



FACULTADE DE MATEMÁTICAS

Trabajo Fin de Grado

Análisis de métricas para evaluación de sistemas de deep learning

Andrés Lorenzo Ramberde

2020/2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Trabajo Fin de Grado

Análisis de métricas para evaluación de sistemas de *deep learning*

Andrés Lorenzo Ramberde

2020/2021

UNIVERSIDAD DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Departamento de Estadística e Investigación Operativa
Título: Análisis de métricas para evaluación de sistemas de <i>deep learning</i>
Breve descripción del contenido
El área del aprendizaje automático ha experimentando un enorme crecimiento en los últimos años. El aumento de los datos disponibles e a capacidad de cómputo para procesarlos trajeron consigo la utilización de nuevas técnicas, entre las que destaca el <i>deep learning</i>. No obstante, a la hora de comparar distintos modelos predictivos se siguen utilizando métricas clásicas de evaluación, que no tienen en cuenta la arquitectura del modelo ni las características poblacionales del conjunto de datos disponibles. En este trabajo se pretende explorar los distintos procedimientos estadísticos que pueden ayudar a realizar una mejor evaluación de los modelos de <i>deep learning</i>, proponiendo una guía de recomendaciones específicas. Este análisis partirá de la exploración del conjunto de datos mediante herramientas no paramétricas e acabará con la presentación de una batería de métricas de evaluación, tanto a nivel de precisión como de complejidad del modelo
Recomendaciones
Otras observaciones
Esta propuesta de TFG se realizará en coordinación con otra propuesta titulada “Comparación de métricas de evaluación de sistemas de <i>deep learning</i> para regresión”, de modo que se proporcionan dos visiones complementarias sobre el mismo tema. En el trabajo orientado a la parte informática, se realizará una experimentación y comparación de distintas métricas para sistemas de <i>deep learning</i> concretos de análisis de imaxes, mostrando el resultado de aplicar las métricas clásicas de evaluación y la mejora que se puede alcanzar con la definición de las nuevas métricas adaptadas a la distribución del conjunto de datos de entrada. El otro TFG sería dirigido por la profesora María José Carreira Nouche.

Índice general

Resumen	VIII
1. Algunos conceptos básicos sobre variables aleatorias continuas	1
1.1. Variable aleatoria continua y función de densidad	1
1.2. Medidas de localización	3
1.2.1. Medidas de localización central	3
1.2.2. Medidas de localización no central	5
1.3. Medidas de dispersión	5
1.4. Medidas de forma	6
1.5. Medidas muestrales	7
2. Contraste de características	11
2.1. Herramientas gráficas	12
2.1.1. Histograma	12
2.1.2. Diagrama de cajas	14
2.1.3. Estimación tipo núcleo	15
2.1.4. Gráfico Q-Q	17
2.2. Test de normalidad	18
2.2.1. Análisis de los test	19
2.2.2. Estudio de simulación	23
2.3. Test de simetría	28
2.3.1. Análisis de los test	29
2.3.2. Estudio de simulación	31
2.4. Test de multimodalidad	32
2.4.1. Análisis de los test	35
2.4.2. Estudio de simulación	38

3. Estudio caso real	45
3.1. Idea del funcionamiento de la red	45
3.2. Caso práctico	47
3.2.1. Recomendaciones prácticas	52
4. Conclusión	55

Resumen

La capacidad predictiva de los sistemas *deep learning* orientados a problemas de regresión han sido evaluados tradicionalmente utilizando medidas clásicas basadas en la media, como el error absoluto medio o el error cuadrático medio. Sin embargo, dichas medidas no tienen en cuenta las distribuciones de los datos ni sus características poblacionales, lo que puede causar una representación poco adecuada de los errores cometidos y del rendimiento obtenido.

En este trabajo, se analizan de diferentes herramientas gráficas y diversos test de contraste de hipótesis de tipo no paramétrico ampliamente conocidos y utilizados en la práctica para poder evaluar las características de normalidad, simetría y multimodalidad de un conjunto de datos. Para los distintos contrastes de hipótesis se realiza un estudio de simulación con el objetivo de analizar su calibrado y potencia, para así seleccionar los test que estando calibrados demuestren mejor potencia. Este conjunto de métodos se utilizan posteriormente para realizar el estudio de un caso real, en donde se analizan los resultados obtenidos por un sistema de regresión basado en técnicas de *deep learning* realizado paralelamente en el TFG en Ingeniería Informática. Dependiendo de las características de normalidad, simetría y modalidad de las distribuciones de los residuos y los errores absolutos de la red, se seleccionará el estadístico muestral más adecuado para la evaluación de cada distribución. El objetivo será elegir la medida más representativa a la hora de evaluar el rendimiento de la red.

Finalmente, se aportan unas recomendaciones prácticas en donde se discuten los estadísticos muestrales más adecuados dependiendo de las características de normalidad, simetría y modalidad detectadas en las distribuciones.

Con este trabajo se pretende mostrar la necesidad de tener en cuenta las características poblacionales de los datos evitando suposiciones habituales como la de simetría y unimodalidad (hipótesis que cumple la distribución normal), para no llegar a conclusiones erróneas derivadas de estas suposiciones iniciales.

Resumo

A capacidade predictiva dos sistemas *deep learning* orientados a problemas de regresión avalíase tradicionalmente empregando medidas clásicas baseadas na media, como o erro absoluto medio ou o erro cuadrático medio. Non obstante, estas medidas non teñen en conta a distribución dos datos nin as súas características de poboación, o que pode provocar unha representación inadecuada dos erros cometidos e do rendemento obtido.

Neste traballo, analízanse e utilízanse na práctica diferentes ferramentas gráficas e varios test de contraste de hipótese non paramétricos amplamente coñecidos para avaliar as características de normalidade, simetría e multimodalidade dun conxunto de datos. Para os diferentes contrastes de hipótese, realízase un estudo de simulación co fin de analizar a súa calibración e potencia, para así seleccionar os test que estando calibrados, mostren mellor potencia. Este conxunto de métodos úsase posteriormente para realizar un estudo de caso real, onde os resultados obtidos por un sistema de regresión baseado en técnicas de *deep learning* son analizados en paralelo no TFG en Enxeñaría Informática. Dependendo das características de normalidade, simetría e modalidade das distribucións dos residuos e dos erros absolutos da rede, seleccionárase o estatístico da mostra máis axeitado para a avaliación de cada distribución. O obxectivo será escoller a medida máis representativa á hora de avaliar o rendemento e rede.

Finalmente, ofrécense algunhas recomendacións prácticas onde se discuten os estatísticos da mostra máis axeitados en función das características de normalidade, simetría e modalidade detectadas nas distribucións.

Este traballo ten como obxectivo amosar a necesidade de ter en conta as características poboacionais dos datos, evitando supostos comúns como a simetría e a unimodalidade (hipótese que cumpre coa distribución normal), para non chegar a conclusións erróneas derivadas destes supostos iniciais.

Abstract

The predictive capacity of *deep learning* systems aimed at regression problems has traditionally been evaluated using classical measures based on the mean, such as mean absolute error or mean squared error. However, these measures do not take into account neither the data distributions nor their population characteristics, which can cause an inadequate representation of the errors committed and the performance obtained.

In this work, different graphic tools and various widely known non-parametric hypothesis contrast test are analyzed and used in a practical real case in order to evaluate the characteristics of normality, symmetry and multimodality of a dataset. For the different hypothesis contrasts, a simulation study is carried out in order to analyze their calibration and power, in order to select the test that, being calibrated, demonstrate the best power. This set of methods is subsequently used to carry out a real case study, where the results obtained by a regression system based on *deep learning* techniques are analyzed in parallel at the TFG in Computer Engineering. Depending on the characteristics of normality, symmetry and modality of the distributions of the residuals and the absolute errors of the network, the most appropriate sample statistic will be selected for the evaluation of each distribution. The objective will be to choose the most representative measure when evaluating network performance.

Finally, some practical recommendations are provided where the most appropriate sample statistics are discussed depending on the characteristics of normality, symmetry and modality detected in the distributions.

This work aims to show the need to take into account the population characteristics of the data, avoiding common assumptions such as symmetry and unimodality (hypothesis that meets the normal distribution), so as not to reach erroneous conclusions derived from these initial assumptions.

Introducción

Las técnicas basadas en el aprendizaje profundo o *deep learning* han ido cobrando importancia a lo largo de los últimos años. En concreto, los modelos predictivos basados en redes neuronales han sido objeto numerosas publicaciones en donde es habitual encontrar estudios de su rendimiento o evaluaciones de su capacidad predictiva. En el caso de que se estudien varias redes con diferentes arquitecturas y un objetivo común, es frecuente la realización estudios comparativos confrontando los resultados obtenidos por cada uno de los modelos para determinar cual de ellos realiza estimaciones más precisas. Para la realización de este tipo de estudios, se siguen utilizando métricas (en el sentido de estadísticos descriptivos, p.e. media) de evaluación tradicionales que no tienen en cuenta la arquitectura del modelo ni las características poblacionales del conjunto de datos suponiendo un comportamiento simétrico unimodal.

El uso de medidas como el error absoluto medio (Mean Absolute Error, MAE):

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (1)$$

o la raíz cuadrática media (Root Mean Square Error, RMSE):

$$RMSE = \frac{\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}}{n} \quad (2)$$

donde y_i representa el valor real, $\hat{y}_i$ el valor estimado por las arquitecturas y n el número de elementos, son muy comunes a la hora de evaluar la magnitud del error cometido por una red. Este tipo de medidas basadas en la media pueden resultar imprecisas si su distribución es asimétrica o multimodal.

La idea de este estudio ha venido motivada la publicación de Vila Blanco et al. (2020), en este artículo se realiza un estudio poblacional de los datos de entrada y salida de la red neuronal presentada en la publicación, con el fin de seleccionar la medida más precisa a la hora de evaluar su rendimiento. La elaboración de este análisis no es lo habitual y se pueden encontrar multitud de ejemplos en los cuales no se realiza. Por ejemplo, en la publicación Li et al. (2018), se diseña y analiza una red neuronal que estima la edad

cronológica a través de radiografías de la pelvis. Cuando se evalúa su rendimiento no se tienen en cuenta las características poblacionales de los datos de entrada y salida de la red ni de las distribuciones de los residuos y errores absolutos. Se utiliza el MAE para medir la magnitud del error de la red sin analizar qué distribución siguen estos errores absolutos que habitualmente resulta ser asimétrica. Ocurre lo mismo en otras publicaciones relacionadas como Xing et al. (2017) y Wang et al. (2015) en las cuales se presentan redes neuronales capaces de predecir la edad de los sujetos a través de fotos faciales. En ambas publicaciones se vuelve a utilizar el MAE para evaluar la red.

En este trabajo, se propondrán diferentes métodos estadísticos además de estadísticos alternativos a la media, que ayuden a mejorar la evaluación de modelos de regresión, concretamente enfocándose en sistemas *deep learning* basados en la utilización de redes neuronales convolucionales, estudiando las distribuciones de los residuos y los errores absolutos.

En primer lugar, se comenzarán recordando algunos conceptos básicos sobre variables aleatorias continuas, función de densidad, medidas de localización, dispersión y forma, etc. A continuación, se presentarán un conjunto de herramientas gráficas que se utilizarán para la exploración de los datos, así como una batería de métodos de contraste con el fin de analizar la normalidad, simetría y modalidad del conjunto de datos. Una vez hecho esto, se realizará un estudio comparativo entre los test que permitan validar las mismas hipótesis y se comprobará su calibrado y potencia través de un estudio de simulación diseñado e implementado utilizando R (R Core Team 2019). Una vez que se haya evaluado que los test se encuentren calibrados, se seleccionarán los que mejor potencia demuestren en cada campo (normalidad, simetría, modalidad). Por último, se realizará un análisis de un caso real, esta sección enlaza con el TFG llevado a cabo en Ingeniería Informática en el cual se implementan varias redes neuronales convolucionales que tratan de solventar un problema de regresión, prediciendo la edad cronológica a partir de radiografías dentales panorámicas. Una de estas redes, en concreto la que se presenta en la publicación que ha inspirado este estudio (Vila Blanco et al. 2020), se utilizará poner en práctica el uso de las herramientas y los métodos que se estudian a lo largo del trabajo para poder elegir las medidas que mejor se adapten a las características de las distribuciones de los residuos y los errores absolutos. Para complementar esta última parte, se aportarán unas recomendaciones prácticas en las que se señalarán los estadísticos muestrales que mejor se adecuen para los diferentes casos posibles atendiendo a características de normalidad, simetría y multimodalidad.

La finalidad del trabajo es mostrar la necesidad del análisis de las características de las distribuciones para evitar el uso casi mecánico de métricas tradicionales a la hora de realizar un análisis de su rendimiento o comparativo de una o más redes.

Capítulo 1

Algunos conceptos básicos sobre variables aleatorias continuas

Dependiendo de las características de las distribuciones y de cómo se repartan los datos que se estén manejando, estos podrán ser interpretados de cierta forma o se podrán hacer determinadas inferencias sobre ellos. En este capítulo, se repasan nociones básicas sobre la variable aleatoria¹ continua para poder entender, cuantificar o resumir de qué manera se distribuyen o agrupan los datos, centrándose en medidas para cuantificar su localización, dispersión y forma. Posteriormente se utilizarán estos conceptos para poder realizar el estudio que se lleva a cabo en este trabajo.

Las nociones expuestas en este capítulo, han sido estudiadas a lo largo del grado en asignaturas como Elementos de Probabilidad y Estadística, Probabilidad y Estadística e Inferencia Estadística.

1.1. Variable aleatoria continua y función de densidad

Definición 1.1. Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathcal{P})$ ², una v.a. continua X es una función que para cada resultado w en Ω , $X(w)$ pertenece a un conjunto no numerable D de números reales. Formalmente $X(\cdot)$ es una función con dominio Ω y rango D , por lo tanto, para cada evento $w \in \Omega$ $X(w) = x \in D$, donde D es un subconjunto no numerable de los números reales.

Se podría ofrecer una versión más formal de la Definición 1.1 utilizando la teoría de la

¹en adelante v.a.

² Ω se denomina espacio muestral y está formado por los posibles resultados elementales del experimento. $\mathcal{F}$ es la colección de todos los sucesos aleatorios. $\mathcal{P}$ es la función probabilidad que asigna una probabilidad a todos los sucesos

medida e introduciendo el concepto de σ -álgebra y σ -álgebra de Borel, pero para llevar a cabo el estudio que queremos realizar nos bastará con esta noción de v.a. continua.

Definición 1.2. La función de distribución o función de distribución acumulada F de la v.a. X , es la función:

$$F(x) = P(A_x) \quad (1.1)$$

donde A_x es el evento $A_x = \{w : X(w) \leq x\}, x \in \mathbb{R}$. Normalmente se denotará (1.1) como:

$$F(x) = P(X \leq x) \quad (1.2)$$

Definición 1.3. Supongamos que tenemos una v.a. continua X con una función de distribución F . Si la derivada,

$$\frac{dF(x)}{dx} = F'(x) \quad (1.3)$$

existe salvo para un número finito de puntos, y la función f definida como

$$f(x) = \begin{cases} F'(x) & \text{si } \exists F'(x), \\ 0 & \text{si } \nexists F'(x) \end{cases} \quad (1.4)$$

satisface

$$F(x) = \int_{-\infty}^x f(x)dx \quad (1.5)$$

entonces X es una v.a. continua con función de densidad f . Dicha función, satisface los siguiente criterios :

- $f(x) \geq 0 \forall x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Definición 1.4 (Muestra aleatoria simple). Una muestra aleatoria simple (m.a.s.), de tamaño n , de una variable aleatoria X con distribución teórica F , son n variables aleatorias $X_1, X_2, \dots, X_n$, independientes e idénticamente distribuidas, con distribución común F .

Definición 1.5 (Estadísticos de orden). Dada una m.a.s. $X_1, X_2, \dots, X_n$ de una v.a. X con una distribución teórica F , los estadísticos de orden se obtienen al colocar los valores de la muestra en orden creciente:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n-1)} \leq X_{(n)}$$

Así se tiene:

$$\begin{array}{ll}
X_{(1)} = \min\{X_1, X_2, \dots, X_n\} & \text{menor valor muestral} \\
X_{(2)} = \min(\{X_1, X_2, \dots, X_n\} - \{X_{(1)}\}) & \text{segundo valor muestral} \\
X_{(3)} = \min(\{X_1, X_2, \dots, X_n\} - \{X_{(1)}, X_{(2)}\}) & \text{tercer valor muestral} \\
& \vdots \\
X_{(n)} = \max\{X_1, X_2, \dots, X_n\} & \text{mayor valor muestral}
\end{array}$$

Hay que hacer notar que, como variables aleatorias, los estadísticos de orden no coinciden con ninguno de los elementos de la muestra, por ejemplo:

$$X_{(1)} = \begin{cases} X_1 & \text{si } X_1 \leq X_2, X_3, \dots, X_n \\ X_2 & \text{si } X_2 \leq X_1, X_3, \dots, X_n \\ \vdots & \\ X_n & \text{si } X_n \leq X_1, X_3, \dots, X_{n-1} \end{cases}$$

Por lo tanto, aunque $X_1, X_2, \dots, X_n$ sean variables aleatorias independientes e igualmente distribuidas, $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ no son ni independientes ni igualmente distribuidas.

1.2. Medidas de localización

Este tipo estadísticos descriptivos, también denominados en muchas ocasiones como medidas de posición, se utilizan para localizar la información más importante dentro el conjunto de datos.

1.2.1. Medidas de localización central

En este apartado se analizarán la esperanza, la mediana y la moda. Las dos primeras, constituyen resúmenes numéricos que ofrecen información acerca de dónde se encuentra ubicado el centro de la distribución. Es decir, los lugares en los que se agrupan los datos. Por otro lado, la moda cuantifica qué valor o valores se repiten con más frecuencia en una distribución de datos.

Definición 1.6 (Esperanza matemática o media poblacional). La esperanza de una v.a. continua con función de densidad f es:

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx \quad (1.6)$$

siempre que se cumpla que,

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty$$

Definición 1.7 (Mediana). Se denotará Me como la mediana de una v.a. X si $P(X \leq Me) \leq 1/2$ y $P(X \geq Me) \geq 1/2$. En el caso que concierne a este capítulo (v.a. continua), basta con que la función de distribución de X sea exactamente $1/2$ en Me .

Definición 1.8 (Moda). Para una v.a. continua con función de densidad f , Mo es una moda si es un máximo local de la función de densidad: $\exists \varepsilon > 0$ tal que, $f(Mo) \geq f(x), \forall x \in (Mo - \varepsilon, Mo + \varepsilon)$.

A las modas que constituyan un máximo absoluto las llamaremos *modas principales*.

La esperanza y la mediana son únicas para cada distribución, pero puede ocurrir que la moda de una distribución no sea única, por ello se introduce también el siguiente concepto.

Definición 1.9 (Modalidad). Diremos que una v.a. X tiene una distribución *unimodal* si tiene una única moda. Si tiene dos o más diremos que X tiene una distribución *multimodal*.

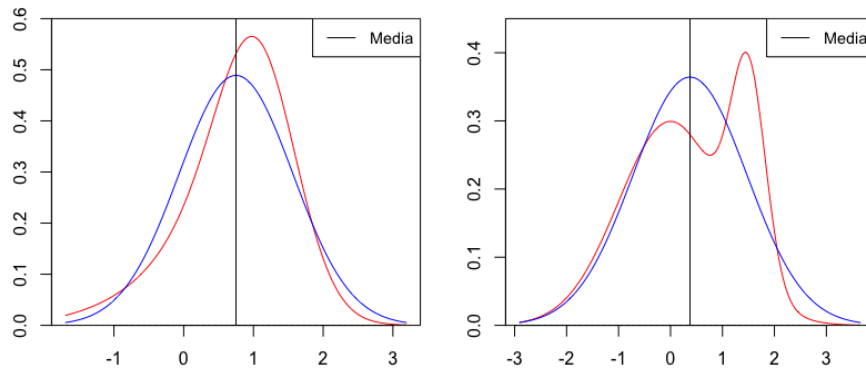


Figura 1.1: A la izquierda, una distribución normal (azul) y una distribución asimétrica (rojo), ambas con $\mu = \frac{3}{4}$. A la derecha, una distribución normal (azul) y una distribución bimodal asimétrica (rojo), ambas con $\mu = \frac{3}{8}$,

La esperanza aporta información acerca del comportamiento de la v.a. y dependiendo de cómo sea la distribución, puede ayudar a localizar dónde se agrupa el grueso de los datos. En otras ocasiones, como se verá a lo largo de este trabajo, será necesario utilizar otras medidas para averiguar en qué punto se agrupan la mayor parte de los elementos de la muestra. De todas formas, puede ocurrir que dos variables aleatorias con distribuciones diferentes tengan la misma esperanza, tal y como se ve en la Figura 1.1. Por otra parte, en la Figura 1.2 también se puede apreciar como estas tres medidas de localización central no siempre coinciden. En distribuciones simétricas y unimodales sí, pero según se introduce asimetría o multimodalidad en las distribuciones la media poblacional, la mediana y la moda dejan de coincidir en el mismo punto.

1.2.2. Medidas de localización no central

Este tipo de medidas estudian características de la distribución de los datos, pero fuera de su centro.

Definición 1.10 (Cuantil). Un cuantil determina la cantidad de datos de una distribución están por encima o por debajo de un cierto límite. El valor x es el k -ésimo q -cuantil de la v.a. X si,

$$P(X < x) \leq k/q \quad (1.7)$$

y además se cumple que,

$$P(X \geq x) \geq 1 - k/q \quad (1.8)$$

Este es el tipo de cuantil más ampliamente utilizado es el cuartil, i.e. $q = 4$ y se denota³ como Q_k , $k \in 1, \dots, 3$. El segundo cuartil o Q_2 se corresponde con la mediana. Otros cuantiles que también son utilizados ampliamente son los deciles ($q = 10$) o los percentiles ($q = 100$)

1.3. Medidas de dispersión

Las medidas de dispersión son valores numéricos que informan sobre el grado de variabilidad de una variable aleatoria.

Definición 1.11 (Varianza). La varianza de una v.a X se define como:

$$Var(X) = E[(X - E(X))^2] \quad (1.9)$$

también se puede escribir de la forma:

$$Var(X) = E(X^2) - E(X)^2 \quad (1.10)$$

en adelante, la varianza en muchas ocasiones será denotada como σ^2 . La raíz cuadrada de la varianza se llama desviación típica:

$$SD(X) = \sqrt{Var(X)} \quad (1.11)$$

en muchas ocasiones será denotada como σ .

Definición 1.12 (Rango intercuartilítico). El rango intercuartilítico se define como,

$$IQR = Q_3 - Q_1 \quad (1.12)$$

en una distribución encontramos el 50% de los datos centrales en este rango.

³Se cometerá un abuso de notación y también se utilizará esta misma notación (Q_1 , Q_2 y Q_3) para los cuantiles en el ámbito muestral.

1.4. Medidas de forma

Este tipo de valores ofrecen información acerca de características especiales de las distribuciones como asimetría, nivel de concentración de datos o apuntamiento. Estos datos son relevantes, ya que cuando se quiere hacer un análisis inferencial, en distintos ámbitos como la inferencia clásica se acostumbra a suponer normalidad. La distribución normal es simétrica y posee un coeficiente de asimetría y de curtosis (apuntamiento) concretos. Por tanto, cuando se analiza una muestra es usual comprobar este tipo de medidas para ver si coinciden con los valores de una distribución normal.

Definición 1.13 (Coeficiente de asimetría). El coeficiente de asimetría de una v.a. X con una media μ y una desviación típica σ se corresponde con el tercer momento estandarizado de X y se define como:

$$CA(X) = E \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] \quad (1.13)$$

Definición 1.14 (Coeficiente de curtosis). El coeficiente de curtosis de una v.a. X con media μ y una desviación típica σ se corresponde con el cuarto momento estandarizado de X y se define como:

$$KU(X) = E \left[\left(\frac{X - \mu}{\sigma} \right)^4 \right] \quad (1.14)$$

Desde un punto de vista analítico podemos decir que una variable aleatoria es simétrica si lo es su función de densidad.

Definición 1.15 (Simetría de una v.a.). Una variable aleatoria X es simétrica respecto a un punto a si $X - a$ tiene la misma distribución que $a - X$. Este número a tiene que ser $E(X)$ si existe la esperanza, ya que,

$$E(X) - a = E(X - a) = E(a - X) = a - E(X) \quad (1.15)$$

Esto se puede simplificar como $E(X) = a$. El valor de a también es la mediana de la distribución, ya que $X - a$ tiene la misma distribución que $a - X$, por tanto:

$$P(X - a \leq 0) = P(a - X \leq 0)$$

así

$$P(X \leq a) = P(X \geq a)$$

que implica:

$$P(X \leq a) = 1 - P(X > a) \geq 1 - P(X \geq a) = 1 - P(X \leq a)$$

mostrando que $P(X \leq a) \geq 1/2$ y $P(X \geq a) \geq 1/2$

1.5. Medidas muestrales

En muchas ocasiones no se conoce de primera mano el comportamiento de la distribución subyacente de una variable aleatoria X . Por tanto, las características de la población deben ser estimadas a partir de una m.a.s. $X_1, \dots, X_n$.

Definición 1.16 (Momentos muestrales y media muestral). Sean $X_1, \dots, X_n$ una m.a.s. de X . El k -ésimo momento muestral es la variable aleatoria:

$$M_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad i \in 1, \dots, n \quad (1.16)$$

En particular nos interesará el primer momento muestral que se corresponde con la media muestral $\bar{X}_n$,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad i \in 1, \dots, n \quad (1.17)$$

Definición 1.17 (Cuasivarianza muestral y cuasidesviación típica muestral). Sean $X_1, \dots, X_n$ una m.a.s. de X , la cuasivarianza muestral se define como:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad i \in 1, \dots, n \quad (1.18)$$

o también,

$$S_n^2 = \sqrt{\frac{n}{n-1}} S_n^2 \quad (1.19)$$

donde S_n^2 se corresponde con la varianza muestral ⁴. La cuasidesviación típica muestral se corresponde con la raíz cuadrada de la cuasivarianza muestral y se denotará como $\mathcal{S}_n$.

Observación 1.18. Se utiliza la cuasivarianza muestral en lugar de la varianza muestral ya que la cuasivarianza es un estimador insesgado de la varianza, i.e.

$$E(S_n^2) = \sigma^2$$

Definición 1.19 (Coeficiente de asimetría muestral). El coeficiente de asimetría muestral de una m.a.s. $X_1, \dots, X_n$ de X se define como:

$$ca = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^3}{S_n^3} \quad (1.20)$$

Definición 1.20 (Coeficiente de curtosis muestral). El coeficiente de curtosis muestral de una m.a.s. $X_1, \dots, X_n$ de X , se define como:

$$ku = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^4}{S_n^4} \quad (1.21)$$

⁴ $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad i \in 1, \dots, n$

Definición 1.21 (Cuartiles muestrales). Existen varios métodos para realizar el cálculo de los cuartiles muestrales. El método que se mostrará a continuación es uno de los más populares y se denomina método de Tukey (Hoaglin et al. 2000):

$$Q_1 = \begin{cases} Y_{(n+3)/4} & \text{si } n \text{ es impar} \\ Y_{(n+2)/4} & \text{si } n \text{ es par} \end{cases}$$

$$Q_2 = \begin{cases} Y_{(n+1)/2} & \text{si } n \text{ es impar} \\ \frac{1}{2}(Y_{n/2} + Y_{(n/2)+1}) & \text{si } n \text{ es par} \end{cases}$$

$$Q_3 = \begin{cases} Y_{(3n+1)/4} & \text{si } n \text{ es impar} \\ Y_{(3n+2)/4} & \text{si } n \text{ es par} \end{cases}$$

donde $Y_1, \dots, Y_n$ son los estadísticos de orden de una muestra aleatoria simple $X_1, \dots, X_n$ de una variable aleatoria X . Como ya se comentó, se cometerá un abuso de notación y se denominará Q_i tanto a los cuartiles muestrales como a los teóricos. Además, la mediana muestral también se denotará como la mediana teórica (Me).

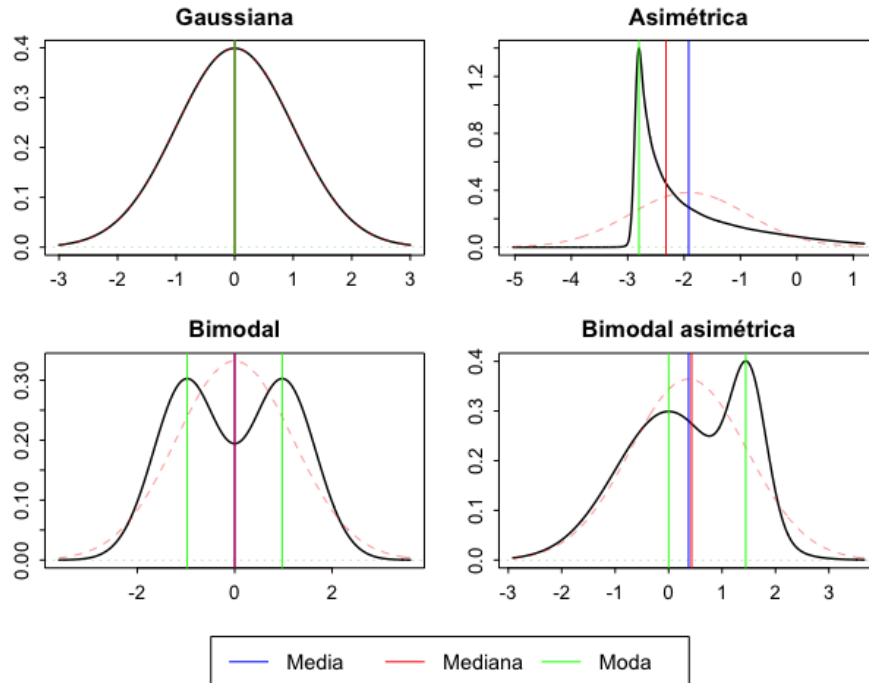


Figura 1.2: Distintos tipos de densidades mostrando su media, mediana y modas. La distribución puntada roja simboliza una distribución normal con la misma media y varianza

	$\bar{X}$	$\mathcal{S}_n^2$	ku	ca
Gaussiana	0.0236	1.0617	3.0029	-0.0104
Asimétrica	-1.9024	1.0953	4.3885	1.3801
Bimodal	0.0256	1.4329	2.0038	-0.0096
Bimodal Asimétrica	0.4047	1.2300	2.2494	-0.2860

Cuadro 1.1: Principales resultados muestrales estudiados de una m.a.s $X_1, \dots, X_n$ con las funciones de densidad de la Figura 1.2 y $n = 2000$

Está claro que para las distribuciones normales o simétricas unimodales, la media es una buena medida del resumen central de los datos, pero tal y como se ve en la Figura 1.2, cuando tenemos una distribución asimétrica unimodal, la media se ve desplazada de la zona central de datos, por lo tanto, la mediana se vuelve un mejor valor para representar el punto donde se produce una mayor agrupación de los datos de la distribución. Por otra parte, si la distribución presenta más de una moda, ni la mediana sirven como un resumen central plausible de la distribución (tal y como se ve en la segunda fila de la Figura 1.2). En ese caso, se propondrá el uso de otra medida relacionada con las modas, siendo esta mucho más dependiente de la muestra en comparación con la media y la mediana. Además, su cálculo acostumbra a ser más costoso que el resto de medidas.

Por otra parte, en el Cuadro 1.1 se muestran los valores de las medidas muestrales presentadas en este capítulo, correspondiéndose con m.a.s. que muestreadas a partir de las distribuciones de la Figura 1.2. La distribución normal estándar tiene un coeficiente de curtosis muy próximo a tres. De hecho, esto ocurre para cualquier distribución normal independientemente de sus parámetros μ y σ . Este valor, puede ser de utilidad para descartar la normalidad de una muestra.

Otra medida útil para tratar de identificar la normalidad de los datos, es el coeficiente de asimetría ya que si su valor no se aproxima a cero, quiere decir que la muestra no es simétrica y por lo tanto, tampoco normal.

En el siguiente capítulo se analizarán herramientas gráficas para observar cómo se comporta una m.a.s. Además, se evaluarán diferentes tipos de test de contrataste no paramétrico de normalidad, asimetría y multimodalidad. De esta forma, se podrán analizar con más exhaustividad las diferentes características de las m.a.s.

Capítulo 2

Contraste de características

Para el estudio que concierne a este trabajo, en primer lugar, se necesitará realizar una exploración de los datos. Esta primera aproximación requiere de diferentes herramientas gráficas para poder observar el comportamiento de los elementos de una muestra. A partir de aquí, se necesitarán de procedimientos más rigurosos para poder realizar un contraste de hipótesis. De esta forma, se podrá evaluar de forma más precisa si los datos se aproximan a un cierto tipo de distribución o si incluso su distribución posee con alguna característica determinada como asimetría o multimodalidad.

En este capítulo, se comenzará presentando un conjunto de herramientas gráficas necesarias para realizar la exploración de los datos. A continuación, se estudiarán un grupo de test para el contraste de hipótesis de tipo no paramétrico. Estos procedimientos se utilizarán para realizar el contraste de hipótesis de normalidad, asimetría y multimodalidad. En la literatura existen muchos más test a mayores de los que se verán en este capítulo, pero se seleccionarán un conjunto de test clásicos ampliamente utilizados en la práctica para cada campo. Se analizará en qué principios estadísticos se basan para evaluar las hipótesis de contraste y por último, se realizará un estudio de simulación diseñado e implementado en R, para cada uno de ellos, midiendo calibrado y comparando su potencia. Entre los test que se encuentren calibrados, se seleccionará el que ofrezca un mejor rendimiento en cada área.

A partir de los resultados que se obtengan en este capítulo y dependiendo de la distribución que tengan los errores de la red neuronal, se podrá evaluar qué medida es más adecuada para evaluar cómo se comporta la red.

2.1. Herramientas gráficas

Como ya se explicó previamente, se comenzará realizando exploración de los datos llevando a cabo un estudio gráfico que permita observar su comportamiento.

Esta sección se centrará la presentación, análisis y definición de cuatro tipos distintos de representaciones gráficas ampliamente conocidas y recurridas, que son: el histograma, el diagrama de cajas, la estimación tipo núcleo y los gráficos Q-Q. Existen muchos otros tipos de herramientas gráficas pero se han seleccionado estas debido a su popularidad y utilidad general.

2.1.1. Histograma

Los histogramas son una representación gráfica de datos agrupados mediante intervalos que se utilizan obtener una primera aproximación de la distribución subyacente de la muestra.

Los resultados que se obtienen del histograma varían en función del número de clases que se escojan. A partir de esto último y unido al rango de los datos es posible determinar el ancho de la clase. El objetivo es tratar de escoger una anchura lo suficientemente grande para capturar características importantes de la muestra pero omitiendo los detalles más concretos derivados de la aleatoriedad de la misma.

Sea $X_1, \dots, X_n$ una m.a.s. de X . Se pueden agrupar los valores de la muestras en J clases: $B_1, \dots, B_J$. Se denotará la anchura de la clase como $|B_j|$. El área bajo una clase B_j es igual a la proporción de elementos en B_j ,

$$f_j = \frac{\text{número de } X_i \text{ en } B_j}{n}$$

La altura de un histograma en B_i deberá ser:

$$\frac{f_j}{|B_j|}$$

2.1.1.1. Elección del número de clases y su anchura

La elección de un número de clases no es trivial. Además, normalmente se utiliza una anchura igual para todas las clases, salvo que los datos sean mucho más escasos en unas zonas que en otras.

Existen varios de criterios aplicables para escoger el número de clases. Por ejemplo, una regla sistemática para determinar el número de clases y que es ampliamente utilizada en

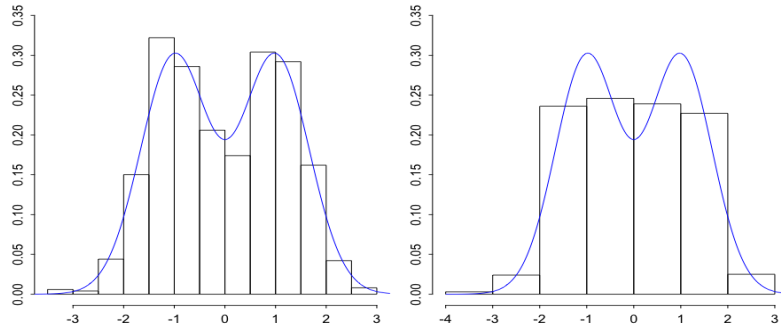


Figura 2.1: Histogramas de una m.a.s. de 1000 elementos con una distribución bimodal simétrica (azul) y con distinto número de clases (12 y 8)

los paquetes de software estadísticos es la de Sturges, donde siendo J el número de clases y n el número de datos, se tiene que:

$$J = \log_2(n) + 1 \quad (2.1)$$

Esta regla utiliza implícitamente una distribución binomial para aproximar la distribución subyacente. Por lo tanto, cuando el número de elementos de la muestra n es grande o la distribución subyacente es no normal puede no ser la regla óptima para la elección del número de clases. Otras reglas muy utilizadas son la de Freedman-Diaconis:

$$J = \frac{2 * IQR}{\sqrt[3]{n}}$$

basada en el rango intercuartílico o la regla de Scott:

$$J = \frac{3.49 * \mathcal{S}_n}{\sqrt[3]{n}}$$

basada en la desviación típica de la muestra.

En la Figura 2.1 se pueden observar dos histogramas contruidos a partir de una m.a.s. $X_1, \dots, X_n$ de X con $n = 1000$, siguiendo X una distribución bimodal simétrica. El histograma de la izquierda cuenta con 12 clases mientras que el de la derecha con 8. Es evidente que mientras el histograma de la izquierda es capaz de reflejar la bimodalidad de la densidad el de la derecha no puede demostrar esta característica debido a que la anchura de las clases no permite reflejar este comportamiento.

En la Figura 2.2 se puede ver una comparativa de cómo resultan los histogramas utilizando las tres reglas vistas. Se utilizarán tres m.a.s. de X con $n = 10$, $n = 100$ y $n = 1000$ respectivamente. Además X tiene una distribución bimodal simétrica, igual a la vista en la Figura 2.1 con $n = 10$, $n = 100$ y $n = 1000$.

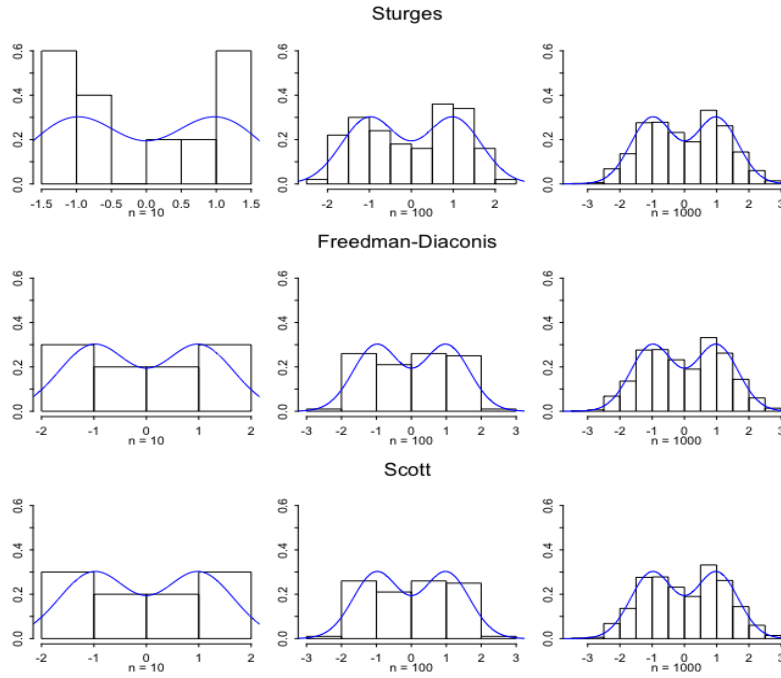


Figura 2.2: Histogramas utilizando las tres reglas vistas de tres m.a.s. de X con $n = 10$, $n = 100$ y $n = 1000$, teniendo X una distribución bimodal simétrica

2.1.2. Diagrama de cajas

Este método de representación de datos resume la información acerca de la dispersión y asimetría de la muestra a través de los cuartiles y el valor máximo y mínimo de los datos. Se representa una caja, los bordes inferior y superior representan Q_1 y Q_3 respectivamente. La línea en el interior de la caja se corresponde con la mediana (Q_2). Los extremos de las líneas salientes de la caja, también llamados bigotes, van desde Q_1 hasta $Q_1 - 1.5 * IQR$ y desde Q_3 hasta $Q_3 + 1.5 * IQR$. Los puntos que se sitúan más allá de estos bigotes se consideran los valores atípicos de la muestra.

Los valores $Q_1 - 1.5 * IQR$ y $Q_3 + 1.5 * IQR$ son escogidos como cota inferior y superior para los valores esperados, ya que si se supone que los datos siguen una distribución normal, desarrollando las expresiones, se estarían permitiendo datos que se encuentren aproximadamente a 3σ de μ . Es interesante señalar que tal y como se ve en la Figura 2.1.2 el diagrama de cajas (también llamado *boxplot*) falla a la hora de mostrar la bimodalidad subyacente de un conjunto de datos, por su construcción.

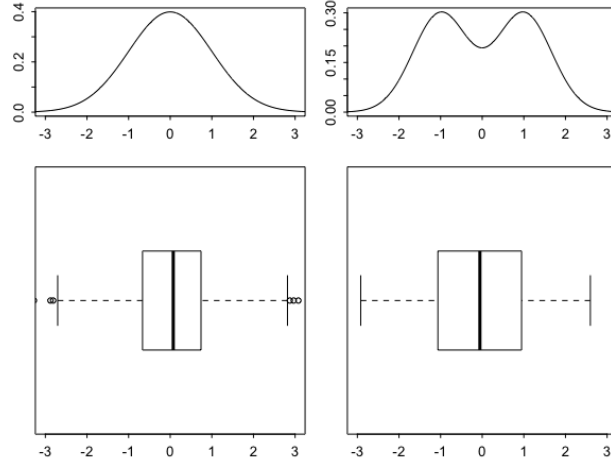


Figura 2.3: Diagramas de caja de dos m.a.s. de 1000 elementos generadas a partir de una distribución normal estándar y una distribución bimodal simétrica

2.1.3. Estimación tipo núcleo

La estimación tipo núcleo de la densidad, produce una función de densidad suave y empírica a partir de los datos de la muestra.

Supongamos de que se tiene una m.a.s. $X_1, \dots, X_n$ de una variable aleatoria X con una función de densidad desconocida f . El estimador tipo núcleo asigna cada i -ésimo punto de la muestra X_i una función K , también llamada función núcleo. Esta función es no negativa y cumple que:

$$\int_{-\infty}^{\infty} K(x)dx = 1 \quad (2.2)$$

y es por tanto, una densidad. En la práctica, la mayoría de las funciones núcleo que se utilizan son simétricas, aunque estas no tienen por qué serlo. El estimador tipo núcleo con función núcleo K se define como:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i) \quad (2.3)$$

La propiedad (2.2) asegura la normalización del estimador $\hat{f}$,

$$\int_{-\infty}^{\infty} \hat{f}(x)dx = \frac{1}{nh} \sum_{i=1}^n \int_{-\infty}^{\infty} K\left(\frac{x - X_i}{h}\right) dx = \frac{1}{nh} \sum_{i=1}^n hK(u)du = 1 \quad (2.4)$$

El parámetro h , también llamado parámetro de suavizado, ventana o amplitud de banda, determina cuánto se expande la influencia de una observación. Un valor muy grande de h puede significar un suavizado excesivo del estimador, mientras que un valor muy pequeño

puede hacer que se muestren detalles dependientes de la muestra aleatoria concreta.

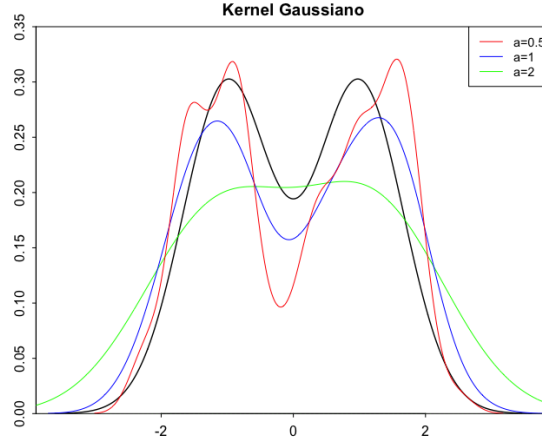


Figura 2.4: Estimador tipo núcleo con núcleo Gaussiano para diferentes escalados de h de la forma $h = a * s$. Siendo s la anchura obtenida utilizando la regla de oro de Silverman. Es este caso concreto $s = 0.468$. El estimador ha sido generado de una m.a.s. de 1000 elementos generada a partir de una distribución bimodal simétrica

En la Figura 2.4 se pueden ver tres estimaciones de densidad kernel de una distribución bimodal. Para realizar estas estimaciones se ha utilizado una función núcleo Gaussiana que utiliza la función de densidad de una distribución normal estándar i.e.

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad (2.5)$$

En las diferentes gráficas se ha utilizado un parámetro de suavizado de la forma $h = a * s$, donde s se calcula utilizando la regla del pulgar de Silverman¹ y a es un parámetro de escalado. Si se hace muy pequeño el valor de h , tal y como se ve en la Figura 2.4 la función de densidad tipo núcleo (rojo) recoge demasiada información de la muestra mostrando detalles no relevantes (fruto del ruido de la muestra). Mientras que si se agranda el valor de h el estimador de la densidad tipo núcleo (verde) falla en detectar la bimodalidad de la distribución subyacente.

El uso diferentes funciones núcleo también puede ser un factor a tener en cuenta, aunque su impacto sea mucho menor que la elección del parámetro de suavizado óptimo. No todas las funciones aproximan la función de densidad de la misma manera. De hecho cuanto más

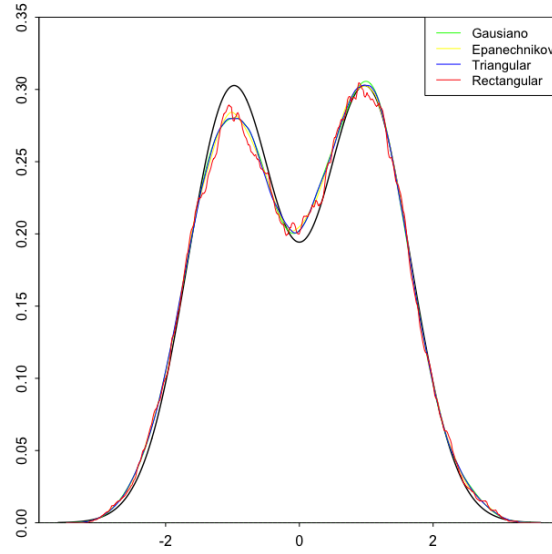


Figura 2.5: Diferentes funciones kernel que aproximan una m.a.s. de 1000 elementos generada a partir de una distribución bimodal simétrica. La ventana se ha calculado utilizando la regla del pulgar de Silverman siendo $h = 0.268$

pequeña es la muestra más aumenta la variabilidad de las estimaciones. En la Figura 2.1.3, se puede apreciar que para una m.a.s. de 1000 elementos de una distribución bimodal simétrica, la diferencia entre las distintas funciones núcleo no es demasiado grande. Sin embargo, la función núcleo Rectangular, no consigue los buenos resultados de las otras tres funciones, dado que no proporciona estimadores continuos.

2.1.4. Gráfico Q-Q

Por último, se estudiará el gráfico Q-Q, este tipo de gráfico se utiliza para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación. Generalmente para la segunda distribución se utiliza una función distribución teórica. De esta forma, se puede comprobar visualmente la bondad de ajuste de los datos de la muestra con la distribución teórica. Esta comparación se efectúa haciendo un gráfico cruzado de los cuantiles de las dos distribuciones. Si las dos distribuciones comparadas son similares, los puntos del gráfico Q-Q se posicionarán aproximadamente sobre la diagonal $y = x$.

La comparación entre una muestra aleatoria simple de 100 elementos muestreadas a partir de una de una función de distribución normal y de una función de distribución asimétrica

¹La regla del pulgar de Silverman es: $h = 0.9 * \min(\mathcal{S}_n, \frac{IQR}{1.34}) n^{-\frac{1}{5}}$

se puede ver en la Figura 2.1.4.

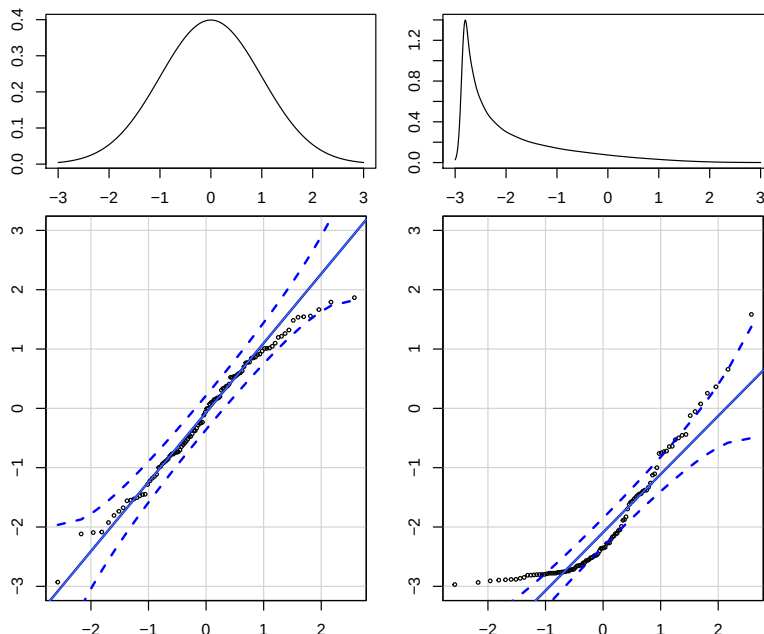


Figura 2.6: Gráfico Q-Q de una m.a.s de 100 elementos muestreada a partir de una distribución normal estándar (izquierda) y una distribución asimétrica (derecha). Ambos comparados con los cuantiles teóricos de una distribución normal estándar (eje x). Se acompaña además, con la recta que se obtiene a partir del método de mínimos cuadrados (azul). Las líneas discontinuas son las bandas de confianza estas muestran donde la zona se deben situar los elementos de la de la muestra un 95 % de las veces.

2.2. Test de normalidad

Determinar si un conjunto de datos proviene de una distribución normal es un proceso requerido en multitud de procedimientos estadísticos. En muchas ocasiones se supone la normalidad de los datos, pero si resulta que esta suposición es errónea, los resultados, conclusiones o inferencias obtenidas a través de los datos resultan inválidos. Por ejemplo, cuando se requiere contrastar la igualdad de medias, es habitual emplear un test T , pero este supone una normalidad de los dos grupos.

Se pueden utilizar diversos métodos para evaluar la normalidad de un conjunto de datos, los métodos gráficos como los analizados en la sección anterior pueden resultar útiles pero no son suficientes para aportar evidencias concluyentes acerca de si la hipótesis de

normalidad es acertada. Por ello, se introducirán los test de normalidad, que consisten en procedimientos más formales para determinar si un conjunto de datos sigue una distribución normal. Existen multitud de test disponibles para efectuar la comprobación de la hipótesis de normalidad, así pues, su elección resulta importante a la hora de realizar un análisis de los datos.

Los test de normalidad que se introducirán en esta sección tienen las siguientes hipótesis nula y alternativa:

$$H_0 : X \sim N(\mu, \sigma)$$

$$H_1 : X \not\sim N(\mu, \sigma)$$

En este estudio se analizarán cuatro test de normalidad distintos separados en dos categorías: test de función de distribución empírica (FDE) y los test específicos.

La primera categoría se basa en test que comparan la función de distribución empírica de la muestra con una función de distribución en concreto (normalmente teórica) y no necesariamente se tiene que usar la distribución normal como distribución de contraste para ver si los datos de una muestra se ajustan a esta distribución. De los múltiples test pertenecientes a esta familia, se analizarán el test de Kolmogorov-Smirnov (Kolmogorov 1933) utilizando la corrección de Lilliefors (1967) y el test de Anderson-Darling (Anderson y Darling 1952).

El segundo tipo de test que se verán en esta sección serán los denominados como test específicos. Esta familia se basa en el estudio de características propias de la distribución normal para comprobar si la muestra comparte características similares. En esta parte, se hablará del test de Shapiro-Wilk (Shapiro y Wilk 1965) y del test de Jarque-Bera (Jarque y Bera 1980). El primero de ellos es uno de los test más populares para contrastes de normalidad, y se basa en el estudio la proporción entre dos estimadores de la varianza de una distribución normal. El segundo, muy utilizado en el ámbito de la econometría, se centra en el análisis de los coeficientes de asimetría y curtosis.

Primero, se introducirán cada uno de los test. A continuación, se explicará el proceso de simulación de Monte Carlo para analizar el calibrado y potencia de los diferentes test. Este estudio de simulación se llevará a cabo utilizando R. Por último, se comentarán los resultados obtenidos y seleccionará el test que mejor potencia demuestre.

2.2.1. Análisis de los test

A continuación, se analizarán cada uno de los test previamente mencionados y se expondrán los fundamentos en los que se basan para tratar de evaluar la normalidad de los datos.

2.2.1.1. Test FDE

Esta familia de test utiliza la función de distribución empírica. En este trabajo, se analizarán dos en concreto: el test de Kolmogorov-Smirnov con la corrección de test de Lilliefors (a partir de ahora se hará referencia a este test como el test de Lilliefors) y el test de Anderson-Darling . Aún así, existen otras pruebas que utilizan este mismo enfoque pero que no se mostrarán, como el test de Crámer-von Mises.

Este tipo de test se basan en comparar la función de distribución empírica de la muestra (F_n) con la función de distribución de la que se supone que proviene la muestra (F_0) midiendo la bondad de ajuste entre ambas distribuciones.

Dada una m.a.s. $X_1, \dots, X_n$ de una v.a. X se quiere comprobar H_0 . O lo que es lo mismo, se quiere saber si la muestra proviene de una población con una distribución normal. Si se desconocen μ y σ , se pueden adoptar dos alternativas diferentes. O bien, estimar la función de distribución normal que se quiere contrastar tomando como parámetros $\mu = \bar{X}_n$ y $\sigma = \mathcal{S}_n$ y utilizar $\Phi_{\bar{X}_n, \mathcal{S}_n}^2$ como función de distribución F_0 . O por otra parte, se puede estandarizar la m.a.s. $X_1, X_2, \dots, X_n$ de la v.a. X de la forma:

$$Y_i = \frac{X_i - \bar{X}_n}{\mathcal{S}_n} \quad (2.6)$$

y comparar esta nueva v.a. con la función de distribución teórica de una normal estándar ($\Phi_{0,1}$) en el caso del test de Kolmogorov-Smirnov la estimación de estos parámetros se conoce como el ajuste de Lilliefors. La función distribución empírica se calcula de la siguiente forma:

$$F_n = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i < x) \quad (2.7)$$

Test de Lilliefors El test de Lilliefors (1967) se basa en el test de Kolmogorov-Smirnov (Kolmogorov 1933). Este último test se utiliza cuando μ y σ son conocidos y la distribución de contraste queda completamente determinada. En cambio en el test de Lilliefors se aproximan estas medidas a través de sus valores muestrales como se explicaba anteriormente y solamente sirve cuando la función de contraste (F_0) es normal o exponencial. Este test se basa en el estudio del siguiente estadístico:

$$D = \max_{x \in \mathbb{R}} |F_n(x) - \Phi_{\bar{X}_n, \mathcal{S}_n}(x)| \quad (2.8)$$

Para su computación, se puede reescribir el estadístico de la siguiente manera: Se supone una m.a.s. $X_1, \dots, X_n$ de una v.a. X , entonces se tiene $D = \max\{D^-, D^+\}$, donde:

$$D^- = \max_{i=1, \dots, n} \left\{ \Phi_{0,1}(Y_{(i)}) - \frac{i-1}{n} \right\} \quad D^+ = \max_{i=1, \dots, n} \left\{ \frac{i}{n} - \Phi_{0,1}(Y_{(i)}) \right\} \quad (2.9)$$

² $\Phi_{\mu, \sigma} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp^{-\frac{u-\mu}{2\sigma^2}} du, x \in \mathbb{R}$

donde $Y_{(i)}$ es el estadístico de orden obtenido tras estandarizar el estadístico de orden $X_{(i)}$ como se veía en (2.6). Valores elevados del estadístico D pueden sugerir un posible rechazo de H_0 . Para saber cuándo se puede rechazar H_0 o no, se observan en los valores críticos tabulados por el propio Lilliefors, aproximados mediante métodos de Monte Carlo. Para realizar el estudio de simulación los p-valores de este test se obtendrán mediante la fórmula desarrollada por Dallal y Wilkinson (1986).

Test de Anderson-Darling Presentado por Anderson y Darling (1952), este test se basa en comparar la diferencia cuadrática entre la función de distribución empírica y la función de distribución de la hipótesis nula. Dadas n observaciones, X_i , con $i = 1, \dots, n$ de una v.a. X , el estadístico de Anderson-Darling mide la distancia entre la función de distribución empírica y la hipotética de la siguiente manera:

$$A_n = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1 - F_0(x))^2} dF_0(x) \quad (2.10)$$

siendo F_n la función de distribución empírica y F_0 la distribución hipotética que se quiere contrastar. La función $w(x) = [F_0(x)(1 - F_0(x))^2]^{-1}$ otorga peso a las colas de las distribuciones. Por ejemplo, en el caso del test de Crámer-von Mises la medida de la distancia cambia ya que la función de ponderación cambia, siendo esta $w(x) = 1$.

Como ya se comentó previamente, la función de contraste que nos interesa en este caso es la función distribución de una normal cuyos parámetros son desconocidos μ y σ ($\Phi_{\mu,\sigma}$). Para resolver esto se hace lo mismo que se vio en el test anterior. Suponiendo una m.a.s $X_1, \dots, X_n$ de una v.a. X , se estandarizan los estadísticos de orden $X_{(1)}, \dots, X_{(n)}$ tal y como se veía en (2.6). De esta manera se puede llegar a reescribir el estadístico A_n como:

$$A_n = -n - \frac{1}{n} \sum_{i=1}^n n(2i-1)(\log(\Phi_{0,1}(Y_{(i)})) + \log(1 - \Phi_{0,1}(Y_{(n+1-i)}))) \quad (2.11)$$

En las simulaciones se utilizará un ajuste de este estadístico utilizado cuando μ y σ son desconocidas, publicado por D'Agostino y Michael A. Stephens (1986). Este ajuste se denominará A_n^* y tiene la siguiente forma:

$$A_n^* = A_n \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2} \right) \quad (2.12)$$

Los valores críticos del estadístico A_n^* se muestran en la tabla 4.7 de D'Agostino y Michael A. Stephens (1986) para diferentes niveles de significación.

2.2.1.2. Test específicos

A diferencia la familia de test vista en la sección anterior que se basa en el estudio de la distribución empírica y se pueden utilizar para medir la bondad de ajuste con otras

distribuciones diferentes a la normal. Se estudiarán dos métodos específicos para tratar de determinar la normalidad de los datos. Estos test se enfocan en el análisis de distintas características propias y distintivas de las distribuciones normales comprobando si los datos de la muestra que se analiza poseen dichas características. Los dos test que se analizarán a continuación, serán: el test de Shapiro-Wilk y el test de Jarque-Bera.

Shapiro-Wilk El estadístico de este test se basa en calcular la proporción entre dos estimadores de la varianza de una distribución normal basados en una m.a.s. de tamaño n . El numerador del estadístico es proporcional al cuadrado del estimador lineal insesgado y con menor varianza de la desviación típica. El denominador es la suma del cuadrado de las observaciones alrededor de la media muestral.

Dada una m.a.s. $X_1, X_2, \dots, X_n$ de una variable aleatoria X , se define el estadístico de contraste de Shapiro-Wilk (Shapiro y Wilk 1965) de la siguiente manera:

$$W = \frac{(\sum_{i=1}^n a_i X_{(i)})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (2.13)$$

donde $X_{(i)}$ es el estadístico de orden i -ésimo. Además,

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

siendo $m = (m_1, \dots, m_n)^T$ la esperanza de los estadísticos de orden de los elementos de una m.a.s. muestreada a partir de una distribución normal estándar. V es la matriz de covarianzas de dimensión $n \times n$ de esos estadísticos de orden.

Para determinar $a \in \mathbb{R}^n$ se necesita de m y V . Los valores de W son muy difíciles de calcular computacionalmente. Debido a la dificultad de calcular los pesos de a , Shapiro y Wilk solo fueron capaces de aproximar estos valores para muestras de hasta 50 elementos (Royston 1992). Finalmente Royston (1995) consiguió aproximarlos para muestras mayores. Aunque, este deja de ser fiable para muestras con más de 5000 elementos.

Test de Jarque-Bera Este test de normalidad fue propuesto por Jarque y Bera (1980). Su estadístico de contraste es una función de los coeficientes de asimetría y curtosis muestrales. El estadístico de Jarque-Bera se define como:

$$JB = \frac{n}{6} \left[ca^2 + \frac{1}{4} (ku - 3)^2 \right] \quad (2.14)$$

donde ca es el coeficiente de asimetría muestral y ku el coeficiente de curtosis muestral. El estadístico JB está asintóticamente distribuido como una chi-cuadrado (χ^2) con dos grados de libertad.

2.2.2. Estudio de simulación

Se diseñará e implementará una simulación de Monte Carlo para medir el calibrado y la potencia de los cuatro test analizados anteriormente. Este tipo de simulación consiste en la repetición sucesiva del mismo experimento repitiendo el muestreo aleatorio. De esta manera, se aleatoriza la prueba haciendo que esta sea más independiente de la muestra escogida. Así, se puede reducir la incertidumbre a la hora de calcular empíricamente el calibrado y potencia de los test.

Se comprobará el comportamiento de los test (en términos de calibrado y potencia). Es decir, el porcentaje de rechazos para H_0 y H_1 para decidir si una muestra de n elementos muestreados a partir de una v.a. X proviene o no de una población con una distribución normal.

2.2.2.1. Procedimiento

Se considerarán los niveles de significación $\alpha = 0.01$, $\alpha = 0.05$ y $\alpha = 0.1$ para ver como afectan al calibrado y al análisis de la potencia de los test. Los tamaños de las muestras que considerarán serán de $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$. De cada experimento se realizarán $M = 1000$ simulaciones independientes repitiendo el muestreo aleatorio cada vez. Además, las pruebas se llevarán a cabo utilizando el lenguaje de programación R. En el Cuadro 2.1 se pueden ver las diferentes funciones utilizadas en R para realizar los test y sus respectivos paquetes.

Para analizar tanto el calibrado como la potencia de los test, se medirá la probabilidad de rechazar H_0 , aproximada a partir de técnicas de Monte Carlo utilizando los p-valores obtenidos a través de las diferentes simulaciones independientes. A este valor lo denotaremos como $\hat{\alpha}$ en el caso de que H_0 sea cierta y $1 - \hat{\beta}$ cuando lo sea H_1 . El método para la obtención de estos dos valores será el siguiente:

1. Simulamos $M = 1000$ muestras $(X_1^m, X_2^m, \dots, X_n^m)$ de tamaño n con $m = 1, \dots, M$.
2. Calculamos $T^m = T(X_1^m, X_2^m, \dots, X_n^m)$, valor del estadístico de interés en cada muestra.
3. Si $T \sim D$ bajo H_0 , consideremos p^m el p-valor de T^m bajo la distribución D .
4. En este paso podemos encontrarnos con los siguientes casos:

Test	Función	Paquete
Lilliefors	<code>lillie.test</code>	nortest
Anderson-Darling	<code>ad.test</code>	nortest
Shapiro-Wilk	<code>shapiro.test</code>	stats
Jarque-Bera	<code>jarque.bera.test</code>	tseries

Cuadro 2.1: Funciones utilizadas en R para llevar a cabo el estudio de simulación de la normalidad y sus paquetes.

(a) Si H_0 es cierta, aproximamos $\mathbb{P}(\text{Rechazar } H_0 \mid H_0 \text{ cierta}) = \alpha$ como:

$$\hat{\alpha} = \frac{1}{M} \sum_{m=0}^M \mathbb{I}(p^m < \alpha)$$

(b) Si H_1 es cierta, aproximamos $\mathbb{P}(\text{Rechazar } H_0 \mid H_0 \text{ falsa}) = 1 - \beta$ como:

$$1 - \hat{\beta} = \frac{1}{M} \sum_{m=0}^M \mathbb{I}(p^m < \alpha)$$

2.2.2.2. Calibrado

Para la comprobación del calibrado de los test, es decir, la medición de la probabilidad de error empírico de tipo I (la probabilidad que se rechaza H_0 siendo cierta), que se denotaba anteriormente como $\hat{\alpha}$, se han considerado las distribuciones normales: $N(0, 1)$, $N(0, 5)$ y $N(0, \frac{1}{4})$.

Los resultados obtenidos se pueden ver en el Cuadro 2.2. Se puede observar que los cuatro test están bien calibrados ya que la probabilidad del error empírico de tipo I siempre es muy próximo al nivel de significación exigido. Además, cuantas más simulaciones se realicen, más próximo estará de α . Con lo cual, rechazamos H_0 en una proporción acorde con los parámetros α establecidos previamente.

2.2.2.3. Potencia

Tras haber comprobado que los test están bien calibrados, se evaluará su potencia, β , es decir, las veces que el test rechaza H_1 siendo esta la hipótesis correcta. Más concretamente, en este análisis de potencia se analiza la probabilidad de rechazar H_0 siendo H_0 falsa ($1 - \beta$). Este es el valor que se pretende aproximar mediante $1 - \hat{\beta}$. Por lo tanto, cuanto mayores sean los valores que se obtienen de $1 - \hat{\beta}$, mayor será la potencia del test.

Distribución H_0	Tamaño de la muestra (n)	Calibrado											
		$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.1$			
		LF	AD	SW	JB	LF	AD	SW	JB	LF	AD	SW	JB
$N(0, 1)$	20	0.008	0.014	0.012	0.010	0.046	0.049	0.051	0.020	0.092	0.108	0.099	0.026
	50	0.010	0.012	0.012	0.021	0.045	0.049	0.057	0.045	0.100	0.102	0.101	0.064
	100	0.014	0.019	0.016	0.013	0.059	0.059	0.054	0.045	0.103	0.104	0.111	0.068
	500	0.006	0.010	0.015	0.016	0.043	0.05	0.053	0.061	0.108	0.103	0.106	0.103
	1000	0.002	0.011	0.019	0.020	0.037	0.057	0.052	0.053	0.084	0.103	0.100	0.090
$N(0, 5)$	20	0.010	0.008	0.008	0.014	0.046	0.046	0.054	0.031	0.090	0.109	0.104	0.042
	50	0.010	0.011	0.004	0.013	0.053	0.043	0.050	0.033	0.111	0.090	0.092	0.048
	100	0.014	0.008	0.008	0.015	0.060	0.048	0.050	0.046	0.103	0.089	0.091	0.068
	500	0.008	0.009	0.012	0.017	0.052	0.048	0.043	0.045	0.106	0.093	0.093	0.075
	1000	0.005	0.011	0.008	0.016	0.038	0.054	0.049	0.050	0.096	0.104	0.099	0.09
$N(0, \frac{1}{4})$	20	0.007	0.006	0.008	0.011	0.046	0.040	0.039	0.022	0.087	0.098	0.087	0.030
	50	0.010	0.016	0.013	0.017	0.051	0.055	0.055	0.035	0.110	0.107	0.107	0.056
	100	0.015	0.011	0.014	0.021	0.058	0.049	0.055	0.053	0.104	0.111	0.104	0.082
	500	0.013	0.015	0.015	0.010	0.057	0.061	0.070	0.050	0.120	0.110	0.118	0.108
	1000	0.010	0.005	0.006	0.007	0.040	0.046	0.051	0.046	0.101	0.099	0.098	0.093

Cuadro 2.2: Calibrado de los test de normalidad. Porcentajes de rechazo de H_0 con 1000 simulaciones y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$.

LF: Lilliefors **AD**: Anderson-Darling **SW**: Shapiro-Wilk **JB**: Jarque-Bera

Para realizar el estudio de la potencia, se utilizarán tres distribuciones no normales obtenidas a partir de mezclas de distribuciones normales. Estas han sido seleccionadas de la lista creada por Marron y Wand (1992), que permite un análisis simple y exacto para una amplia variedad de tipos de densidades. Elegiremos una distribución unimodal asimétrica³, una bimodal simétrica⁴ y una bimodal asimétrica⁵. Además se utilizarán tres distribuciones *skewed normal* (normales asimétricas) con parámetros ξ , ω y λ que controlan su localización, escala y forma respectivamente. Los efectos de estos parámetros se pueden observar en la Figura 2.7. Para las tres distribuciones se tendrá que $\xi = 0$, $\omega = 1$ y se variará el parámetro que controla la forma, siendo este $\lambda = 1$, $\lambda = 2$ o $\lambda = 4$. En la Figura 2.8 se pueden ver gráficamente qué forma tienen estas distribuciones. Los resultados obtenidos de la simulación se muestran en el Cuadro 2.3. A mayores, se muestra gráficamente la potencia de los test (con $\alpha = 0.05$) para que sea más sencillo comparar su potencia.

En la Figura 2.9 se puede ver como varía la potencia de los tres test frente a las seis distribuciones consideradas según aumenta el tamaño de la muestra n y con un nivel de

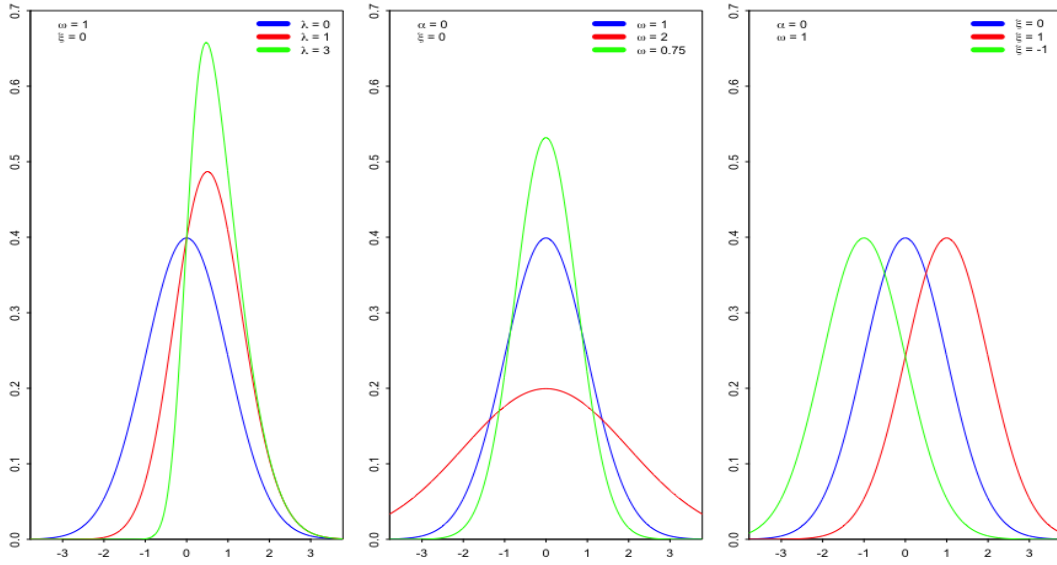


Figura 2.7: Efecto de los parámetros λ , ω y ξ en las distribuciones *skewed-normal*

significación $\alpha = 0.05$.

En primer lugar, se observa que el test con la mejor potencia demostrada es el de Shapiro-Wilk, que independientemente de la distribución muestra resultados similares o superiores al resto de test. El test de Anderson-Darling es el segundo más potente en general. Aunque, por ejemplo, para las distribuciones normales asimétricas con $\lambda = 1$ y $\lambda = 2$, muestra resultados ligeramente inferiores al resto de test (excluyendo el de Lilliefors). El test de Jarque-Bera es inferior a los dos test previamente mencionados en las distribuciones bimodales, siendo el test menos potente entre todos en las distribuciones bimodales simétrica y asimétrica. En la distribución normal asimétrica con $\lambda = 1$ demuestra ser ligeramente superior al test de Anderson-Darling. Por último, el peor test en cuanto a potencia es el de Lilliefors que demuestra unos resultados inferiores al resto test en las distribuciones normales asimétricas y en la distribución asimétrica de Marron y Wand. Aun así, en las dos distribuciones bimodales demuestra ser superior al test de Jarque-Bera.

A la vista de los resultados observados tanto en la Tabla 2.3 como en la Figura 2.9 y tras haber comprobado que todos los test de normalidad están bien calibrados. El test que demuestra una mejor potencia para prácticamente todos los tipos de distribuciones es el test de Shapiro-Wilk. Por lo tanto este será el test que se seleccione para evaluar la

³Unimodal asimétrica: $\frac{1}{5}N(0, 1) + \frac{1}{5}N\left(\frac{1}{2}, \frac{2}{3}\right) + \frac{3}{5}N\left(\frac{13}{12}, \frac{5}{9}\right)$

⁴Bimodal simétrica: $\frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$

⁵Bimodal asimétrica: $\frac{3}{4}N(0, 1) + \frac{1}{4}N\left(\frac{3}{2}, \frac{1}{3}\right)$

Distribución H_1		Tamaño de la muestra (n)	Análisis de potencia											
			$\alpha = 0.01$				$\alpha = 0.05$				$\alpha = 0.1$			
			LF	AD	SW	JB	LF	AD	SW	JB	LF	AD	SW	JB
Unimodal asimétrica		20	0.057	0.060	0.081	0.088	0.117	0.177	0.201	0.139	0.187	0.270	0.294	0.180
		50	0.107	0.181	0.240	0.252	0.257	0.369	0.439	0.381	0.376	0.498	0.549	0.467
		100	0.265	0.408	0.485	0.493	0.491	0.631	0.683	0.647	0.618	0.716	0.770	0.744
		500	0.965	0.998	0.999	0.999	0.995	0.999	1	1	0.998	1	1	1
		1000	1	1	1	1	1	1	1	1	1	1	1	1
Bimodal simétrica		20	0.029	0.038	0.020	0	0.126	0.157	0.134	0.004	0.219	0.268	0.266	0.005
		50	0.097	0.189	0.108	0	0.327	0.429	0.366	0	0.462	0.599	0.548	0.01
		100	0.293	0.570	0.427	0	0.648	0.828	0.765	0.106	0.768	0.907	0.876	0.450
		500	1	1	1	0.999	1	1	1	1	1	1	1	1
		1000	1	1	1	1	1	1	1	1	1	1	1	1
Bimodal asimétrica		20	0.021	0.039	0.035	0.014	0.103	0.149	0.153	0.025	0.185	0.237	0.265	0.036
		50	0.106	0.154	0.129	0.005	0.269	0.369	0.382	0.026	0.404	0.522	0.530	0.053
		100	0.312	0.442	0.422	0.019	0.581	0.714	0.720	0.135	0.708	0.830	0.822	0.363
		500	1	1	1	0.968	1	1	1	0.998	1	1	1	1
		1000	1	1	1	1	1	1	1	1	1	1	1	1
Skew-normal	$\lambda=1$	20	0.017	0.015	0.013	0.018	0.056	0.063	0.066	0.032	0.104	0.118	0.112	0.051
		50	0.013	0.016	0.015	0.020	0.077	0.063	0.079	0.049	0.139	0.128	0.138	0.071
		100	0.015	0.020	0.022	0.033	0.063	0.067	0.077	0.066	0.123	0.116	0.130	0.102
		500	0.029	0.045	0.073	0.092	0.109	0.158	0.216	0.195	0.200	0.249	0.310	0.310
		1000	0.067	0.097	0.158	0.162	0.197	0.250	0.339	0.335	0.319	0.371	0.445	0.452
	$\lambda=2$	20	0.018	0.018	0.025	0.036	0.078	0.087	0.097	0.059	0.144	0.156	0.169	0.073
		50	0.021	0.062	0.071	0.094	0.112	0.172	0.218	0.158	0.203	0.274	0.312	0.206
		100	0.078	0.134	0.181	0.173	0.213	0.310	0.352	0.281	0.325	0.395	0.446	0.38
		500	0.530	0.786	0.880	0.849	0.766	0.928	0.960	0.951	0.865	0.960	0.985	0.976
		1000	0.918	0.999	1	1	0.980	1	1	1	0.993	1	1	1
	$\lambda=4$	20	0.051	0.080	0.090	0.063	0.151	0.207	0.225	0.115	0.252	0.297	0.339	0.149
		50	0.136	0.255	0.318	0.214	0.331	0.478	0.567	0.339	0.482	0.614	0.692	0.457
		100	0.355	0.602	0.699	0.492	0.626	0.805	0.873	0.711	0.764	0.864	0.922	0.828
		500	0.997	1	1	1	1	1	1	1	1	1	1	1
		1000	1	1	1	1	1	1	1	1	1	1	1	1

Cuadro 2.3: Potencia de los test de normalidad. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$

LF: Lilliefors **AD**: Anderson-Darling **SW**: Shapiro-Wilk **JB**: Jarque-Bera

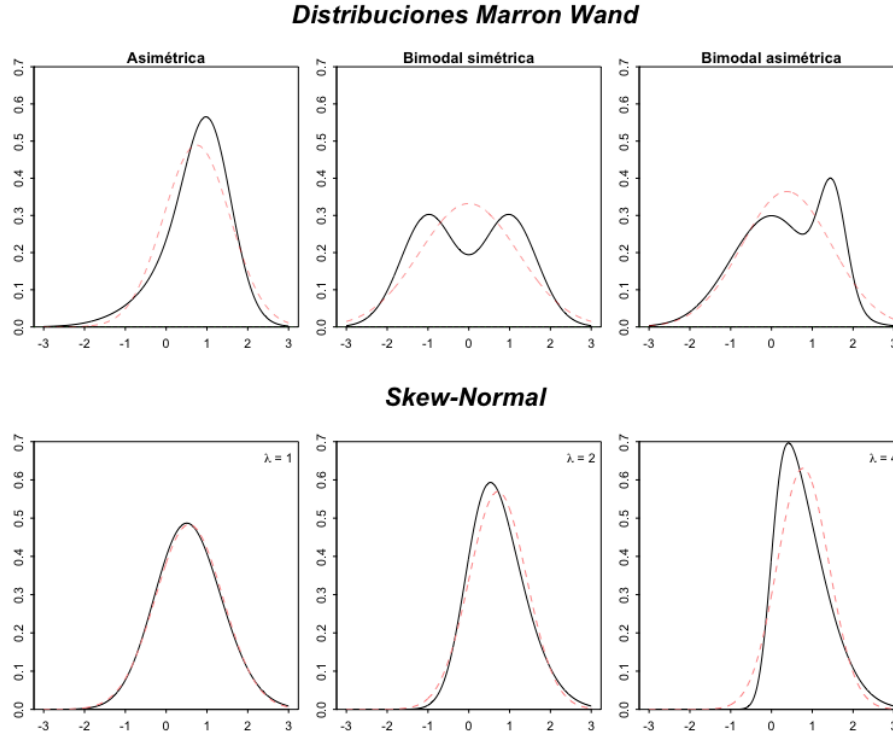


Figura 2.8: Distribuciones utilizadas para el estudio de la potencia superpuestas a una distribución normal (punteado en rojo) que tiene una media y varianza similares.

normalidad de los datos cuando se realice el análisis del caso real posteriormente.

2.3. Test de simetría

En caso de que los datos no sigan una distribución normal se necesitará evaluar su simetría. Como ya se explicó, si los elementos de la muestra siguen una distribución simétrica y unimodal la media sigue siendo una buena función resumen. En cambio, si no existe simetría se puede optar por la mediana como una potencial medida de resumen.

Se considera una m.a.s. $X_1, X_2, \dots, X_n$ de una v.a. continua X con una función de distribución F y una función de densidad f . Se necesita averiguar si f es simétrica alrededor de una mediana desconocida Me , i.e.

$$\begin{aligned} H_0 : f(Me - x) &= f(Me + x) \\ H_1 : f(Me - x) &\neq f(Me + x) \end{aligned} \quad (2.15)$$

Los test de simetría se realizarán y evaluarán del mismo modo que el seguido en los test de normalidad. Se utilizará el procedimiento de la simulación de Monte Carlo para poder

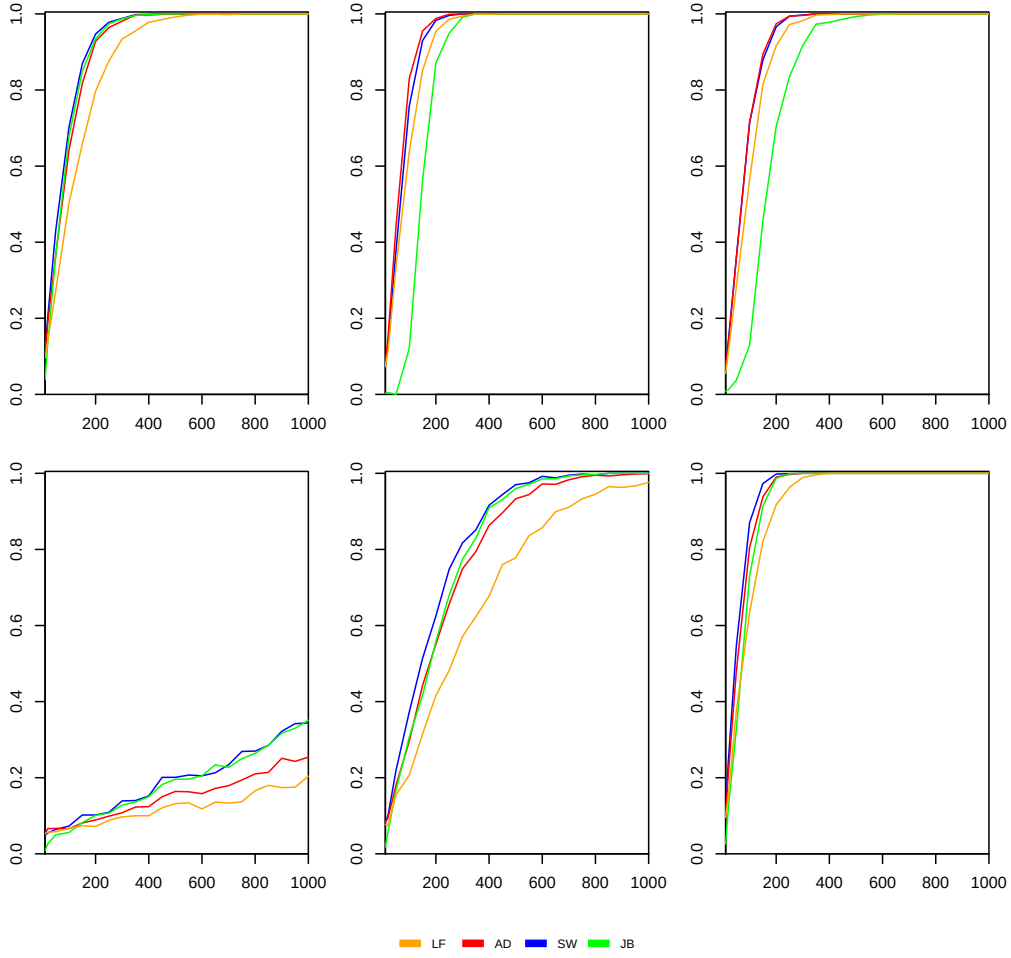


Figura 2.9: Potencia simulada de los test de normalidad frente a las diferentes distribuciones de la Figura 2.8, por el mismo orden y con $\alpha = 0.05$

analizar el calibrado y la potencia de los test de simetría que se expondrán a continuación. Posteriormente se estudiarán los resultados para comprobar qué test resulta más adecuado de acuerdo al calibrado y la potencia mostrada en las simulaciones.

2.3.1. Análisis de los test

En este apartado, se analizarán los dos test de simetría que se considerarán en el estudio. La región de rechazo es la misma para los dos test. Sea T el estadístico de contraste de uno de los test que se verán a continuación. La región de rechazo para T es:

$$\text{Rechazar } H_0 \text{ si } |T| \geq \frac{z_{\alpha/2}\sigma_T}{\sqrt{n}} \quad (2.16)$$

donde σ_T^2 es la varianza asintótica de T y $z_{\alpha/2}$ es el percentil asociado a $1 - \alpha/2$ en una distribución normal estándar.

El valor de σ_T^2 depende de la distribución subyacente de F y es desconocido, por lo que Cabilio y Masaro (1996) proponen el uso de la varianza asintótica 0.5708, derivada de la asunción de que X sigue una distribución normal estándar. Afirmaron que el efecto del error de especificación de la varianza asintótica es menor. Por lo que la nueva región de rechazo que se propondrá a partir de ahora es:

$$\text{Rechazar } H_0 \text{ si } |T| \geq \frac{z_{\alpha/2}\sqrt{0.5708}}{\sqrt{n}} \quad (2.17)$$

La idea de ambos test es simple y consiste en hacer uso del hecho de que si una distribución es simétrica la mediana y la media coinciden. Los estadísticos tratan de evaluar la diferencia de estos dos valores y si este es grande puede llevar a rechazar la hipótesis nula.

2.3.1.1. Test de Cabilio-Masaro

El primer método que se estudiará, será el propuesto por Cabilio y Masaro (1996), basado en el estadístico:

$$C = \frac{\bar{X} - Me}{\mathcal{S}_n} \quad (2.18)$$

en donde, Me se corresponde con la mediana muestral, $\bar{X}$ con la media muestral y $\mathcal{S}_n$ con la cuasivarianza muestral.

El estadístico $\sqrt{n}C$ sigue una distribución asintóticamente normal. En la Figura 2.10 se puede observar el comportamiento de este estadístico

2.3.1.2. Test de Miao et al.

Es una modificación del método de Cabilio y Masaro y fue propuesto por Miao et al. (2006). Este test propone el siguiente estadístico de contraste:

$$G = \frac{\bar{X} - Me}{J_n} \quad (2.19)$$

en donde,

$$J_n = \sqrt{\frac{\pi}{2n^2}} \sum_{i=1}^n |X_i - Me|$$

al igual que en el método anterior $\sqrt{n}G$ también tiene una distribución asintóticamente normal. Un ejemplo de esta distribución asintóticamente normal se puede ver en la Figura 2.10

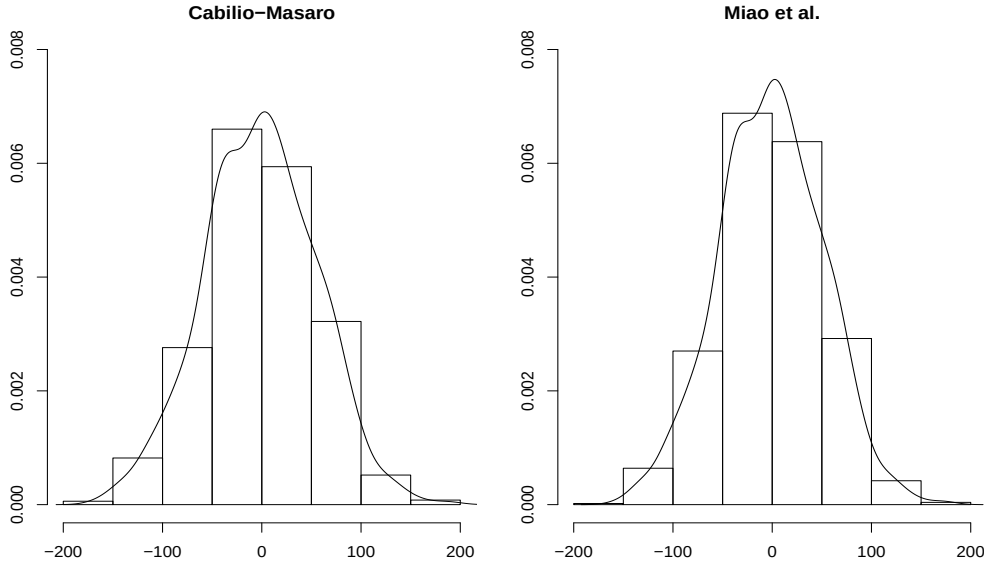


Figura 2.10: Histograma de los estadísticos de Cabilio-Masaro y Miao et al. para 1000 remuestreos de una m.a.s. $X_1, \dots, X_n$ y $n = 1000$, de una v.a. X con una distribución bimodal simétrica.

2.3.2. Estudio de simulación

Se seguirán los mismos procedimientos utilizados en la Sección 2.2.2 para medir el calibrado y potencia de los test y así evaluar su desempeño utilizando diferentes distribuciones de contraste.

2.3.2.1. Procedimiento

Para la evaluación de los test de simetría que se acaban de presentar se llevará a cabo un procedimiento similar al visto en la Sección 2.2.2.1, utilizando el porcentaje de rechazo de H_0 . Por último, para las simulaciones se utilizará también el lenguaje de programación de R. En concreto se usará el paquete *symmetry* y la función `symmetry_test`, con la opción *stat* igual a “CM” cuando se efectúe la prueba de simetría de Cabilio-Masaro y “MGG” cuando se trate de la prueba de Miao et al.

2.3.2.2. Calibrado

Para la medición del calibrado de los test se utilizarán cuatro distribuciones simétricas. Las tres primeras, serán las mismas distribuciones normales que se seleccionaron para el estudio del calibrado de los test de normalidad. Estas son: $N(0, 1)$, $N(0, 5)$ y $N(0, \frac{1}{4})$. También se considera la distribución bimodal simétrica de Marron y Wand (1992).

Los resultados obtenidos se pueden ver en el Cuadro 2.4. Se puede observar que ambos test se encuentran bien calibrados, ya que el error porcentaje de rechazo de H_0 está muy próximo al nivel de significación determinado.

2.3.2.3. Potencia

Para analizar la potencia de los dos test, se utiliza un enfoque similar al visto en los test de normalidad. Se aproxima la potencia de los test β tratando de hallar $1 - \hat{\beta}$ mediante las simulaciones.

Para realizar el estudio de la potencia se utilizan las mismas distribuciones que en el la Sección 2.2.2.3, a excepción de la distribución bimodal simétrica que se incluye en el estudio del calibrado ya que no cumple la hipótesis H_1 .

Los resultados obtenidos se muestran en el Cuadro 2.5 donde se puede ver que ambos test tienen potencias muy similares. Ambos resultan ser excesivamente conservadores para muestras pequeñas independientemente de la distribución alternativa. Esto se puede ver claramente en la Figura 2.11 donde se comparan ambos test frente a las distribuciones previamente mencionadas y con un nivel de significación $\alpha = 0.05$. Además para la distribución bimodal simétrica y la distribución skew-normal con $\lambda = 1$ ambos test demuestran no ser muy buenos a la hora de detectar la asimetría de las muestras. En el resto de distribuciones analizadas para muestras con un tamaño grande, la potencia mejora notablemente. A la vista de los resultados obtenidos es difícil decantarse por uno u otro test ya que ambos se encuentran calibrados y la potencia que muestran es muy similar. De todas formas, se seleccionará el test de Miao et al. para evaluar la simetría en el posterior estudio del caso real, ya que J_n , utilizado para obtener el estadístico de Miao et al. (ecuación 2.19), representa un estimador robusto (menos sensible a valores atípicos) de la desviación típica ($\mathcal{S}_n$) (Kendall et al. 1987).

2.4. Test de multimodalidad

Por último, para cerrar el capítulo de contrastes de características, se presentarán los test de multimodalidad. Un test de hipótesis formal, con una hipótesis nula (H_0) de un cierto número de modas se puede formular de la siguiente manera. Sea f la función de densidad de una variable aleatoria X y sea j el número de modas. Con $k \in \mathbb{Z}^+$, el test para identificar el número de modas puede ser formulado como:

$$H_0 : j = k \quad H_1 : j > k \quad (2.20)$$

Distribución H_0	Tamaño de la muestra (n)	Calibrado					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		CM	MGG	CM	MGG	CM	MGG
$N(0, 1)$	20	0.007	0.006	0.046	0.047	0.098	0.095
	50	0.004	0.004	0.047	0.047	0.085	0.086
	100	0.009	0.008	0.053	0.054	0.104	0.113
	500	0.011	0.010	0.053	0.054	0.102	0.100
	1000	0.009	0.012	0.051	0.047	0.103	0.100
$N(0, 5)$	20	0.008	0.008	0.042	0.040	0.085	0.090
	50	0.006	0.009	0.050	0.052	0.100	0.098
	100	0.006	0.004	0.042	0.045	0.103	0.104
	500	0.008	0.008	0.049	0.046	0.099	0.102
	1000	0.005	0.005	0.034	0.033	0.081	0.079
$N\left(0, \frac{1}{4}\right)$	20	0.007	0.006	0.039	0.043	0.092	0.094
	50	0.004	0.003	0.035	0.036	0.079	0.079
	100	0.009	0.011	0.033	0.036	0.073	0.087
	500	0.010	0.008	0.051	0.046	0.101	0.102
	1000	0.007	0.006	0.048	0.053	0.088	0.088
Bimodal simétrica	20	0.026	0.027	0.098	0.099	0.186	0.178
	50	0.022	0.024	0.093	0.083	0.155	0.143
	100	0.026	0.025	0.098	0.096	0.153	0.149
	500	0.023	0.023	0.063	0.063	0.110	0.108
	1000	0.012	0.013	0.058	0.054	0.119	0.119

Cuadro 2.4: Calibrado de los test de simetría. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$

CM: Cabilio-Masaro, **MGG:** Miao et al.

Distribución H_1		Tamaño de las muestras (n)	Análisis de potencia					
			$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
			CM	MGG	CM	MGG	CM	MGG
Unimodal asimétrica		20	0.014	0.019	0.077	0.09	0.163	0.179
		50	0.047	0.054	0.185	0.197	0.281	0.289
		100	0.174	0.188	0.365	0.370	0.483	0.480
		500	0.863	0.863	0.955	0.952	0.979	0.979
		1000	0.992	0.991	0.999	0.999	1	1
Bimodal asimétrica		20	0.020	0.020	0.082	0.085	0.153	0.141
		50	0.035	0.035	0.101	0.102	0.165	0.160
		100	0.041	0.039	0.121	0.114	0.189	0.187
		500	0.106	0.107	0.249	0.249	0.352	0.347
		1000	0.214	0.207	0.410	0.409	0.528	0.526
Skew-normal	$\lambda = 1$	20	0.007	0.009	0.044	0.053	0.098	0.098
		50	0.008	0.010	0.049	0.056	0.104	0.111
		100	0.012	0.012	0.072	0.070	0.125	0.131
		500	0.032	0.033	0.115	0.112	0.180	0.180
		1000	0.065	0.066	0.172	0.174	0.256	0.255
	$\lambda = 2$	20	0.007	0.009	0.060	0.059	0.118	0.123
		50	0.021	0.027	0.115	0.112	0.215	0.212
		100	0.068	0.071	0.221	0.226	0.321	0.321
		500	0.484	0.485	0.720	0.718	0.815	0.818
		1000	0.862	0.854	0.951	0.951	0.978	0.980
	$\lambda = 4$	20	0.024	0.030	0.125	0.132	0.229	0.228
		50	0.123	0.134	0.324	0.330	0.444	0.437
		100	0.306	0.315	0.570	0.571	0.674	0.673
		500	0.989	0.992	1	1	1	1
		1000	1	1	1	1	1	1

Cuadro 2.5: Potencia de los test de simetría. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$

CM: Cabilio-Masaro, MGG: Miao et al.

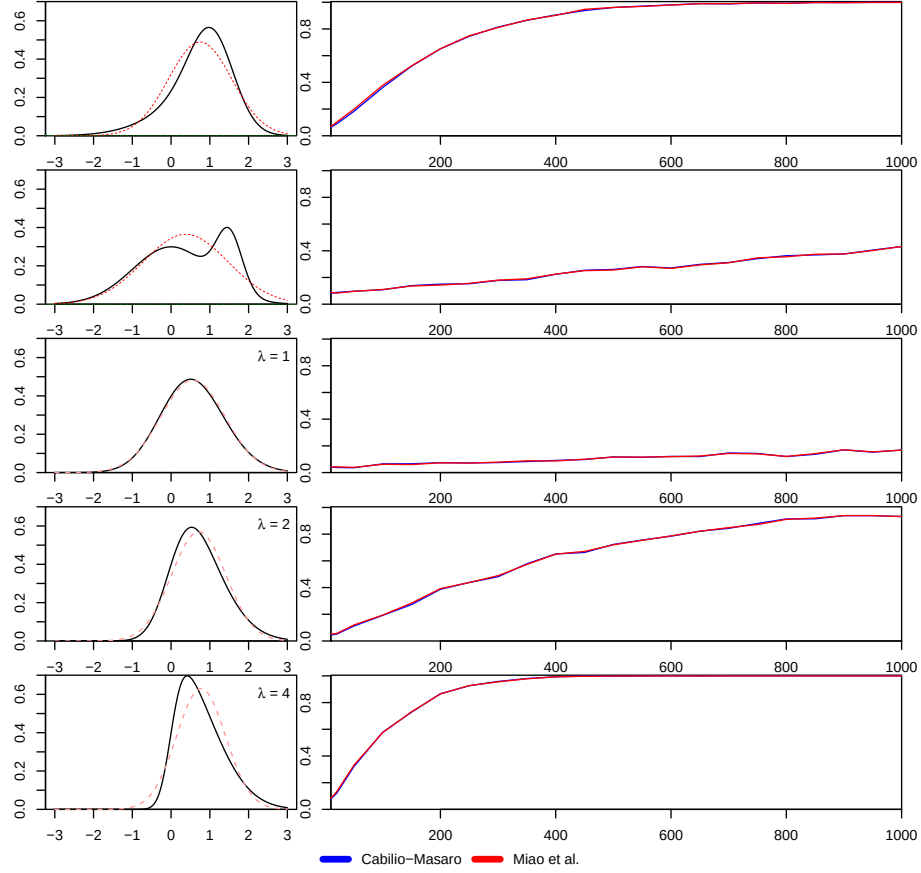


Figura 2.11: Potencia simulada de los test para el estudio de la simetría frente a las diferentes distribuciones del Cuatro 2.5, por el mismo orden de aparición y con $\alpha = 0.05$

Han existido multitud de propuestas para resolver (2.20), pudiendo estas ser separadas en dos grupos: los test que utilizan el ancho de banda crítico y los test basados en el *exceso de masa*. En este trabajo, se estudiará el subcaso de (2.20) en el que $k = 1$, por lo que los test que se analicen servirán para evaluar la unimodalidad vs multimodalidad.

2.4.1. Análisis de los test

A continuación, se presentarán tres test diferentes para evaluar la multimodalidad de una m.a.s. X . Estos tienen diferentes enfoques a la hora de contrastar H_0 , por ejemplo, el test de Silverman utiliza el ancho de banda crítico, mientras que el test basado en *exceso de masa*, mide las zonas dónde se “concentra la probabilidad”. Los test que se analizarán en esta sección serán: el test de Silverman y el test basado en el *exceso de masa*.

2.4.1.1. Test de Silverman

Este test se basa en el uso del ancho de banda crítico para un $k \in \mathbb{Z}^+$ fijo (en el estudio que concierne a este trabajo $k = 1$) y fue formulado por Silverman (1981). Para llevar a cabo el test, se utiliza estimador de tipo núcleo $\hat{f}_h$ que se veía en (2.3). El ancho de banda crítico para el contraste de la unimodalidad es el mínimo valor de h para el cual el estimador de tipo núcleo tiene una moda:

$$h_1 = \inf\{h : \hat{f}_h \text{ tiene una moda}\}$$

Silverman (1981) propone el uso de un estimador tipo núcleo, utilizando un núcleo gaussiano (2.5) como estadístico para contrastar H_0 . Se utiliza este tipo de núcleo ya que el número de modas de $\hat{f}_h$ es una función monótona decreciente en h . Por lo tanto, $\hat{f}_h$ tiene más de una moda si y solo si $h < h_1$. Esto permite rechazar H_0 para valores grandes de h_1 y su distribución se aproxima a partir de un remuestreo por *bootstrap* suavizado.

Calibrado de Hall y York: Peter Hall y York (2001) demostraron que el test de Silverman está mal calibrado. En el Cuadro 2.7 se puede ver un ejemplo de una simulación realizada para comprobar el calibrado del test para una distribución $N(0, 1)$ como la distribución de los p-valores no se aproxima a una distribución $U(0, 1)$. Como se puede ver los porcentajes de rechazo teóricos y empíricos son considerablemente distintos. Peter Hall y York (2001) proponen un calibrado para las hipótesis de (2.20). La propuesta consiste en acotar el soporte de la v.a. y por lo tanto cambiar las hipótesis de contraste de (2.20). De esta forma, dada una m.a.s. $X_1, \dots, X_n$ de X , si j es el verdadero número de modas en un intervalo cerrado I , el nuevo contraste de hipótesis será el siguiente:

$$H_0 : j = 1 \text{ y no tiene mínimos locales en } I$$

Así también es necesario reescribir la ventana crítica:

$$h_{HY} : \inf\{h : \hat{f}_h \text{ tiene exactamente una moda en } I\}$$

Restringiendo el soporte se puede evitar que valores atípicos de la muestra afecten a propiedades de h_1 . Al cambiar la ventana crítica es que ahora $\hat{f}_h$ ya no tiene por qué ser una función monótona decreciente de h . Para solucionar esto se utiliza un núcleo gaussiano para la estimación de $\hat{f}_h$, la probabilidad de que el número de modas de $\hat{f}_h$ sea una función monótona converge a uno (ver Peter Hall y York 2001).

Peter Hall y York (2001) proponen el uso del estadístico de contraste h_{HY} . La distribución nula del estadístico se aproxima mediante procedimientos de *bootstrap*. El problema viene de que los anchos de banda generados mediante estos procedimientos son más pequeños

que los valores reales de h_{HY} . Por lo tanto, para computar el p-valor en un test con significación α se propone el ajuste por un factor λ_α . Para computar este valor se sugirieron dos métodos: el primero basado en una aproximación polinomial y el segundo basado en técnicas de Monte Carlo considerando una distribución unimodal.

2.4.1.2. Test basado en *exceso de masa*

Presentado por Muller y Sawitzki (1991), este test toma un enfoque diferente a lo que se veía en el apartado anterior. Partiendo de una función de densidad f interpreta la moda como la zona donde se concentra la masa de probabilidad. Por lo tanto, teniendo f y una constante λ , el exceso de masa se define como:

$$E(\mathbb{P}_X, \lambda) = \mathbb{P}_X(C(\lambda)) - \lambda \|C(\lambda)\| = \int_{C(\lambda)} f(x) dx - \lambda \|C(\lambda)\| \quad (2.21)$$

donde $C(\lambda) = \{x : f(x) \geq \lambda\}$, $\|C(\lambda)\|$ denota una medida de $C(\lambda)$. Si f tiene k modas independientemente de λ , puede ser dividida en k conjuntos conexos disjuntos en el soporte de f , llamados λ -clusters. Si f tiene k λ -clusters, entonces la definición (2.21) puede ser reescrita como:

$$E_k(\mathbb{P}_X, \lambda) = \sup_{C_1(\lambda), \dots, C_k(\lambda)} \left\{ \sum_{m=1}^k \mathbb{P}_X(C_m(\lambda)) - \lambda \|C_m(\lambda)\| \right\} \quad (2.22)$$

el supremo se toma de entra las familias $\{C_m(\lambda) : m = 1, \dots, k\}$ de λ -clusters. Bajo la hipótesis de que f tiene k λ -clusters, el exceso de masa definido en (2.4.1.2) puede ser estimado empíricamente con $E_{n,k}(\mathbb{P}_n, \lambda)$ de la siguiente manera:

$$E_{n,k}(\mathbb{P}_n, \lambda) = \sup_{\hat{C}_1(\lambda), \dots, \hat{C}_k(\lambda)} \left\{ \sum_{m=1}^k \mathbb{P}_n(\hat{C}_m(\lambda)) - \lambda \|\hat{C}_m(\lambda)\| \right\} \quad (2.23)$$

los conjuntos $\{\hat{C}_m(\lambda) : m = 1, \dots, k\}$ son intervalos cerrados y sus puntos extremos son datos de la muestra. Además, $\mathbb{P}_n(C_m(\lambda)) = (1/n) \sum_{i=1}^n \mathbb{I}(X_i \in \hat{C}_m(\lambda))$. La diferencia $D_{n,k+1}(\lambda) = E_{n,k+1}(\mathbb{P}_n, \lambda) - E_{n,k}(\mathbb{P}_n, \lambda)$, sirve para poder evaluar la veracidad o no de la hipótesis nula. Valores grandes de $D_{n,k+1}(\lambda)$ indican que H_0 es falso. Desarrollando lo anterior, Muller y Sawitzki (1991) propusieron el siguiente estadístico:

$$\Delta_{n,k+1} = \max_{\lambda} \{D_{n,k+1}(\lambda)\}$$

rechazando H_0 si el valor de $\Delta_{n,k+1}$ es muy grande.

Existe otro test conocido como *dip* test que se utiliza únicamente para contrastar la unimodalidad de un conjunto de datos desarrollado por J. A. Hartigan y P. M. Hartigan (1985).

Este test no se estudiará en este trabajo ya que el test de *exceso de masa* demuestra ser una extensión del *dip* test. Los valores obtenidos utilizando ambos test son proporcionales (los valores del estadístico de contraste del *dip* test son el doble de los obtenidos mediante el test basado en *exceso de masa*).

Calibrado de Cheng y Hall: A la vista del comportamiento excesivamente conservativo de la propuestas de exceso de masa, Cheng y P. Hall (1998) diseñaron un procedimiento de calibrado basado en el siguiente resultado: para tamaños grandes de muestras bajo la hipótesis de que f sea unimodal, la distribución de $\Delta_{n,2}$ es independiente de la distribución desconocida excepto por el factor $c = (f^3(x_0)/|f''(x_0)|)^{1/5}$, donde Mo denota la única moda de f .

Cheng y P. Hall (1998) proponen aproximar $\Delta_{n,2}$ empleando la distribución de $\Delta_{n,2}^*$ obtenida mediante las muestras generadas a partir de una distribución de calibración paramétrica $\Psi(\cdot, \beta)$, siendo β un cierto parámetro. Dependiendo de $d = c^{-5}$ diferentes distribuciones paramétricas son propuestas por los autores: si $d = 2\pi$, se utiliza una normal, $d < 2\pi$ se utiliza una distribución beta y si $d > 2\pi$ una t de Student reescalada. Por último d también tendrá que ser aproximado a partir de los valores de la muestra. Suponiendo $\hat{Mo}$ la moda predominante de $\hat{f}_h$, se propone $\hat{d} = |\hat{f}_h''(\hat{Mo})|/\hat{f}_h^3(\hat{Mo})$ siendo $\hat{f}$ y $\hat{f}''$ estimaciones tipo núcleo gaussiano de f y f'' . Donde, h y h' son los anchos de banda asintóticamente óptimos.

La metodología propuesta para realizar el calibrado es la siguiente: suponiendo una m.a.s $X_1, \dots, X_n$ de X , se aproxima $\hat{d}$. A partir de este valor, se estima $\hat{\beta}$ y se escoge una función de densidad $\Psi(\cdot, \hat{\beta})$ asociada. Los valores críticos del estadístico $\Delta_{n,2}^*$ se obtienen utilizando procedimientos de *bootstrap*.

2.4.2. Estudio de simulación

Para el estudio de simulación de los test de multimodalidad se utilizarán los mismos procedimientos que se explicaban en la Sección 2.2.2 y que se utilizaron también en el estudio de simulación de los test de simetría.

2.4.2.1. Procedimiento

Se procederá de la misma forma que se veía en la sección 2.2.2.1 para evaluar el calibrado y potencia de los test de multimodalidad. Los test que se utilizarán para este estudio comparativo serán: el test de Hall y York que se trata de la versión calibrada del test de Silverman y el test de Cheng y Hall, que como ya se explicó, es la modificación calibrada de los test de *exceso de masa*. El estudio se realizará utilizando R y la función que se utilizará

será `modetest` perteneciente al paquete *multimode* con la opción del método siendo “HY” en el caso del test de Hall y York o “CH” en el caso del test de Cheng y Hall.

2.4.2.2. Calibrado

Para analizar el calibrado de los test se hará uso de siete distribuciones diferentes que serán las distribuciones normales: $N(0, 1)$, $N(0, 5)$ y $N(0, \frac{1}{4})$. Además de una distribución bimodal simétrica de Marron y Wand (1992) y tres distribuciones *skew-normal* con $\lambda \in \{1, 2, 4\}$. Los resultados se pueden observar en la Cuadro 2.6. Se aprecia en general que el test de Hall y York es muy conservativo mostrando porcentajes de rechazo inferiores a los niveles de significación. Además, para $\alpha = 0.01$ no consigue rechazar en casi ningún caso salvo para tamaños de muestra grande (y no siempre). Por otra parte, el test de Cheng y Hall se encuentra mejor calibrado que el anterior siendo los niveles de significación empíricos muy próximos a los teóricos. De todas formas tiende a ser conservativo, ya que los porcentajes de rechazo acostumbra a ser menores que los niveles de significación en las distribuciones unimodales asimétricas y *skew-normal* con $\lambda = 1$ y $\lambda = 2$ y sobretodo para tamaños de muestra grandes $n \geq 100$

Por último, también se muestran en el Cuadro 2.7 los porcentajes de rechazo del test de Silverman y del test de *exceso de masa*. Aquí se puede ver la mala calibración de los dos test utilizando como distribución de contraste una $N(0, 1)$. Como se observa en los resultados de la misma es incapaz de rechazar H_0 en ningún caso, apreciándose que los p-valores de los test no siguen en ningún caso una distribución $U(0, 1)$.

2.4.2.3. Potencia

Para el análisis de la potencia de los dos test se han utilizado tres distribuciones bimodales seleccionadas de la lista de Marron y Wand (1992). La llamada distribución bimodal simétrica y la bimodal asimétrica del Cuadro 2.8 se corresponden con las dos distribuciones bimodales ya vistas en los análisis de la potencia y calibrado de los test anteriores (Figura 2.8). A mayores, en esta ocasión, se ha añadido otra distribución a la que se denominará como bimodal simétrica separada⁶ y que se puede ver en la Figura 2.12. Se trata de una distribución bimodal cuyas modas se encuentran a más distancia que las de la otra distribución bimodal simétrica.

⁶Bimodal simétrica separada: $\frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$

Distribución H_0		Tamaño de la muestra (n)	Calibrado					
			$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
			HY	CH	HY	CH	CH	HY
N(0,1)		20	0	0.010	0	0.051	0.004	0.094
		50	0	0.010	0.008	0.046	0.029	0.096
		100	0	0.010	0.009	0.038	0.042	0.078
		500	0.002	0.012	0.010	0.039	0.070	0.084
		1000	0	0.010	0.015	0.037	0.094	0.085
N(0,5)		20	0	0.007	0.001	0.051	0.016	0.099
		50	0	0.011	0.003	0.040	0.031	0.082
		100	0	0.013	0.009	0.042	0.040	0.082
		500	0	0.005	0.006	0.033	0.063	0.085
		1000	0.001	0.008	0.017	0.039	0.082	0.087
N(0, $\frac{1}{4}$)		20	0	0.013	0.002	0.039	0.007	0.092
		50	0	0.01	0.005	0.045	0.032	0.088
		100	0.001	0.008	0.009	0.043	0.039	0.089
		500	0.002	0.008	0.01	0.038	0.070	0.080
		1000	0	0.005	0.011	0.037	0.062	0.088
Unimodal asimétrica		20	0	0.007	0.001	0.039	0.006	0.079
		50	0	0.009	0.001	0.038	0.035	0.076
		100	0	0.003	0.006	0.020	0.059	0.046
		500	0	0.004	0.006	0.037	0.068	0.064
		1000	0	0.009	0.010	0.031	0.072	0.053
Skew-normal	$\lambda=1$	20	0	0.005	0	0.053	0.005	0.096
		50	0	0.011	0.004	0.042	0.035	0.077
		100	0	0.011	0.007	0.035	0.046	0.076
		500	0.002	0.007	0.013	0.041	0.068	0.082
		1000	0.002	0.006	0.011	0.039	0.085	0.072
	$\lambda=2$	20	0	0.013	0	0.051	0.008	0.094
		50	0	0.012	0.001	0.034	0.031	0.08
		100	0	0.003	0.008	0.034	0.048	0.077
		500	0	0.006	0.006	0.035	0.067	0.072
		1000	0	0.004	0.013	0.033	0.073	0.068
	$\lambda=4$	20	0	0.012	0.002	0.058	0.019	0.11
		50	0	0.009	0.004	0.054	0.035	0.116
		100	0	0.010	0.007	0.062	0.050	0.102
		500	0	0.013	0.010	0.048	0.072	0.098
		1000	0.001	0.011	0.018	0.045	0.083	0.101

Cuadro 2.6: Calibrado de los test de multimodalidad. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$

HY: Hall y York, **CH**: Cheng y Hall

Distribución H_0	Tamaño de la muestra (n)	Calibrado					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		MS	SI	MS	SI	MS	SI
$N(0, 1)$	20	0	0	0.007	0	0.019	0
	50	0	0	0.001	0	0.008	0
	100	0	0	0.001	0	0.004	0
	500	0	0	0	0	0	0.001
	1000	0	0	0	0	0	0

Cuadro 2.7: En esta tabla se muestra la mala calibración del test de Silverman y del test de *exceso de masa*. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$

MS: Müller y Sawitzki (*Dip* test), **SI:** Silverman

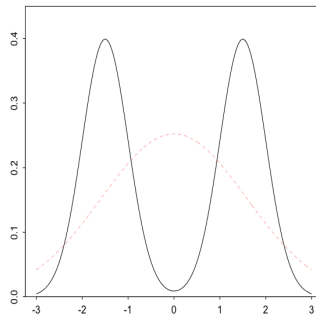


Figura 2.12: Distribución bimodal simétrica separada de Marron y Wand

Los resultados obtenidos se pueden ver en la Cuadro 2.8. Para la distribución bimodal simétrica ambos test muestran una potencia baja para tamaños de muestra $n \leq 100$, aunque cuando $\alpha = 0.1$ ambos test son capaces de rechazar H_0 más de un 60% de las veces. Con tamaños de muestra $n = 500$ y $n = 1000$, la potencia mejora notablemente. En el caso de la distribución bimodal simétrica separada, los resultados mejoran notablemente frente a la distribución anterior, esto es normal ya que debido a la separación de las modas es más fácil detectar la multimodalidad. En este caso Cheng y Hall demuestra una potencia mucho mayor cuando los tamaño de muestra son pequeños ($n \leq 50$) y $\alpha = 0.01$. Lo mismo ocurre cuando $n = 20$ y $\alpha = 0.05$. En el resto de casos ambos test demuestran ser muy potentes. Por último, en el caso de la distribución bimodal asimétrica, los dos test tienen un porcentaje de rechazo de H_0 inferior al del resto de distribuciones. Hall y York es incapaz de rechazar ni la mitad de las veces para el nivel de significación $\alpha = 0.01$, según aumenta el nivel de significación y los tamaños de las muestra, el test mejora notablemente

sus resultados. En el caso de Cheng y Hall ocurre algo similar aunque con porcentajes de rechazo más altos que el test de Hall y York. Cheng y Hall ofrece en general una buena potencia para esta distribución para muestras con $n \geq 500$. Para muestras con un tamaño $n \leq 100$ la potencia no es muy buena y nunca superior al 50 % aunque siempre superior al test de Hall y York.

En la Figura 2.13 se puede ver gráficamente la potencia de los test para $\alpha = 0.05$ y diferentes tamaños de muestra para las diferentes distribuciones. La distribución bimodal simétrica separada es dónde los test demuestran una potencia considerablemente mejor al resto de distribuciones. Además, el test de Cheng y Hall es más potente que el de Hall y York en las tres distribuciones. Por lo tanto seleccionará el test de Cheng y Hall para llevar a cabo los contrastes de unimodalidad durante el estudio del caso real.

En el caso de que los resultados permitan descartar la unimodalidad, se ha incluido un nuevo test para contrastar la hipótesis nula $H_0 : k \text{ modas y } k > 1$ y se ha utilizado el test de Ameijeiras-Alonso et al. (2018) Este test combina los enfoques del ancho de banda crítico de Silverman (1981) y el *exceso de masa* de Muller y Sawitzki (1991), la función para realizar el test será al igual que para los otros tests de multimodalidad `modetest` pero la opción elegida será “*ACR*”.

Distribución H_1	Tamaño de las muestras (n)	Análisis de potencia					
		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
		HY	CH	HY	CH	HY	CH
Bimodal simetrica	20	0	0.083	0.015	0.18	0.102	0.267
	50	0.017	0.165	0.167	0.313	0.358	0.425
	100	0.085	0.296	0.409	0.504	0.603	0.608
	500	0.806	0.894	0.946	0.952	0.974	0.972
	1000	0.981	0.99	0.997	1	0.999	1
Bimodal simetrica separada	20	0.002	0.76	0.287	0.89	0.766	0.938
	50	0.455	0.996	0.995	0.998	1	0.999
	100	0.991	1	1	1	1	1
	500	1	1	1	1	1	1
	1000	1	1	1	1	1	1
Bimodal asimetrica	20	0	0.047	0.011	0.143	0.062	0.227
	50	0.001	0.092	0.064	0.214	0.184	0.329
	100	0.012	0.162	0.183	0.322	0.35	0.441
	500	0.285	0.52	0.574	0.724	0.689	0.82
	1000	0.43	0.715	0.709	0.857	0.808	0.917

Cuadro 2.8: Potencia de los test de multimodalidad. Porcentajes de rechazo de H_0 con 1000 simulaciones y y tamaños de muestra $n = 20$, $n = 50$, $n = 100$, $n = 500$ y $n = 1000$
HY: Hall y York, **CH**: Cheng y Hall

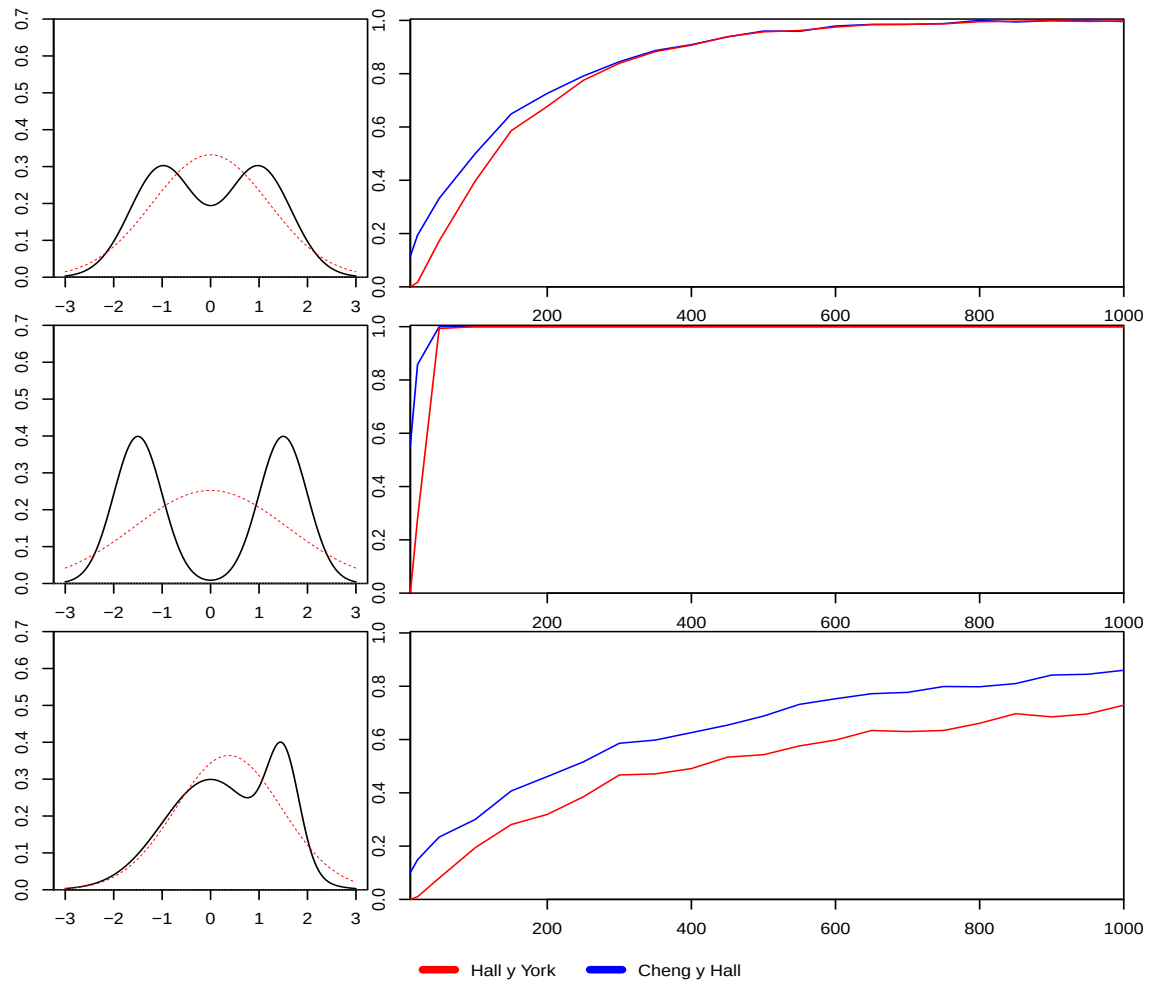


Figura 2.13: Porcentajes de rechazo de los test de modalidad para $\alpha = 0.05$

Capítulo 3

Estudio caso real

En este último capítulo se presentará un estudio de un caso real donde se hace uso de las herramientas gráficas y los contrastes analizados en el Capítulo 2. Con este análisis se pretende, por una parte, comparar las características poblacionales de los datos de entrada y de salida de la red y por otra, analizar las distribuciones tanto de los residuos como de los errores absolutos de las predicciones de la red para seleccionar la métrica que mejor se adapte a cada una dependiendo de sus características.

Este estudio viene enlazado con el TFG desarrollado paralelamente en Ingeniería Informática en el cual se utilizarán los diferentes contrastes analizados anteriormente para evaluar el funcionamiento de distintas arquitecturas de redes neuronales y realizar un análisis comparativo entre ellas.

Por último, como el estudio de una situación real de cada caso queda fuera del alcance de este trabajo, se presentará un apartado de recomendaciones prácticas, en el cual se exponen métricas adecuadas dependiendo de las características de las distribuciones de los errores y errores absolutos de la red para cada caso.

3.1. Idea del funcionamiento de la red

La red que se utilizará es una implementación en el *framework* de *PyTorch* del diseño presentado en Vila Blanco et al. (2020). Esta red está diseñada y entrenada para predecir la edad cronológica y el sexo a través del análisis de ortopantomografías de maduración dental (Figura 3.1) etiquetadas con el sexo y la edad real del sujeto (en días).

Para medir el rendimiento de la red, en primer lugar, se realizará una exploración del conjunto total de datos seleccionados. A partir de aquí se dividirá el conjunto de partida en tres subconjuntos diferentes: el conjunto de entrenamiento, validación y test. El sub-



Figura 3.1: Ejemplo de una ortopantomografía de maduración dental

conjunto de entrenamiento es sobre el que se entrena la red y se efectúa el algoritmo de retropropagación. La red itera sobre los datos de entrenamiento un cierto número de veces (*épocas*). El conjunto de validación mide la pérdida de la red en cada sucesiva *época*. Es decir, en una vez se entra en una *época*, se actualizan los pesos de la red utilizando el subconjunto de entrenamiento y se prueba la red sobre el subconjunto de validación sin actualizar ningún peso. Si la función de coste de la red no disminuye en un cierto número de *épocas* tras aplicar como entrada el conjunto de validación, se pasa directamente a la fase de test. De no ser así la red itera sobre el conjunto de entrenamiento ajustando sus pesos el número de *épocas* predeterminadas. Por último, se evalúa el rendimiento de la red sobre el conjunto de test aplicando los contrastes de hipótesis y realizando una inspección gráfica en busca de sus características.

Sea a la edad real del sujeto (en días) y $\hat{a}$ la edad predicha por la red, además sea $s \in \{0, 1\}$ el sexo real y $\hat{s} \in [0, 1]$ el sexo predicho. La función de coste de la red $\mathcal{L}$ se define como:

$$\mathcal{L} = \omega \sqrt{\frac{\sum_{i=0}^n (a_i - \hat{a}_i)^2}{n}} - \frac{1}{n} \sum_{i=0}^n [s_i * \log(\hat{s}_i) + (1 - s_i) * \log(1 - \hat{s}_i)] \quad (3.1)$$

donde $\omega = 10^{-3}$ para evitar que la red solamente mejore en función a la edad, reduciéndose el impacto de la pérdida de la edad por ese factor. Esta función de coste es una combinación de la función RMSE escalada por el factor ω que mide la pérdida de la edad y una función de entropía cruzada binaria que mide la precisión a la hora de predecir el sexo. Es sobre la función $\mathcal{L}$ sobre la que se aplica el algoritmo de retropropagación, comprobando la contribución de cada parte de la red a la función de pérdida. Los pesos de la red entonces son actualizados de acuerdo con esa contribución y el *learning rate*, buscando minimizar $\mathcal{L}$.

3.2. Caso práctico

Las edades de los sujetos seleccionados para realizar el análisis del caso real irán de 0 a 30 años. A continuación, se muestra la configuración de los subconjuntos de entrenamiento, validación y test. En este caso los conjuntos se han seleccionado de tal manera que el muestreo de los elementos del conjunto total para cada conjunto es aleatorio. El orden de selección de los subconjuntos fue primero el de entrenamiento, validación y test. A mayores, se ha realizado una prueba χ^2 de Pearson en la variable discreta del sexo en los subconjuntos de entrenamiento y validación. El test de Pearson mide la bondad de ajuste entre la distribución observada y una teórica. En este caso, ha usado la distribución uniforme discreta como distribución teórica y se ha establecido un umbral de $\alpha = 0.1$. De esta forma, el número de hombres y mujeres en los subconjuntos de entrenamiento y validación están equilibrados evitando sesgos en el aprendizaje de la red. En la Figura 3.2 se pueden ver los histogramas de los diferentes conjuntos junto con su estimador tipo núcleo.

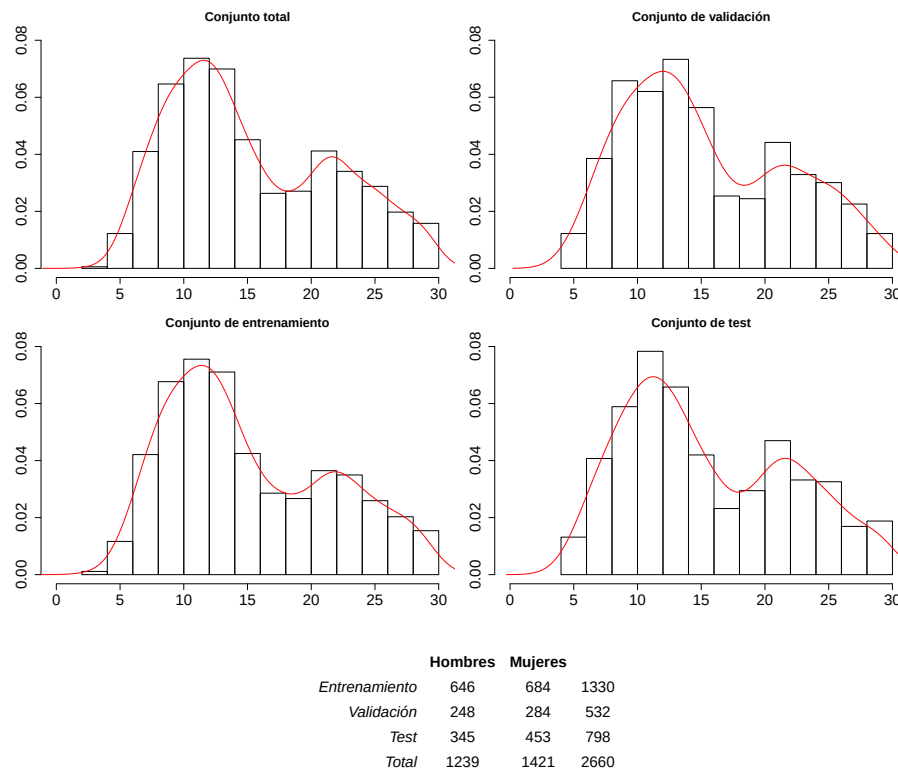


Figura 3.2: Histogramas y estimaciones tipo núcleo (rojo) del conjunto de datos total y los subconjuntos entrenamiento, validación y test.

	$\mu \pm \sigma$	Q_2	IQR	$p99$	ca	k	R^2
Datos reales	15.54 ± 6.49	13.73	10.91	29.35	0.43	2.06	0.88
Predicciones	16.26 ± 6.23	14.92	10.91	28.28	0.29	1.85	

Cuadro 3.1: Principales medidas de las distribuciones vistas en la Figura 3.3

3.2.0.1. Análisis poblacional de los datos de reales y los estimados por la red

Ahora se mostrarán los resultados reales que se han introducido en la red y los obtenidos por esta, estos resultados reales se corresponden con los datos del subconjunto de test. Con esta información se realizará un estudio de las características de los datos reales y los datos estimados por la red por separado para ver sus posibles diferencias. Como se puede ver en la Figura 3.3 tanto la distribución de los datos reales como los de las predicciones son parecidas, aún así es obvio que no son iguales y aunque la forma sea similar no es exactamente igual esto se puede ver tanto en los histogramas como en las estimaciones tipo núcleo de las densidades. El diagrama de dispersión en la parte inferior de la figura se aproxima a una función $y = x$ hasta aproximadamente los 15 años. Después, a medida que avanza la edad de los sujetos aumenta la dispersión de las predicciones. Esto es debido a que a una temprana edad el desarrollo dental es más evidente y según se crece la dentadura no muestra signos tan distintivos que permitan fácilmente discernir la edad con exactitud. Este diagrama se ha incluido porque podría ocurrir que aunque los dos histogramas tengan distribuciones similares, la red podría estar prediciendo erróneamente pero no se reflejase en los histogramas. Utilizando el diagrama de dispersión se pueden detectar este tipo de situaciones. A mayores, los diagramas de cajas muestran características de dispersión de los datos similares. En el Cuadro 3.1 se resumen las medidas más importantes de las distribuciones empíricas de los datos reales y los datos predichos. Se observa que tanto la media como la mediana son superiores en las predicciones mostrando en este conjunto un rango intercuartílico similares así como un percentil 99, un coeficiente de asimetría y una curtosis levemente inferiores. Por último, el valor del R^2 (muy utilizado en este tipo de estudios) nos indica una dependencia lineal alta entre los datos reales y los estimados por la red con un valor de 0.88.

En el Cuadro 3.2 se pueden ver los resultados de los test seleccionados del capítulo anterior. Se ha omitido el test de Shapiro-Wilk debido a que realizando la exploración es evidente que las distribuciones no siguen una distribución normal. Se rechaza la hipótesis de simetría con un nivel de significación $\alpha = 0.05$ para ambos conjuntos de datos. Además, también se

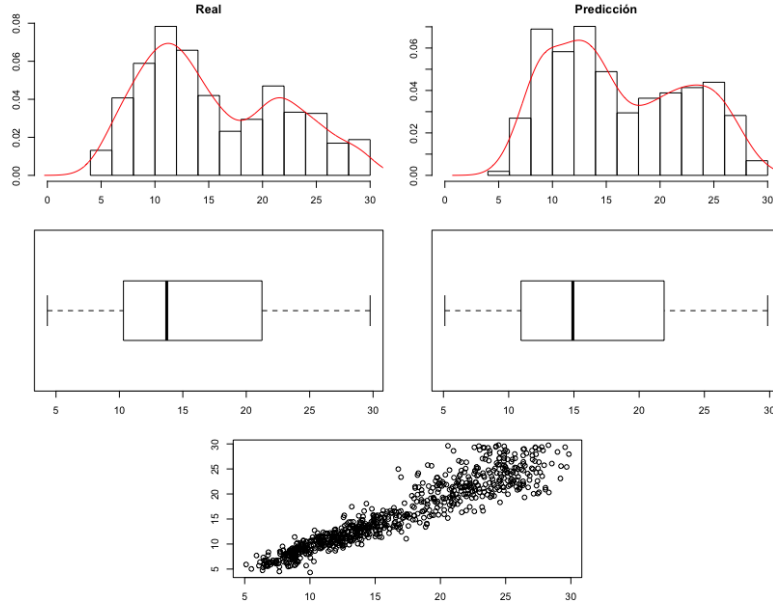


Figura 3.3: Histogramas y estimador tipo núcleo de las distribuciones los datos reales y predichos por la red. Un diagrama de dispersión y un *boxplot* de cada uno

Test		Datos Reales	Predicciones
Simetría (MGG)		$<10^{-6}$	$<10^{-6}$
Modalidad	H_0 : Unimodal (CH)	$<10^{-6}$	$<10^{-6}$
	H_0 : Bimodal (AM)	0.076	0.888

Cuadro 3.2: P-valores obtenidos al ejecutar los test en cada uno de los conjuntos.

MGG: Miao et al. **CH**: Cheng y Hall. **AM**:Ameijeiras et al.

rechaza la hipótesis de unimodalidad. Por otros lado, los resultados del test de Ameijeiras et al. no aportan evidencias significativas que permitan descartar la bimodalidad de ambos conjuntos. Esto concuerda con lo observando la Figura 3.3 en donde las estimaciones tipo núcleo de los datos reales y de las predicciones se observan dos modas.

3.2.0.2. Análisis de los errores

Ahora que se ha comprobado que no hay resultados que contradigan que las distribuciones del conjunto de datos reales y de las predicciones son similares, se analizarán las distribuciones de los errores o residuos ($real - pred$) y de los errores absolutos ($|real - pred|$). En la Figura 3.4 se pueden observar el histograma y estimador tipo núcleo de ambas distribuciones acompañadas de un gráfico Q-Q comparando la distribución muestral con los cuantiles teóricos de una distribución normal y un diagrama de cajas. En los Cuadros 3.4

	$\mu \pm \sigma$	Q_2	IQR	$p99$	ca	k
E	-0.72 ± 2.23	-0.71	2.30	6.01	0.43	4.95
AE	1.75 ± 1.55	1.30	1.90	7.09	1.48	5.33

Cuadro 3.3: Principales medidas de las distribuciones vistas en la Figura 3.4. **E**: Residuos. **AE**: Errores absolutos.

Test	E	AE
Normalidad (SW)	$<10^{-6}$	$<10^{-6}$
Simetría (MGG)	0.90	$<10^{-6}$
Unimodalidad (CH)	0.78	0.11

Cuadro 3.4: P-valores obtenidos al ejecutar los test en los datos de los errores y los errores absolutos. **E**: Residuos. **AE**: Errores absolutos. **SW**: Shapiro-Wilk. **MGG**: Miao et al. **CH**: Cheng y Hall.

y 3.3 se pueden ver los resultados de los test y las medidas muestrales más importantes. A la vista de las gráficas y de los resultados de los test, una medida que se ajusta bien a la distribución de los residuos sería la media, ya que la distribución sin ser normal (esto se puede comprobar observando el p-valor del test de Shapiro-Wilk en el Cuadro 3.4 y el gráfico Q-Q izquierdo de la Figura 3.4), es claramente unimodal y no se puede rechazar la hipótesis nula de simetría para un nivel de significación $\alpha = 0.05$. Además, esto es comprobable visualmente, observando el *boxplot* y viendo que el coeficiente de asimetría es muy bajo. Por tanto, no se encuentran pruebas suficientes que permitan rechazar la simetría. Por otro lado, para la distribución de los errores absolutos la mediana resulta ser una medida más adecuada que la media, esto se debe principalmente a la asimetría y unimodalidad de la muestra. Gracias al test MGG (Cuadro 3.4) se obtienen evidencias significativas que permiten rechazar la hipótesis de simetría. Este comportamiento asimétrico puede verse claramente en el histograma y comprobando el coeficiente de asimetría que aparece en el Cuadro 3.3. Como además no existen evidencias que nos permitan rechazar la unimodalidad (visualmente tampoco se observa un comportamiento no unimodal), se selecciona la mediana como una medida que resume mejor que la media el comportamiento del error absoluto.

En conclusión, para este caso tras haber realizado la exploración los datos reales y de las predicciones ofrecidas por la red comprobando que las características de ambas distribuciones eran similares y tras estudiar las características distribuciones de los residuos y errores

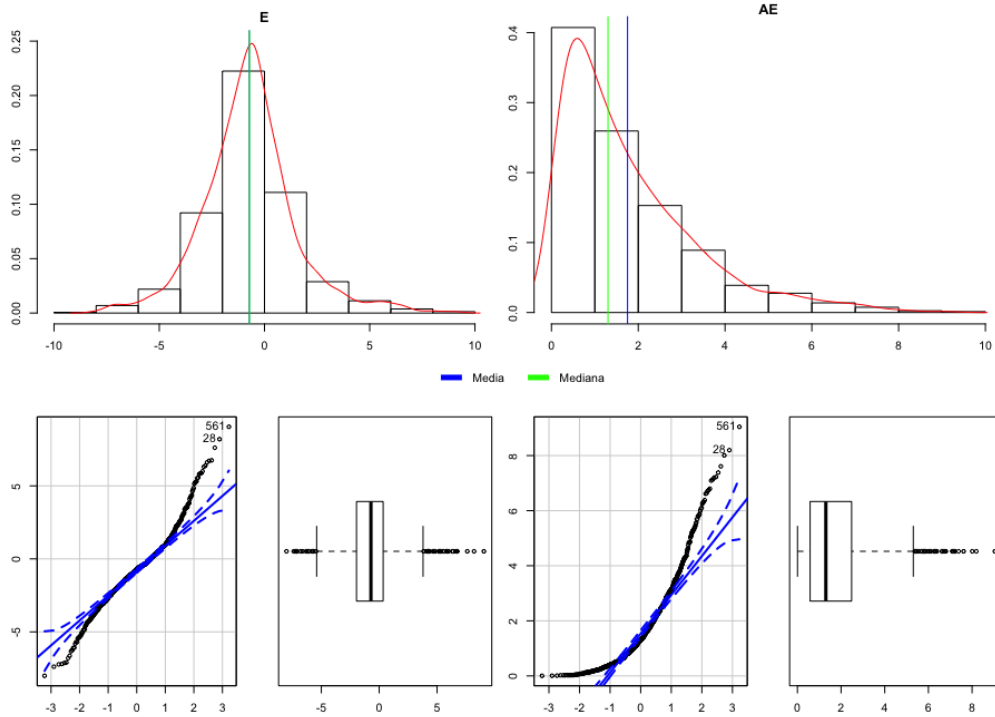


Figura 3.4: Histogramas y estimador tipo núcleo de las distribuciones de los errores y errores absolutos. Se marcan la media y la mediana muestrales de los datos. Se acompaña de un diagrama de caja y un gráfico QQ con los cuantiles teóricos de una distribución normal estándar. **E**: Residuos. **AE**: Errores asolutos

absolutos de las predicciones. Se puede concluir, que para los residuos el uso de la media resulta justificado debido a las características unimodales y simétricas observadas tras realizar los test de contraste de hipótesis y mediante la exploración gráfica. En cambio, para el error absoluto la media deja de ser una buena medida ya que aunque no hay evidencias que contradigan la unimodalidad de los datos, sí las hay para rechazar la hipótesis de simetría. Basta con observar los resultados del test MGG (muy inferior a 0.05) en el Cuadro 3.4 y el valor del coeficiente de asimetría (1.48) en el Cuadro 3.3. Por lo tanto, la asimetría de la distribución hace que la media se vea más afectada y sea preferible el uso de la mediana. Esto se puede ver en la Figura 3.4, la media se ve desplazada hacia la cola derecha de la distribución de los errores absolutos. Por lo tanto, la mediana se considera que la mediana localiza con más precisión donde se agrupan la mayoría de los datos.

3.2.1. Recomendaciones prácticas

En la sección anterior presenté una propuesta de como llevar a cabo un análisis de los datos de una red para estudiar su comportamiento y para seleccionar las métricas (características de la muestra) adecuadas para evaluar el error y el error absoluto. En el primer paso, se han comparado las características de normalidad, simetría y multimodalidad de las distribuciones tanto de los datos de entrada como los de salida para tratar de observar posibles discrepancias entre ellas. En el siguiente paso se han analizado las mismas características de las distribuciones de los residuos y errores absolutos. Se analiza cada una de ellas por separado, lo que justifica el uso de diferentes métricas para cada distribución dependiendo de cómo sea su comportamiento.

En primer lugar, se debe señalar que el estudio de las distribuciones de los datos reales y los datos predichos por la red es útil para obtener una visión preliminar de su comportamiento y de cómo predice la red, pero no aporta evidencias significativas de su verdadera precisión. En el ejemplo expuesto, se utiliza para dar una primera idea de como aproxima la red la edad de los sujetos comparándola con su edad real.

A continuación, a la hora de analizar los residuos y errores absolutos, las medidas relacionadas con los residuos aportan información sobre si la red tiende a sobre o infraestimar la edad de los sujetos pero al cancelarse los valores positivos con los negativos las medidas no aportan información sobre la verdadera magnitud del error de la red más allá de lo mencionado. Es en este punto es donde entra el error absoluto. Las medidas de esta distribución sí nos dan una idea del rendimiento real de la red. Se podría haber considerado otras transformaciones alternativas como el error cuadrático que penaliza los valores más extremos pero siguiendo la publicación de Vila Blanco et al. (2020) este estudio se centrará solamente en el error absoluto. En todo caso, las pautas a la hora de seleccionar una métrica serían las mismas.

Ahora bien, ¿qué medidas se considerarán adecuadas para cada caso? Como ya se explicó anteriormente en muchas ocasiones se tiende a utilizar la media sin analizar ni las características de las distribuciones de los datos de entrada ni de los errores de la red. Esto es debido a que es una medida común y conocida por prácticamente todo el mundo. Su uso queda justificado cuando la distribución que se analice sea normal (una suposición habitual) o simétrica y unimodal. La mediana también se podría utilizar en este caso, ya que para distribuciones continuas perfectamente simétricas ambas medidas coinciden. En este caso de estudio estamos tratando de inferir si las distribuciones son simétricas y es muy complicado obtener datos que sigan una distribución perfectamente simétrica. Por ejemplo, en el caso práctico visto en la sección anterior, no hay pruebas concluyentes que permitan descartar la hipótesis de simetría para el error y como se observa en el Cuadro

3.3, la media y la mediana son prácticamente iguales. Por lo tanto, aunque la distribución de nuestros datos no sea perfectamente simétrica, pero no se obtienen pruebas que permitan descartar las hipótesis de unimodalidad y simetría (si fuese normal, también cumpliría con estas características) la media y la mediana serán muy parecidas por lo que el uso de cualquiera de estas medidas estará justificado.

En el caso que la distribución sea asimétrica y unimodal la media deja de ser una medida adecuada. Esto se debe a que la media se ve muy afectada por la asimetría de la distribución. Por tanto, es preferible utilizar la mediana ya que es mucho más robusta que la media en el sentido de que se ve menos afectada por valores extremos de la muestra. Se puede observar perfectamente en la Figura 3.4 como la asimetría izquierda de la distribución de los residuos hace que la media se vea desplazada hacia la cola derecha de la distribución y la mediana se posiciona más hacia la izquierda. En este tipo de casos se utilizará la mediana como medida.

En el caso de que exista multimodalidad el modo de estudiar la distribución será diferente y este ha sido un problema que no se ha estudiado en la práctica, y que excede los contenidos de este TFG, pero para el que hacemos la siguiente propuesta preliminar. Primero se hallarán por medio de la inspección gráfica y mediante el uso de los test de multimodalidad el número de modas de la distribución, supongamos k . A continuación, se realiza una aproximación de la distribución de los datos a partir de k componentes normales a partir de una aproximación por un modelo de mixtura de normales utilizando el algoritmo de maximización de la esperanza (EM) (Dempster et al. 1977). Esta aproximación, fijando k componentes para k modas, puede ser adecuada si estas son simétricas y suficientemente separadas. A través de las herramientas gráficas se comprueba y tras la obtención de la ubicación de las componentes, se comprueba que dichas componentes se hallan suficientemente separadas, de ser así, se pueden aproximar las modas de las distribuciones mediante las medias de aproximadas de las componentes. Dicha mixtura de normales tiene la siguiente forma: sea f la función de densidad subyacente de los residuos o del error absoluto, y k el número de modas observadas mediante las herramientas gráficas y contrastadas mediante el uso de los contrastes de hipótesis:

$$f \approx \sum_{i=1}^k \lambda_i N(\mu_i, \sigma_i) \quad (3.2)$$

donde μ_i y σ_i son las medias y desviaciones típicas de las componentes normales aproximadas por el algoritmo. Además, los valores μ_i representarán las aproximaciones de las modas de la muestra original. Por otra parte, los valores λ_i representarán el peso que tiene cada componente sobre la distribución estimada siendo además $\sum_{i=1}^m \lambda_i = 1$. Estas componentes se estudiarán por separado. En este contexto, para obtener una magnitud que

ayude a representar la concentración de los datos de la distribución se ha optado por usar los pesos (λ_i) y las medias (μ_i) dadas por el algoritmo EM para medir la contribución de cada componente y obtener un valor dependiente de cada uno de las componentes. Por lo tanto, en el caso del estudio de una distribución con soporte no acotado, a mayores del estudio de cada componente por separado la medida que se propone, la ponderación de cada una de las modas (WM), que mide la contribución global de cada moda y se define como:

$$WM = \sum_{i=1}^m \lambda_i \cdot \mu_i \quad (3.3)$$

en el caso del error absoluto, como el soporte es acotado $[0, +\infty)$, habrá que tener en cuenta la masa de probabilidad que queda por debajo de 0 utilizando el algoritmo EM (ya que este se realiza suponiendo un soporte no acotado). Por tanto, sea Φ_i la función de distribución de la componente normal i -ésima y si denotamos $m_i = \Phi_i(0)$, entonces WM se redefine como:

$$WM = \sum_{i=1}^m \lambda_i \cdot \mu_i \cdot (m_i + 1) \quad (3.4)$$

Nótese que esta propuesta tiene la limitación de que, el número de componentes de una mixtura de normales no tiene porqué coincidir con el número de modas de la distribución (pensemos, por ejemplo, en la combinación de dos normales con la misma media pero distinta desviación, donde la mixtura de dos componentes presentará únicamente una moda). En todo caso, podría ser una aproximación inicial a la distribución. Otra opción sería considerar la estimación tipo núcleo con ventana crítica para k modas, y obtener la localización de las mismas, pero se haría necesario asignarle a cada una un cierto peso, de manera similar a las ecuaciones 3.3 y 3.4.

Capítulo 4

Conclusión

En este trabajo se realiza un análisis de diferentes herramientas gráficas y un estudio de simulación de distintos contrastes de hipótesis de tipo no paramétrico con el fin de detectar características de normalidad, simetría y multimodalidad en un conjunto de datos. El objetivo final ha sido utilizar estas herramientas para analizar las estimaciones de una red neuronal desarrollada paralelamente en el TFG en Ingeniería Informática. Como conclusión de este trabajo, se presentarán algunas consideraciones sobre la utilización práctica en el ejemplo expuesto de las técnicas explicadas a lo largo de la memoria, y validadas mediante estudios de simulación.

Primeramente, se han comparado las distribuciones de los datos reales y los estimados por la red para comprobar que se conserven las características poblacionales. A continuación, se han analizado las distribuciones de los residuos y los errores absolutos. Para estas distribuciones qué estadístico muestral resulta idóneo a la hora de dar un valor que “resuma” cada distribución (evalúe las estimaciones y el rendimiento de la red) dependiendo de sus características (normalidad, simetría y multimodalidad). Es interesante resaltar, que para el caso multimodal, a mayores del uso del algoritmo EM para el caso de soporte acotado inferiormente (errores absolutos) se evaluaron alternativas distintas siendo una de ellas el uso de una transformación Box-Cox de los datos, el uso posterior del algoritmo EM y por último deshacer la transformación pero el resultado final dejaba de ser una mixtura de normales y no se encontró una forma clara de ponderar las componentes, por lo tanto, se descartó esta opción.

En definitiva, este estudio contribuye a mostrar la importancia del análisis de las distribuciones de los datos para no hacer asunciones erróneas de los sobre ellas que puedan llevar a realizar inferencias que lleven a conclusiones erróneas. La utilización de una metodología similar a la presentada en el caso práctico puede ayudar a mejorar la forma de evaluar todo tipo sistemas orientados a la resolución de problemas de regresión (no solamente sistemas

basado en *deep learning*) reforzando las conclusiones obtenidas a partir de los resultados desde un punto de vista estadístico.

Bibliografía

- Ameijeiras-Alonso, J. (2014). *Test de multimodalidad: ventana crítica y exceso de masa* (Tesis de maestría). Departamento de Estadística e Investigación Operativa. Universidad de Santiago de Compostela. Universidad de Vigo. Universidad de la Coruña.
- Ameijeiras-Alonso, J., Crujeiras, R. M. & Rodríguez-Casal, A. (2018). Mode testing, critical bandwidth and excess mass. *TEST*, 28(3), 900-919. <https://doi.org/10.1007/s11749-018-0611-5>
- Anderson, T. W. & Darling, D. A. (1952). Asymptotic theory of certain «goodness of fit» criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2), 193-212. <https://doi.org/10.1214/aoms/1177729437>
- Anderson, T. W. & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268), 765-769. <https://doi.org/10.1080/01621459.1954.10501232>
- Cabilio, P. & Masaro, J. (1996). A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(3), 349-361. <http://www.jstor.org/stable/3315744>
- Cheng, M.-Y. & Hall, P. [P.]. (1998). Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3), 579-589. <https://doi.org/https://doi.org/10.1111/1467-9868.00141>
- D'Agostino, R. B. & Stephens, M. A. [Michael A.]. (1986). *Goodness-of-Fit Techniques* (Vol. 68). Marcel Dekker, Inc.
- Dallal, G. E. & Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors's test statistic for normality. *The American Statistician*, 40(4), 294-296. <https://doi.org/10.1080/00031305.1986.10475419>
- Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1-22.

- Hall, P. [Peter] & York, M. (2001). On the calibration of Silverman's test for multimodality. *Statistica Sinica*, 11(2), 515-536. <http://www.jstor.org/stable/24306875>
- Hartigan, J. A. & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, 13(1), 70-84. <http://www.jstor.org/stable/2241144>
- Hoaglin, D. C., Mosteller, F. & (Editor), J. W. T. (2000). *Understanding Robust and Exploratory Data Analysis* (1.^a ed.). Wiley-Interscience.
- Ibarrola, R. & Pérez, A. (2012). *Principios de Inferencia Estadística*. Universidad Nacional de Educación a Distancia. <https://books.google.es/books?id=rVEqmAEACAAJ>
- Jarque, C. M. & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259. [https://doi.org/https://doi.org/10.1016/0165-1765\(80\)90024-5](https://doi.org/https://doi.org/10.1016/0165-1765(80)90024-5)
- K. Blitzstein, J. & Hwang, J. (2019). *Introduction to Probability Second Edition*. <https://drive.google.com/file/d/1VmkAAGOYCTORq1wxSQqy255qLJjTNvBI/view>
- Kendall, M. G., Stuart, A. & Ord, J. K. (1987). *Kendall's Advanced Theory of Statistics*. Oxford University Press, Inc.
- Kolmogorov, A. L. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4, 83-91. <https://ci.nii.ac.jp/naid/10030673552/en/>
- Li, Y., Huang, Z., Dong, X., Liang, W., Xue, H., Zhang, L., Zhang, Y. & Deng, Z. (2018). Forensic age estimation for pelvic X-ray images using deep learning. *European Radiology*, 29. <https://doi.org/10.1007/s00330-018-5791-6>
- Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399-402. <http://www.jstor.org/stable/2283970>
- Marron, J. S. & Wand, M. P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, 20(2), 712-736. <https://doi.org/10.1214/aos/1176348653>
- Miao, W., Gel, Y. & Gastwirth, J. (2006). A new test of symmetry about an unknown median. *Random Walk, Sequential Analysis and Related Topics-A Festschrift in Honor of Yuan-Shih Chow*. https://doi.org/10.1142/9789812772558_0013
- Muller, D. W. & Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415), 738-746. <http://www.jstor.org/stable/2290406>
- Peña, D. (2014). *Fundamentos de Estadística*. Alianza Editorial. <https://books.google.es/books?id=z0mUBQAAQBAJ>
- R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>

- Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2(3), 117-119. <https://doi.org/10.1007/bf01891203>
- Royston, P. (1995). Remark AS R94: a remark on algorithm AS 181: the W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4), 547-551. <http://www.jstor.org/stable/2986146>
- Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (Complete Samples). *Biometrika*, 52(3/4), 591-611. <http://www.jstor.org/stable/2333709>
- Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1), 97-99. <https://doi.org/https://doi.org/10.1111/j.2517-6161.1981.tb01155.x>
- Stephens, M. A. [M. A.]. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347), 730-737. <http://www.jstor.org/stable/2286009>
- Stirzaker, D. (2003). *Elementary Probability* (2.^a ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511755309>
- Vila Blanco, N., Carreira, M., Varas-Quintana, P., Balsa-Castro, C. & Tomás Carmona, I. (2020). Deep Neural networks for chronological age estimation from OPG images. *IEEE Transactions on Medical Imaging, PP*, 1-1. <https://doi.org/10.1109/TMI.2020.2968765>
- Wang, X., Guo, R. & Kambhamettu, C. (2015). Deeply-learned feature for age estimation. *2015 IEEE Winter Conference on Applications of Computer Vision*, 534-541.
- Weglarczyk, S. (2018). Kernel density estimation and its application. *ITM Web of Conferences*, 23, 00037. <https://doi.org/10.1051/itmconf/20182300037>
- Xing, J., Li, K., Hu, W., Yuan, C. & Ling, H. (2017). Diagnosing deep learning models for high accuracy age estimation from a single image. *Pattern Recognition*, 66, 106-116. <https://doi.org/https://doi.org/10.1016/j.patcog.2017.01.005>