estimador. En el caso de este estudio se utilizará la función normal, ya que es una función con un comportamiento conocido que no afectará al estimador y que nos facilitará la estimación de cuantiles a través del concepto del ENFD como una mixtura de normales, del que se hablará más adelante.

Estimación núcleo de la función de densidad

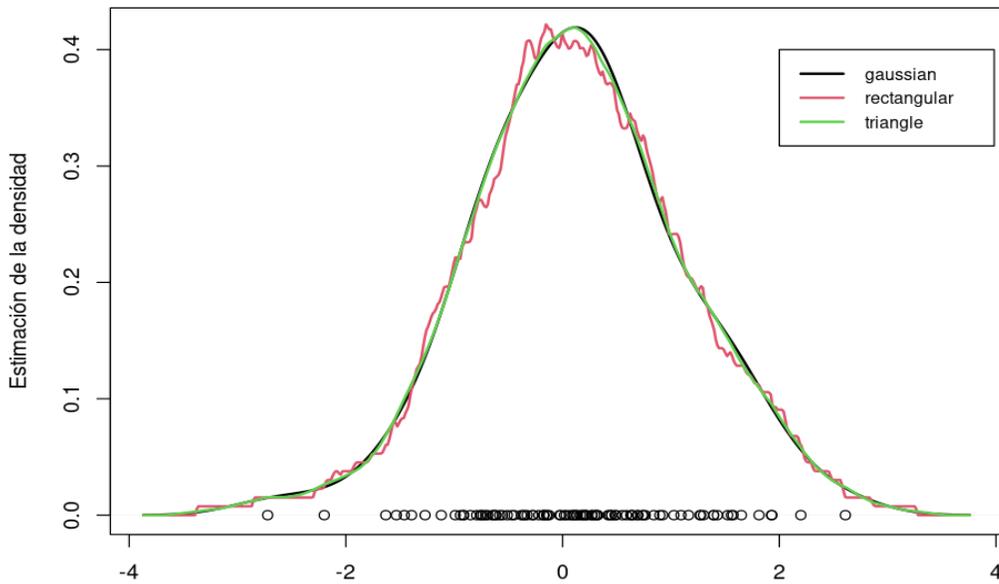


Figura 2.2: Comparación de la estimación a través de diferentes funciones núcleo

Debido a que nuestra investigación tiene como objetivo detectar las diferencias con la distribución normal, utilizaremos una función núcleo que trate de acercar los datos a la distribución normal, para poder detectar así las diferencias entre ambas de forma más precisa. Usualmente, esta función es elegida entre las funciones de densidad simétricas con media cero y varianza finita, para suavizar de esta forma el efecto de los valores aislados. Aún tomando esto como la práctica habitual, está generalmente aceptado que la forma exacta no es del todo relevante. Es común utilizar como núcleo funciones reales, no negativas, integrables con $\int w = 1$, unimodal y simétrica respecto al origen. En nuestra investigación se utilizará una distribución Normal centrada en 0 y con desviación típica h . De esta forma obtendríamos que:

$$w(y - y_i; h) = \phi(y - y_i; h) \tag{2.18}$$

donde $\phi(y - y_i; h)$ es el valor de la función de densidad $N(y_i, h)$ calculada en y .

Esto se traduce en que el estimador de la función densidad a través de funciones núcleo $N(0, h)$ se podrá considerar una mixtura de n distribuciones normales, centradas en cada una de las observaciones presentes en la muestra y con varianza igual al parámetro de suavizado h escogido.

En resumen, la función núcleo es un componente cuya importancia radica en que, al no ser decisivo, se puede escoger de forma que mejor convenga al planteamiento del estudio, pero no es crucial a la hora de estimar una función de distribución a través de este procedimiento. La elección una función núcleo suave nos llevará a estimar una función de densidad con esta misma característica.

Métodos de elección del parámetro de suavizado

Con el objetivo de construir un estimador de la función de densidad de los datos observados, es necesario escoger el valor del parámetro de suavizado h . En esta sección explicaremos asintóticamente la elección óptima y tres de los procedimientos más comunes tratados en *Applied Smoothing Techniques for Data Analysis* [7] con el fin de buscar un valor consistente para h .

Suavizado óptimo Utilizaremos como indicador de la eficacia de \hat{f} a la hora de estimar f el error cuadrático medio integrado, comentado anteriormente. En el caso unidimensional se mide como:

$$MISE(\hat{f}) = E \int [\hat{f}(y) - f(y)]^2 dy = \int [E\hat{f}(y) - f(y)]^2 dy + \int var \hat{f}(y) \quad (2.19)$$

Esta combinación de sesgo al cuadrado y varianza ha sido el centro fundamental de la mayor parte del trabajo que se ha llevado a cabo en estas estimaciones. En particular, la aproximación de series de Taylor permiten aproximar el MISE como:

$$MISE(\hat{f}) \approx \frac{1}{4} h^4 \sigma_w^4 \int f''(y)^2 dy + \frac{1}{nh} \alpha(w) \quad (2.20)$$

De esta aproximación se extrae que el valor de h que minimiza el MISE en un sentido asintótico es:

$$h_{opt} = \left\{ \frac{\gamma(w)}{\beta(f)n} \right\}^{\frac{1}{5}}, \quad (2.21)$$

donde $\gamma(w) = \alpha(w)/\alpha_w^4$, y $\beta(f) = \int f''(y)^2 dy$. Este valor óptimo de h no puede ser inmediatamente usado en la práctica ya que incluye el parámetro de densidad desconocido f . Sin embargo, es muy útil ya que muestra como el parámetro de suavizado debe decrecer con el tamaño de la muestra, proporcional a $n^{1/5}$, y cuantificando el efecto de la curvatura de f a través del factor $\beta(f)$

Suavizado normal óptimo Evaluando la fórmula óptima de h cuando f es una densidad normal se llega a la simple fórmula de h

$$h_{norm} = \left(\frac{4}{3n} \right)^{\frac{1}{5}} \sigma \quad (2.22)$$

donde σ representa la desviación típica de la distribución. La asunción de normalidad es potencialmente contraproducente a la hora de estimar una densidad no normal pero, al menos para las distribuciones unimodales, ofrece una elección útil del parámetro de suavizado que no requiere mucho cálculo. Para su cálculo en R se puede emplear la función $hnorm$ del paquete *sm* [9].

Validación cruzada Se aplican las ideas de la validación cruzada a la estimación del parámetro de suavizado a través de la estimación del Error Cuadrático Integrado, ISE en sus siglas en inglés.

$$\int \{\hat{f}(y) - f(y)\}^2 dy = \int \hat{f}(y)^2 dy - 2 \int f(y)\hat{f}(y) dy + \int f(y)^2 dy \quad (2.23)$$

El único término en el que no interviene h es en el final, el resto pueden ser estimados como

$$\frac{1}{n} \sum_{i=1}^n \int \hat{f}_{-i}^2(y) dy - \frac{2}{n} \sum_{i=1}^n \hat{f}_{-i}^2(y_i) \quad (2.24)$$

donde $\hat{f}_{-i}(y)$ hace referencia al estimador construido a partir de los datos sin la observación y_i

Test de normalidad basado en el estimador núcleo

Un clásico para la comparación de distribuciones empíricas con la distribución normal se basa en los gráficos de probabilidad como se muestra en la Figura 2.3. La comparación clásica se realiza a través de los gráficos de cuantiles o Q-Q plot, en los que se puede observar cuán similares son las distribuciones que queremos comparar y la distribución normal.

En el caso de los Q-Q plot, lo que se busca es que los puntos se posicionen alineados sobre una recta, lo que se verá traducido en que existe una relación entre ambos cuantiles. Este argumento sirve de sustento para la hipótesis nula (H_0), ya que apoya que los datos de la muestra proviene de una distribución normal. Para que este sustento en favor de H_0 no sea de carácter visual, sino numérico, se utiliza como indicador el cuadrado del coeficiente de correlación lineal de los puntos del Q-Q plot.

Este procedimiento no deja de ser una aproximación visual y, por tanto, no formal, pero sí que aporta una información útil. La transformación de este supuesto a un método más formal se puede realizar a través de varias alternativas.

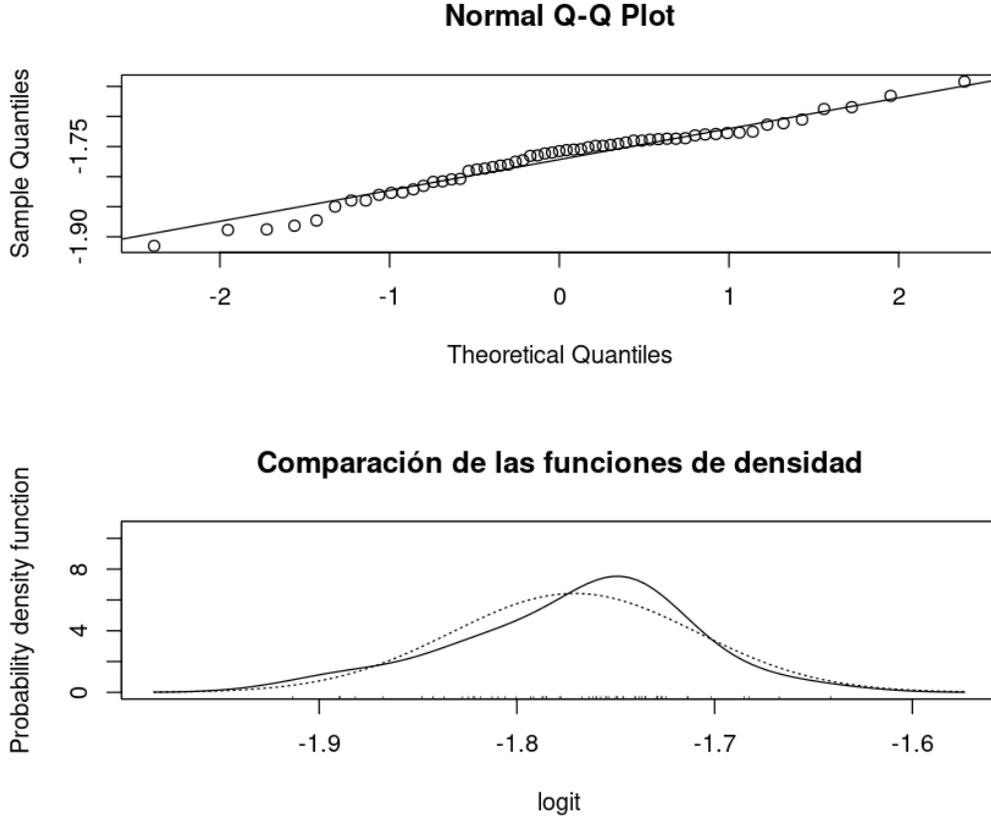


Figura 2.3: Ejemplo de QQ plot y de comparación de dos densidades

Hay una gran multitud de posibilidades para medir las diferencias entre dos densidades. En nuestro caso, el objetivo es comparar el ENFD con una distribución Normal. En Bowmann y Azzalini [7] se estudia el test basado en el error cuadrático integrado, que denominamos *END*:

$$END = \int \{\hat{f}(y) - \phi(y - \hat{\mu}, \sqrt{\hat{\sigma}^2 + h^2})\}^2 dy \quad (2.25)$$

Hay que tener en cuenta que la densidad estimada no se compara con la densidad normal propuesta $\phi(y - \hat{\mu}; \hat{\sigma})$. Debido a que el parámetro de suavizado produce un sesgo, es más apropiado comparar la densidad estimada con su valor esperado bajo la asunción de normalidad, que es el que se muestra en la expresión 2.26. Para eliminar los efectos de localización y escala se estandarizan los datos a media 0 y desviación típica 1. De esta forma el estadístico final del test, nombrado a partir de ahora como *END* en este documento, se puede expresar como:

$$END = \int \{\hat{f}(y) - \phi(y, \sqrt{1 + h^2})\}^2 dy \quad (2.26)$$

Este estadístico se calcula a través de la función *nise* del paquete *sm* [9]

La Región Crítica del test viene marcada por valores altos en las diferencias de ambas distribuciones, por la definimos como:

$$RC = \{NISE > C.END_{n,\alpha}\} \quad (2.27)$$

2.2. Test propuesto

El Test que se propone en este estudio comparte muchas de las características del test de Shapiro-Francia, ya que está basado en el cuadrado del coeficiente de correlación lineal entre cuantiles del ENFD y cuantiles de su valor esperado bajo la hipótesis de normalidad. Es decir, que sustituye los cuantiles muestrales por los cuantiles del Estimador Núcleo de la Función de Densidad, y los cuantiles de la distribución normal, por los cuantiles del valor esperado del ENFD bajo la hipótesis de normalidad.

El objetivo del test es aprovechar el buen comportamiento del Estimador Núcleo de la Función de Densidad.

2.2.1. Definición formal

Considerar X_1, \dots, X_n v.a. i.i.d. con función de densidad desconocida f , suave. El problema es contrastar las hipótesis:

H_0 : f es la función de densidad de una distribución normal.

H_1 : f no es la función de densidad de una distribución normal

Considerar el Estimador Núcleo de la Función de Densidad f , con núcleo $N(0, 1)$, con el parámetro de suavizado $h > 0$.

$$\hat{f}(x, h) = \frac{1}{n} \sum_{i=1}^n \phi(x - X_i, h); \forall x \in R \quad (2.28)$$

donde

$$\phi(z, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right) \quad (2.29)$$

es la función de densidad en z de una distribución normal con media cero y desviación típica σ .

Es fácil ver que si $f \rightarrow N(\mu, \sigma)$ la esperanza de $\hat{f}(x, h)$ es:

$$E\hat{f}(x, h) = \phi(x - \mu, \sqrt{\sigma^2 + h^2}). \quad (2.30)$$

Sean $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ y $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2$ los estimadores máximo verosímiles de μ y σ , bajo la hipótesis nula, se observa que $f \rightarrow N(\hat{\mu}, \hat{\sigma})$ y, por lo tanto, el valor esperado de $\hat{f}(\cdot, h)$ es:

$$E_0 \hat{f}(x, h) = \phi(x - \hat{\mu}, \sqrt{\hat{\sigma}^2 + h^2}) \quad (2.31)$$

Sin pérdida de generalidad, podemos suponer que $\hat{\mu} = 0$ y $\hat{\sigma} = 1$, por lo que el valor esperado bajo H_0 de $\hat{f}(\cdot, h)$ es:

$$E_0 \hat{f}(x, h) = \phi(x, \sqrt{1 + h^2}) \quad (2.32)$$

El estadístico test que consideramos es el cuadrado del coeficiente de correlación lineal entre los n cuantiles de $\hat{f}(\cdot, h)$ en los puntos $\frac{i - \frac{3}{8}}{n + \frac{1}{4}}, i = 1, \dots, n$ y los correspondientes cuantiles de $E_0 \hat{f}(\cdot, h)$. Es decir:

$$NEW = Cor^2[(\hat{F}^{-1}(\frac{i - \frac{3}{8}}{n + \frac{1}{4}}), G^{-1}(\frac{i - \frac{3}{8}}{n + \frac{1}{4}})), i = 1, \dots, n] \quad (2.33)$$

donde \hat{F} es la función de distribución correspondiente al ENFD y G es la función de distribución correspondiente a $E_0 \hat{f}(x, h)$.

El cálculo de los cuantiles de la ENFD se facilita gracias a la observación de que el ENFD es la mixtura de n distribuciones Normales, y cada una de estas Normales centrada en cada una de las observaciones, con desviación típica igual al parámetro de suavizado. Estos cuantiles se pueden obtener en R fácilmente gracias a la librería *norImix* [10], que implementa la función *qnorMix* con este fin.

$$\hat{f}(x, h) = \sum_{i=1}^n \frac{1}{n} \phi(x - X_i; h) \quad (2.34)$$

Como parámetro de suavizado h se utiliza el valor de referencia a la distribución normal, cuyo valor se obtiene como se muestra en la expresión 2.22, ya que si H_0 es cierta, es el valor adecuado y, cuando no es cierta, se obtiene un test conservador.

2.2.2. Región crítica

De forma análoga al test de Shapiro-Francia, el test rechaza la hipótesis nula para valores "pequeños" del estadístico.

$$RC = \{NEW < C.NewTest_{n,\alpha}\} \quad (2.35)$$

Los valores críticos, $C.NewTest_{n,\alpha}$, obtenidos a través de una simulación de Monte-Carlo, se muestran en la tabla 2.1. Estos valores, como se comenta detalladamente en la sección 3.2, se obtienen a partir de la simulación de 100.000 muestras de cada uno de los tamaños:

Se muestran a continuación unos ejemplos en los que se puede observar la diferencia entre el Q-Q plot empleado en el test de Shapiro-Francia (gráfico derecho), donde se utilizan los

| α | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.01 | 0.94965 | 0.96285 | 0.96339 | 0.96818 | 0.97012 | 0.97971 | 0.98419 | 0.98872 | 0.99074 |
| 0.05 | 0.97098 | 0.97499 | 0.97831 | 0.97962 | 0.98360 | 0.98822 | 0.99017 | 0.99231 | 0.99391 |

Cuadro 2.1: Valores críticos para los distintos tamaños muestrales

cuantiles muestrales, y el Q-Q plot empleado en el test que se propone, donde se utilizan los cuantiles obtenidos de la distribución estimada a través del Estimador Núcleo de la Función de Densidad para las distribuciones Normal Fig. 2.4, Exponencial Fig. 2.5 y T student con 3 grados de libertad Fig. 2.6.

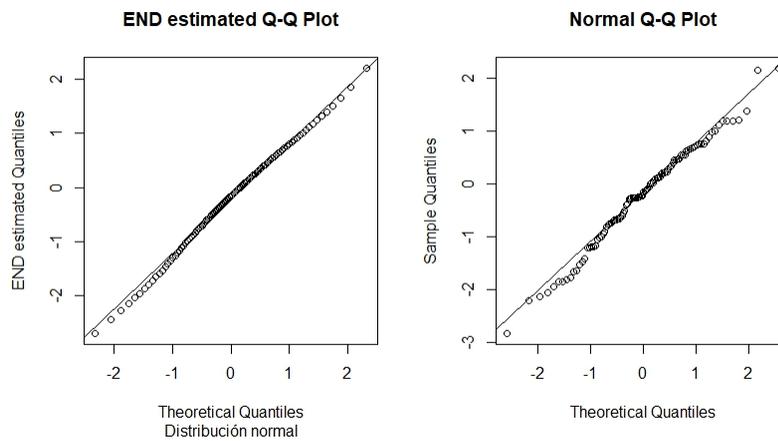


Figura 2.4: Comparación Q-Q Plot - Distribución Normal

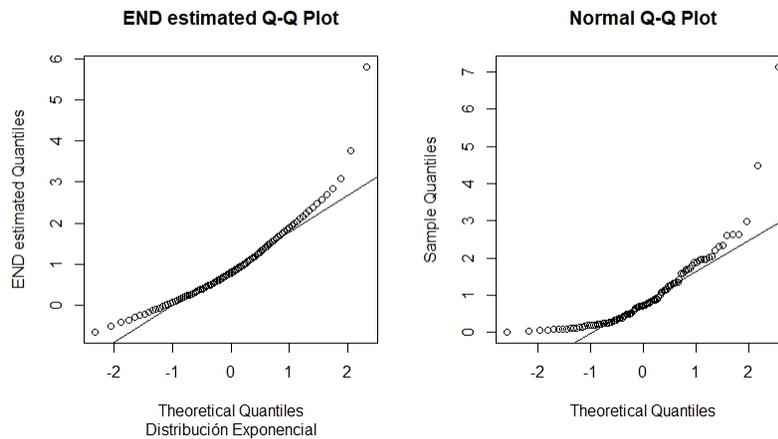


Figura 2.5: Comparación Q-Q Plot - Distribución Exponencial

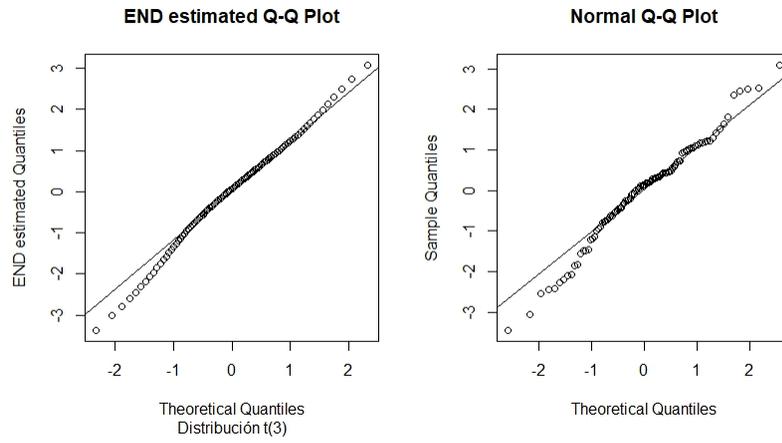


Figura 2.6: Comparación Q-Q Plot - Distribución Tstudent 3GL

Capítulo 3

Comparación de los tests de normalidad

En esta sección vamos a estudiar la capacidad de detección de los tests de normalidad en términos de la potencia, es decir, la capacidad del test para detectar una diferencia real entre la distribución Normal y una distribución alternativa. La potencia es la probabilidad de rechazar la hipótesis nula cuando la hipótesis alternativa es verdadera, por lo que un test es tanto mejor cuanto mayor es su potencia.

La potencia depende fundamentalmente de dos factores: el tamaño de la muestra y el nivel de significación. Para un test dado, una potencia alta respecto de una distribución alternativa indica que el test es capaz de detectar una diferencia real y por lo tanto, se considera una prueba más fiable. En contra, una potencia baja para un determinado test respecto de una distribución alternativa, indica que el test tiene poca capacidad de detectar esta distribución alternativa.

Hay dos métodos que podemos utilizar para aumentar la potencia de un test frente a una alternativa. Una posibilidad es aumentar el tamaño de la muestra que se utiliza en el test, lo que produce un aumento de la información y por lo tanto una mayor capacidad de detección. El otro procedimiento que se puede emplear si se quiere una potencia mayor en un test, es aumentar el nivel de significación.

El estudio planteado trata de comparar la potencia de los tests considerados frente a diferentes alternativas, de forma que aquellos tests que obtengan unos mejores resultados en términos de potencia podrán ser considerados más eficaces a la hora de detectar cada hipótesis alternativa. Esto quiere decir que, aunque un test obtenga una potencia alta frente a una distribución alternativa, no tiene por que obtenerla frente a otras alternativas. De hecho, esta circunstancia es la que impide que exista un test de ajuste óptimo para todas las hipótesis, y lo que provoca que exista una gran cantidad de tests disponibles con el mismo objetivo.

La comparación se realizará sobre cada una de las distribuciones, aplicando todos los tests comentados sobre distintos tamaños muestrales de la distribución en cuestión. Se realizarán 5000 repeticiones del experimento sobre tamaños entre 10 y 200 observaciones. Para aquellos casos en los que los resultados no sean representativos, o no aporten la información que se busca, se tomarán medidas ad hoc.

Para todos los cálculos realizados se ha utilizado el software estadístico R [11], aprovechando las posibilidades que ofrece de paralelización del cómputo para tratar de maximizar

al máximo los recursos disponibles.

3.1. Tests incluidos en la comparación

Se incluye un listado con los tests que se han evaluado así como la abreviatura utilizada a partir de ahora de cada uno de ellos.

- Test de Lilliefors (Lillie)
- Test de Anderson-Darling (AD)
- Test de Shapiro-Wilks (SW)
- Test de Shapiro-Francia (SF)
- Test de Cramer Von Mises (CVM)
- Test basado en el Estimador Núcleo de la Función de Densidad (ENFD)
- Test propuesto (NEW)

3.2. Obtención de las Regiones Críticas

La región crítica de en un test de hipótesis es el conjunto de valores para los que se rechaza la hipótesis nula. En otras palabras, es la región del espacio en la que los resultados del estadístico son tan extremos que no pueden ser considerados procedentes de la hipótesis nula, debido a que, si se realiza la prueba bajo las condiciones de la hipótesis nula, no se obtendrían estos valores en el mayor número de las veces.

El nivel de 'extremo' que consideramos se define a través del nivel de significación (α) del test. Este nivel α quiere decir que si la hipótesis nula es cierta y replicamos un alto número de veces la prueba, acertaremos el $(1 - \alpha) \times 100\%$ de las veces. El proceso para calcular estos valores a través de la simulación es el siguiente:

En primer lugar se se simulan B muestras bajo la hipótesis nula de tamaño n ,

$$[(x_{11}, \dots, x_{1n}), (x_{21}, \dots, x_{2n}), \dots, (x_{B1}, \dots, x_{Bn})] \quad (3.1)$$

siendo n el tamaño de la muestra sobre la que se realiza el test. Sobre cada una de las B muestras se calcula el valor del estadístico D en cuestión, obteniendo así B valores del estadístico: $[D_1, D_2, \dots, D_B]$, siendo el valor crítico el cuantil α o $(1 - \alpha)$ en función de si el test rechaza para valores bajos o altos del estadístico respectivamente.

En este caso se han simulado 100.000 muestras de la distribución normal para cada uno de los tamaños sobre los que se plantea el experimento. Cada una de las muestras normales simuladas se utiliza para construir los estadísticos test. De esta forma habremos construido 100.000 veces el estadístico test a partir de una muestra de cada uno de los tamaños considerados, pudiendo considerar que los valores inferiores al q_α no son valores típicos del estadístico

para una muestra Normal. En el caso de los tests cuya región crítica está definida para valores grandes el estadístico procedemos de forma análoga a considerar que los valores del estadístico superiores al cuantil $q_{1-\alpha}$ nos llevan a rechazar la hipótesis nula de normalidad.

Se muestran a continuación los valores críticos a nivel $\alpha = 0,05$ para cada uno de los tests y tamaños muestrales considerados. Se marcan con (*) aquellos tests cuya región crítica está definida para valores pequeños del estadístico, es decir los que rechazarán la hipótesis nula si el valor obtenido es inferior al mostrado en la tabla

| n | Lillie | AD | SW* | SF* | CVM | END | NEW* |
|-----|--------|--------|--------|--------|--------|--------|--------|
| 10 | 0.2620 | 0.6869 | 0.8449 | 0.8429 | 0.1194 | 0.0156 | 0.9783 |
| 15 | 0.2185 | 0.7117 | 0.8822 | 0.8806 | 0.1222 | 0.0142 | 0.9781 |
| 25 | 0.1724 | 0.7226 | 0.9199 | 0.9190 | 0.1228 | 0.0116 | 0.9815 |
| 30 | 0.1590 | 0.7322 | 0.9300 | 0.9294 | 0.1244 | 0.0108 | 0.9825 |
| 50 | 0.1241 | 0.7371 | 0.9541 | 0.9537 | 0.1243 | 0.0083 | 0.9864 |
| 80 | 0.0991 | 0.7443 | 0.9691 | 0.9689 | 0.1248 | 0.0065 | 0.9896 |
| 100 | 0.0890 | 0.7479 | 0.9747 | 0.9745 | 0.1255 | 0.0057 | 0.9909 |
| 150 | 0.0732 | 0.7484 | 0.9823 | 0.9822 | 0.1257 | 0.0045 | 0.9931 |
| 200 | 0.0633 | 0.7475 | 0.9863 | 0.9863 | 0.1260 | 0.0038 | 0.9942 |

Cuadro 3.1: Regiones críticas con 100.000 simulaciones para $\alpha = 0.05$

3.3. Simulación de la potencia

La obtención de los valores de la potencia en una distribución alternativa para cada uno de los tests y tamaños muestrales incluidos en el estudio se realiza a través del mismo procedimiento. En primer lugar se genera una muestra aleatoria del tamaño en cuestión procedente de la distribución alternativa. A partir de esta muestra, se construyen todos los estadísticos test y se comparan con los valores obtenidos a través de la simulación de la región crítica. Se indica si el valor pertenece o no a la región crítica o de rechazo, significando esto que el estadístico test es lo suficientemente extremo como para no poder ser considerado procedente de una distribución Normal. En este caso, el test habrá logrado detectar las diferencias entre las dos distribuciones, lo que se traduce en un aumento de la potencia.

En el estudio planteado este procedimiento se realiza 5000 veces para cada tamaño muestral, siendo la potencia el promedio de veces que el valor del estadístico test pertenece a la región crítica para cada uno de los tests.

3.4. Comparación de la potencia de los tests de normalidad

Para realizar la comparación de los tests en términos de la potencia se mostrarán los resultados obtenidos para distintos tamaños, desde muestras pequeñas de tamaño 10 hasta muestras mayores de tamaño 200. Para realizar las comparaciones se han dividido las distribuciones en tres grupos: Simétricas, asimétricas y mixturas de normales. Dentro de cada grupo se ha tratado de elegir un conjunto de distribuciones que representen distintas carac-

terísticas, como puede ser distintos niveles de kurtosis o una mayor diferencia en el peso de las colas. Esta clasificación se ha realizado a partir de la lectura del artículo 'Comparison of Tests for Univariate Normality' de Edith Seier [5].

Para todas las distribuciones incluidas en el experimento, se proporcionará una tabla con los valores de la potencia obtenida con cada uno de los tests, donde se resaltarán los valores de la potencia más alta para cada tamaño muestral, y tres gráficos, en los que se mostrarán un histograma de una muestra de tamaño 100 junto con la función de densidad obtenida a través del ENFD, la densidad teórica alternativa y los valores de la potencia obtenidos para cada tamaño de muestra y test.

3.4.1. Distribuciones simétricas

A priori deberían ser el grupo que presente una mayor dificultad a la hora de diferenciarlo de una normal ya que comparten esta propiedad. Se han escogido para la comparación las siguientes distribuciones: T(3), Cauchy(0, 1), Laplace y U(0, 1)

T-student con 3 G.L.

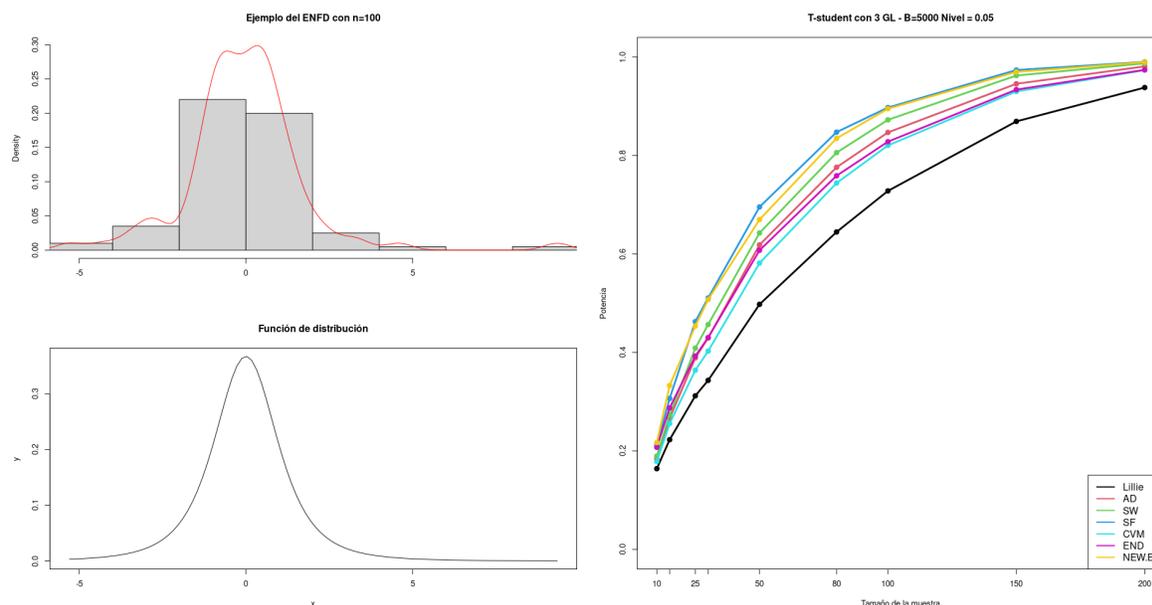


Figura 3.1: Distribución Tstudent(3)

La T-student con 3 g.l. tiene una forma más aplanada en el centro y un mayor peso en las colas. Este mayor peso se debe a que tiene una varianza mayor, que conlleva una mayor dispersión en los datos. Observamos como el test propuesto logra detectar estas diferencias con mayor precisión que la mayoría de los tests, obteniendo los mejores resultados junto al test de Shapiro-Francia. Obtiene una potencia superior a 0.8 para muestras de tamaño 80 y de 0.9 para muestras de tamaño 100. Logra detectar las diferencias en casi el 100 % de las muestras cuando superamos el tamaño 200.

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.16 | 0.22 | 0.31 | 0.34 | 0.50 | 0.64 | 0.73 | 0.87 | 0.94 |
| AD | 0.18 | 0.27 | 0.39 | 0.43 | 0.62 | 0.78 | 0.85 | 0.95 | 0.98 |
| SW | 0.19 | 0.27 | 0.41 | 0.46 | 0.64 | 0.81 | 0.87 | 0.96 | 0.99 |
| SF | 0.21 | 0.31 | 0.46 | 0.51 | 0.70 | 0.85 | 0.90 | 0.97 | 0.99 |
| CVM | 0.18 | 0.26 | 0.36 | 0.40 | 0.58 | 0.74 | 0.82 | 0.93 | 0.97 |
| END | 0.21 | 0.29 | 0.39 | 0.43 | 0.61 | 0.76 | 0.83 | 0.93 | 0.97 |
| NEW.E0 | 0.22 | 0.33 | 0.45 | 0.51 | 0.67 | 0.83 | 0.90 | 0.97 | 0.99 |

Cuadro 3.2: Potencia obtenida con $\alpha = 0.05$

Cauchy

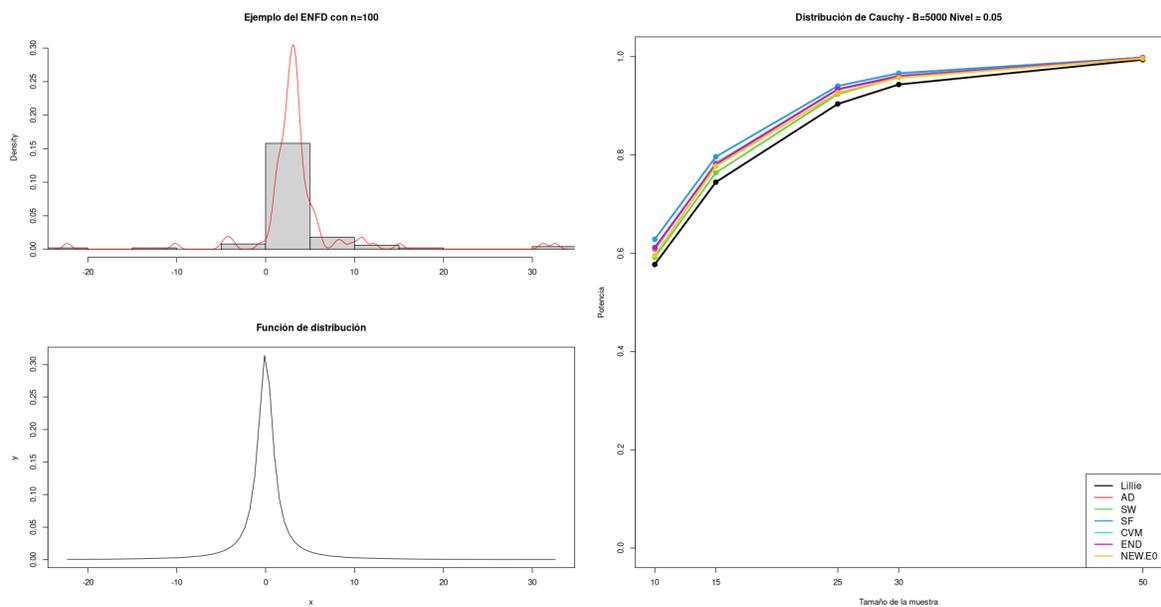


Figura 3.2: Distribución Cauchy(0,1)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.58 | 0.74 | 0.90 | 0.94 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| AD | 0.61 | 0.78 | 0.93 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SW | 0.59 | 0.76 | 0.92 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SF | 0.63 | 0.80 | 0.94 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| CVM | 0.61 | 0.78 | 0.93 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| END | 0.61 | 0.78 | 0.93 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| NEW.E0 | 0.59 | 0.78 | 0.93 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Cuadro 3.3: Potencia obtenida con $\alpha = 0.05$

La distribución de Cauchy se diferencia de la normal en el peso de sus colas. Como se puede apreciar en la Fig 3.2 esta distribución tiene una dispersión de los datos muy superior a la normal, convirtiéndose así en una distribución con un peso muy alto en las colas. Al

no tener una varianza definida la dispersión de los datos es infinita, lo que conlleva obtener valores extremos con mucha más facilidad que en la distribución Normal.

Estas características descritas son de gran relevancia a la hora de detectar las diferencias con la distribución bajo la hipótesis nula, por lo que todos los tests obtienen un gran rendimiento. A partir de un tamaño de 50 se comienzan a obtener valores extremos de forma habitual en las muestras estudiadas, lo que hace que todos los tests consigan detectar siempre la hipótesis alternativa. A pesar de compartir la característica de simetría, todos los tests obtienen un gran rendimiento para esta distribución.

No se incluyen en el gráfico los resultados para tamaños de muestra superiores a 50 ya que no aporta más información y no permite apreciar la evolución en los tamaños significativos.

Laplace

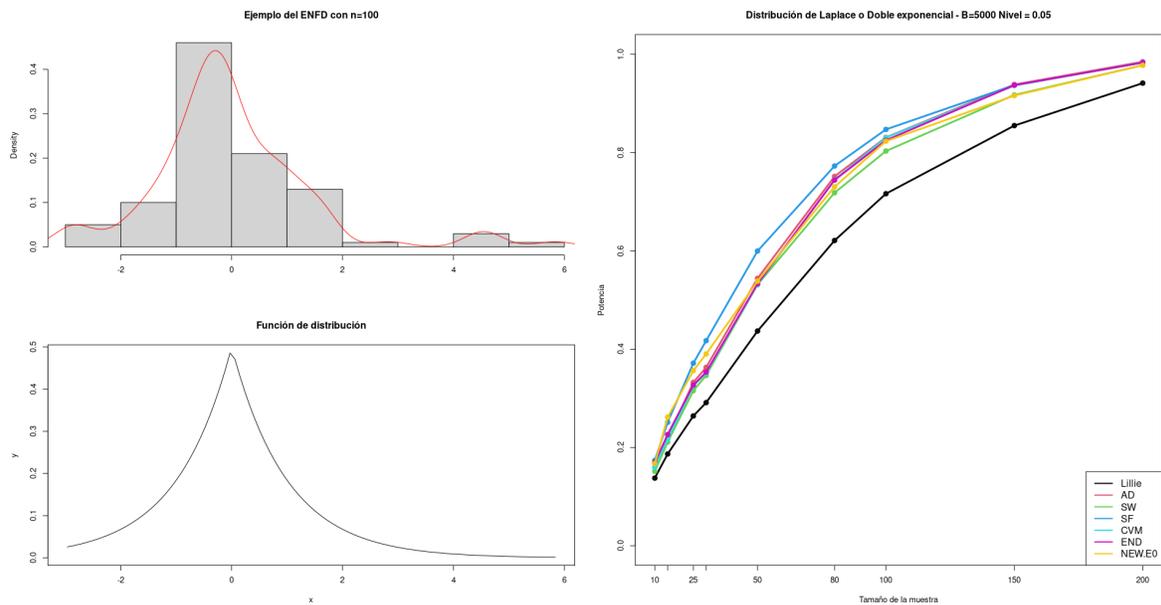


Figura 3.3: Distribución Laplace(0, 1)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.14 | 0.19 | 0.26 | 0.29 | 0.44 | 0.62 | 0.72 | 0.85 | 0.94 |
| AD | 0.16 | 0.22 | 0.33 | 0.36 | 0.54 | 0.75 | 0.83 | 0.94 | 0.98 |
| SW | 0.15 | 0.21 | 0.32 | 0.35 | 0.53 | 0.72 | 0.80 | 0.92 | 0.98 |
| SF | 0.17 | 0.25 | 0.37 | 0.42 | 0.60 | 0.77 | 0.85 | 0.94 | 0.98 |
| CVM | 0.16 | 0.22 | 0.33 | 0.35 | 0.53 | 0.74 | 0.83 | 0.94 | 0.98 |
| END | 0.17 | 0.23 | 0.33 | 0.35 | 0.53 | 0.74 | 0.82 | 0.94 | 0.98 |
| NEW.E0 | 0.17 | 0.26 | 0.36 | 0.39 | 0.54 | 0.73 | 0.82 | 0.92 | 0.98 |

Cuadro 3.4: Potencia obtenida con $\alpha = 0.05$

La principal característica de la Distribución de Laplace o Doble Exponencial es la forma exponencial de sus colas. Esta característica es suficiente para que todos los tests logren un

buen rendimiento. El test propuesto también consigue unos buenos resultados, destacando sobre todo en muestras de menor tamaño, donde logra detectar un mayor número de veces, aunque no es significativo.

Uniforme(0, 1)

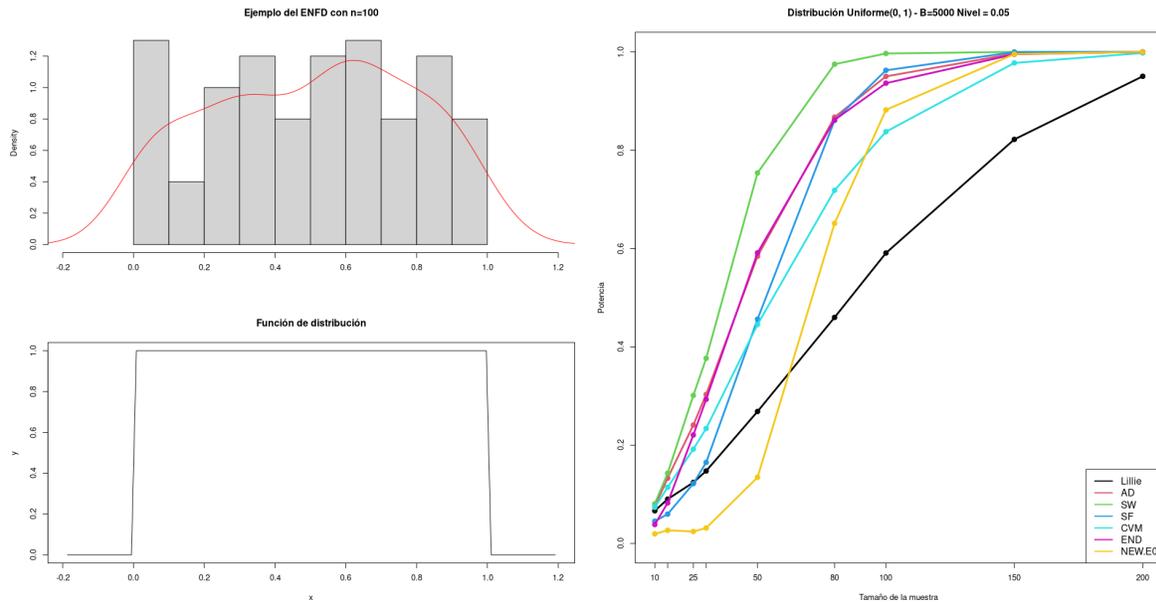


Figura 3.4: Distribución Uniforme(0, 1)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.07 | 0.09 | 0.12 | 0.15 | 0.27 | 0.46 | 0.59 | 0.82 | 0.95 |
| AD | 0.08 | 0.13 | 0.24 | 0.30 | 0.58 | 0.87 | 0.95 | 1.00 | 1.00 |
| SW | 0.08 | 0.14 | 0.30 | 0.38 | 0.75 | 0.97 | 1.00 | 1.00 | 1.00 |
| SF | 0.05 | 0.06 | 0.12 | 0.16 | 0.46 | 0.86 | 0.96 | 1.00 | 1.00 |
| CVM | 0.07 | 0.11 | 0.19 | 0.23 | 0.45 | 0.72 | 0.84 | 0.98 | 1.00 |
| END | 0.04 | 0.08 | 0.22 | 0.29 | 0.59 | 0.86 | 0.94 | 1.00 | 1.00 |
| NEW.E0 | 0.02 | 0.03 | 0.02 | 0.03 | 0.13 | 0.65 | 0.88 | 1.00 | 1.00 |

La distribución Uniforme en un intervalo se caracteriza porque su función de densidad es constante en el intervalo y cero en el resto. Con tamaños de muestra pequeños resulta muy complicado diferenciar los datos una distribución Normal, y hasta muestras de tamaño 50 no encontramos ningún test que alcance unos buenos resultados, pero a partir de este tamaño la mayoría de los tests logran detectar la distribución alternativa en un gran número de ocasiones.

Respecto al test propuesto, obtiene muy malos resultados hasta un tamaño considerablemente grande en comparación con algunos de los otros tests. El motivo por el que este test no logra unos buenos resultados es porque el ENFD no tiene buenas propiedades en distribuciones no continuas, debido al conocido efecto frontera. Sería interesante considerar

modificaciones del ENFD para corregir este efecto frontera, de forma que se evite distribuir la densidad hacia las zonas fuera del soporte de la distribución, pero esto podría ser objeto de trabajos futuros.

3.4.2. Distribuciones Asimétricas

En esta sección se describirá el comportamiento con funciones asimétricas. A priori se espera un desempeño superior al obtenido con las funciones simétricas, ya que sus características difieren en mayor grado de la normalidad.

Distribución exponencial

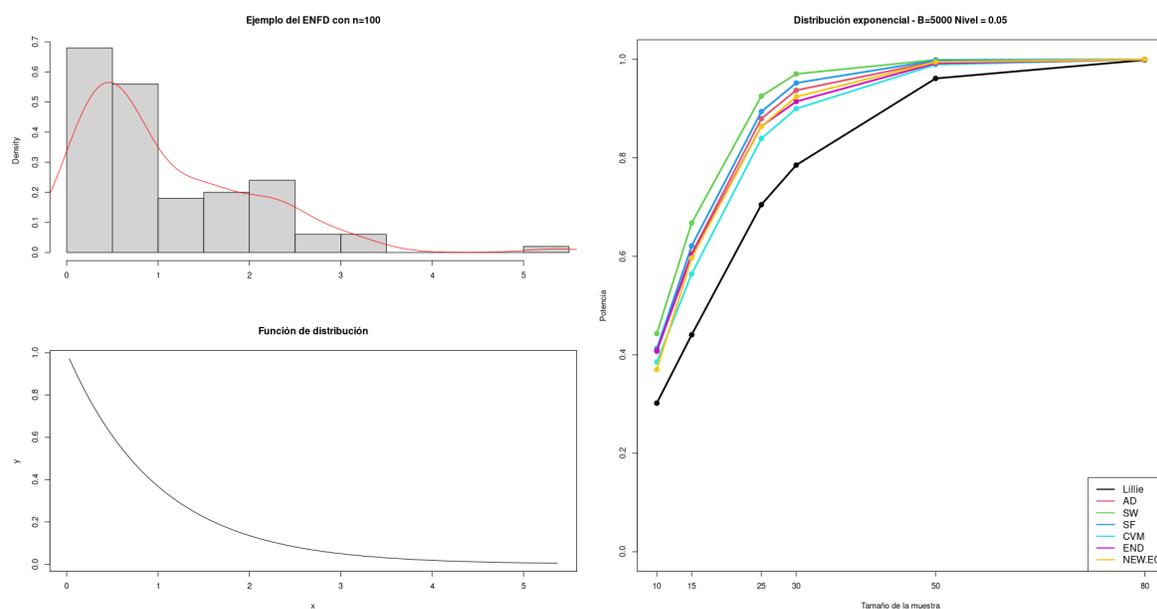


Figura 3.5: Distribución Exponencial

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.30 | 0.44 | 0.70 | 0.79 | 0.96 | 1.00 | 1.00 | 1.00 | 1.00 |
| AD | 0.41 | 0.61 | 0.88 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SW | 0.44 | 0.67 | 0.93 | 0.97 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SF | 0.41 | 0.62 | 0.89 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| CVM | 0.39 | 0.56 | 0.84 | 0.90 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| END | 0.41 | 0.60 | 0.86 | 0.91 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| NEW.E0 | 0.37 | 0.60 | 0.86 | 0.92 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |

Cuadro 3.5: Potencia obtenida con $\alpha = 0.05$

La distribución exponencial está definida solo para $x \geq 0$, lo que ya supone una gran ventaja a la hora de detectar las diferencias con la distribución de la H_0 . Ninguno de los tests

incluidos en el estudio necesita muestras de tamaño superior a 50 para discriminar de forma correcta todas las muestras hacia su distribución real. En esta situación volvemos a encontrar el ya comentado efecto frontera, en este caso con una discontinuidad en $x=0$.

Distribución Weibull(4, 1)

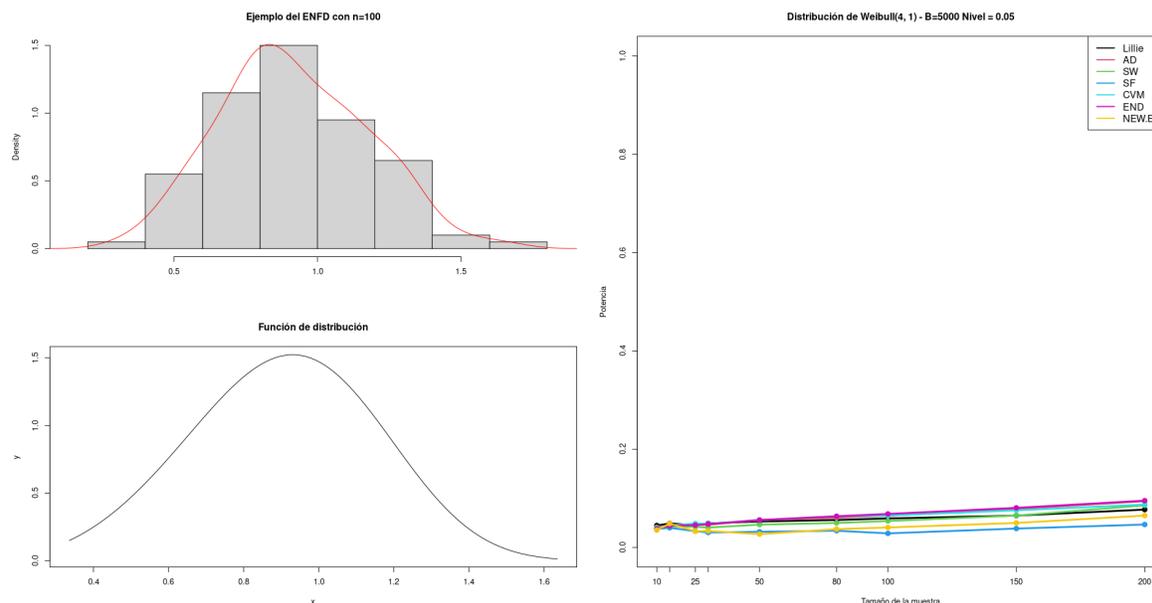


Figura 3.6: Distribución Weibull(4, 1)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.06 | 0.06 | 0.06 | 0.08 |
| AD | 0.04 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.06 | 0.08 | 0.09 |
| SW | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.05 | 0.05 | 0.06 | 0.09 |
| SF | 0.04 | 0.04 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.04 | 0.05 |
| CVM | 0.04 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.06 | 0.08 | 0.09 |
| END | 0.04 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.07 | 0.08 | 0.10 |
| NEW.EO | 0.04 | 0.05 | 0.03 | 0.03 | 0.03 | 0.04 | 0.04 | 0.05 | 0.06 |

Cuadro 3.6: Potencia obtenida con $\alpha = 0.05$

La distribución de Weibull es altamente asimétrica, sobre todo con parámetros de forma pequeños. Para este estudio se ha incluido la distribución de Weibull(4, 1) que se asemeja más a una distribución Normal. Podemos observar tanto en el gráfico con los resultados como en la tabla que ninguno de los tests es capaz de detectar estas diferencias con los tamaños propuestos para el estudio, no superando en ningún caso una potencia superior a 0.1.

En la Fig. 3.7 se puede observar el porqué de estos resultados. Al superponer la función de densidad de una Normal con la media y varianza estimadas de una muestra de tamaño $n = 200$ proveniente de esta distribución se puede ver que apenas hay diferencias notables entre las dos distribuciones.

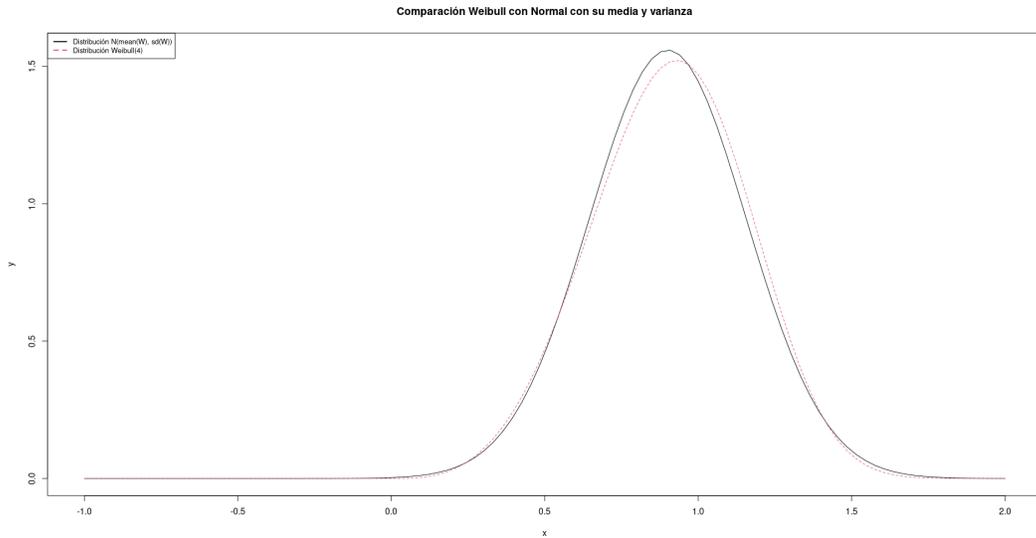


Figura 3.7: Distribución Weibull(4, 1) junto a una distribución Normal con su misma media y varianza

Con el objetivo de tratar de cuantificar la potencia de los tests hasta unos resultados significativos, se ha repetido el proceso de cálculo de la potencia para tamaños superiores, probando con muestras de tamaños 400, 600, 800 y 1000.

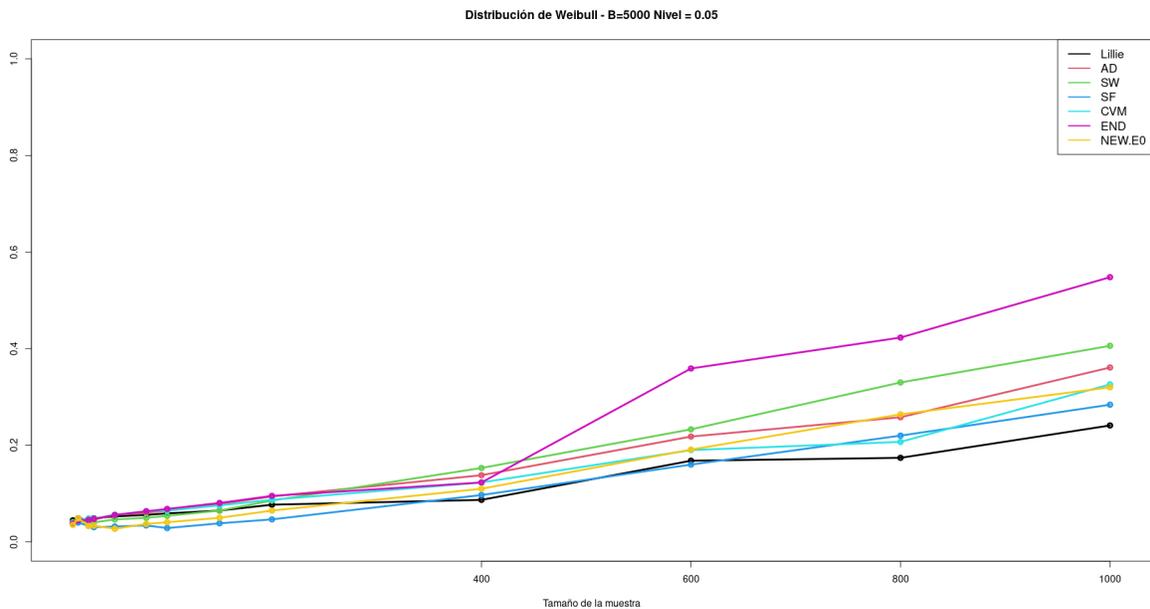


Figura 3.8: Distribución Weibull(4, 1)

Se ve como al aplicar una de las técnicas comentadas anteriormente para aumentar la potencia ésta se ve incrementada, pero no lo suficiente como para considerar fiables a los métodos para esta distribución. Tan solo el test basado en el Estimador Núcleo de la Función de Densidad consigue despegarse de la tendencia general pero sin lograr unos resultados que se puedan considerar fiables para tamaños de muestra de 1000 observaciones.

Distribución Weibull(1, 1.5)

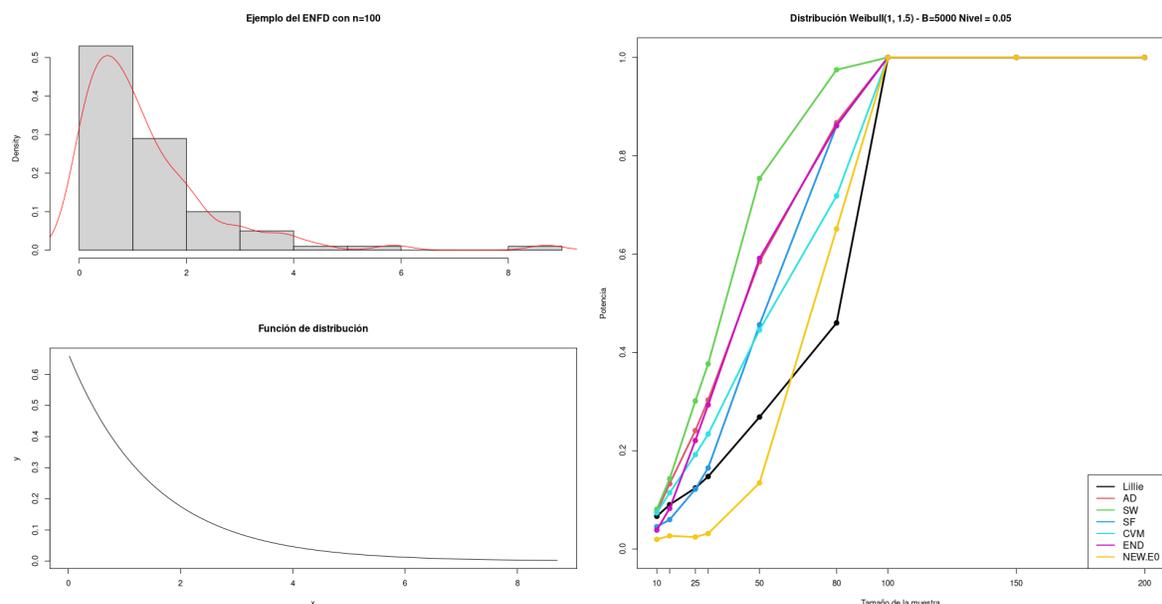


Figura 3.9: Distribución Weibull(1, 1.5)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.07 | 0.09 | 0.12 | 0.15 | 0.27 | 0.46 | 1.00 | 1.00 | 1.00 |
| AD | 0.08 | 0.13 | 0.24 | 0.30 | 0.58 | 0.87 | 1.00 | 1.00 | 1.00 |
| SW | 0.08 | 0.14 | 0.30 | 0.38 | 0.75 | 0.97 | 1.00 | 1.00 | 1.00 |
| SF | 0.05 | 0.06 | 0.12 | 0.16 | 0.46 | 0.86 | 1.00 | 1.00 | 1.00 |
| CVM | 0.07 | 0.11 | 0.19 | 0.23 | 0.45 | 0.72 | 1.00 | 1.00 | 1.00 |
| END | 0.04 | 0.08 | 0.22 | 0.29 | 0.59 | 0.86 | 1.00 | 1.00 | 1.00 |
| NEW.E0 | 0.02 | 0.03 | 0.02 | 0.03 | 0.13 | 0.65 | 1.00 | 1.00 | 1.00 |

Con un parámetro de forma menor, podemos observar como todos los tests tienden a detectar en un mayor número de veces esta distribución alternativa. Si bien el resultado del test propuesto no es bueno para tamaños de muestra pequeños, sí que logra alcanzar buenos resultados a partir de muestras con 100 observaciones.

El motivo por que el test propuesto vuelve a no tener un buen comportamiento en tamaños pequeños es, otra vez, debido al efecto frontera provocado por la discontinuidad de la distribución que se plantea como alternativa.

Distribución Chi-Cuadrado

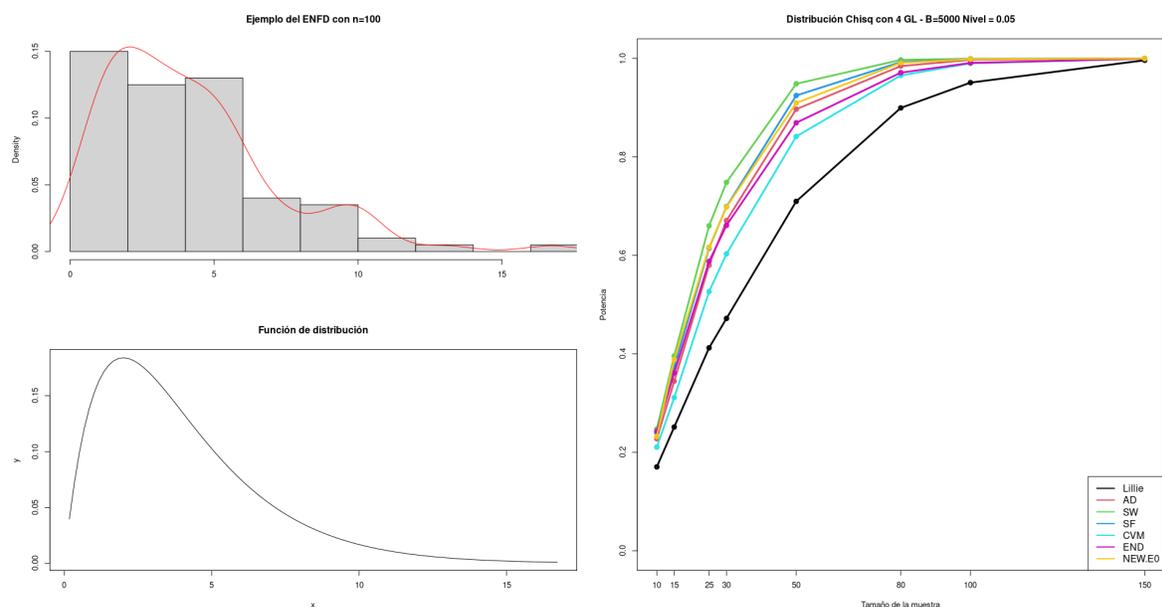


Figura 3.10: Distribución Chi-Cuadrado con 4 GL

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.17 | 0.25 | 0.41 | 0.47 | 0.71 | 0.90 | 0.95 | 1.00 | 1.00 |
| AD | 0.23 | 0.34 | 0.58 | 0.67 | 0.90 | 0.98 | 1.00 | 1.00 | 1.00 |
| SW | 0.25 | 0.40 | 0.66 | 0.75 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 |
| SF | 0.24 | 0.37 | 0.61 | 0.70 | 0.92 | 0.99 | 1.00 | 1.00 | 1.00 |
| CVM | 0.21 | 0.31 | 0.53 | 0.60 | 0.84 | 0.97 | 0.99 | 1.00 | 1.00 |
| END | 0.24 | 0.36 | 0.59 | 0.66 | 0.87 | 0.97 | 0.99 | 1.00 | 1.00 |
| NEW.E0 | 0.23 | 0.39 | 0.62 | 0.70 | 0.91 | 0.99 | 1.00 | 1.00 | 1.00 |

Cuadro 3.7: Potencia obtenida con $\alpha = 0.05$

La distribución Chi-cuadrado con 4 grados de libertad es una distribución con una fuerte asimetría. Esto es una gran diferencia fácilmente detectable por todos los tests, lo que conlleva que la gran mayoría logren valores de la potencia muy altos a partir de tamaños muestrales superiores a las 100 observaciones. Nada reseñable respecto al test propuesto, que se comporta de manera similar a los demás.

Lognormal

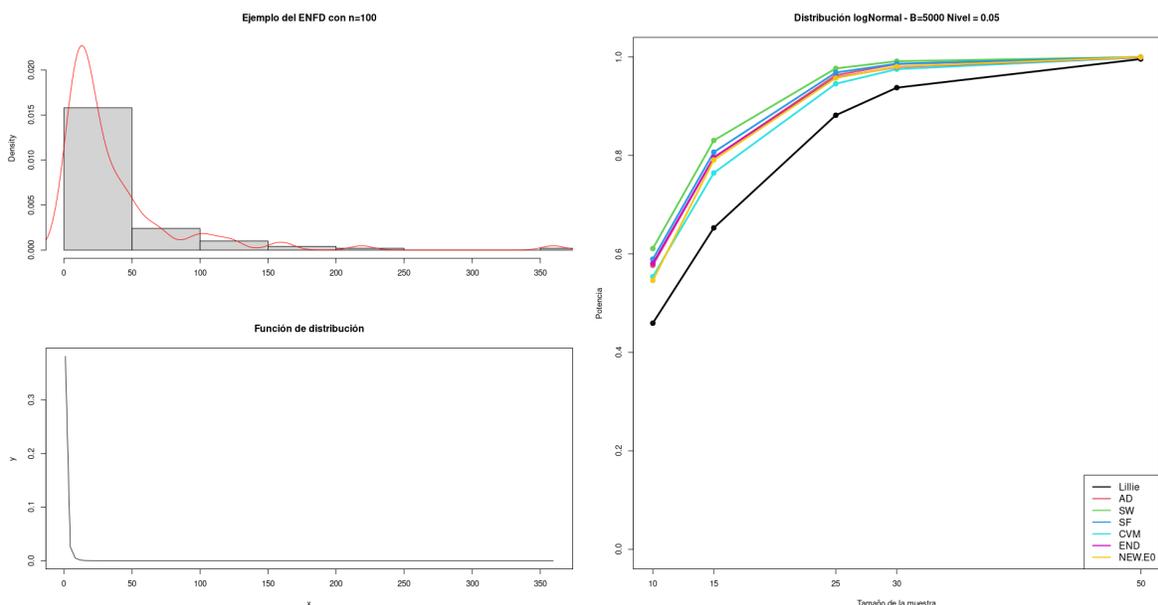


Figura 3.11: Distribución LogNormal(0, 1)

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.46 | 0.65 | 0.88 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| AD | 0.58 | 0.80 | 0.96 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SW | 0.61 | 0.83 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| SF | 0.59 | 0.81 | 0.97 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| CVM | 0.55 | 0.76 | 0.95 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| END | 0.58 | 0.80 | 0.96 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| NEW.E0 | 0.55 | 0.79 | 0.96 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Cuadro 3.8: Potencia obtenida con $\alpha = 0.05$

La función lognormal está definida como una distribución de probabilidad continua de una variable aleatoria cuyo logaritmo está normalmente distribuido. Agrava aún más las diferencias que se han destacado de la distribución Chi-cuadrado por lo que, como es natural, los comportamientos de los tests son superiores. Con muestras de tamaños superiores a las 80 observaciones ya se logra que todos los tests rondan el 100 % de acierto en la diferenciación con la distribución normal.

3.4.3. Mixturas

Una mixtura es un modelo de probabilidad que se utiliza para representar una población de datos que puede tener diferentes subpoblaciones con distintas distribuciones. En el caso de nuestro estudio se incluyen mixturas normales con distintas características. Los datos observados son generados por varias distribuciones normales, cada una con su media y varianza,

donde cada una de las distribuciones tiene un peso en la mixtura.

$$f(x) = p\phi(x, \mu_1, \sigma_1) + (1 - p)\phi(x, \mu_2, \sigma_2) \quad (3.2)$$

Mixtura normal

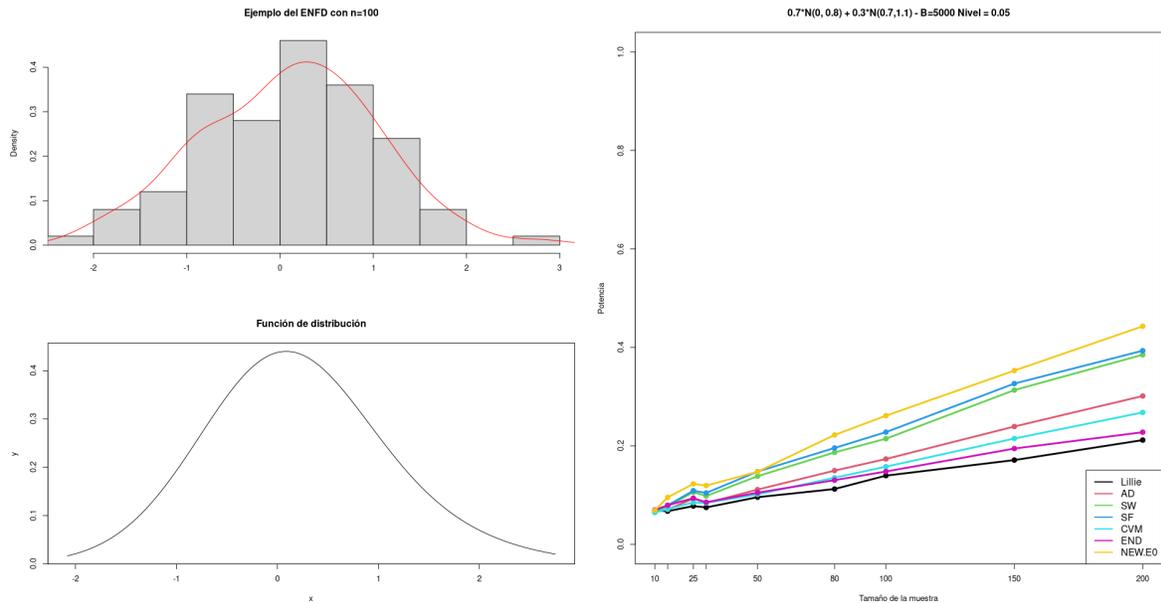


Figura 3.12: Mixtura con aspecto Normal

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.07 | 0.07 | 0.08 | 0.07 | 0.10 | 0.11 | 0.14 | 0.17 | 0.21 |
| AD | 0.07 | 0.07 | 0.09 | 0.08 | 0.11 | 0.15 | 0.17 | 0.24 | 0.30 |
| SW | 0.07 | 0.08 | 0.11 | 0.10 | 0.14 | 0.19 | 0.21 | 0.31 | 0.39 |
| SF | 0.07 | 0.08 | 0.11 | 0.10 | 0.15 | 0.20 | 0.23 | 0.33 | 0.39 |
| CVM | 0.06 | 0.07 | 0.09 | 0.08 | 0.10 | 0.14 | 0.16 | 0.21 | 0.27 |
| END | 0.07 | 0.08 | 0.09 | 0.09 | 0.10 | 0.13 | 0.15 | 0.19 | 0.23 |
| NEW.EO | 0.07 | 0.10 | 0.12 | 0.12 | 0.15 | 0.22 | 0.26 | 0.35 | 0.44 |

Cuadro 3.9: Potencia obtenida con $\alpha = 0.05$

La primera mixtura incluida es una combinación entre una $N(0, 0.8)$ y una $N(0.7, 1.1)$, donde la primera distribución tiene un peso de 0.7 y la segunda el 0.3 restante. Como se puede observar las medias no están muy alejadas y su función de densidad teórica se asemeja a una distribución normal. Para los tamaños de muestra incluidos a priori en el estudio, ninguno de los tests obtiene muy buenos resultados, siendo el test propuesto el que mejores resultados obtiene.

Con objeto de resolver si estos resultados se deben a algún factor externo o si simplemente es necesario aumentar los tamaños muestrales para aumentar la potencia, se realiza el estudio para muestras de hasta tamaño 1000.

Como se comentó anteriormente, otro posible mecanismos para aumentar la potencia es el aumento del nivel de significación, pero no se contempla en este estudio modificar este parámetro.

Los resultados obtenidos se muestran en la Fig. 3.13 y se observa con claridad que, al aumentar los tamaños de la muestra, los resultados sufren una notable mejoría.

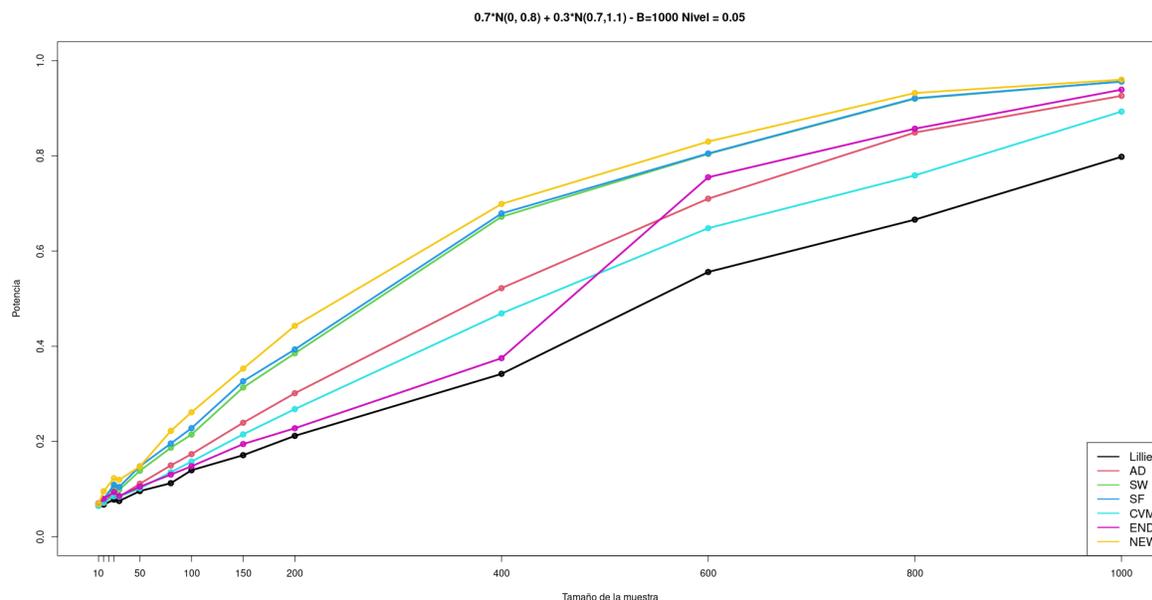


Figura 3.13: Aumento de tamaño en la mixtura $0.7*N(0, 0.8) + 0.3*N(0.7, 1.1)$

| | 400 | 600 | 800 | 1000 |
|--------|-------------|-------------|-------------|-------------|
| Lillie | 0.34 | 0.56 | 0.67 | 0.80 |
| AD | 0.52 | 0.71 | 0.85 | 0.93 |
| SW | 0.67 | 0.80 | 0.92 | 0.96 |
| SF | 0.68 | 0.81 | 0.92 | 0.96 |
| CVM | 0.47 | 0.65 | 0.76 | 0.89 |
| END | 0.38 | 0.76 | 0.86 | 0.94 |
| NEW.E0 | 0.70 | 0.83 | 0.93 | 0.96 |

Se demuestra así que los valores de la potencia bajos simplemente se deben a que el tamaño de las muestras no son suficientemente grandes para poder detectar estas diferencias. Con tamaños muestrales de 1000 observaciones los tests logran un muy buen rendimiento, siendo el rendimiento del test propuesto superior para todos los tamaños comparados sobre esta distribución

Mixtura bimodal

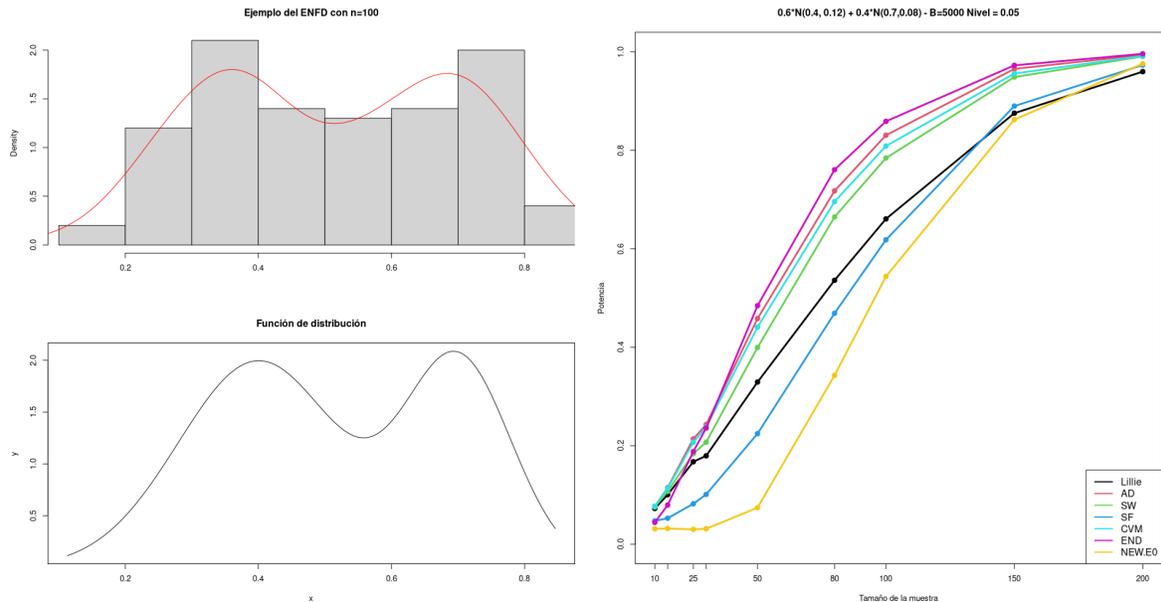


Figura 3.14: Mixtura de Normales con aspecto Bimodal

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.07 | 0.10 | 0.17 | 0.18 | 0.33 | 0.54 | 0.66 | 0.88 | 0.96 |
| AD | 0.08 | 0.11 | 0.21 | 0.24 | 0.46 | 0.72 | 0.83 | 0.97 | 0.99 |
| SW | 0.08 | 0.11 | 0.18 | 0.21 | 0.40 | 0.66 | 0.78 | 0.95 | 0.99 |
| SF | 0.05 | 0.05 | 0.08 | 0.10 | 0.22 | 0.47 | 0.62 | 0.89 | 0.97 |
| CVM | 0.08 | 0.11 | 0.21 | 0.24 | 0.44 | 0.70 | 0.81 | 0.96 | 0.99 |
| END | 0.04 | 0.08 | 0.19 | 0.24 | 0.48 | 0.76 | 0.86 | 0.97 | 1.00 |
| NEW.E0 | 0.03 | 0.03 | 0.03 | 0.03 | 0.07 | 0.34 | 0.54 | 0.86 | 0.98 |

Cuadro 3.10: Potencia obtenida con $\alpha = 0.05$

Se considera interesante también incluir en el estudio una mixtura bimodal, definida por $0,6 * N(0,4, 0,12) + 0,4 * N(0,7, 0,08)$. A priori, al ser una distribución bimodal, los tests deberían ser capaces de detectar las diferencias con la distribución Normal.

Como se observa tanto en la tabla como en la figurada aportadas, los resultados de la mayoría de los tests cumplen con las expectativas, pero sorprende ver como el test propuesto, que había logrado unos resultados notablemente superiores para otras distribuciones no los obtenga sobre una distribución que podría considerarse más sencilla.

Se puede observar en la Fig 3.15 que, para el parámetro de suavizado considerado, el óptimo con referencia a la normal [2.22], el Estimador Núcleo de la Función de Densidad (ENFD) está sobreesuavizando, es decir que se obtendrían estimadores más precisos con los menores valores de h .

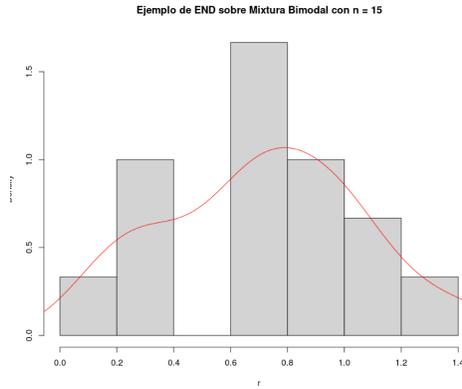


Figura 3.15: ENFD en mixtura bimodal

Con objeto de demostrar el porqué del mal, o peor, comportamiento del test propuesto en estos casos de sobreesuavización, se ha realizado una simulación con diferentes valores en el entorno del parámetro de suavizado. Para la construcción del estadístico se ha utilizado la aproximación normal óptima del parámetro [2.22], descrita anteriormente.

Como se ha explicado en varias ocasiones durante este documento, el parámetro de suavizado es el encargado de regular el número de sumando no nulos que se tienen en consideración a la hora de realizar la estimación de la densidad en cada uno de los puntos, o dicho al revés, la zona de influencia en la estimación de la densidad que tiene cada uno de los puntos.

En la figura 3.16 se puede apreciar el aumento en la potencia usando valores menores de h , obteniendo para $h = 0,1 * hnorm(x)$ valores de la potencia parecidos a los obtenidos con el test de SF, incluso en algunos casos superiores.

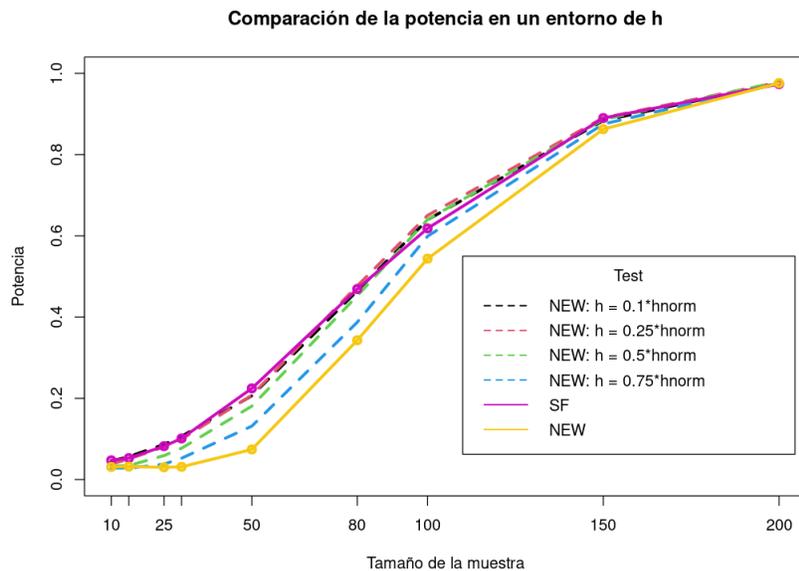


Figura 3.16: Estudio en el entorno del parámetro de suavizado

Mixtura contaminada por media

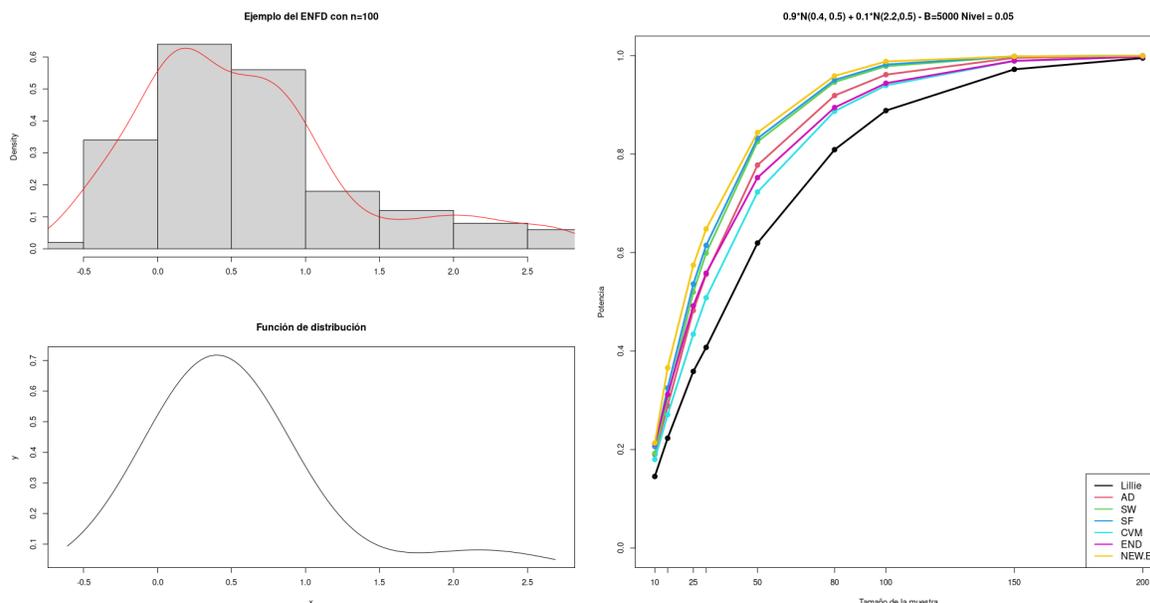


Figura 3.17: Mixtura contaminada por la media

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.15 | 0.22 | 0.36 | 0.41 | 0.62 | 0.81 | 0.89 | 0.97 | 1.00 |
| AD | 0.19 | 0.29 | 0.48 | 0.56 | 0.78 | 0.92 | 0.96 | 1.00 | 1.00 |
| SW | 0.19 | 0.31 | 0.52 | 0.60 | 0.83 | 0.95 | 0.98 | 1.00 | 1.00 |
| SF | 0.21 | 0.33 | 0.54 | 0.61 | 0.83 | 0.95 | 0.98 | 1.00 | 1.00 |
| CVM | 0.18 | 0.27 | 0.43 | 0.51 | 0.72 | 0.89 | 0.94 | 0.99 | 1.00 |
| END | 0.21 | 0.31 | 0.49 | 0.56 | 0.75 | 0.89 | 0.94 | 0.99 | 1.00 |
| NEW.EO | 0.21 | 0.37 | 0.57 | 0.65 | 0.84 | 0.96 | 0.99 | 1.00 | 1.00 |

Cuadro 3.11: Potencia obtenida con $\alpha = 0.05$

Una mixtura contaminada es una mixtura en la que la mayor parte de los datos, el 90 % en este caso, de la muestra, provienen de una distribución, y una pequeña parte, el 10 % en este ejemplo, de otra, siendo esta segunda distribución la considerada como contaminación.

Una mixtura contaminada por media se compone de dos Normales en las que hay una diferencia considerable entre las medias, logrando así una distribución bimodal pero con una clara diferencia en el peso de las dos modas. Al igual que en la mixtura bimodal comentada en la sección anterior, los tests obtienen una potencia considerablemente alta. En este caso el test propuesto sí es capaz de detectar correctamente la hipótesis alternativa y obtiene los mejores resultados.

Mixtura contaminada por varianza

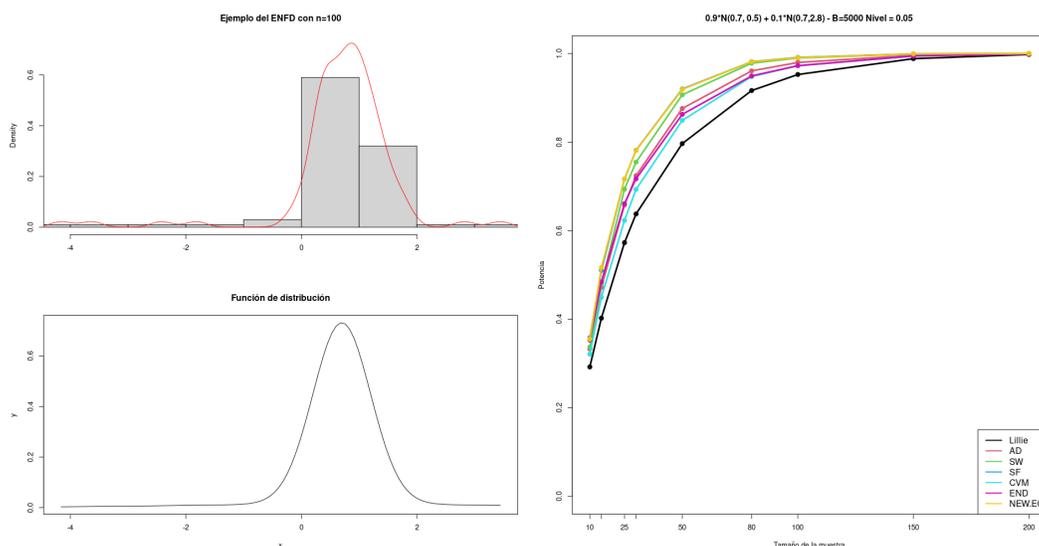


Figura 3.18: Mixtura contaminada por la varianza

| | 10 | 15 | 25 | 30 | 50 | 80 | 100 | 150 | 200 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Lillie | 0.29 | 0.40 | 0.57 | 0.64 | 0.80 | 0.92 | 0.95 | 0.99 | 1.00 |
| AD | 0.33 | 0.47 | 0.66 | 0.72 | 0.88 | 0.96 | 0.98 | 1.00 | 1.00 |
| SW | 0.34 | 0.48 | 0.69 | 0.75 | 0.91 | 0.98 | 0.99 | 1.00 | 1.00 |
| SF | 0.36 | 0.51 | 0.72 | 0.78 | 0.92 | 0.98 | 0.99 | 1.00 | 1.00 |
| CVM | 0.32 | 0.45 | 0.62 | 0.69 | 0.85 | 0.95 | 0.97 | 0.99 | 1.00 |
| END | 0.35 | 0.48 | 0.66 | 0.72 | 0.86 | 0.95 | 0.97 | 0.99 | 1.00 |
| NEW.E0 | 0.36 | 0.52 | 0.72 | 0.78 | 0.92 | 0.98 | 0.99 | 1.00 | 1.00 |

Cuadro 3.12: Potencia obtenida con $\alpha = 0.05$

El caso de una Mixtura contaminada por varianza es análogo a la contaminación por media, con la diferencia de que en este caso las dos distribuciones Normales tiene un valor común en la media, y una diferencia notable en la varianza. A simple vista, la función de distribución es muy similar a la función de distribución de una Normal, pero al realizar la estimación a través del ENFD se puede observar mejor esta contaminación.

Si bien la vista no es fiable, los procedimientos numéricos sí lo son, y estos sí son capaces de detectar esta diferencia en un gran número de veces. Cabe destacar que con muestras de tan solo 15 observaciones muchos de los tests logran detectar la mitad de las muestras como no normales, lo que es un gran resultado. En este caso, el test propuesto obtiene de nuevo los mejores resultados, lo que certifica su buen comportamiento.

Capítulo 4

Conclusiones y trabajos futuros

A la vista de los resultados obtenidos en la comparación de los tests de Normalidad se pueden extraer algunas conclusiones firmes sobre el nuevo test propuesto.

Para el caso de las Distribuciones Simétricas, podemos concluir que el nuevo test obtiene muy buenos resultados a la hora de detectar este tipo de distribuciones cuando se plantean como alternativa, siendo superior a muchos de los otros procedimientos empleados y obteniendo resultados parecidos a los del test de SF, salvo en el caso de la distribución $U(0, 1)$, debido al ya comentado efecto frontera del ENFD cuando el soporte está acotado y la función de densidad no es continua en los puntos frontera.

A la hora de detectar como alternativa Distribuciones Asimétricas basta con comentar que los resultados son acordes al resto, ya que este tipo de distribuciones no presentan demasiados problemas a la hora de distinguirlos de una distribución Normal ya que tienen unas características ampliamente distintas, por lo que no supone un reto para ninguno de los tests más utilizados.

El punto más interesante de debate y donde resultan más llamativos los resultados del nuevo test es a la hora de detectar Mixturas de Normales como alternativa. El mayor problema que presentan las mixturas es que, en el caso unimodal, la mayoría de los test presentan dificultades para detectar la alternativa. Es en estos casos precisamente donde el test propuesto logra obtener los mejores resultados, siendo este un resultado muy significativo debido a la dificultad que esto supone. Los peores resultados obtenidos son para el caso de la mixtura bimodal planteada, donde su rendimiento es claramente inferior debido al parámetro de suavizado empleado. Como se ha visto, este rendimiento mejora con la siminución de h .

Tras analizar todos los resultados obtenidos, se puede concluir que el test propuesto obtiene unos resultados similares al test de Shapiro-Francia, siendo superior en algunos casos y, para aquellos en los que obtiene un rendimiento inferior, se ha podido encontrar la causa en la elección del parámetro de suavizado.

Es por ello que se plantea como trabajo futuro a partir de este documento el análisis de diferentes técnicas de elección del parámetro de suavizado, así como considerando el ENFD que mitigue el efecto frontera para funciones de densidad con soporte acotado. También puede resultar de interés ampliar el estudio para el caso de la mixturas, incluyendo mixturas de más de dos normales.

Bibliografía

- [1] Nusrat Khatun. «Applications of Normality Test in Statistical Analysis». En: *Open Journal of Statistics* 11 (2021), págs. 113-122. DOI: [10.4236/ojs.2021.111006](https://doi.org/10.4236/ojs.2021.111006). URL: <https://doi.org/10.4236/ojs.2021.111006>.
- [2] B. W. Yap y C. H. Sim. «Comparisons of various types of normality tests». En: *Journal of Statistical Computation and Simulation* 81.12 (2011), págs. 2141-2155. DOI: [10.1080/00949655.2010.520163](https://doi.org/10.1080/00949655.2010.520163). URL: <https://doi.org/10.1080/00949655.2010.520163>.
- [3] Justine Rochon, Matthias Gondan y Meinhard Kieser. «To test or not to test: Preliminary assessment of normality when comparing two independent samples». En: *BMC Med Res Methodol* 12.1 (2012), pág. 81. DOI: [10.1186/1471-2288-12-81](https://doi.org/10.1186/1471-2288-12-81). URL: <https://doi.org/10.1186/1471-2288-12-81>.
- [4] Björn Lantz. «The impact of sample non-normality on ANOVA and alternative methods». En: *British Journal of Mathematical and Statistical Psychology* 65.2 (2012), págs. 295-313. DOI: [10.1111/j.2044-8317.2012.02047.x](https://doi.org/10.1111/j.2044-8317.2012.02047.x). URL: <https://doi.org/10.1111/j.2044-8317.2012.02047.x>.
- [5] Edith Seier. «Comparison of tests for univariate normality». En: *Interstat* 1 (2002), págs. 1-17.
- [6] J. Gross y U. Ligges. *nortest: Tests for Normality*. R package version 1.0-4. 2015. URL: <https://cran.r-project.org/web/packages/nortest/index.html>.
- [7] Adrian Bowman y Adelchi Azzalini. *Applied Smoothing Techniques for Data Analysis: the kernel approach with S-Plus illustrations*. Oxford University Press, 1997.
- [8] Amalia. *Tesis Amalia*. 2005. URL: http://dm.udc.es/profesores/ricardo/Archivos/Tesis_amalia.pdf.
- [9] Adelchi Azzalini y Adrian W. Bowman. *A Package for Smooth Univariate Density Estimation*. R package version 2.2.7. 2019. URL: <https://cran.r-project.org/package=sm>.
- [10] Klaus Nordhausen, David A. van Dyk, and Esa J. Kärkäinen. *norlmix: Normal Mixture Models*. R package version 1.2-4. 2021. URL: <https://cran.r-project.org/package=norlmix>.
- [11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: <https://www.R-project.org/>.

Apéndice A

Funciones desarrolladas en R

A.1. Cálculo del estadístico para el test propuesto

```
library(norlmix)
library(sm)
newEstadistico.E0 <- function(data){

  x <- (data - mean(data))/sd(data)
  h <- hnorm(x)
  n <- length(data)

  # Generacion de la mixtura que necesitamos
  nmix <- norMix(mu = x, sigma = rep(h, n), w = rep(1/n, n))

  # Generacion de los puntos en los que queremos los cuantiles
  i <- seq(1, n, by = 1)

  # Ajuste de los puntos
  i <- (i - (3/8)) / (n + (1/4))

  # Cuantiles de la esperanza del estimador  $\sim N(0, \text{sqrt}(h^2 + 1))$  bajo H0
  QH0 <- qnorm(i, sd = sqrt(h^2 + 1))

  # Cuantiles teoricos del ENFD
  QH1 <- qnorMix(i, nmix)

  W <- cor(QH0, QH1)^2

  return(W)
}
```

A.2. Simulacion de Valores Críticos

```

library('sm')
library('noritest')

calculaFrontera_test <- function(nsim, tam){

  sim_Lillie <- matrix(nrow = length(tam), ncol = nsim)
  sim_AD <- matrix(nrow = length(tam), ncol = nsim)
  sim_SW <- matrix(nrow = length(tam), ncol = nsim)
  sim_SF <- matrix(nrow = length(tam), ncol = nsim)
  sim_CVM <- matrix(nrow = length(tam), ncol = nsim)
  sim_nise <- matrix(nrow = length(tam), ncol = nsim)

  aux <-1

  for(nn in tam){

    print('-----')
    print(nn)
    print('-----')

    for(i in 1:nsim){
      # Generamos una distribucion normal
      sim <- rnorm(nn)

      # Calculamos los estadisticos
      lil <- lillie.test(sim)$statistic
      ad <- ad.test(sim)$statistic
      sw <- shapiro.test(sim)$statistic
      sf <- sf.test(sim)$statistic
      cvm <- cvm.test(sim)$statistic
      nise <- nise(sim)

      if(i%%1000 == 0){
        print(i)
      }

      # Anadimos al resultado al conjunto de la simulacion
      sim_Lillie[aux, i] <- lil
      sim_AD[aux, i] <- ad
      sim_SW[aux, i] <- sw
      sim_SF[aux, i] <- sf
      sim_CVM[aux, i] <- cvm
      sim_nise[aux, i] <- nise

    }

    aux <- aux + 1
  }

  filename <- paste('./data/sim-', nsim, '-test_', tam[1], '-', tam[length(tam)
]), '.RData', sep = '')
  tam_sim_test <- tam

```

```

save(sim_Lillie , sim_AD, sim_SW, sim_SF, sim_CVM, sim_nise , tam_sim_
test , file = filename)

}

calculaFronteraNew <- function(nsim , tam){

sim_new.E0 <- matrix(nrow = length(tam), ncol = nsim)
aux <-1

for(nn in tam){
  print('-----')
  print(nn)
  print('-----')
  for(i in 1:nsim){
    # Generamos una distribucion normal
    sim <- rnorm(nn)

    ## new test
    new.E0 <- newEstadistico.E0(sim)
    if(i%%1000 == 0){
      print(i)
    }

    sim_new.E0[aux , i] <- new.E0

  }

  aux <- aux + 1
}

filename <- paste('./data/sim-', nsim , '-new01E0-', tam[1] , '-', tam[length(
tam)], '.RData', sep = '')
tam_sim_new <- tam
save(sim_new.E0, tam_sim_new, file = filename)

}

B <- 100000
n <- c(10, 15, 25, 30, 50, 80, 100, 150, 200)

calculaFrontera_test(nsim = B, tam = n)

calculaFronteraNew(B, n)

```

A.3. Cálculo de la Potencia

```

## Calculo de la potencia de un test de normalidad:

library(sm)
library(nortest)
library(norlmix)
library(extraDistr)

##### 1. Definicion de las condiciones del test ####
test <- c('Lillie', 'AD', 'SW', 'SF', 'CVM', 'END', 'NEW.E0' )

##### 2. Cargo los datos de los Valores Criticos ####
load('./data/ValoresCriticos/c_005_L_AD_SW_SF_CVM_END_NEWE0.RData')
c.005

##### 4. Calculo de la potencia ####
calculaPotencia <- function(n, B, test, c.alfa, H1, potencia){

  potencia <- data.frame(matrix(nrow = length(test), ncol = length(n)))

  rownames(potencia) <- test
  colnames(potencia) <- n

  for(nn in n ){

    P.l <- 0
    P.ad <- 0
    P.sw <- 0
    P.sf <- 0
    P.cvm <- 0
    P.end <- 0

    P.new.E0 <- 0

    print(nn)
    for(i in 1:B){
      # Comportamiento de los test en funcion de la distribucion

      # SIM TRICAS
      if(H1 == 'tstudent_3'){
        simdata <- rt(nn, 3)
      }else if(H1 == 'cauchy'){
        simdata <- rcauchy(nn)
      }else if(H1 == 'laplace'){
        # Laplace o doble exponencial
        simdata <- rlaplace(nn)
      }else if(H1 == 'lognormal'){
        simdata <- rlnorm(nn)

      }else if(H1 == 'uniforme'){
        simdata <- runif(nn)
      }else

      # MIXTURAS
      if(H1 == 'mix.bimodal'){

```

```

    nmix <- norMix(mu = c(0.4, 0.7), sigma = c(0.12,0.08), w = c(0.6,
0.4))
    simdata <- rnorMix(n = nn, nmix)
  }else if(H1 == 'mix.norm'){
    nmix <- norMix(mu = c(0, 0.7), sigma = c(0.8,1.1), w = c(0.7,
0.3))
    simdata <- rnorMix(n = nn, nmix)
  }else if(H1 == 'mix.contam'){
    nmix <- norMix(mu = c(0.4, 0.7), sigma = c(0.5,1), w = c(0.9,
0.1))
    simdata <- rnorMix(n = nn, nmix)
  }else if(H1 == 'mix.contam.media'){
    nmix <- norMix(mu = c(0.4, 2.2), sigma = c(0.5, 0.5), w = c(0.9,
0.1))
    simdata <- rnorMix(n = nn, nmix)
  }else if(H1 == 'mix.contam.var'){
    nmix <- norMix(mu = c(0.7, 0.7), sigma = c(0.5, 2.8), w = c(0.9,
0.1))
    simdata <- rnorMix(n = nn, nmix)
  }else

  # ASIMTRICAS
  if(H1 == 'weibull'){
    simdata <- rweibull(nn, 4)
  }else if(H1 == 'weibull1_15'){
    simdata <- rweibull(nn, 1, 1.5)
  }else if(H1 == 'exp'){
    simdata <- rexp(nn)
  }else if(H1 == 'chisq'){
    simdata <- rchisq(nn, 4)
  }else{
    warning('Distribucion no disponible')
    return()
  }

  # construimos el estadistico
  t.lillie <- lillie.test(simdata)$statistic
  t.ad <- ad.test(simdata)$statistic
  t.sw <- shapiro.test(simdata)$statistic
  t.sf <- sf.test(simdata)$statistic
  t.cvm <- cvm.test(simdata)$statistic
  t.end <- nise(simdata)

  t.new.E0 <- newEstadistico.E0(simdata)

  # Comparamos con cada uno de los valores cr ticos con alfa 0.05
  P.l <- P.l + (t.lillie > c.alfa[as.character(nn), 'c.alfa.lillie'])
  P.ad <- P.ad + (t.ad > c.alfa[as.character(nn), 'c.alfa.AD'])
  P.sw <- P.sw + (t.sw < c.alfa[as.character(nn), 'c.alfa.SW'])
  P.sf <- P.sf + (t.sf < c.alfa[as.character(nn), 'c.alfa.SF'])
  P.cvm <- P.cvm + (t.cvm > c.alfa[as.character(nn), 'c.alfa.CVM'])
  P.end <- P.end + (t.end > c.alfa[as.character(nn), 'c.alfa.end'])

  P.new.E0 <- P.new.E0 + (t.new.E0 < c.alfa[as.character(nn), 'c.alfa.
new.E0'])

```

```

}

potencia[ 'Lillie' ,as.character(nn)] <- P.l/B
potencia[ 'AD' ,as.character(nn)] <- P.ad/B
potencia[ 'SW' ,as.character(nn)] <- P.sw/B
potencia[ 'SF' ,as.character(nn)] <- P.sf/B
potencia[ 'CVM' ,as.character(nn)] <- P.cvm/B
potencia[ 'END' ,as.character(nn)] <- P.end/B
potencia[ 'NEW.E0' ,as.character(nn)] <- P.new.E0/B

}
return(potencia)
}

```

A.4. Generacion de gráficos

```

plotPotencia3 <- function(df, n, B, alfa, title, H1, ylim = c(0,1),
  legend.coords = 'bottomright'){

  layout(matrix(c(1,2, 3, 3), ncol = 2))

  l <- 100
  # SIMETRICAS
  if(H1 == 'tstudent_3'){
    r <- rt(l, 3)
    d <- density(r)

    x <- seq(min(r), max(r), length = l)
    y <- dt(x, 3)

    title <- 'T-student con 3 GL'
  }else if(H1 == 'cauchy'){

    r <- rcauchy(l, 3)
    d <- density(r)

    x <- seq(min(r), max(r), length = l)
    y <- dcauchy(x)

    title <- 'Distribucion de Cauchy'
  }else if(H1 == 'laplace'){
    # Laplace o doble exponencial

    r <- rlaplace(l)
    d <- density(r)

    x <- seq(min(r), max(r), length = l)
    y <- dlaplace(x)

    title <- 'Distribucion de Laplace o Doble exponencial'
  }
}

```

```

}else if(H1 == 'lognormal'){

  r <- rlnorm(1, 3)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dlnorm(x)

  title <- 'Distribucion logNormal'
}else

# MIXTURAS
if(H1 == 'mix.bimodal'){
  nmix <- norMix(mu = c(0.4, 0.7), sigma = c(0.12,0.08), w = c(0.6,
0.4))

  r <- rnorMix(1, nmix)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dnorMix(x, nmix)

  title <- '0.6*N(0.4, 0.12) + 0.4*N(0.7,0.08)'
}else if(H1 == 'mix.norm'){
  nmix <- norMix(mu = c(0, 0.7), sigma = c(0.8,1.1), w = c(0.7, 0.3))

  r <- rnorMix(1, nmix)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dnorMix(x, nmix)

  title <- '0.7*N(0, 0.8) + 0.3*N(0.7,1.1)'
}else if(H1 == 'mix.contam'){
  nmix <- norMix(mu = c(0.4, 0.7), sigma = c(0.5,1), w = c(0.9, 0.1))
  r <- rnorMix(1, nmix)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dnorMix(x, nmix)

  title <- '0.9*N(0.4, 0.5) + 0.1*N(0.7,1)'
}else if(H1 == 'mix.contam.media'){
  nmix <- norMix(mu = c(0.4, 2.2), sigma = c(0.5, 0.5), w = c(0.9,
0.1))
  r <- rnorMix(1, nmix)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dnorMix(x, nmix)

  title <- '0.9*N(0.4, 0.5) + 0.1*N(2.2,0.5)'
}else if(H1 == 'mix.contam.var'){
  nmix <- norMix(mu = c(0.7, 0.7), sigma = c(0.5,2.8), w = c(0.9,
0.1))

```

```

r <- rnorMix(1, nmix)
d <- density(r)

x <- seq(min(r), max(r), length = 1)
y <- dnorMix(x, nmix)
title <- '0.9*N(0.7, 0.5) + 0.1*N(0.7,2.8)'
}else

# ASIMETRICAS
if(H1 == 'weibull'){
  r <- rweibull(1, 4)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dweibull(x, 4)

  title <- 'Distribucion de Weibull(4, 1)'
}else if(H1 == 'exp'){
  r <- rexp(1)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dexp(x)

  title <- 'Distribucion exponencial'
}else if(H1 == 'chisq'){
  r <- rchisq(1, 4)
  d <- density(r)

  x <- seq(min(r), max(r), length = 1)
  y <- dchisq(x, 4)
  title <- 'Distribucion Chisq con 4 GL'
}else

  if(H1 == 'uniforme'){

    r <- runif(1)
    d <- density(r)

    x <- seq(min(r) -0.2, max(r)+0.2, length = 1)
    y <- dunif(x)
    title <- 'Distribucion Uniforme(0, 1)'
  }else if(H1 == 'weibull_15'){
    r <- rweibull(1, 1, 1.5)
    d <- density(r)

    x <- seq(min(r), max(r), length = 1)
    y <- dweibull(x, 1, 1.5)
    title <- 'Distribucion Weibull(1, 1.5)'
  }else{
    warning('Distribucion no disponible')
    return()
  }
}

```

```
hist(r, main = 'Ejemplo del ENFD con n=100', freq = F,
```

```

        ylim = c(0, max(d$y, hist(r, plot = F)$density)),
        xlim = c(min(r) - 0.2, max(r) + 0.2),
        xlab = '')
lines(d, col = 'red')

plot(x, y, type = 'l', main = 'Funcion de distribucion')

plot(colnames(df), df['Lillie',], col = 1, ylim = ylim, lwd = 3, type =
'o',
      xlab = 'Tamaño de la muestra', ylab = 'Potencia',
      xaxt = 'n', main = paste(title, ' - B=', B, ' Nivel = ', alfa, sep
= ''))

# Eje x
axis(1, at = n)

# Title
lines(colnames(df), df['AD',], col = 2, lwd = 3, type = 'o')
lines(colnames(df), df['SW',], col = 3, lwd = 3, type='o')
lines(colnames(df), df['SF',], col = 4, lwd = 3, type='o')
lines(colnames(df), df['CVM',], col = 5, lwd = 3, type='o')
lines(colnames(df), df['END',], col = 6, lwd = 3, type='o')
lines(colnames(df), df['NEW.E0',], col = 7, lwd = 3, type='o')

legend( legend.coords, legend = test, col = 1:length(test), lwd = 2,
       cex = 1.2)
}

```