



Universidad de Oviedo

ANÁLISIS COMPARATIVO DE DISTINTOS  
TEST DE NORMALIDAD

Lorena González Rodríguez

Dirigido por Raúl Pérez Fernández

UNIVERSIDAD DE OVIEDO  
Facultad de Ciencias  
Grado en Matemáticas

Mayo de 2022

# Índice general

<b>1. Introducción</b>	<b>1</b>
<b>2. Preliminares</b>	<b>3</b>
2.1. Variable aleatoria . . . . .	3
2.2. Función de distribución y función de densidad . . . . .	5
2.3. Esperanza y varianza . . . . .	6
2.4. Distribuciones de interés . . . . .	7
2.4.1. Distribución normal . . . . .	7
2.4.2. Distribución uniforme . . . . .	8
2.4.3. Distribución exponencial . . . . .	10
2.4.4. Doble exponencial . . . . .	11
2.4.5. Mixtura de dos normales . . . . .	12
<b>3. Gráficos para ilustrar la normalidad</b>	<b>14</b>
3.1. Histogramas . . . . .	15
3.2. Estimación de la función de densidad . . . . .	18
3.2.1. El histograma como estimador de la función de densidad y la poligonal de frecuencias . . . . .	18
3.2.2. Estimador naïve o histograma móvil . . . . .	19
3.2.3. Estimador núcleo . . . . .	20
3.3. Diagrama de cajas . . . . .	22
3.4. Diagrama de violín . . . . .	27
3.5. QQ-plots . . . . .	28
3.6. PP-plots . . . . .	32
3.7. Gráficos de la función de distribución empírica acumulada . . . . .	35
<b>4. Tests de normalidad</b>	<b>38</b>
4.1. Test de Lilliefors . . . . .	38
4.2. Test de Cramer Von Mises . . . . .	40
4.3. Test de Anderson Darling . . . . .	42
4.4. Test de Shapiro-Wilk . . . . .	43
4.5. Test de Shapiro-Francia . . . . .	44
4.6. Test de Jarque-Bera . . . . .	45

4.7. Test de Jarque-Bera robusto . . . . .	47
4.8. Adjusted Jarque Bera . . . . .	48
4.9. Test de D'Agostino-Pearson . . . . .	49
4.10. Test chi-cuadrado . . . . .	50
4.10.1. En el caso de hipótesis simples . . . . .	50
4.10.2. En el caso de hipótesis compuestas . . . . .	51
4.11. Test de Filliben . . . . .	52
4.12. Test de Chen-Shapiro . . . . .	53
4.13. Test de Frosini . . . . .	54
4.14. Test de Geary . . . . .	54
<b>5. Comparación de la potencia de los distintos tests presentados</b>	<b>56</b>
5.1. Potencia bajo la hipótesis nula . . . . .	57
5.1.1. Distribución normal . . . . .	57
5.2. Potencia bajo la hipótesis alternativa . . . . .	59
5.2.1. Distribución exponencial . . . . .	59
5.2.2. Distribución uniforme . . . . .	61
5.2.3. Distribución doble exponencial . . . . .	63
5.2.4. Distribución mixtura gausiana . . . . .	65
5.2.5. Distribución t de Student . . . . .	76
5.2.6. Distribución chi-cuadrado . . . . .	86
<b>6. Conclusiones</b>	<b>96</b>
<b>Bibliografía</b>	<b>98</b>
<b>A. Apéndice</b>	<b>101</b>

# Capítulo 1

## Introducción

A la hora de realizar el análisis estadístico de un conjunto de datos, con el fin de conocer su comportamiento y sacar las conclusiones más correctas del estudio, es de gran utilidad conocer la distribución de probabilidad subyacente. Siempre y cuando no se encuentren claras evidencias de lo contrario, uno de los caminos que más se siguen cuando se plantea una situación de este tipo consiste en suponer que los datos que se manejan provienen de una distribución normal. Esta suposición, denominada hipótesis de normalidad, nos permite realizar una gran cantidad de pruebas paramétricas al conjunto de datos. Afortunadamente, para la comprobación de la hipótesis de normalidad, disponemos de una gran variedad de opciones. En este trabajo vamos a estudiar algunas de ellas. En particular, nos centraremos en dos tipos de pruebas: las pruebas gráficas y los test estadísticos de normalidad. En el primer caso, tenemos como objetivo distinguir los métodos gráficos que nos dan más información acerca del conjunto de datos, así como aquellos que a simple vista nos permitan decidir si existe la normalidad que buscamos comprobar. Con respecto a los test de normalidad, el objetivo es conocer cuáles son más fiables a la hora de aplicarse a un experimento de estudio de la normalidad, por lo que realizaremos también un estudio de la potencia de los mismos.

En el Capítulo 2 se enunciarán los preliminares necesarios para el desarrollo de este trabajo. En concreto, se recordarán diversas nociones relativas a la Teoría de la Probabilidad y, en particular, aquellos aspectos relevantes para el estudio de variables aleatorias unidimensionales. Además, se repasarán los conceptos de función de distribución y de densidad, así como sus características. Se tratarán con mayor profundidad cinco distribuciones dando, para cada una de ellas, sus expresiones, propiedades y el cómo afecta la variación de sus correspondientes parámetros a la expresión gráfica de su función de densidad.

En el Capítulo 3 se tratarán algunos de los métodos gráficos más importantes para comprobar la normalidad en los datos. Se aplicarán estos

métodos a muestras aleatorias generadas a partir de las distribuciones explicadas en el capítulo anterior, y se comentarán los resultados obtenidos.

Por otro lado, el Capítulo 4 se centrará en presentar los principales test de normalidad, algunas modificaciones de estos, y a mencionar otros menos utilizados. Se introducirán el tipo de hipótesis con las que se trabajará, se explicarán los estadísticos de cada test, las regiones de rechazo, los valores críticos de cada región de rechazo y se aplicará cada test a modo ilustrativo para cada una de las muestras del anterior capítulo.

Por último, en el Capítulo 5 se llevará a cabo una estimación por simulación de la potencia de los test anteriores para distintos tamaños muestrales. Por medio de simulaciones por el método de MonteCarlo, se realizarán los experimentos (tanto bajo la hipótesis nula de normalidad, como bajo distintas hipótesis alternativas) buscando la obtención de conclusiones acerca de cuáles son los test de mayor y menor potencia. De esta manera se concluirá cuáles son los más y menos recomendados a la hora de contrastar la normalidad de un conjunto de datos dado.

## Capítulo 2

# Preliminares

### 2.1. Variable aleatoria

De forma intuitiva una variable aleatoria es una función medible, la cuál consiste en dado un experimento aleatorio asignar un valor numérico a cada resultado del mismo. No es posible conocer de forma previa el valor que va a tomar dicha variable, por eso se le asigna el adjetivo aleatoria, sin embargo sí que podemos conocer su distribución de probabilidad sobre el conjunto de valores que puede tomar.

Para definir formalmente el concepto de variable aleatoria (unidimensional) primero debemos introducir una serie de conceptos que explicaremos a continuación y que dan lugar al modelo matemático a partir del cual construiremos nuestra variable:

- El **espacio muestral**  $\Omega$  o conjunto de los posibles resultados del experimento en el que estemos trabajando.
- La  **$\sigma$ -álgebra de sucesos** asociados al experimento aleatorio: la definimos como el conjunto de todas las cuestiones relativas a la realización de un experimento aleatorio, de manera que una vez se haya realizado el mismo podemos saber de manera inequívoca si dichas cuestiones son ciertas o no, es decir que podemos decidir sin miedo a equivocarnos si han sucedido o no. Se denotará por  $\mathcal{A}$  y debe cumplir las propiedades siguientes:
  1. El vacío es un elemento de la  $\sigma$ -álgebra:  $\emptyset \in \mathcal{A}$ ;
  2. Si tomamos un elemento en la  $\sigma$ -álgebra también encontramos en ella a su complementario: sea  $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$ ;
  3. La unión numerable de elementos de la  $\sigma$ -álgebra está en la  $\sigma$ -álgebra : sean  $A_1, A_2, \dots, A_n, \dots \in \mathcal{A}$ , entonces  $A_1 \cup A_2 \cup \dots \cup A_n \cup \dots \in \mathcal{A}$ .

Puesto que nuestro objetivo es poder llegar a definir el concepto de variable aleatoria (unidimensional) debemos de tener en cuenta una  $\sigma$ -álgebra en concreto, y es la  $\sigma$ -álgebra de Borel  $\mathcal{B}_{\mathbb{R}}$  para el caso  $\Omega = \mathbb{R}$  y que se define como la menor  $\sigma$ -álgebra que contiene a todos los conjuntos que pueden expresarse según intervalos reales mediante complementación y unión/intersección numerables.

Con estos dos elementos  $(\Omega, \mathcal{A})$  podemos formar un espacio medible o probabilizable, al que añadiendo el tercer elemento que explicamos a continuación obtenemos nuestro espacio probabilístico.

- Según la definición axiomática de probabilidad de Kolmogorov, una **probabilidad**  $\mathcal{P}$  es una aplicación  $\mathcal{P} : \mathcal{A} \rightarrow [0, 1]$  que cumple las siguientes tres condiciones:
  1. **Axioma de no negatividad:**  $\forall A \in \mathcal{A} : P(A) \geq 0$ ;
  2. **Axioma de normalización:**  $P(\Omega) = 1$ ;
  3. **Axioma de  $\sigma$ -aditividad:**  $A_1, A_2, \dots, A_n, \dots \in \mathcal{A}$  con  $A_i \cap A_j = \emptyset$  si  $i \neq j$ , se tiene:

$$P(A_1 \cup A_2 \cup \dots \cup A_n \cup \dots) = \sum_{i=1}^{\infty} P(A_i).$$

Con estos tres elementos tenemos definido un espacio probabilístico  $(\Omega, \mathcal{A}, \mathcal{P})$ , que nos permite definir a continuación el concepto de variable aleatoria. Una **variable aleatoria** (unidimensional) se define entonces a partir de un espacio probabilístico  $(\Omega, \mathcal{A}, \mathcal{P})$  como una función del espacio muestral  $\Omega$  a otro espacio medible  $E$ , que en general se va a corresponder con el conjunto de los números reales  $\mathbb{R}$ , es decir es una aplicación  $X : \Omega \rightarrow \mathbb{R}$  de forma que dado un  $B \in \mathcal{B}_{\mathbb{R}}$  la antiimagen del mismo es un suceso de interés contenido en  $\mathcal{A}$ , por lo que se cumple que

$$\text{Dado } B \in \mathcal{B}_{\mathbb{R}} : X^{-1}(B) = \{ \omega \in \Omega \mid X(\omega) \in B \} \in \mathcal{A}.$$

Debido a la definición de la sigma-álgebra de Borel  $\mathcal{B}_{\mathbb{R}}$  a partir de intervalos y a las propiedades de la antiimagen, podemos obtener una definición equivalente: **una variable aleatoria (unidimensional)** es cualquier aplicación  $X : \Omega \rightarrow \mathbb{R}$  tal que  $\forall x \in \mathbb{R}$  se tiene que

$$X^{-1}((-\infty, x]) = (X \leq x) = \{ \omega \in \Omega \mid X(\omega) \leq x \} \in \mathcal{A}.$$

El **rango o recorrido** de una variable aleatoria  $X$  se define como el conjunto de valores que toma  $X$ , es decir,  $R_X = \{x \in \mathbb{R} \mid \exists \omega \text{ con } X(\omega) = x\}$ . En función de esto vamos a distinguir dos tipos de variables aleatorias:

- Variables aleatorias discretas:  $R_X$  es un conjunto finito o infinito numerable.

- Variables aleatorias continuas:  $R_X$  es un conjunto infinito no numerable.

De ahora en adelante, nos interesaremos exclusivamente en variables aleatorias continuas e introduciremos la noción de función de distribución. En particular, nos interesaremos en aquellas variables aleatorias cuya función de distribución es absolutamente continua y definiremos el concepto de función de densidad.

## 2.2. Función de distribución y función de densidad

La función de distribución (acumulada) nos permite describir la distribución que sigue una variable aleatoria. Formalmente, la función de distribución (acumulada), o CDF del inglés Cumulative Distribution Function, de una variable aleatoria  $X$  definida en su correspondiente espacio probabilístico  $(\Omega, \mathcal{A}, \mathcal{P})$  se define como:

$$F_X(x) = P(X \leq x), \text{ para cualquier } x \in \mathbb{R}.$$

Es decir que la función de distribución acumulada nos devuelve la probabilidad de que nuestra variable aleatoria tome valores menores o iguales que cualquier valor real dado  $x$ . Al ser una probabilidad, el espacio de llegada será el intervalo unidad  $[0, 1]$ , es decir,  $F_X : \mathbb{R} \rightarrow [0, 1]$ . El subíndice lo utilizamos para indicar explícitamente que nos referimos a la variable aleatoria  $X$ , sin embargo muchas veces se omite y se escribe únicamente  $F(x)$  siempre y cuando no haya lugar posible a confusión. La función de distribución nos permite fácilmente calcular la probabilidad de que nuestra variable tome valores en un cierto intervalo  $(a, b]$  con  $a < b$ , en ese caso se expresará como  $P(a < X \leq b) = F(b) - F(a)$ .

Una función  $F : \mathbb{R} \rightarrow [0, 1]$  debe cumplir una serie de propiedades para poder ser una función de distribución:

1. Al ser una probabilidad, toma valores entre 0 y 1 y por lo tanto está acotada;
2. Debe ser una función monótona no decreciente, es decir que dados dos valores  $x, y$  con  $x < y$  se cumple que  $F_X(x) \leq F_X(y)$ ;
3. Debe ser continua por la derecha, es decir que

$$\lim_{x \rightarrow x_0^+} F_X(x) = \lim_{x \rightarrow x_0, x_0 > x} F_X(x) = F_X(x_0);$$

4. Se cumple que  $\lim_{x \rightarrow \infty} F_X(x) = 1$  y  $\lim_{x \rightarrow -\infty} F_X(x) = 0$ .



En este trabajo, nos interesaremos en variables aleatorias continuas para las que la función de distribución es  $F$  es absolutamente continua, es decir,  $F_X$  se puede expresar para todo  $x \in \mathbb{R}$  como

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f(u) \, du,$$

para cierta función  $f : \mathbb{R} \rightarrow \mathbb{R}$  tal que  $f(u) \geq 0$  para todo  $u \in \mathbb{R}$  y  $\int_{-\infty}^{+\infty} f(u) \, du = 1$ . A esta función se la denomina función de densidad y hablaremos de ella a continuación.

De manera análoga, haciendo uso del Teorema Fundamental del Cálculo, tenemos que:

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Es decir que la función de densidad se define como la derivada de la función de distribución con respecto a  $x$ .

Una función de densidad  $f$  debe cumplir las siguientes propiedades:

- $f_X(x) \geq 0 \, \forall x$ ;
- $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$ ;
- $\int_a^b f_X(x) \, dx = F(b) - F(a)$ .

Algo importante a tener en cuenta es que en los puntos en los que la función de distribución no es derivable, se considera que la función de densidad se anula.

## 2.3. Esperanza y varianza

Sea  $(\Omega, \mathcal{A}, \mathcal{P})$  un espacio probabilístico y  $X$  una variable aleatoria definida en dicho espacio, con su correspondiente función de densidad  $f_X$ . Definimos la **esperanza o media** de la variable aleatoria  $X$  como el ‘valor medio’ que se espera como resultado de nuestro experimento aleatorio cuando se repite varias veces y la probabilidad de cada suceso no varía. Viene dada por:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx.$$

De igual forma definimos la **varianza** como la variabilidad que presentan los datos de un experimento con respecto a la media del mismo. Viene dada por:

$$Var[X] = E[X^2] - E[X]^2 = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx - \left( \int_{-\infty}^{\infty} x f_X(x) \, dx \right)^2.$$

A la raíz cuadrada de la varianza se la conoce como **desviación típica**.

Destacamos brevemente algunas de las propiedades más importantes y utilizadas de la esperanza y varianza:

- Linealidad de la esperanza:  $E(aX + bY) = aE(X) + bE(Y)$ .
- Si tenemos  $X$  e  $Y$  variables aleatorias independientes:

$$E(XY) = E(X)E(Y) \text{ y } Var(X + Y) = Var(X) + Var(Y)$$

## 2.4. Distribuciones de interés

En esta sección discutiremos las principales distribuciones de interés para este trabajo.

### 2.4.1. Distribución normal

La distribución normal, que también se conoce habitualmente como distribución gaussiana, es un tipo de distribución de probabilidad continua para variables aleatorias que toman valores en  $\mathbb{R}$ .

Vamos a representar el hecho de que una variable aleatoria  $X$  sigue una distribución normal como  $X \sim N(\mu, \sigma)$ , donde  $\mu$  se refiere a la media de la distribución, que determina el centro de la gráfica de la función de densidad, y  $\sigma$  se refiere a la desviación típica. Su función de densidad viene dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

La función de distribución vendrá dada entonces por:

$$F(x) = \int_{-\infty}^x f(u) du,$$

no siendo posible dar una fórmula explícita.

Esta distribución es simétrica con respecto a su media, lo que quiere decir que la mitad de los datos quedarán por debajo de la media y la mitad quedarán por encima de la misma. Además la media coincidirá con la mediana y la moda.

También cuenta con la propiedad de reproductividad, es decir, que la suma de variables aleatorias independientes que sigan distribuciones normales es una nueva variable aleatoria que también sigue una distribución normal cuya media será la suma de las medias de las variables, y la varianza la suma de las varianzas:

$$X_1 + X_2 + \dots + X_n \sim N(\mu_1 + \dots + \mu_n, \sqrt{\sigma_1^2 + \dots + \sigma_n^2})$$

siempre que  $X_i \sim N(\mu_i, \sigma_i)$  independientes para todo  $i$ .

El caso más sencillo de este tipo de distribuciones es conocido como distribución normal estándar y viene dado por los valores de los parámetros  $\mu = 0$  y  $\sigma = 1$ , con lo que  $f(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$  es su función de densidad.

Podemos pasar de tener una distribución normal con unos parámetros cualesquiera  $N(\mu, \sigma)$ , a una distribución normal estándar  $N(0, 1)$  por medio de un proceso al que vamos a llamar tipificación de la variable, el cuál consiste en tomando nuestra variable aleatoria  $X \sim N(\mu, \sigma)$ , restarle su media y dividir dicha resta entre la desviación típica, obteniendo así la nueva variable tipificada  $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ . Este proceso nos va a ser útil puesto que para la normal estándar podemos siempre consultar tablas de valores para su función de distribución.

Vamos a tomar como ejemplo visual de la función de densidad la Figura 2.1

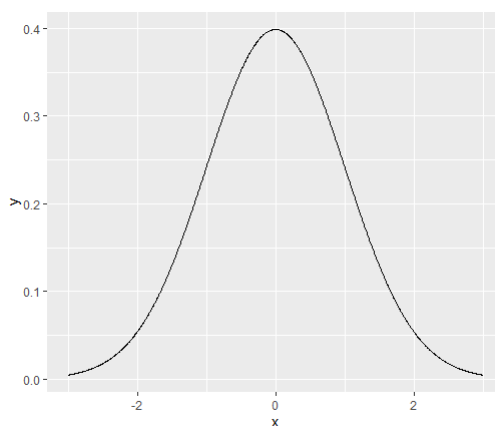


Figura 2.1: Función de densidad para la  $N(0,1)$ .

Podemos apreciar cómo tiene forma de campana, y es simétrica respecto a la media que en este caso toma el valor 0, al ser la normal estándar. Para los diversos valores de  $\mu$  y  $\sigma$  la forma de la campana irá variando, volviéndose más o menos achatada en función de la desviación típica y desplazándose hacia derecha o izquierda en función del valor que tome la media.

#### 2.4.2. Distribución uniforme

Esta es también una distribución de probabilidad continua y además simétrica. El hecho de que una cierta variable aleatoria  $X$  sigue una distribución uniforme se denota por  $X \sim U(a, b)$  donde  $a$  y  $b$ , con  $a < b$ , son los parámetros que nos indican lo que valen el mínimo y el máximo del soporte de la variable aleatoria.

En este caso la función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{b-a} , & \text{si } x \in [a, b] \\ 0 , & \text{en el resto} \end{cases}$$

mientras que la función de distribución vendrá dada por:

$$F(x) = \begin{cases} 0, & \text{si } x < a \\ \frac{x-a}{b-a}, & \text{si } x \in [a, b) \\ 1, & \text{si } x \geq b \end{cases}$$

En este caso con respecto a la media y la varianza de la variable  $X \sim U(a, b)$  tenemos que:

1. La esperanza viene dada por

$$E(X) = \frac{a+b}{2}.$$

2. La varianza viene dada por

$$Var(X) = \frac{(b-a)^2}{12}.$$

Al igual que en la distribución normal existía lo que llamábamos distribución normal estándar, con la uniforme ocurre lo mismo y tenemos la llamada distribución uniforme estándar que se caracteriza por los parámetros  $a = 0$  y  $b = 1$ . Luego, si nuestra variable  $X$  sigue una distribución uniforme estándar, escribiremos  $X \sim U(0, 1)$  cuya función de densidad será:

$$f(x) = \begin{cases} 1, & \text{si } x \in [0, 1] \\ 0, & \text{en el resto} \end{cases}$$

Este caso concreto de uniforme es un caso especial de la conocida distribución beta, con parámetros  $p = 1$  y  $q = 1$ , es decir  $U(0, 1) = \beta(1, 1)$ .

Con respecto al gráfico de la función de densidad tomamos de referencia la Figura 2.2.

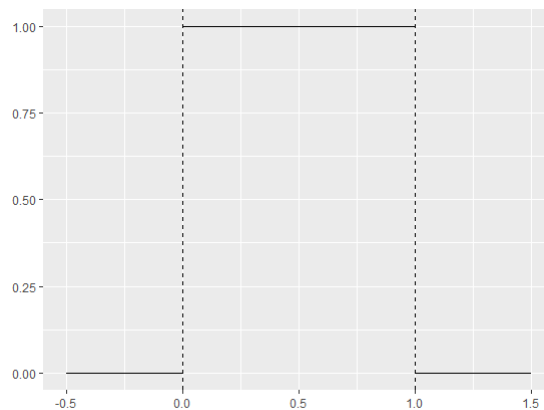


Figura 2.2: Función de densidad para la  $U(0, 1)$ .

Observamos claramente en la representación cómo la densidad de los valores que están fuera del intervalo unidad es 0, y los que están dentro de dicho intervalo unidad tienen una densidad constante de valor 1, en este caso. La modificación de los parámetros de esta función de densidad solamente va a modificar el intervalo de valores en los que la función no sea nula, y el valor constante que toma la función en los puntos en los que no lo es.

### 2.4.3. Distribución exponencial

El hecho de que una cierta variable aleatoria  $X$  sigue una distribución exponencial se denota por  $X \sim \text{exp}(\lambda)$ . Su función de densidad viene dada por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x \geq 0 \\ 0, & \text{si } x < 0 \end{cases}$$

donde  $\lambda$  es el parámetro de escala asociado a este tipo de distribuciones y siempre es positivo. Para la gráfica de la función de densidad de la exponencial, en función del valor que tome el parámetro  $\lambda$ , cambiará la escala de la función. En la Figura 2.3 representamos como ejemplo la función de densidad de una variable aleatoria con distribución exponencial de parámetro  $\lambda = 1$ .

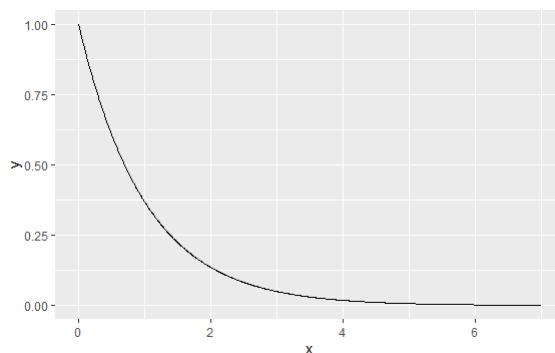


Figura 2.3: Función de densidad para la exponencial de parámetro  $\lambda = 1$ .

La función de distribución viene dada pues por:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{si } x \geq 0 \\ 0, & \text{si } x < 0 \end{cases}$$

En este caso con respecto a la media y la varianza tenemos que:

1. La esperanza viene dada por:

$$E(X) = \frac{1}{\lambda}$$

2. La varianza viene dada por:

$$Var(X) = \frac{1}{\lambda^2}$$

#### 2.4.4. Doble exponencial

La distribución doble exponencial es también conocida como distribución de Laplace. Se puede obtener a partir de dos variables que siguen distribuciones exponenciales, idénticamente distribuidas e independientes, pero con distinto signo.

Denotaremos el hecho de que una variable aleatoria sigue una distribución doble exponencial por  $X \sim Laplace(\mu, b)$ . Su función de densidad viene dada por:

$$f(x) = \frac{1}{2b} \begin{cases} e^{-\frac{\mu-x}{b}}, & \text{si } x < \mu \\ e^{-\frac{x-\mu}{b}}, & \text{si } x \geq \mu \end{cases}$$

La función de distribución será entonces:

$$F(x) = \begin{cases} \frac{1}{2}e^{-\frac{\mu-x}{b}}, & \text{si } x < \mu \\ 1 - \frac{1}{2}e^{-\frac{x-\mu}{b}}, & \text{si } x \geq \mu \end{cases}$$

donde  $\mu$  y  $b$  son los parámetros asociados a esta distribución, con  $b$  siempre positivo.

La distribución doble exponencial estándar se presenta cuando los parámetros toman los valores  $\mu = 0$  y  $b = 1$  y su gráfica de la función de densidad para estos parámetros se representa en la Figura 2.4.

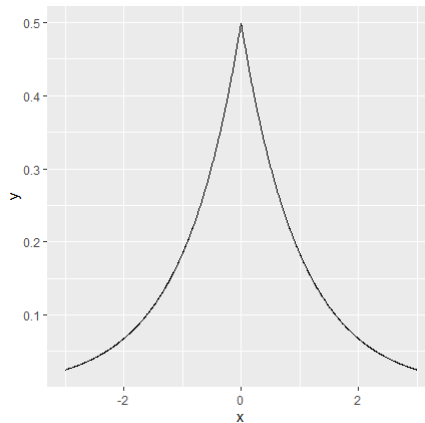


Figura 2.4: Función de densidad de Laplace estándar.

Al variar el primer parámetro haremos que la gráfica se mueva a derecha o izquierda, y la variación del segundo parámetro hará que la función sea más o menos achatada.

### 2.4.5. Mixtura de dos normales

En general, la función de densidad de una variable aleatoria  $X$  cuya distribución es una distribución mixtura finita de  $n$  componentes [23] viene dada por:

$$f(x) = \sum_{i=1}^n p_i f_i(x)$$

donde

- $f_i(x)$  son las funciones de densidad de cada una de las  $i$  componentes por las cuáles está formada la mixtura.
- $p_i$  son los pesos que se les da a cada una de dichas componentes, debiendo cumplirse que  $\sum_{i=1}^n p_i = 1$ .

En esta ocasión, la distribución que vamos a estudiar en concreto es la mixtura finita de dos normales: es decir las componentes de nuestra mixtura van a ser dos distribuciones normales o gaussianas generadas de manera aleatoria. Para asignar un peso a cada una de las dos distribuciones normales, seleccionaremos una probabilidad  $p \in (0, 1)$  y para ver si nos encontramos en una de las dos distribuciones normales o en la otra, utilizaremos una distribución binomial que tome valores 0 y 1 (Bernoulli) de forma aleatoria también. En particular, la función de densidad vendrá dada por:

$$\begin{aligned} f(x) &= \lambda f_1(x, \mu_1, \sigma_1) + (1 - \lambda) f_2(x, \mu_2, \sigma_2) \\ &= \lambda \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + (1 - \lambda) \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}} \end{aligned}$$

con  $f_1$  la función de densidad correspondiente a la primera de las componentes normales de la mixtura, y  $f_2$  la función de densidad correspondiente a la segunda de las componentes normales de la mixtura. Además, al tener solo dos componentes realizamos una combinación lineal convexa, donde asignamos como pesos a cada una de las distribuciones  $\lambda$  y  $1 - \lambda$  respectivamente, cumpliendo que la suma de los pesos sea 1.

Tenemos que:

- La esperanza vendrá dada por :

$$E(X) = \lambda\mu_1 + (1 - \lambda)\mu_2$$

- La varianza vendrá dada por :

$$Var(X) = \lambda\sigma_1^2 + (1 - \lambda)\sigma_2^2 + \lambda(1 - \lambda)(\mu_1 - \mu_2)^2$$

Veamos en la Figura 2.5 un ejemplo de la función de densidad de una variable aleatoria con distribución mixtura de dos normales representada en color negro, una de parámetros media 3 y desviación típica 5 cuya densidad podemos ver en color rojo, y la otra de parámetros media 3 igual que la anterior, pero con distinta desviación típica: 30 en este caso y representada de color azul. La  $\lambda$  utilizada en este caso toma el valor de 0.5. En función del valor que tome la  $\lambda$  la densidad de la mixtura cambiará asemejándose más o menos a la densidad de alguna de las dos normales que la componen.

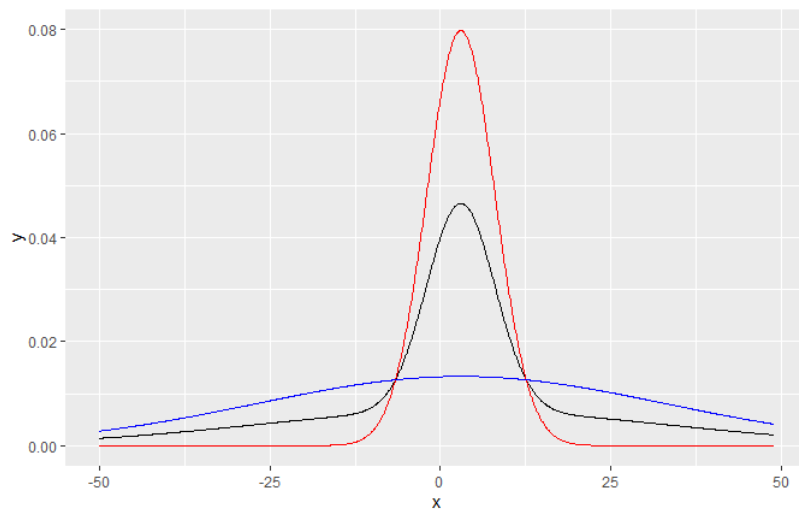


Figura 2.5: Función de densidad para la mixtura de dos normales.



## Capítulo 3

# Gráficos para ilustrar la normalidad

Los gráficos son una parte importante en el estudio de la normalidad de una muestra. A continuación, vamos a explicar algunos de los procedimientos gráficos más utilizados para analizar si una muestra proviene o no de una distribución normal. Para ello vamos a utilizar cinco muestras de tamaño 100 provenientes de cinco distribuciones distintas, generadas de manera aleatoria:

- Distribución normal;
- Distribución uniforme;
- Distribución exponencial;
- Distribución doble exponencial;
- Distribución mixtura de dos normales.

Mostramos todos los valores que conforman cada una de las muestras en las Tablas A.1, A.2 y A.3, incluidas en el apéndice. Antes de empezar a explicar los procedimientos gráficos, vamos a describir brevemente las muestras mostradas en la tabla anterior mediante un resumen estadístico, mostrado en la Tabla 3.1. Este resumen aportará datos importantes como las medidas de posición de tendencia central, los cuartiles, o la asimetría y la curtosis de las mismas. Nótese que la asimetría y la curtosis son especialmente útiles para identificar las distribuciones normales, pues están asociadas a asimetría 0 y curtosis 3.

	Normal	Uniforme	Exponencial	Doble exp.	Mixtura
Min.	-8.550	0.010	0.020	-19.490	-25.740
Q1	0.528	2.455	1.570	-1.862	-9.255
Mediana	3.310	4.665	4.240	3.185	0.070
Media	3.452	4.986	5.229	2.947	1.613
Q3	6.457	7.550	7.348	6.470	10.448
Max.	13.940	9.940	36.060	39.060	49.070
SD	4.564	2.849	5.194	7.388	14.571
IQR	5.930	5.095	5.778	8.332	19.703
Asimetría	0.06	0.079	2.79	-0.595	0.543
Curtosis	2.893	1.745	15.123	4.161	3.271

Tabla 3.1: Tabla resumen de las muestras consideradas.

### 3.1. Histogramas

Los histogramas sirven para visualizar la distribución de un conjunto de datos. Son gráficos en forma de barras, por lo general verticales, en los cuales se representa la distribución de frecuencias de un conjunto de datos discretizado en intervalos, es decir, las barras reflejan las frecuencias observadas de cada una de las clases en las que discretizamos nuestra variable de estudio. Dicha variable por fuerza ha de ser cuantitativa, puesto que de lo contrario estaríamos hablando simplemente de un diagramas de barras. Por medio de un histograma nos es sencillo identificar la simetría de los datos, su dispersión, si hay presencia de valores atípicos o no en la muestra, o la multimodalidad (es decir, la función de densidad subyacente tendrá más de un pico). A continuación, vamos a ver ejemplos concretos de histogramas de nuestras cinco muestras y procederemos a dar una breve explicación de los aspectos más significativos que observamos en ellos.

En la Figura 3.1, se observa el histograma de la muestra proveniente de la distribución normal. La forma del histograma se asemeja a una campana de Gauss, lo que es característico de las distribuciones normales. Podemos observar que el histograma es bastante simétrico. Además, observamos que no presenta grandes valores atípicos. Con respecto a la multimodalidad: no se encuentra presente en este histograma, pues se intuye una densidad subyacente con una única moda.

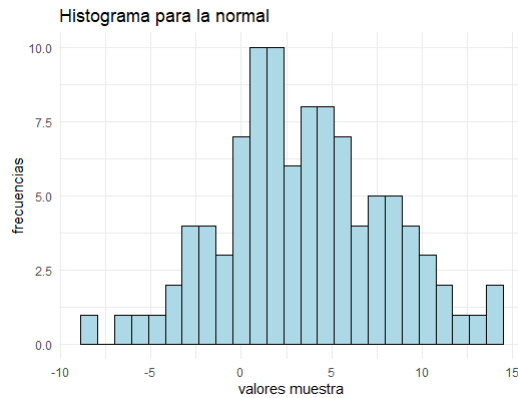


Figura 3.1: Histograma de una muestra de una distribución normal generada aleatoriamente.

En la Figura 3.2 observamos el histograma de la muestra que proviene de la distribución exponencial. La forma se ajusta más o menos a la esperada para una gráfica de este tipo. La asimetría es una característica propia de las distribuciones exponenciales y aquí se puede observar una clara asimetría positiva, así como varios valores atípicos en la cola derecha. Otra vez nos encontramos sin presencia de multimodalidad debido al tipo de distribución de la que procede la muestra.

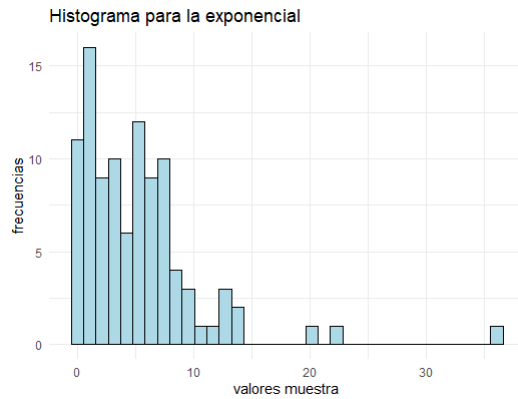


Figura 3.2: Histograma de una muestra de una distribución exponencial generada aleatoriamente.

En la Figura 3.3 tenemos la representación de un histograma de una distribución uniforme, en el que no hay presencia de valores atípicos, y en el que podemos observar simetría. Una de las características de la distribución uniforme es su multimodalidad, al ser todas las barras del histograma de una altura parecida.

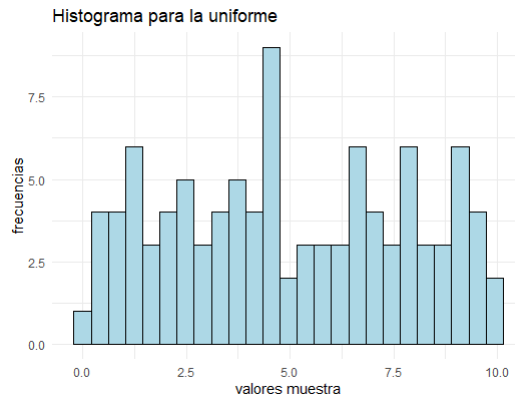


Figura 3.3: Histograma de una muestra de una distribución uniforme generada aleatoriamente.

En la Figura 3.4, correspondiente a la distribución doble exponencial, nuestro histograma presenta una simetría bastante clara. Aunque a priori se podría observar una forma acampanada en este histograma, la presencia de una gran cantidad de valores atípicos, nos permite inferir que no nos encontramos ante una distribución normal.

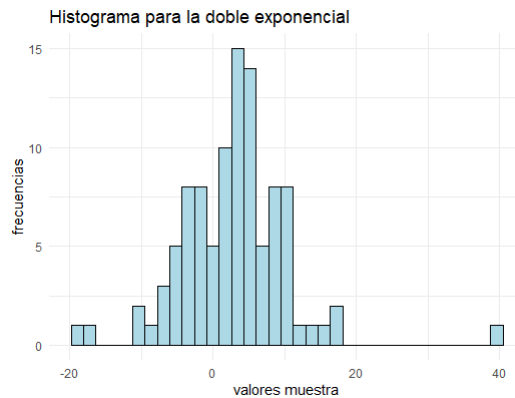


Figura 3.4: Histograma de una muestra de una distribución doble exponencial generada aleatoriamente.

En la Figura 3.5, correspondiente a la mixtura de normales, se puede ver la presencia de algún valor atípico. Observamos que la forma del histograma presenta cierta simetría, queriendo parecerse a una campana de Gauss. Esto se debe a la elección de una mixtura de gaussianas en el que todas las dos componentes presentan la misma media.

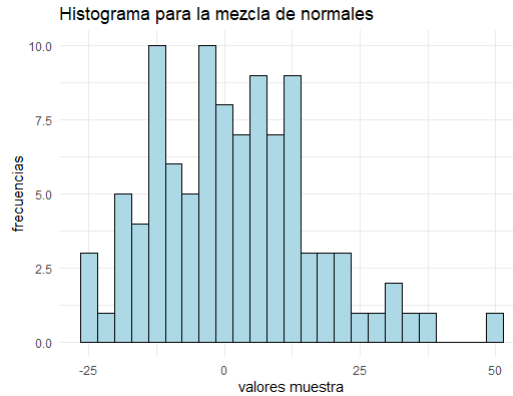


Figura 3.5: Histograma de una muestra de una distribución mezcla de normales generada aleatoriamente.

## 3.2. Estimación de la función de densidad

Las variables aleatorias continuas vienen determinadas por su función de densidad. En numerosas ocasiones cuando disponemos simplemente de una muestra, no vamos a conocer la función de densidad subyacente. Es aquí cuando entra el juego el papel de los estimadores de la función de densidad. A continuación vamos a tratar los estimadores de la función de densidad más comunes.

### 3.2.1. El histograma como estimador de la función de densidad y la poligonal de frecuencias

En la sección anterior hemos estado hablando del histograma, el gráfico de representación de distribuciones de frecuencias. El histograma se puede ver como un estimador de la función de densidad de una variable continua, que consiste en separar en clases a las que denotaremos  $C_j$  la recta real, y cada una de estas clases va a tener una cierta amplitud  $a_j$  que hará variar la forma que tome el histograma. Sobre cada una de estas clases se elevan las frecuencias relativas de las mismas, como ya hemos visto anteriormente. Es por tanto que podemos construir la función:

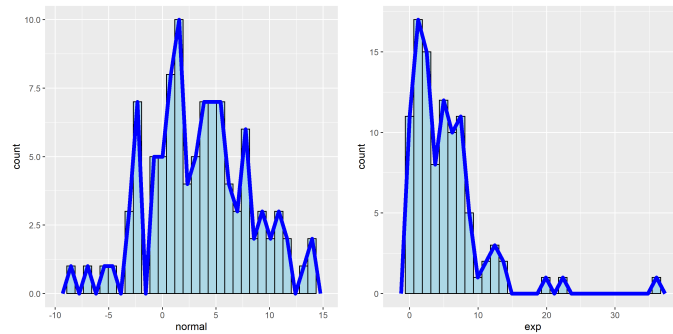
$$\widehat{f}_X(x) = \sum_{i=1}^n \frac{1}{na_j} I_{C_j}(X_i),$$

donde  $C_j$  es la clase tal que  $x \in C_j$ . Nótese que esta función  $\widehat{f}_X$  cumple las condiciones de la función de densidad.

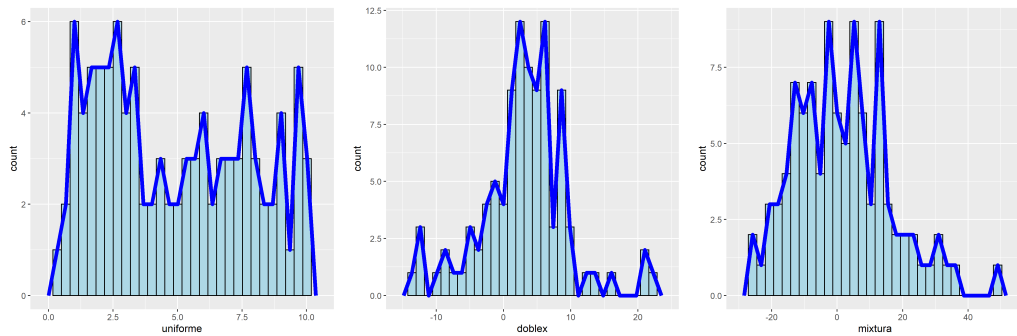
Este estimador presenta un pequeño problema y es que resulta en una función escalonada, por lo que no se ajusta muy bien a modelos conocidos. Una forma de solventar este problema sería por medio de otro estimador:

la **poligonal de frecuencias**, que consiste en unir los puntos medios de la parte superior de cada celda del histograma.

A continuación, vamos a mostrar las poligonales de frecuencias sobre los histogramas anteriores:



(a) Poligonal de frecuencias normal (b) Poligonal de frecuencias exponencial



(c) Poligonal de frecuencias uniforme (d) Poligonal de frecuencias doble exponencial (e) Poligonal de frecuencias mezcla

Figura 3.6: Poligonal de frecuencias sobre los histogramas de las distintas muestras consideradas.

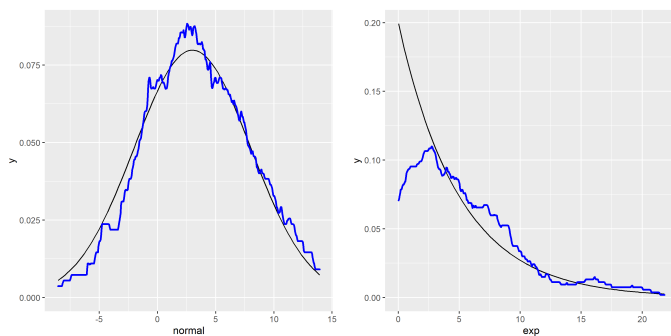
### 3.2.2. Estimador naïve o histograma móvil

Con el objetivo de evitar los problemas que da el histograma como estimador de la función de densidad, se crea la idea de histograma móvil, el cuál se basa en la idea de  $f_X = F'_X = \lim_{h \rightarrow 0} \frac{F_X(x+h) - F_X(x-h)}{2h}$ . Usando intervalos de la forma  $(x - h, x + h)$ , donde  $h$  se denomina *amplitud de la ventana* y está fijado, el histograma móvil viene dado por:

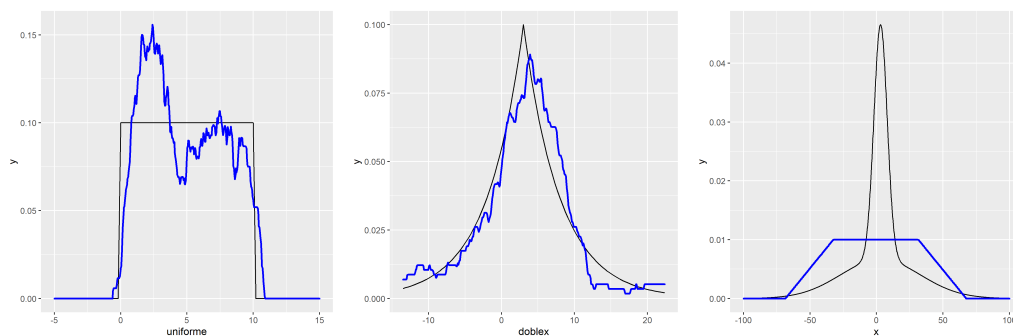
$$\widehat{f}_X(x) = \frac{1}{2nh} \sum_{i=1}^n I_{[-1,1]} \left( \frac{x - X_i}{h} \right).$$

La función definida de esta forma cumple las propiedades de la función de

densidad, aunque también tiene el inconveniente de ser una función definida a trozos. En la Figura 3.7 se ilustra el histograma móvil para las cinco muestras consideradas, superponiendo la función de densidad subyacente para permitir la comparación.



(a) Histograma móvil muestra normal (b) Histograma móvil muestra exponencial



(c) Histograma móvil muestra uniforme (d) Histograma móvil muestra doble exponencial (e) Histograma móvil mixtura

Figura 3.7: Histogramas móviles para las distintas muestras consideradas.

### 3.2.3. Estimador núcleo

En la definición del histograma móvil, podemos plantearnos que no tiene mucho sentido el dar a cada una de las observaciones  $x_i$  el mismo peso en la estimación de la función de densidad en un cierto punto  $x$ , sino que sería más lógico que las observaciones más cercanas al punto tengan más peso, así como que las observaciones que se alejan del punto tengan menos peso. Los estimadores tipo núcleo [20] tienen como misión dar a cada observación un peso lógico en función de lo alejada que esté del punto  $x$ .

Un estimador núcleo está asociado a una función núcleo  $K$ , que debe cumplir ser una función no negativa tal que:

1.  $\int_{-\infty}^{+\infty} K(z) dz = 1$ ;

2.  $\int_{-\infty}^{+\infty} zK(z) dz = 0;$
3.  $\int_{-\infty}^{+\infty} z^2K(z) dz = c < +\infty.$

Entonces, la estimación núcleo de la función de densidad basada en la función núcleo  $K$  será:

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right).$$

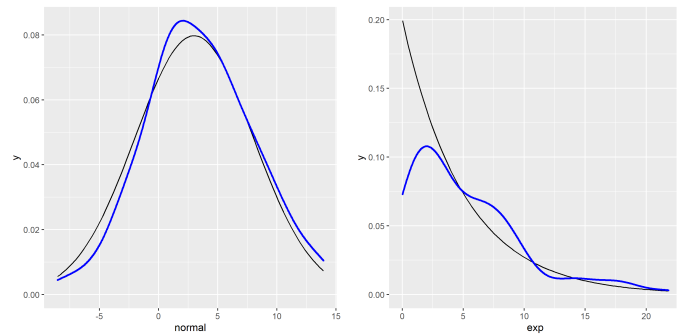
Algunas de las funciones núcleo más utilizadas son:

- Gaussiano:  $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- Triangular:  $(1 - |x|)\mathbb{I}(1 + |x|)$
- Epanechnikov:  $\frac{3}{4}(1 - x^2)\mathbb{I}(|x| - 1)$
- Uniforme:  $\frac{1}{2}\mathbb{I}(|x| < 1)$

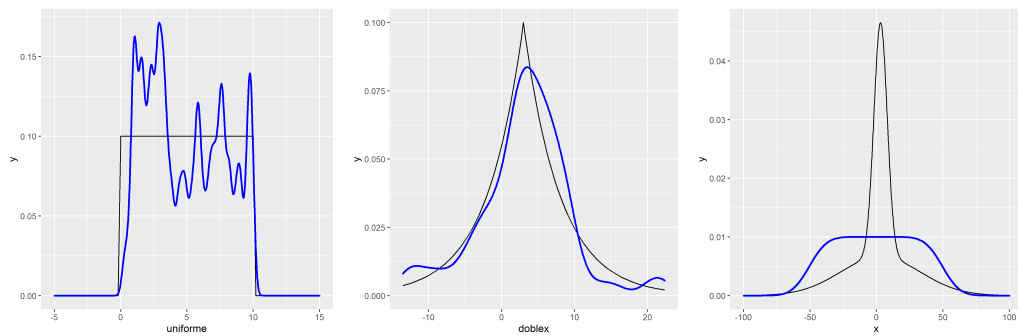
Nótese que el estimador de tipo naïve o histograma móvil es en realidad la estimación núcleo con función núcleo uniforme.

Vamos a ver a continuación en la Figura 3.8, ejemplos de estimación núcleo de la función de densidad para el caso del núcleo gaussiano a partir de las cinco muestras hemos generado.





(a) Histograma núcleo muestra normal (b) Histograma núcleo muestra exponencial



(c) Histograma núcleo muestra uniforme (d) Histograma núcleo muestra doble exponencial (e) Histograma núcleo muestra mixtura

Figura 3.8: Estimación de la función de densidad con núcleo gaussiano para las distintas muestras consideradas.

### 3.3. Diagrama de cajas

Este tipo de representación gráfica, introducido por Tukey [32, 31], nos da menos información que las anteriores, sin embargo sigue siendo muy útil para estudiar características como la asimetría o dispersión de los datos, así como para la identificación de los outliers. Estos gráficos son de especial utilidad cuando se quieren comparar distintos grupos de muestras sobre una misma gráfica. Los diagramas de cajas tienen efectivamente una forma cuadrada o rectangular, como su propio nombre indica. Un diagrama de caja tiene como elementos los siguientes:

- **Primer y tercer cuartiles:** el primero viene dado por la base de la caja y el tercero viene dado por la tapa de la caja.
- **La mediana,** segundo cuartil o valor central del conjunto de datos, que viene dada por la línea horizontal que se encuentra dentro de la

caja, y que nos da información acerca de la simetría o asimetría, según dónde esté situada.

- **El rango intercuartílico**, que es la diferencia entre el tercer y primer cuartiles y viene representado gráficamente por la altura de la caja.
- **Los ‘bigotes’** o líneas que salen por encima y por debajo de la caja, que se extienden hasta el menor y el mayor valor, no considerado atípico. Los llamados valores atípicos son los que están por encima o por debajo de los límites de los ‘bigotes’ y se representan con puntos. La definición más común de valor atípico es un punto que dista del primer/tercer cuartil más de 1.5 veces el rango intercuartílico.
- **El mínimo**, que es el dato más pequeño de la muestra. Este valor será el extremo inferior del bigote, a menos que existan datos atípicos, en cuyo caso será el valor atípico representado más abajo.
- **El máximo**, que es el dato más grande de la muestra. Este valor será el extremo superior del bigote, a menos que existan datos atípicos, en cuyo caso será el valor atípico representado más arriba.

En la Figura 3.9 se representan ejemplos de diagramas de cajas provenientes de las distintas muestras. Explicamos a continuación lo que se observa en cada una de las cajas, de izquierda a derecha.

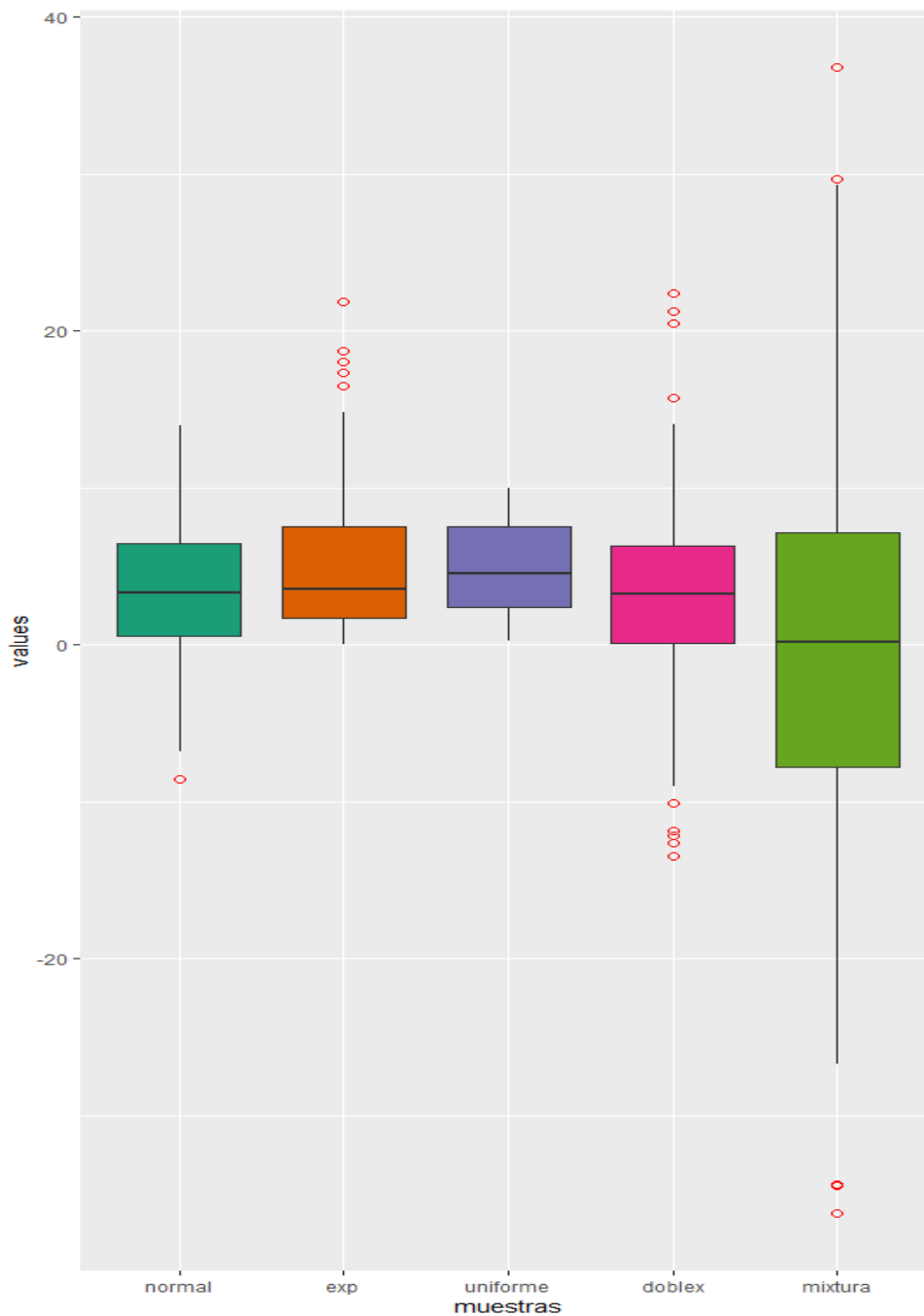


Figura 3.9: Diagrama de cajas para nuestras muestras.

La primera caja que aparece en el gráfico se corresponde con un ejemplo prototípico de **distribución normal**. Lo que esperamos de la misma es que sea simétrica, es decir que la mediana esté en el centro de la caja y coincida

su valor con el de la media de la muestra. También se espera que los bigotes de la caja sean simétricos y no haya apenas presencia de valores atípicos. En nuestro caso tenemos que la mediana está prácticamente en el centro exacto de la caja, con lo que se comprueba una cierta presencia de simetría. Con respecto a los bigotes, podemos observar también que son simétricos, lo que cabe esperar de una normal, como hemos indicado previamente. Observamos la presencia de un único outlier representado en el gráfico por medio de un círculo rojo. Para la muestra que estamos utilizando podemos comparar los valores poblacionales y muestrales de los siguientes parámetros en la Tabla 3.2:

	Media	SD	Q1	Q2	Q3	IQR
Muestral	3.452	4.564	0.528	3.310	6.457	5.929
Poblacional	3.000	5.000	-0.372	3.000	6.372	6.745

Tabla 3.2: Tabla comparación de parámetros muestrales y poblacionales para la normal.

Pasamos ahora a fijarnos en la segunda caja, que se corresponde con **una distribución exponencial**: con respecto a la mediana o Q2, observamos que no se encuentra para nada en el centro, al contrario que en el caso anterior, lo que indica asimetría. Los bigotes en esta ocasión tampoco son simétricos, podemos ver como el bigote inferior es mucho más corto que el superior. Con respecto a los valores atípicos, en esta ocasión sí tenemos varios. Estas tres cosas que hemos descrito son todo lo contrario a lo que cabría esperar de un diagrama de cajas de una distribución normal. El mínimo de la muestra se corresponde con el valor 0 (lo que tiene sentido ya que las exponenciales nunca toman valores negativos), y el máximo viene dado por uno de los outliers. Vamos a fijarnos ahora en los valores de los cuartiles y rango intercuartílico, tanto muestrales como poblacionales por medio de la Tabla 3.3. Cabe destacar que existen modificaciones del diagrama de caja para representar distribuciones claramente asimétricas [14].

	Q1	Q2	Q3	IQR
Muestral	1.570	4.240	7.348	5.778
Poblacional	1.438	3.465	6.931	5.493

Tabla 3.3: Tabla comparación de cuartiles e IQR muestral y poblacional para la exponencial.

La tercera caja correspondería a **la distribución uniforme**: podemos ver como la mediana de la caja se encuentra bastante centrada, con lo que habrá simetría. Los bigotes también se presentan simétricos aunque más cor-

tos que en el caso de la normal, y no hay presencia de valores atípicos. Con estas características, nos recuerda a la caja de una distribución normal prototípica. El valor mínimo de la muestra está en el punto 0 y el máximo en el punto 10 (estos valores son los que caracterizan las distribuciones uniformes, con lo que al provenir esta de una con esos parámetros tiene sentido que el mínimo y máximo se alcancen en dichos puntos). Vamos a fijarnos ahora en los cuartiles y rango intercuartílico de esta distribución, comparándolos de nuevo por medio de la Tabla 3.4.

	Q1	Q2	Q3	IQR
Muestral	2.455	4.667	7.550	5.095
Poblacional	2.500	5.000	7.500	5.000

Tabla 3.4: Tabla comparación de cuartiles e IQR muestral y poblacional para la uniforme.

Con respecto a **la doble exponencial o Laplace** podemos observar que la caja presenta una cierta simetría debido a la posición de la mediana en la misma. Con respecto a los bigotes no son completamente simétricos, podemos observar cómo el superior es ligeramente más corto. Hay presencia de muchos datos atípicos, característica contraria a la de la caja de una distribución normal. De hecho, podemos observar que tanto el mínimo como el máximo de la muestra vienen dados por datos outlier. Con respecto a los cuartiles y el rango intercuartílico, vienen recogidos en la Tabla 3.5. Podemos ver una gran diferencia en el Q1 y el rango intercuartílico muestrales y poblacionales.

	Q1	Q2	Q3	IQR
Muestral	-1.862	3.185	6.470	8.332
Poblacional	-0.465	3.000	6.466	6.931

Tabla 3.5: Tabla comparación de cuartiles e IQR muestral y poblacional para la doble exponencial.

Por último, con respecto a la caja correspondiente a **la mixtura de dos normales**, para empezar se ve claramente que el tamaño de la misma es mucho mayor que el del resto, esto se deberá a que su IQR toma un valor mayor, habrá más diferencia entre los valores que toman los cuartiles. Observamos puntos outlier a ambos extremos de los bigotes, aunque no muchos como en los casos de la exponencial o de la doble exponencial, los cuales nos proporcionarán los datos del mínimo y máximo de la muestra. Con respecto a la mediana, no se encuentra exactamente en el centro de la caja pero sí que nos permite afirmar una cierta simetría en los datos. Los

bigotes son simétricos al igual que en la distribución normal. Con respecto a los cuartiles y el rango intercuartílico muestrales, se pueden consultar las Tablas 3.6 y 3.7.

	Q1	Q2	Q3	IQR
Muestral	-9.255	0.070	10.448	19.703

Tabla 3.6: Tabla donde se recogen los cuartiles muestrales de la mixtura gaussiana.

	Media	SD
Muestral	1.613	14.571
Poblacional	3.000	21.500

Tabla 3.7: Tabla comparación de parámetros muestrales y poblacionales usando 2.4.5 para la mixtura

### 3.4. Diagrama de violín

Al igual que los diagramas de caja, los diagramas de violín [12] sirven para visualizar la distribución por grupos de variables numéricas. La mayor diferencia con estos últimos es en la forma de la representación, ya que los diagramas de cajas representan cajas (tal y como su nombre indica), mientras que en el caso de los diagramas de violín las formas varían para representar una estimación de la función de densidad. Estos diagramas muestran más información que un diagrama de cajas convencional, ya que ilustran toda la distribución de los datos y no solamente valores como los cuartiles o la mediana. A continuación en la Figura 3.10, podemos ver cómo son los diagramas de violín para nuestras muestras de referencia anteriores (con los anteriores diagramas de caja representados en el interior).

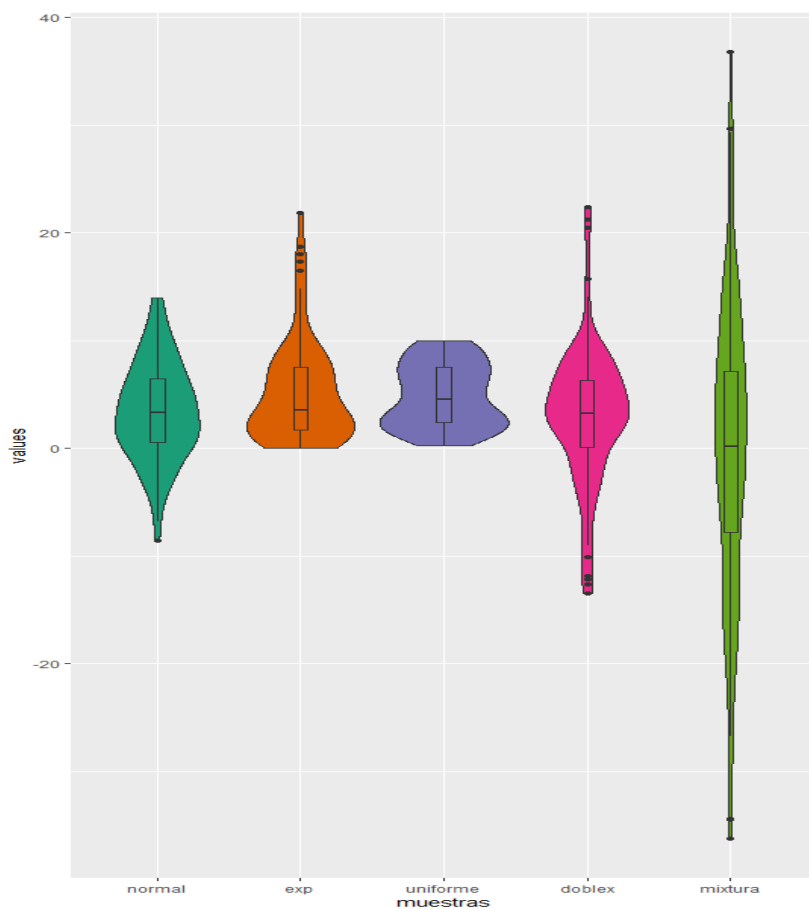


Figura 3.10: Diagrama de violín para nuestras muestras.

Comparando ambos tipos de gráfico, podemos llegar a la conclusión de que los diagramas de violín son más informativos que los diagramas de caja, debido a que muestran la distribución completa de nuestro conjunto de datos y no solamente un resumen estadístico de los mismos como hacen los diagramas de cajas. Esto es uno de los motivos por los que cada vez se utilizan más los primeros.

### 3.5. QQ-plots

Los QQ-plots [35] son un método gráfico que consiste en la comparación de una distribución empírica, la de los datos que tenemos, con una distribución teórica, en nuestro caso una normal que va a ser a la que queremos comprobar si se ajusta o no nuestra muestra. Esto lo haremos por medio de la comparación de cuantiles frente a cuantiles, de donde viene el nombre de estos gráficos.

Para construir este gráfico, en primer lugar tomamos nuestra muestra y la ordenamos de forma creciente, para así obtener los estadísticos de orden de la muestra:  $x_{(i)}$ . Ahora necesitamos identificar cuáles son los cuantiles teóricos [3] y eso lo haremos a partir de los valores  $p_i = \frac{i-0.5}{n}$  y de la inversa de la función de distribución. Nótese que, aunque en la propuesta original se consieraron los valores  $p_i = \frac{i}{n}$ , esta opción supone un problema a la hora de trabajar con ciertas distribuciones como la normal o la exponencial donde el correspondiente cuantil no es finito. Otras correcciones también son utilizadas como, por ejemplo:

- La corrección de Bloom [1], que define  $p_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$ ;
- La corrección de Tukey [30], dada por  $p_i = \frac{i - \frac{1}{3}}{n + \frac{1}{3}}$ ;
- La corrección dada por  $p_i = \frac{i}{n+i}$ .

En concreto, nuestros cuantiles teóricos vienen dados por  $F^{-1}(p_i)$ . Para realizar el gráfico QQ-plot, representamos el par  $(F^{-1}(p_i), x_{(i)})$ . En caso de que queramos estudiar la normalidad, se sustituye  $F^{-1}$  por la función cuantil de una distribución normal con los parámetros estimados por la media y desviación típica muestrales.

La idea de este método gráfico se basa en la linealidad con respecto a la bisectriz del primer cuadrante, es decir que, una vez representados todos los pares, el dibujo que estos crean debe ser una línea superpuesta a dicha bisectriz o que se asemeje mucho a ella. Por el contrario, cuando hay desviaciones de la linealidad, nos encontramos ante un claro indicio de que la distribución teórica no se corresponde con la observada. En el caso del estudio de la normalidad, significaría que la muestra con la que trabajamos no parece provenir de una distribución normal. A continuación, veremos algunos ejemplos de QQ-plots para las cinco distribuciones que hemos generado de forma aleatoria.

En la Figura 3.11 podemos observar que la linealidad se mantiene perfectamente, lo que nos lleva a concluir que la muestra bien podría provenir de una distribución normal. Al mantenerse la linealidad, podemos también concluir que no hay presencia de asimetría.



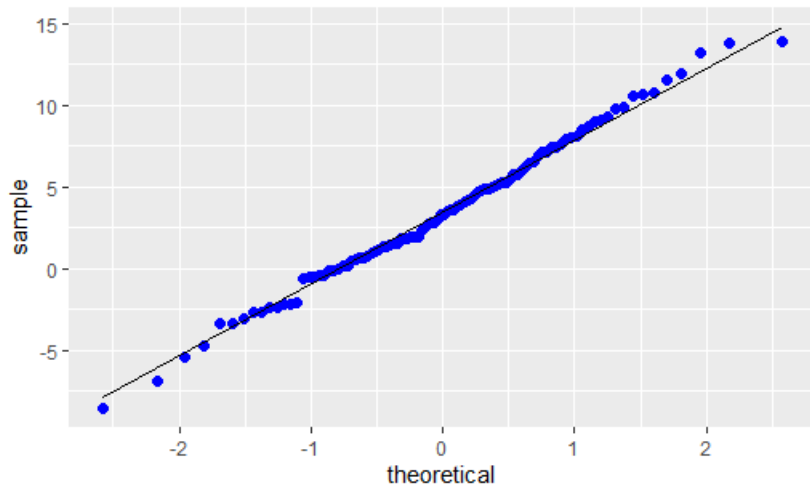


Figura 3.11: QQ-plot para la muestra normal considerada.

En la Figura 3.12 la parte central mantiene la linealidad, sin embargo en ambos extremos se produce una desviación clara de la bisectriz del primer cuadrante. Podemos observar que es asimétrica hacia la izquierda debido a la forma de la gráfica, lo que tiene sentido al provenir de una exponencial, que de por sí es una distribución asimétrica.

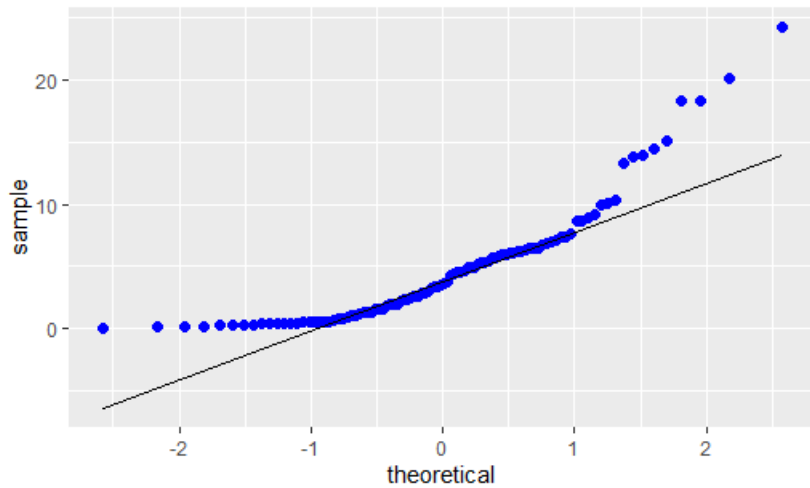


Figura 3.12: QQ-plot para la muestra exponencial considerada.

En el caso de la Figura 3.13 se observa una cierta forma de onda. La linealidad se mantiene en la zona central, pero a medida que nos vamos alejando comienzan las desviaciones: en la parte izquierda de la gráfica los datos se encuentran por encima de la línea de referencia y en la parte derecha por debajo. Debido a esta desviación de la linealidad, es razonable concluir que

no nos encontramos ante una muestra normal. Con respecto a la simetría: ambas colas se desvían de la misma forma con lo que hay simetría, lo que tiene sentido al provenir de una distribución uniforme, que es simétrica.

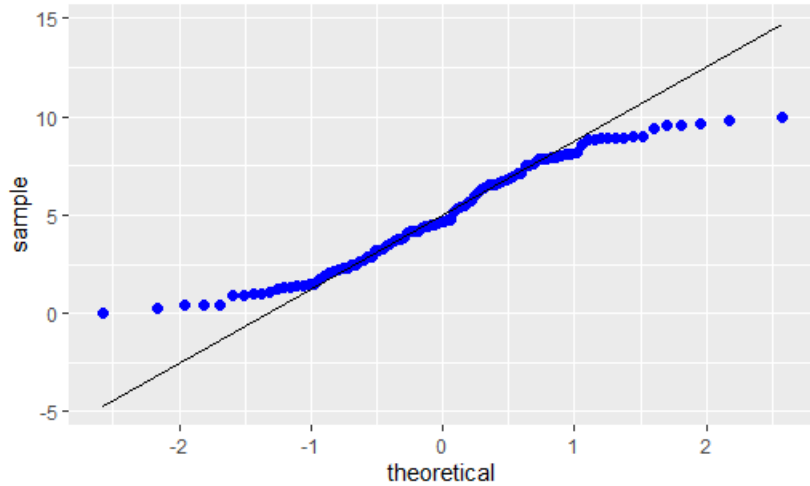


Figura 3.13: QQ-plot para la muestra uniforme considerada.

En la Figura 3.14 claramente hay presencia de valores atípicos en ambos extremos del gráfico. Observamos una cierta linealidad lo cuál podría llevarnos a cometer el error de asumir la normalidad cuando en realidad los datos no provienen de una distribución de ese tipo.

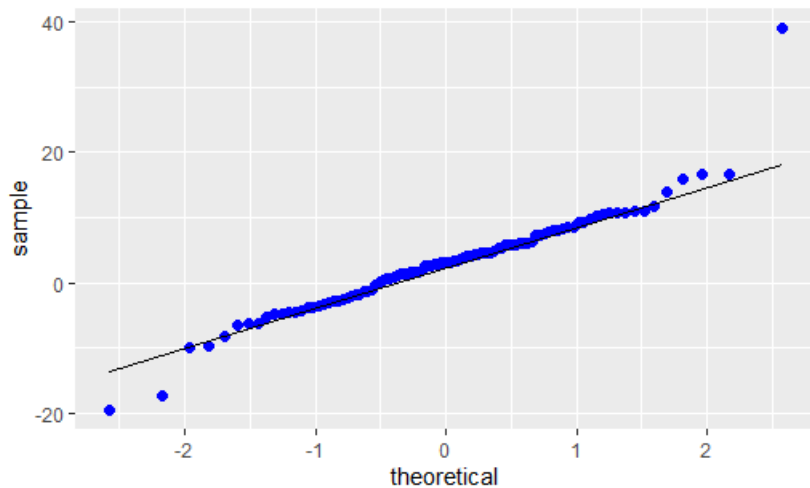


Figura 3.14: QQ-plot para la muestra doble exponencial considerada.

Por último, en la Figura 3.15 podemos ver cómo los datos mantienen en gran medida la linealidad, aunque se observa también la presencia de varios

outliers. Con respecto a la simetría, no hay desviación de datos en ninguna parte del gráfico por tanto en esta ocasión sí hay simetría.

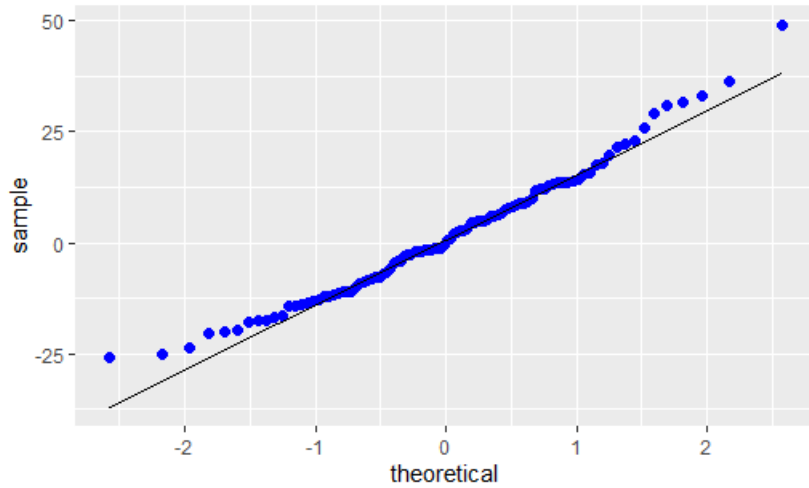


Figura 3.15: QQ-plot para la muestra mixtura de normales considerada.

### 3.6. PP-plots

Los PP-plots son gráficos análogos a los anteriores, pero en este caso se representan las probabilidades teóricas frente a las probabilidades muestrales, haciendo uso de las funciones de distribución acumulada teórica y empírica. Para los PP-plots [35], los pares que representamos ya no son cuantiles, sino probabilidades, y vienen dados por:  $(p_i, F(x_{(i)}))$ . De nuevo, el método se basa en estudiar si los pares que representamos mantienen o no la linealidad con respecto a la bisectriz del primer cuadrante. Su uso es menos habitual que el de los QQ-plots. Veamos también los ejemplos gráficos para las cinco muestras que hemos generado de manera aleatoria.

En la Figura 3.16 la linealidad se mantiene de forma perfecta, mostrando un ejemplo prototípico de PP-plot para una distribución normal.

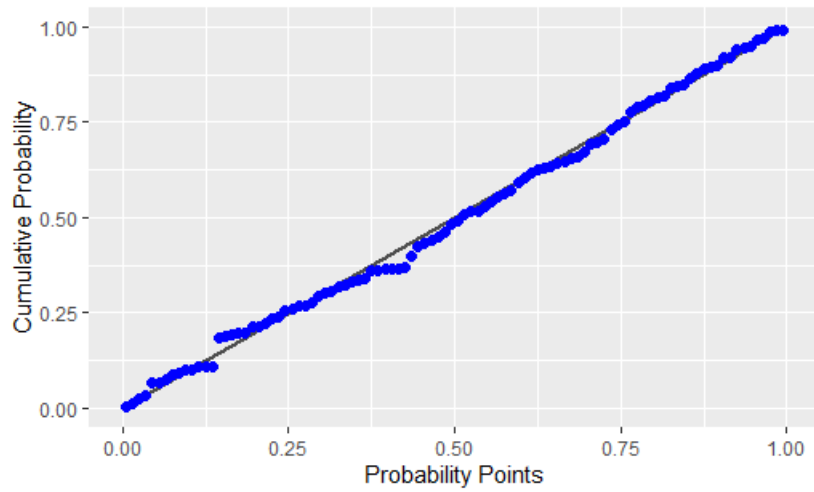


Figura 3.16: PP-plot para la muestra normal considerada.

En la Figura 3.17 podemos observar que la linealidad no se mantiene en ningún momento, sino que la mayor parte de los pares representados se desvían por debajo de la línea de referencia formando una curva cóncava hacia arriba, y cuando llegamos a los extremos de la gráfica comienzan a desviarse por encima de la línea de referencia. Con respecto a valores atípicos en este gráfico no podemos distinguirlos.

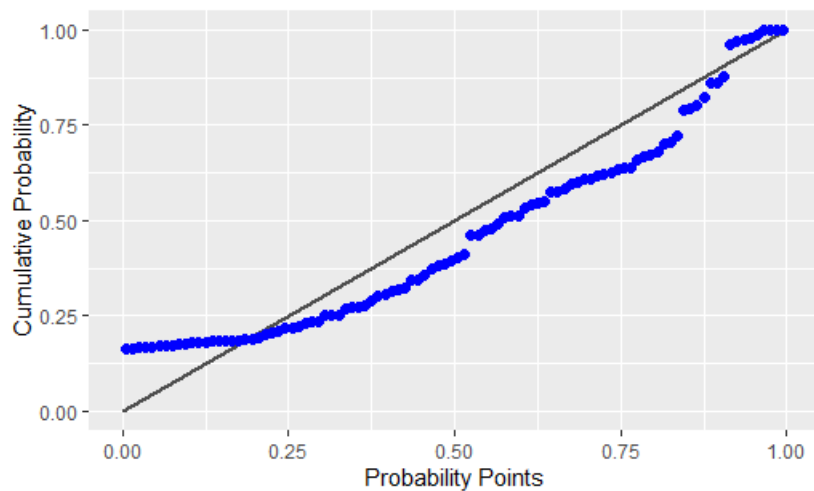


Figura 3.17: PP-plot para la muestra exponencial considerada.

En la Figura 3.18 de nuevo podemos observar que la linealidad no se conserva, en la mitad izquierda los puntos representados están por debajo de la línea de referencia y en la mitad derecha están por arriba, observándose un comportamiento simétrico.

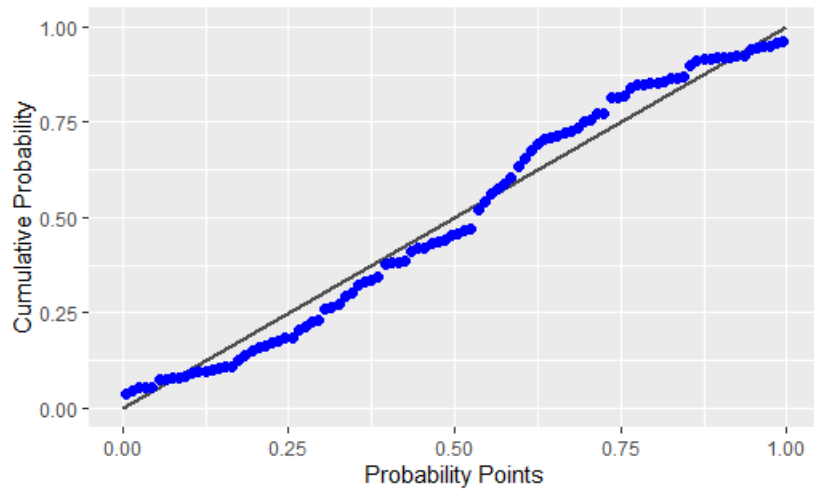


Figura 3.18: PP-plot para la muestra uniforme considerada.

En la Figura 3.19 también se observaría un comportamiento similar, encontrándonos en la mitad izquierda con los puntos representados por encima de la línea de referencia y en la mitad derecha por abajo, observándose también un comportamiento simétrico. La diferencia entre esta gráfica y en la anterior, donde el rol de la mitad izquierda/derecha se intercambian, es debido a la diferencia en curtosis entre ambas distribuciones.

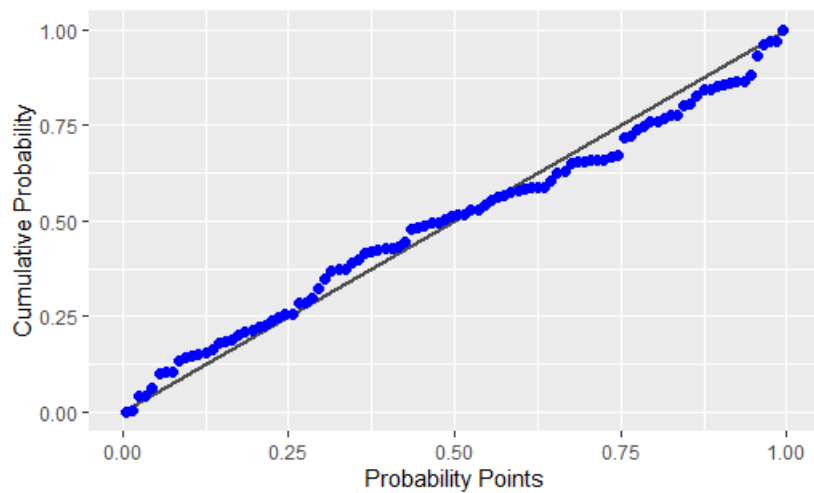


Figura 3.19: PP-plot para la muestra doble exponencial considerada.

Por último en la Figura 3.20 podemos ver que se mantiene la linealidad, con este gráfico no podemos apreciar los outliers pero sí vemos que los valores representados se comportan de forma simétrica.

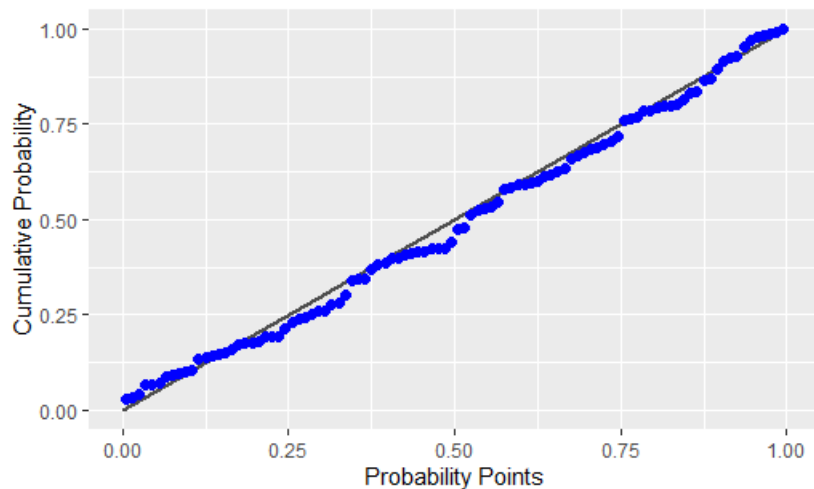


Figura 3.20: PP-plot para la muestra mixtura de normales considerada.

### 3.7. Gráficos de la función de distribución empírica acumulada

En este apartado vamos a dibujar las funciones de distribución empírica acumulada de cada una de nuestras muestras junto con las curvas de aceptación-rechazo de Lilliefors [15]. Como se verá más adelante, estas curvas nos van a permitir saber si rechazamos o no la hipótesis nula de normalidad cuando estemos realizando el test de Lilliefors de normalidad, comprobando si la función de distribución empírica cae dentro o fuera de las mismas. Estas curvas se construyen a partir del valor  $L_n$  que se obtiene por medio de la tabla de Lilliefors para los valores de tamaño de muestra y nivel de significación correspondientes (en este caso  $n = 100$  y  $\alpha = 0.05$ ) y de los datos estandarizados de la muestra.

En la Figura 3.21 podemos ver cómo la función de distribución empírica asociada a la muestra normal se encuentra dentro de los límites dados por las curvas de Lilliefors, en cambio en la Figura 3.22 (subfigura superior izquierda) se observa como claramente nos encontramos fuera de estos límites para la muestra exponencial. En las otras tres subfiguras de la Figura 3.22 vemos como nos encontramos también dentro de los límites marcados por Lilliefors, debido a que el test de Lilliefors es en general un test poco potente para detectar desviaciones de la normalidad en estos tres tipos de distribuciones.

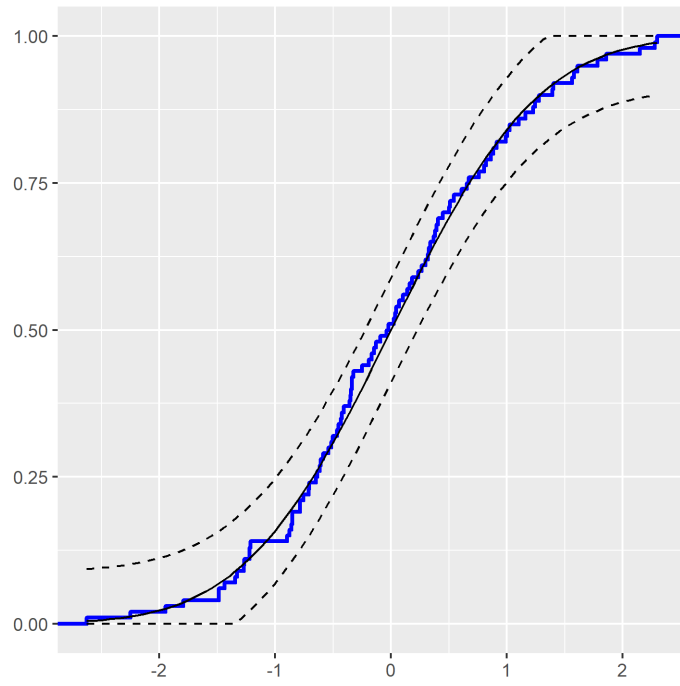


Figura 3.21: Curvas de aceptación-rechazo de Lilliefors para la muestra normal.

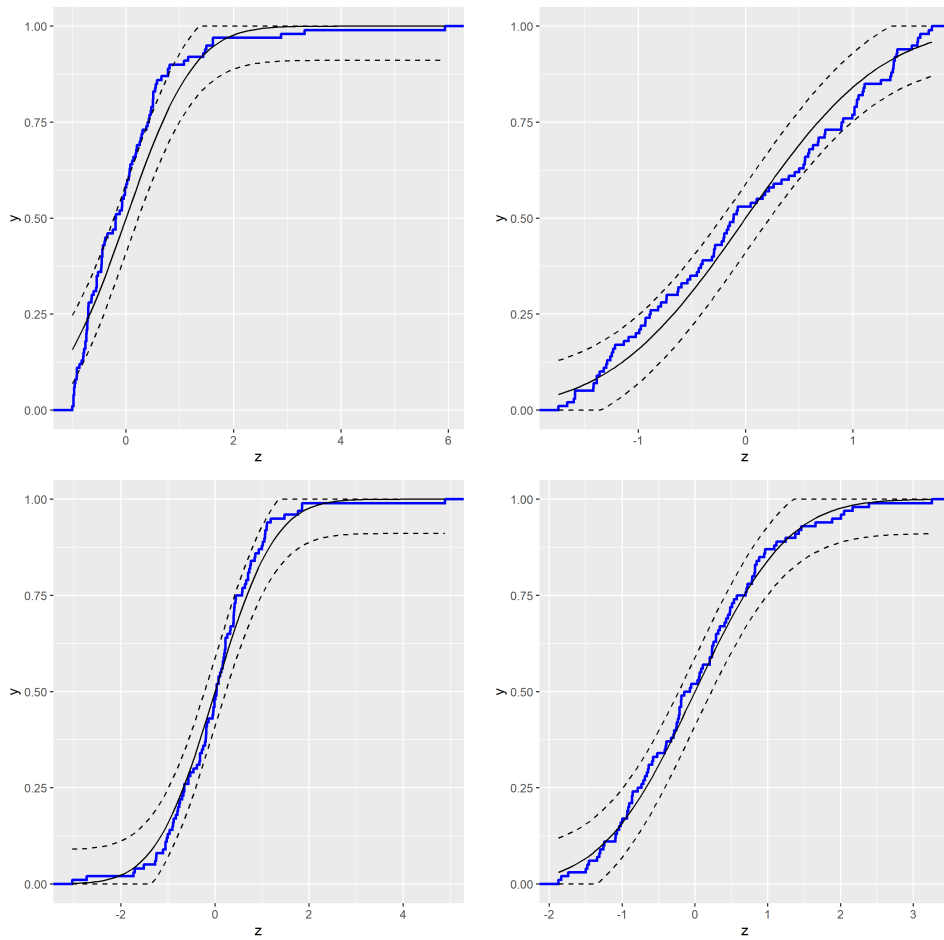


Figura 3.22: Curvas de aceptación-rechazo de Lilliefors para las cuatro muestras que no son normales.



## Capítulo 4

# Tests de normalidad

Los test de normalidad tienen como objetivo comprobar si un cierto conjunto de datos dado se ajusta o no a una distribución normal (con parámetros desconocidos). Vamos a trabajar con una muestra aleatoria simple  $(X_1, \dots, X_n)$  de una variable aleatoria continua  $X$ , es decir un conjunto de variables independientes, aleatorias e idénticamente distribuidas obtenidas a partir de dicha variable aleatoria  $X$ .

Las hipótesis nula y alternativa para cualquiera de los test de normalidad que se describen en esta sección serán las siguientes:

$$\begin{cases} H_0 = \text{la muestra proviene de una distribución normal} \\ H_1 = \text{la muestra no proviene de una distribución normal} \end{cases}$$

Nótese que estamos ante una hipótesis nula compuesta, pues ambos parámetros de la normal son desconocidos.

A continuación, vamos a describir los test de normalidad más utilizados y conocidos, así como algunos que no son tan conocidos. También veremos modificaciones de dichos test con el objetivo de que sean más eficientes o más sencillos de utilizar.

### 4.1. Test de Lilliefors

Para poder hablar del test de normalidad de Lilliefors en primer lugar vamos a explicar el test de bondad de ajuste de Kolmogorov-Smirnov para hipótesis nulas simples. La hipótesis nula en este caso es si la muestra proviene o no de una distribución normal de parámetros completamente especificados. La idea del test consiste en la comparación de la función de distribución empírica de la muestra con la función de distribución teórica a la que queremos ver si se ajusta. Si ambas coinciden o se encuentran bastante próximas no podremos rechazar la hipótesis nula con la que trabajamos, sin embargo en caso de que difieran en gran medida supondrá un rechazo de la

hipótesis nula. La comparación se hace por medio del siguiente estadístico:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|,$$

donde  $F_0$  es la función de distribución de referencia y  $F_n$  es la función de distribución empírica de la muestra. Bajo la hipótesis nula de que la muestra proviene de la distribución  $F_0$  totalmente especificada, este estadístico es de libre distribución, lo que quiere decir que tiene la misma distribución independientemente de la distribución de referencia considerada bajo la hipótesis nula. La distribución del estadístico bajo la hipótesis nula está tabulada [17, 27]. La región crítica vendrá dada por valores grandes del estadístico  $D_n$ , es decir,  $RC = \{D_n > c\}$ , para un cierto  $c$  dependiente del tamaño muestral y del nivel de significación. Nótese que este test se centra mayoritariamente en la parte central de la distribución, haciendo que, si existen diferencias en la parte de las colas, éstas sean difíciles de detectar.

Ahora ya podemos explicar el que es uno de los test de normalidad más usados: **el test de Lilliefors** [19], el cuál es una modificación del test de Kolmogorov-Smirnov para el caso de hipótesis nulas compuestas (en la que los parámetros son estimados, y generalmente esta estimación viene dada por la estimación de los parámetros máximo-verosímil a partir de la muestra). La idea con la que se trabaja es exactamente la misma: la comparación entre las funciones de distribución empírica y teórica para una muestra aleatoria simple de una variable continua  $X$ . Las hipótesis nula y alternativa son las explicadas al principio del capítulo.

El estadístico del contraste tiene la siguiente forma:

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - \hat{F}_{N(\bar{x}, s)}(x)|$$

donde se mide la diferencia máxima entre la función de distribución empírica  $F_n$  y la función de distribución teórica de una normal con los parámetros estimados  $\hat{F}_{N(\bar{x}, s)}$ . Es por ello que se rechaza la hipótesis nula de normalidad en caso de que valores del estadístico sean suficientemente grandes. Es decir, se define una región crítica tal que  $RC = \{D_n > c\}$ . Este estadístico también es de libre distribución bajo la hipótesis nula, significando en este caso que la distribución del estadístico no depende de los parámetros de la normal.

En la Tabla 4.1 recogemos los valores del estadístico de Lilliefors y los p-valores correspondientes para las cinco muestras de referencia utilizadas en este trabajo. Se puede observar que el test de Lilliefors a un nivel de significación  $\alpha = 0.05$  solamente rechaza la normalidad para la muestra procedente de una distribución exponencial. Como se verá en la experimentación del trabajo, esto se debe a que el test de Lilliefors es un test por norma general poco potente. En RStudio podemos implementarlo por medio de la función `lillie.test()` del paquete `nortest`.

	Valor del estadístico	p-valor
Normal	0.058	0.5558
Exponencial	0.158	1.863e-06
Uniforme	0.084	0.0784
Doble exp.	0.079	0.1315
Mixtura	0.064	0.4038

Tabla 4.1: Tabla donde se recogen los valores del estadístico y p-valores de muestras muestras para el test de Lilliefors.

Con respecto a los límites de la región crítica, tenemos, para tamaño de muestra  $n = 100$  y en función de nuestro nivel de significación, los valores recogidos en la Tabla 4.2. Los podemos encontrar tabulados [26] o calcularlos por simulación.

$\alpha$	0.01	0.05	0.1
c	0.1031	0.0886	0.0805

Tabla 4.2: Tabla donde se recogen los límites de la región crítica del test de Lilliefors para los distintos niveles de significación y tamaño de muestra 100.

## 4.2. Test de Cramer Von Mises

Este test [6], [34] es un contraste no paramétrico de bondad de ajuste para variables continuas unidimensionales, en el que de nuevo se utiliza la idea de la comparación entre las funciones de distribución empírica y teórica. Es una alternativa más potente al test de Kolmogorov-Smirnov. En particular, ahora ya no nos vamos a fijar en un único punto como en el test anterior, sino que tendremos en cuenta la totalidad de los puntos.

El estadístico de este test viene dado por:

$$W^2 = \int_{-\infty}^{\infty} [F_0(x) - F_n(x)]^2 \cdot dF_0(x)$$

donde  $F_0$  representa la distribución de una normal con los parámetros estimados por la media y la desviación típica muestrales. Al igual que pasaba con los tests de Kolmogorov-Smirnov y su variante para hipótesis nulas compuesta (test de Lilliefors), la distribución del estadístico será distinta para el caso de la hipótesis nula simple y la hipótesis nula compuesta. En esta subsección nos centramos exclusivamente en el segundo caso. Es fácil observar que efectivamente se basa en una diferencia entre estas dos funciones,

pero en esta ocasión es una diferencia ponderada. De manera análoga, el estadístico se puede expresar también de la siguiente manera:

$$T = n \cdot W^2 = \frac{1}{12n} + \sum_{i=1}^n \left[ \frac{2i-1}{2n} - F_0(x_{(i)}) \right]^2$$

donde  $x_{(i)}$  son los datos de la muestra en orden creciente. Esta forma de expresar el estadístico permite ver que éste es también un estadístico de libre distribución. Las hipótesis con las que trabajaremos son las explicadas al principio del capítulo, donde los parámetros son desconocidos pero los estimamos a partir de la muestra. Definimos la región crítica por los valores grandes de nuestro estadístico:  $RC = \{T > c\}$ . De nuevo con este criterio es más sencillo encontrar discrepancias en la parte central de la distribución que en la parte de las colas.

En la Tabla 4.3 se recogen los p-valores y valores del estadístico de CVM para las muestras con las que trabajamos. Tenemos que tener cuidado, ya que RStudio, en la implementación de `cvm.test()` del paquete `nortest`, nos proporciona el valor del estadístico  $nW^2$  a pesar de que indique solamente  $W$ . En este caso podemos observar cómo el test de CVM rechaza la normalidad a un nivel de significación  $\alpha = 0.05$  para todas las muestras a excepción de las que provienen de la normal y de la mixtura gaussiana.

	Valor del estadístico $T=nW^2$	Valor del estadístico $W^2$	p-valor
Normal	0.029	0.00029	0.8628
Exponencial	0.670	0.0067	9.831e-08
Uniforme	0.180	0.0018	0.0093
Doble exp.	0.134	0.00134	0.0382
Mixtura	0.056	0.00056	0.4266

Tabla 4.3: Tabla donde se recogen los valores del estadístico y p-valores de muestras muestras para el test de Cramer Von Mises.

En la Tabla 4.4 se muestran los límites de la región crítica para tamaño de muestra  $n = 100$  y distintos niveles de significación[5].

$\alpha$	0.01	0.05	0.1
c	0.1777	0.1256	0.1032

Tabla 4.4: Tabla donde se recogen los límites de la región crítica del test de CVM para los distintos niveles de significación y tamaño de muestra 100.

### 4.3. Test de Anderson Darling

El test de Anderson Darling [18], [28] es otro de los test utilizados para ver si una muestra proviene o no de una cierta distribución dada (que, en nuestro caso, será una distribución normal con parámetros estimados).

El estadístico de este test viene dado por:

$$A^2 = n \int_{-\infty}^{\infty} [F_0(x) - F_n(x)]^2 \cdot w(x) \cdot dF_0(x)$$

donde  $F_0$  representa la distribución de una normal con los parámetros estimados por la media y la desviación típica muestrales, y

$$w(x) = \frac{1}{F_0(x)(1 - F_0(x))}$$

Nótese que, al igual que pasaba con el test de Cramer-Von Mises, la distribución del estadístico será distinta para el caso de la hipótesis nula simple y la hipótesis nula compuesta. En esta subsección nos centramos, de nuevo, exclusivamente en el segundo caso.

Véase por otro lado que este estadístico tiene relación con el del test de Cramer-Von Mises, con la diferencia de que en el anterior la función  $w(x)$  tomaba el valor 1. En este caso, sustituyendo la expresión correspondiente de  $w(x)$  en la fórmula del estadístico se obtiene:

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_0(x) - F_n(x)]^2}{F_0(x)(1 - F_0(x))} dF_0(x)$$

Una forma más sencilla de expresarlo es:

$$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n} (\ln(F_0(x_{(i)}))) + \ln(1 - F_0(x_{(n-i+1)}))$$

De nuevo, la idea principal del estadístico se basa en estudiar las distancias entre función de distribución empírica y función de distribución teórica, utilizando todos los puntos de la muestra. La región crítica o valores para los que se rechaza la hipótesis nula viene dada por los valores grandes del estadístico:  $RC = \{A^2 > c\}$ .

En este caso se da mayor importancia a la zona de las colas, por lo que, si existe una diferencia entre las funciones en la parte central, va a ser más difícil de detectar, al contrario por ejemplo del test de Lilliefors.

En la Tabla 4.5 recogemos los datos de los p-valores y valores del estadístico para nuestras muestras. Al igual que pasaba con el test de Cramer-Von Mises, tenemos que tener en cuenta que RStudio nos proporciona el valor  $A^2$  del estadístico, a pesar de que lo denote como  $A$ . El test de AD para un nivel de significación de  $\alpha = 0.05$  de nuevo rechaza la hipótesis de

normalidad para todas las muestras a excepción de las que provienen de las distribuciones normal y mixtura gaussiana.

	Valor del estadístico $A^2$	p-valor
Normal	0.182	0.911
Exponencial	4.404	5.225e-11
Uniforme	1.269	0.002539
Doble exp.	1.002	0.01168
Mixtura	0.4321	0.299

Tabla 4.5: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Anderson Darling.

En la Tabla 4.6 se muestran los límites de la región crítica para tamaño de muestra  $n = 100$  y distintos niveles de significación para este test [7].

$\alpha$	0.01	0.05	0.1
c	1.0262	0.7453	0.6262

Tabla 4.6: Tabla donde se recogen los límites de la región crítica del test de Anderson Darling para los distintos niveles de significación y tamaño de muestra 100.

#### 4.4. Test de Shapiro-Wilk

Este es un test de normalidad [25] para hipótesis nulas compuestas, utilizado para ver si una muestra dada proviene de una población con distribución normal. En este caso tenemos la muestra  $(x_1, \dots, x_n)$  y la ordenamos obteniendo los estadísticos de orden, los cuales comparamos con los estadísticos de orden de una muestra aleatoria independiente e idénticamente distribuida de tamaño  $n$ , obtenida a partir de la distribución  $N(0, 1)$ . El estadístico por medio del cual realizamos dicha comparación es el siguiente:

$$W = \frac{(\sum_{i=1}^n a_i x_{(i)})^2}{ns^2}$$

donde:

- $x_{(i)}$  es el  $i$ -ésimo estadístico de orden de la muestra;
- $s^2$  es la desviación típica muestral;

- $a_i = (a_1, \dots, a_n) = \frac{m^t V^{-1}}{\|V^{-1}m\|} = \frac{V^{-1}m^t}{\sqrt{m^t V^{-1} V^{-1} m}}$ . Estos coeficientes se obtienen por medio de una simulación de Montecarlo o también podemos utilizar tablas donde se recogen los valores de los mismos.
- $m = (m_1, \dots, m_n)^t$  valores esperados de los estadísticos de orden de las muestras a partir de  $N(0, 1)$ .
- $V$  la matriz de covarianzas de los estadísticos de orden.

En este caso la región crítica se define para valores pequeños del estadístico:  $RC = \{W < c\}$ . Como es de esperar en un test de normalidad, el estadístico es de libre distribución bajo  $H_0$ .

A continuación en la Tabla 4.7 tenemos recogidos los valores del estadístico del test y los correspondientes p-valores para las muestras consideradas. El test de Shapiro-Wilk para un nivel de significación  $\alpha = 0.05$  rechaza la hipótesis de normalidad para las muestras provenientes de una exponencial, uniforme y doble exponencial. En RStudio se puede implementar por medio de `shapiro.test()` del paquete `stats`.

	Valor del estadístico	p-valor
Normal	0.994	0.9359
Exponencial	0.761	1.851e-11
Uniforme	0.952	0.0012
Doble exp.	0.928	3.772e-05
Mixtura	0.978	0.0975

Tabla 4.7: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Shapiro-Wilk.

El valor del límite de la región crítica se puede buscar en tablas, aunque al ser un test que se suele usar para tamaños de muestra menores a 50 las tablas solamente muestran los límites de la región crítica hasta ese tamaño, o también por simulación.

## 4.5. Test de Shapiro-Francia

El test de Shapiro-Francia [24] se basa en el anterior test, pero tiene una modificación que nos simplifica los cálculos. Hacemos de nuevo uso de los estadísticos de orden  $x_{(i)}$ , los valores esperados de los estadísticos de orden  $m$  y la matriz de covarianzas, pero ahora podremos prescindir del vector de coeficientes  $a_i$ , y añadimos el uso de los coeficientes de correlación de

Pearson. En este caso el estadístico del test viene dado por [33]:

$$W^* = \frac{\text{cov}(x, m)}{\sigma_x \sigma_m} = \frac{\sum_{i=1}^n (x_{(i)} - \bar{x})(m_i - \bar{m})}{\sqrt{(\sum_{i=1}^n (x_{(i)} - \bar{x})^2)(\sum_{i=1}^n (m_i - \bar{m})^2)}}$$

La región crítica viene de nuevo dada por los valores del estadístico pequeños:  $RC = \{W^* < c\}$ , esto se debe a que el estadístico viene dado por un cociente y cuanto más se asemejen los datos teóricos a los empíricos este se aproximará a 1.

En la Tabla 4.8 hemos recogido los valores del estadístico del test y los correspondientes p-valores para nuestras muestras. En este test, a un nivel de significación  $\alpha = 0.05$  ocurre lo mismo que en el anterior, rechazando la hipótesis de normalidad para todas las muestras a excepción de las provenientes de la normal y de la mixtura. En RStudio se puede implementar por medio de `sf.test()` del paquete `nortest`.

	Valor del estadístico	p-valor
Normal	0.995	0.9456
Exponencial	0.751	3.536e-10
Uniforme	0.960	0.0054
Doble exp.	0.915	2.658e-05
Mixtura	0.9785	0.0928

Tabla 4.8: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Shapiro Francia.

En este caso en RStudio se implementa por medio de `sf.test()` del paquete `nortest`.

## 4.6. Test de Jarque-Bera

Este test [16] estudia la normalidad de una variable haciendo uso del estudio de la simetría y de la curtosis. El estadístico de este test viene dado por:

$$JB = \frac{n}{6}(S^2 + \frac{1}{4}(K - 3)^2)$$

donde:

- $S$  es el coeficiente de asimetría para la muestra. En particular, es una medida que indica si la distribución de una variable es simétrica o no respecto a su media, donde valores positivos indican asimetría positiva, valores negativos indican asimetría negativa y las muestras simétricas



tendrán un valor para este coeficiente de cero. Formalmente, este coeficiente se define como:

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

- $K$  es el coeficiente de curtosis para la muestra. En particular, es una medida del apuntamiento/achatamiento, es decir, de la forma de la distribución. Formalmente, este coeficiente se define como:

$$K = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2\right)^2}$$

Asintóticamente, el estadístico  $JB$  sigue una distribución  $\chi_2^2$  bajo  $H_0$ . La región crítica viene dada por los valores del estadístico grandes (hay que tener en cuenta que este estadístico toma solo valores positivos y muestras normales tenderán a tomar valores del estadístico cercanos a cero, pues una distribución normal tiene asimetría igual a cero y curtosis igual a tres):  $RC = \{JB > \chi_{1-\alpha,2}^2\}$ .

Vamos a ver a continuación en la Tabla 4.9 los valores que toma el estadístico de Jarque-Bera para las muestras consideradas, así como los p-valores. Observamos como para una significación  $\alpha = 0.05$  las únicas muestras para las que no rechazamos la normalidad son las provenientes de la normal y de la mixtura. Con un nivel de significación  $\alpha = 0.01$  tampoco podríamos rechazar la normalidad de la muestra proveniente de una distribución uniforme. En RStudio tenemos la implementación `jb.norm.test()` del paquete `normtest`.

	Valor del estadístico	p-valor
Normal	0.1679	0.909
Exponencial	742.1	<2.2e-16
Uniforme	6.177	0.042
Doble exp.	122.88	<2.2e-16
Mixtura	5.2214	0.0505

Tabla 4.9: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Jarque-bera.

En la Tabla 4.10 recogemos los valores del límite de la región crítica para los distintos niveles de significación.

$\alpha$	0.01	0.05	0.1
$c$	9.21034	5.9915	4.6052

Tabla 4.10: Tabla donde se recogen los límites de la región crítica del test de Jarque Bera para los distintos niveles de significación y tamaño de muestra 100.

## 4.7. Test de Jarque-Bera robusto

Este test es una variante [11] (véase también [29]) del test de Jarque-Bera presentado en la sección anterior. En particular, se propone como una modificación del test de Jarque-Bera tradicional debido a que por la forma en la que está construido el anterior es muy sensible a la presencia de valores atípicos (outliers). La modificación consiste en que en este caso se utilizarán coeficientes de asimetría y curtosis más robustos, es decir, menos sensibles a outliers.

El estadístico en esta ocasión viene dado por :

$$RJB = \frac{n}{C_1} \left( \frac{m_3}{J_n^3} \right)^2 + \frac{n}{C_2} \left( \frac{m_4}{J_n^4} - 3 \right)^2$$

donde

- $C_1$  y  $C_2$  son constantes positivas, calculadas típicamente por simulación.
- $\frac{m_3}{J_n^3}$  es un coeficiente robusto de la asimetría de la muestra.
- $\frac{m_4}{J_n^4}$  es un coeficiente robusto de la curtosis de la muestra.
- 

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

y  $m_3$  y  $m_4$  son los momentos centrados de orden tres y cuatro, respectivamente.

La distribución asintótica del estadístico es también una  $\chi_2^2$ . La región crítica viene dada para valores grandes del estadístico,  $RC = \{RJB > \chi_{1-\alpha,2}^2\}$ .

A continuación mostramos en la Tabla 4.11 los valores del estadístico y los p-valores correspondientes a nuestras muestras. En la Tabla 4.10 (véase sección anterior), vienen dados los valores del límite de la región crítica para diferentes niveles de significación. Observamos cómo para el nivel de significación  $\alpha = 0.05$  únicamente se rechaza la hipótesis de normalidad para las muestras provenientes de las distribuciones exponencial y doble

exponencial. En RStudio tenemos la implementación `rjb.test()` del paquete `lawstat`.

	Valor del estadístico	p-valor
Normal	0.13768	0.9335
Exponencial	1703	<2.2e-16
Uniforme	4.5628	0.1021
Doble exp.	176.33	<2.2e-16
Mixtura	5.0554	0.07984

Tabla 4.11: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Jarque-bera robusto.

## 4.8. Adjusted Jarque Bera

Es otro test de normalidad basado en el siguiente estadístico [33]:

$$AJB = \frac{(\sqrt{b_1})^2}{Var(\sqrt{b_1})} + \frac{(b_2 - E(b_2))^2}{Var(b_2)}$$

donde

- los elementos  $\sqrt{b_1} = S$  y  $b_2 = K - 3$  del estadístico se corresponden con los coeficientes de asimetría y curtosis (trasladado) de la muestra.
- $Var(\sqrt{b_1}) = \frac{6(n-2)}{(n+1)(n+3)}$
- $E(b_2) = \frac{3(n-1)}{n+1}$
- $Var(b_2) = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$

La distribución asintótica del estadístico es también una  $\chi_2^2$ . La región crítica viene dada para valores grandes del estadístico,  $RC = \{AJB > \chi_{1-\alpha,2}^2\}$ .

En la Tabla 4.12 tenemos los datos del estadístico y los p-valores de este test para nuestras muestras de referencia. Como ocurría para el test anterior, los límites de la región crítica se muestran en la Tabla 4.10 (véanse secciones anteriores). En este caso para un nivel de significación  $\alpha = 0.05$  ocurre como con el test anterior y rechazamos las hipótesis de normalidad para las muestras que vienen de las distribuciones exponencial y doble exponencial. En RStudio lo implementamos como `ajb.norm.test()` del paquete `normtest`.

	Valor del estadístico	p-valor
Normal	0.11361	0.937
Exponencial	855.39	<2.2e-16
Uniforme	6.4856	0.052
Doble exp.	144.74	<2.2e-16
Mixtura	5.7455	0.054

Tabla 4.12: Tabla donde se recogen los valores del estadístico y p-valores de muestras muestras para el test de adjusted Jarque-bera.

## 4.9. Test de D'Agostino-Pearson

Este test [22, 33] estudia también una combinación de la simetría y la curtosis. Su mayor inconveniente es que es necesario que el tamaño de la muestra sea considerable, puesto que para muestras pequeñas presenta menos fiabilidad al estar basado en distribuciones asintóticas.

El estadístico de este test viene dado por:

$$K^2 = [Z(\sqrt{b_1})]^2 + [Z(b_2)]^2$$

donde :

- $Z(\sqrt{b_1})$  [33] es una función del coeficiente de asimetría.
- $Z(b_2)$  [33] es una función del coeficiente de curtosis.

La distribución asintótica de este estadístico es una distribución  $\chi_2^2$ . Además, la región crítica viene dada para valores grandes del estadístico:  $RC = \{K^2 > c\}$ .

Mostramos en la Tabla 4.13 los p-valores y valores del estadísticos para muestras cinco muestras. Consúltese la Tabla 4.10 (véanse secciones anteriores) para obtener los valores del límite de la región crítica para diferentes niveles de significación. Para un nivel de significación de  $\alpha = 0.05$  con el test de Agostino-Pearson rechazamos las hipótesis de normalidad para todas las muestras a excepción de las provenientes de las distribuciones normal y mixtura. La implementación en RStudio viene dada por `dagoTest()` del paquete `fBasics`. Tenemos que tener cuidado puesto que esta implementación nos da tres valores de estadísticos, y sus correspondientes p-valores. El que nos interesa es el primero (Omnibus Test), así como su respectivo p-valor. Hay que tener en cuenta que esta implementación solamente nos permite trabajar con muestras de tamaño 20 o mayor.

	Valor del estadístico	p-valor
Normal	0.0717	0.9648
Exponencial	86.1142	< 2.2e-16
Uniforme	36.1677	1.4e-08
Doble exp.	27.969	8.445e-07
Mixtura	5.7464	0.05652

Tabla 4.13: Tabla donde se recogen los valores del estadístico y p-valores de muestras muestras para el test de Agostino Pearson.

## 4.10. Test chi-cuadrado

Los test chi-cuadrado son test de bondad de ajuste, que se puede utilizar para estudiar si una muestra procede o no de una distribución normal. Reciben este nombre puesto que cuando se cumple la hipótesis nula de normalidad el estadístico del test sigue una distribución  $\chi^2$  con ciertos grados de libertad. Típicamente se utilizan para variables aleatorias discretas y con un número finito de observaciones, pero en caso de no contar con este tipo de variables es posible discretizar la variable que se tenga, agrupando los valores de la misma en un número finito de clases.

Estos tests abarcan tanto el caso de hipótesis nulas simples como el de hipótesis nulas compuestas, con ciertas diferencias a la hora de la aplicación debido a la necesidad de utilizar estimaciones en el segundo caso. Hay varios test dentro de este grupo pero uno de los más conocidos y en el que nos vamos a centrar es el llamado **test chi-cuadrado de Pearson** [13] [21]. La idea general que utiliza este test es, dada una variable aleatoria  $X$  cuyos valores se agruparán en clases y una muestra aleatoria simple a partir de la misma, se realiza la comparación entre frecuencias observadas y frecuencias esperadas. Cuanto más diferentes sean las frecuencias observadas y las esperadas, más evidencias en contra de la hipótesis nula de normalidad tendremos.

### 4.10.1. En el caso de hipótesis simples

Dada una variable aleatoria  $X$  y una muestra aleatoria simple  $(X_1, \dots, X_n)$  a partir de  $X$ , definimos  $k$  clases disjuntas  $C_1, \dots, C_k$ , de tal manera que denotamos por  $n_i$  el número de observaciones en cada clase, cumpliendo que  $\sum_{i=1}^k n_i = n$  es el número total de observaciones de la muestra. A las frecuencias  $n_i$  se las denomina las frecuencias observadas de la muestra. Por otro lado,  $\forall i \in \{1, \dots, k\}$ , tenemos  $p_i = P(X \in C_i)$  la probabilidad de estar en la clase  $C_i$ . Gracias a esta probabilidad podemos definir las frecuencias esperadas como  $E_i = n p_i^0$  donde  $p_i^0$  es la probabilidad de estar en la clase

$C_i$  bajo la hipótesis nula. Entonces el estadístico de Pearson viene dado por:

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

Al estar basado en la diferencia entre frecuencias observadas y esperadas, si ambas son similares, el cociente implicado en el estadístico tiende a 0 con lo que tenemos que la región crítica viene dada para valores grandes del estadístico:  $RC = \{D > c\}$ . Bajo  $H_0$ , se tiene que  $D \xrightarrow{n \rightarrow \infty} \chi_{k-1}^2$ , donde  $k-1$  es la dimensión del espacio paramétrico (número de clases menos 1).

#### 4.10.2. En el caso de hipótesis compuestas

En este caso, la distribución no está completamente especificada con lo que habrá un cambio a la hora de definir las frecuencias esperadas. De nuevo contamos con una variable aleatoria  $X$  que toma valores en  $k$  clases disjuntas  $C_1, \dots, C_k$  y que tienen  $n_i$  observaciones en cada clase (frecuencias observadas). Por medio del método de máxima verosimilitud, se calcula la estimación del parámetro o parámetros desconocidos (en el caso de la normal, la media y la desviación típica), es decir  $\hat{\theta}$ , y entonces podremos definir las frecuencias esperadas como  $\hat{E}_i = n P_i(\hat{\theta})$ . Esa es la única diferencia con respecto a una distribución completamente especificada puesto que de nuevo se define el estadístico de Pearson como:

$$D = \sum_{i=1}^k \frac{(n_i - \hat{E}_i)^2}{\hat{E}_i}$$

Se sigue cumpliendo que la  $RC = \{D > c\}$ , y en este caso bajo  $H_0$  el estadístico cumple  $D \xrightarrow{n \rightarrow \infty} \chi_{k-1-q}^2$  donde  $q$  es el número de parámetros que se estiman. Es importante mencionar que  $P_i(\hat{\theta})$  puede no ser sencillo de estimar para los datos agrupados y en ocasiones se recurre a la estimación con los datos sin agrupar. En dicho caso, la distribución asintótica se encontrará entre una  $\chi_{k-1}^2$  y una  $\chi_{k-1-q}^2$ .

A continuación se recogen en la Tabla 4.14 los p-valores, así como los valores del estadístico de Pearson para nuestras muestras. En la Tabla 4.15 vienen recogidos los límites de la región crítica para algunos niveles de significación, calculados a partir de la distribución asintótica del estadístico. Este test con un nivel de significación  $\alpha = 0.05$  rechaza la hipótesis de normalidad únicamente para las muestras provenientes de las distribuciones exponencial y uniforme. En el caso de  $\alpha = 0.01$  solamente se rechazaría para la muestra que viene de la exponencial. En RStudio se implementa por medio de `pearson.test()` del paquete `nortest`.

	Valor del estadístico	p-valor
Normal	5.3	0.8703
Exponencial	54.44	4.017e-08
Uniforme	21.68	0.01682
Doble exp.	7.9	0.6386
Mixtura	7.9	0.6386

Tabla 4.14: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test  $\chi^2$  de Pearson.

$\alpha$	0.01	0.05	0.1
c	23.20925	18.30704	15.98718

Tabla 4.15: Tabla donde se recogen los límites de la región crítica del test  $\chi^2$  de Pearson para los distintos niveles de significación.

#### 4.11. Test de Filliben

El test de Filliben [8] utiliza la correlación entre los estadísticos de orden de la muestra dados por  $X_i$  y la mediana estimada de los estadísticos de orden teóricos. Filliben propuso usar esta correlación para comprobar la normalidad de una muestra. Filliben trabaja con las siguientes medianas estimadas de los estadísticos de orden de una distribución uniforme estándar:

$$m_{(i)} = \begin{cases} 1 - 0.5\left(\frac{1}{n}\right), & \text{Si } i = 1 \\ \frac{i-0.3175}{n+0.365}, & \text{Si } 1 < i < n \\ 0.5\left(\frac{1}{n}\right), & \text{Si } i = n \end{cases}$$

A continuación, se ha de usar el método de la transformación inversa para a partir de cada  $m_{(i)}$  obtener  $M_{(i)} = \phi^{-1}(m_{(i)})$ , que son estimaciones de la mediana de los estadísticos de orden en una normal estándar.

El coeficiente de correlación entre los  $x_{(i)}$  y los  $M_{(i)}$  viene dado por:

$$r = \frac{\sum_{i=1}^n x_{(i)} M_{(i)}}{\sqrt{\sum_{i=1}^n M_{(i)}^2} \sqrt{(n-1)s^2}}$$

La región crítica para este test viene dada para valores pequeños del coeficiente de correlación  $RC = \{r < c\}$ .

A continuación en la Tabla 4.16 vienen recogidos los p-valores y valores del estadístico para nuestras muestras, y vemos que para una signifi-

ción  $\alpha = 0.05$  rechazamos la hipótesis de normalidad para todas las muestras a excepción de las provenientes de las distribuciones normal y mixtura. En RStudio su implementación es por medio de `ppccTest(x, "qnorm", ppos="Filliben")` del paquete `ppcc`, donde `x` es la muestra sobre la que estamos estudiando la normalidad.

	Valor del estadístico	p-valor
Normal	0.99755	0.9574
Exponencial	0.86594	< 2.2e-16
Uniforme	0.98048	0.0064
Doble exp.	0.95556	1e-04
Mixtura	0.98916	0.096

Tabla 4.16: Tabla donde se recogen los valores del estadístico y p-valores de muestras muestras para el test de Filliben.

Con respecto a los límites de la región crítica para este test los podemos encontrar tabulados [8]. Mostramos los límites de la región crítica para los niveles de significación más usados y tamaño de muestra 100 en la Tabla 4.17.

$\alpha$	0.01	0.05	0.1
$c$	0.981	0.987	0.989

Tabla 4.17: Tabla donde se recogen los límites de la región crítica del test de Filliben para los distintos niveles de significación y tamaño de muestra 100.

## 4.12. Test de Chen-Shapiro

El test de Chen-Shapiro [2], [4] también sirve para comprobar la hipótesis de si unos datos se distribuyen o no de forma normal. Se recomienda aplicarlo para distribuciones simétricas, es presentado como una buena alternativa a muchos test de normalidad. Este test utiliza los estadísticos de orden de la muestra dada para el estadístico del test, que viene dado por:

$$QH = \frac{1}{(n-1)s} \sum_{i=1}^{n-1} \frac{x_{(i+1)} - x_{(i)}}{H_{i+1} - H_i}$$

donde como hemos indicado los  $x_{(i)}$  son estadísticos de orden y además tenemos que  $H_i = \varphi^{-1}((i-3/8)/(n+1/4))$  donde  $\varphi^{-1}$  es la función inversa de la función de distribución de una normal estándar. La región crítica viene de nuevo dada por los valores del estadístico pequeños:  $RC = \{QH < c\}$ .



En este caso, para la última versión de RStudio no disponemos de la implementación del test.

### 4.13. Test de Frosini

El test de Frosini [9], [33], [36] se basa en el estadístico:

$$B_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left| \phi(Y_i) - \frac{i - 0.5}{n} \right|$$

donde:

- $Y_i$  viene dado por  $\frac{X_{(i)} - \bar{X}}{s}$ .
- $X_{(i)}$  son los estadísticos de orden.
- $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2$  es la varianza de la muestra, por lo que  $s$  denotará la desviación típica.

La región crítica viene dada para valores grandes del estadístico:  $RC = \{B_n > c\}$ . A continuación mostramos la Tabla 4.18 con los p-valores y valores del estadístico para nuestras muestras, observando cómo para una significación  $\alpha = 0.05$  rechazamos la normalidad en todos los casos a excepción de las muestras que vienen de las distribuciones normal y mixtura. Para  $\alpha = 0.01$  tampoco podríamos rechazar la normalidad de la muestra proveniente de la doble exponencial. En RStudio la implementación viene dada por `frosini.norm.test()` del paquete `normtest`.

	Valor del estadístico	p-valor
Normal	0.13171	0.8825
Exponencial	0.72508	<2.2e-16
Uniforme	0.38547	0.004
Doble exp.	0.31086	0.0305
Mixtura	0.19734	0.371

Tabla 4.18: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Frosini.

### 4.14. Test de Geary

El test de Geary [10] para comprobar la normalidad de una muestra se basa en la comparación de la desviación media de nuestra muestra y la

desviación estándar de la misma. En particular, el estadístico para este test viene dado por:

$$d = \frac{\sum_{i=1}^n |X_i - \bar{X}|}{ns^2}$$

Tenemos que en esta ocasión, sabiendo que las hipótesis nula y alternativa son las mismas con las que hemos trabajado en los test anteriores, la región crítica para la primera de ellas viene dada por valores mayores a uno, ya que con este test el valor esperado del estadístico en caso de que nuestra muestra se ajuste a una distribución normal es de  $\sqrt{\frac{2}{\pi}} \approx 0.7979$ .

Es decir  $RC = \{d > 0.7979 + c\}$  o  $RC = \{d < 0.7979 - c\}$ , cuando  $d$  es un valor muy por encima o por debajo del de referencia.

En la Tabla 4.19 tenemos los p-valores y valores del estadístico para nuestras muestras. Para un nivel de significación  $\alpha = 0.05$  rechazamos la normalidad de las muestras provenientes de la distribución uniforme únicamente, con lo que podemos deducir que este test tiene una potencia muy baja. En RStudio viene implementado por `geary.norm.test()` del paquete `normtest`.

	Valor del estadístico	p-valor
Normal	0.80295	0.442
Exponencial	0.67441	1
Uniforme	0.87007	5e-04
Doble exp.	0.71155	1
Mixtura	0.79977	0.495

Tabla 4.19: Tabla donde se recogen los valores del estadístico y p-valores de nuestras muestras para el test de Geary.

## Capítulo 5

# Comparación de la potencia de los distintos tests presentados

En este capítulo vamos a comparar la potencia de distintos test de normalidad, en concreto de los que aparecen en el capítulo anterior a excepción de los tres últimos, puesto que para estos se nos presentan ciertas dificultades: en el caso de Chen-Shapiro no disponemos en la versión de RStudio con la que se trabaja de una implementación directa, mientras que en los casos de Frosini y de Geary lo que ocurre es que dentro del propio test se recurre a simulación de MonteCarlo con un gran número de repeticiones, lo que hace que no fuese factible realizar las simulaciones con los medios disponibles en un tiempo razonable.

Realizaremos el estudio de la potencia en primer lugar con las distribuciones con las que llevamos trabajando durante todos los capítulos, a nivel de significación  $\alpha = 0.05$  y distintos tamaños muestrales. También estudiaremos la potencia de los test bajo distribuciones conocidas como, por ejemplo, la distribución t de Student o la distribución chi-cuadrado. Este estudio se realizará por medio de simulaciones de MonteCarlo con 10000 repeticiones, siendo la estimación de la potencia el número de veces que en estas repeticiones se rechaza la hipótesis nula de normalidad.

## 5.1. Potencia bajo la hipótesis nula

### 5.1.1. Distribución normal

Test de normalidad	10	20	30	40	50	60	70	80	90	100
Lilliefors	0.0511	0.0488	0.0498	0.0500	0.0502	0.0489	0.0485	0.0520	0.0524	0.0497
Cramer Von Mises	0.0489	0.0531	0.0517	0.0519	0.0481	0.0497	0.0496	0.0516	0.0531	0.0493
Anderson Darling	0.0499	0.0518	0.0518	0.0516	0.0476	0.0506	0.0462	0.0503	0.0513	0.0489
Shapiro-Wilk	0.0498	0.0504	0.0523	0.0519	0.0521	0.0530	0.0482	0.0494	0.0506	0.0489
Shapiro Francia	0.0564	0.0542	0.0512	0.0521	0.0559	0.0543	0.0478	0.0522	0.0532	0.0497
Jarque Bera	0.0089	0.0245	0.0312	0.0345	0.0393	0.0411	0.0364	0.0415	0.0412	0.0490
JB Robusto	0.0607	0.0644	0.0622	0.0610	0.0610	0.0622	0.0532	0.0603	0.0593	0.0545
Adjusted JB	0.0718	0.0608	0.0600	0.0591	0.0604	0.0604	0.0517	0.0568	0.0540	0.0535
Agostino-Pearson	-	0.0579	0.0591	0.0583	0.0573	0.0606	0.0537	0.0582	0.0563	0.0534
Chi Pearson	0.0660	0.0524	0.0511	0.0575	0.0550	0.0603	0.0481	0.0517	0.0533	0.0501
Filliben	0.0513	0.0519	0.0495	0.0504	0.0539	0.0520	0.0472	0.0505	0.0504	0.0477

Tabla 5.1: Potencias de las pruebas de normalidad para una distribución  $N(3,5)$  con distintos tamaños de muestra.

En el caso de la potencia bajo la hipótesis nula (Tabla 5.1, Figura 5.1), tenemos un comportamiento prácticamente idéntico de la potencia de todos los test, con valores muy bajos de las mismas, cercanos al nivel de significación considerado. Se observa como algunos tests dan potencias ligeramente superiores para tamaños muestrales pequeños, debido al uso de distribuciones asintóticas.

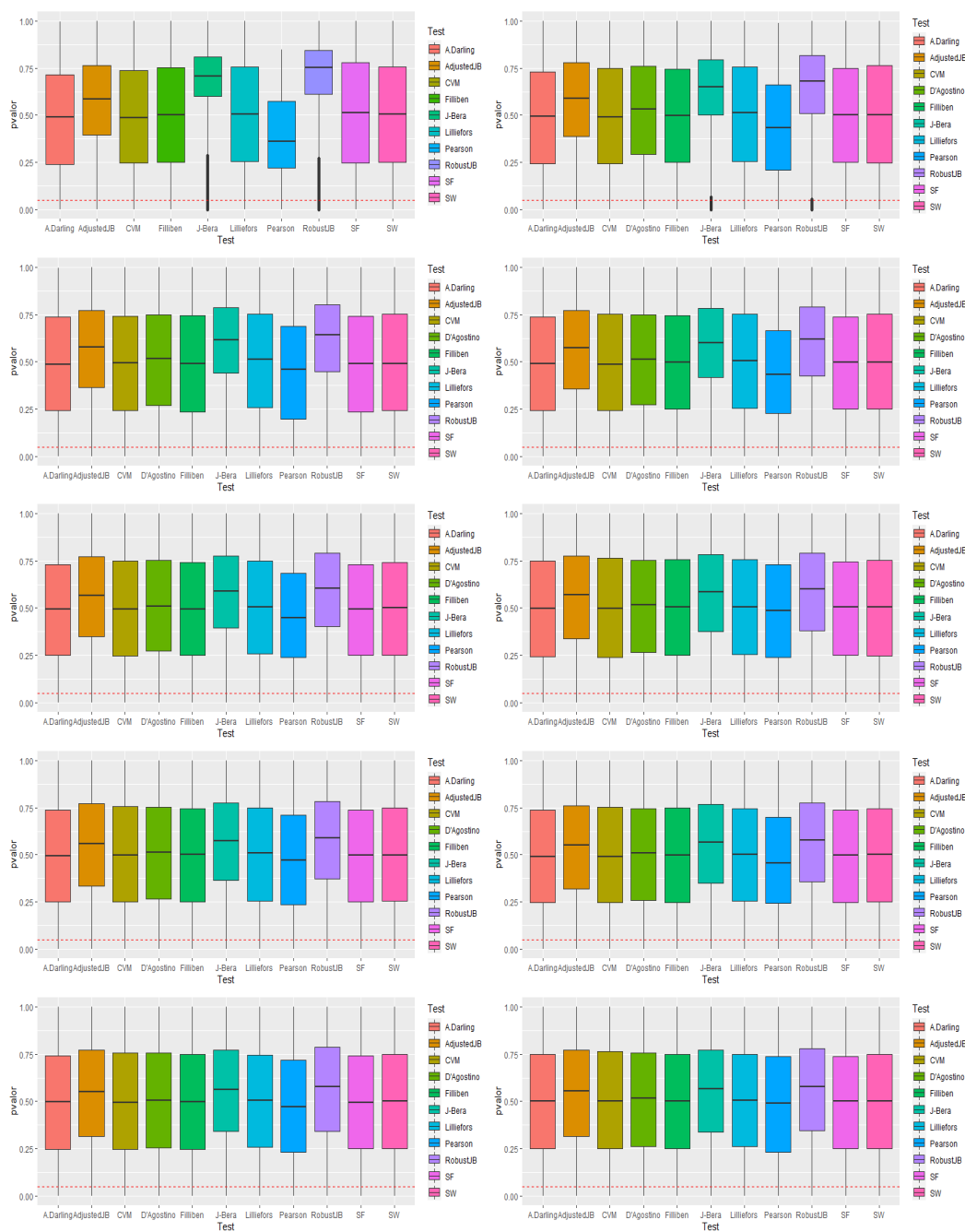


Figura 5.1: Distribución de los p-valores de las pruebas de normalidad para una distribución  $N(3,5)$  con distintos tamaños de muestra.

## 5.2. Potencia bajo la hipótesis alternativa

### 5.2.1. Distribución exponencial

Test de normalidad	10	20	30	40	50	60	70	80	90	100
Lilliefors	0.3033	0.5759	0.7768	0.9027	0.9598	0.9858	0.9966	0.9988	0.9998	1.0000
Cramer Von Mises	0.3837	0.7265	0.9013	0.9670	0.9908	0.9971	0.9996	1.0000	0.9999	1.0000
Anderson Darling	0.4149	0.7772	0.9337	0.9833	0.9968	0.9994	0.9999	1.0000	1.0000	1.0000
Shapiro-Wilk	0.4475	0.8367	0.9699	0.9951	0.9989	0.9998	1.0000	1.0000	1.0000	1.0000
Shapiro Francia	0.5342	0.7948	0.9499	0.9901	0.9983	0.9998	1.0000	1.0000	1.0000	1.0000
Jarque Bera	0.1472	0.4800	0.7273	0.8815	0.9553	0.9878	0.9971	0.9997	0.9998	1.0000
JB Robusto	0.3211	0.5954	0.7808	0.8920	0.9479	0.9796	0.9931	0.9973	0.9997	1.0000
Adjusted JB	0.3400	0.6011	0.7975	0.9146	0.9687	0.9915	0.9979	0.9998	0.9998	0.9998
Agostino-Pearson	-	0.5997	0.7900	0.9040	0.9632	0.9914	0.9976	1.0000	0.9999	1.0000
Chi Pearson	0.3968	0.6527	0.6527	0.9524	0.9831	0.9951	0.9991	0.9997	0.9998	1.0000
Filliben	0.4148	0.7777	0.9415	0.9887	0.9981	0.9998	0.9999	1.0000	1.0000	1.0000

Tabla 5.2: Potencias de las pruebas de normalidad bajo distribución alternativa  $\exp(0.2)$  con distintos tamaños de muestra.

Para la potencia bajo la hipótesis alternativa de que los datos provienen de una distribución exponencial (Tabla 5.2, Figura 5.2) se obtienen potencias muy altas en todos los test, incluso en aquellos que por lo general tienen potencias bajas bajo otras hipótesis alternativas. Nótese que se comienzan a alcanzar tamaños de potencia cercanos a 0.9 en la totalidad de los tests a partir de un tamaño de muestra de  $n = 40$ .

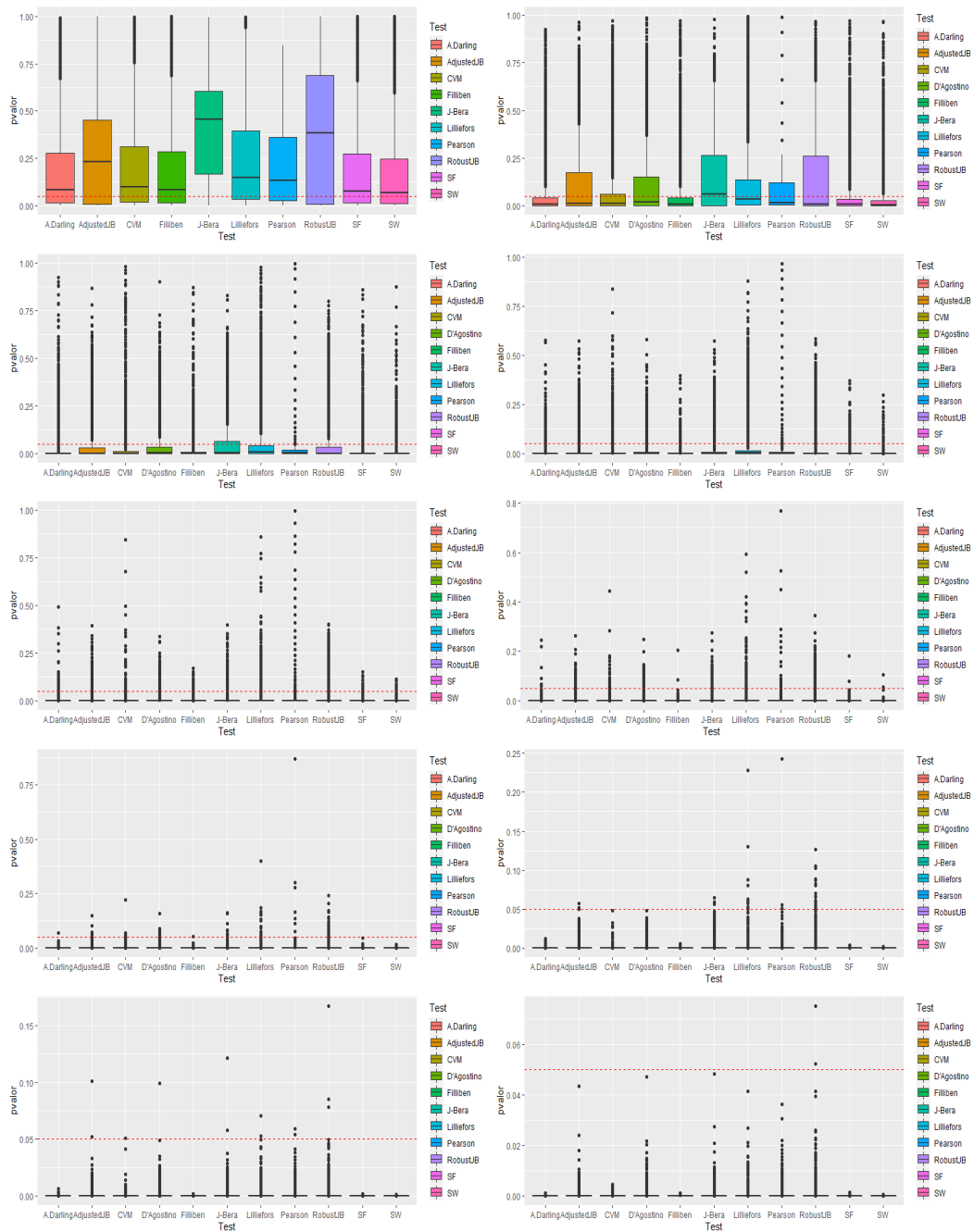


Figura 5.2: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa  $\exp(0.2)$  con distintos tamaños de muestra.

### 5.2.2. Distribución uniforme

Test de normalidad	10	20	30	40	50	60	70	80	90	100
Lilliefors	0.0646	0.0951	0.1408	0.1975	0.2633	0.3257	0.3947	0.4647	0.5265	0.5881
Cramer Von Mises	0.0772	0.1389	0.2278	0.3355	0.4423	0.5378	0.6306	0.7104	0.7781	0.8363
Anderson Darling	0.0823	0.1708	0.2938	0.4396	0.5722	0.6910	0.7919	0.8575	0.9118	0.9439
Shapiro-Wilk	0.0863	0.1984	0.3812	0.5852	0.7457	0.8633	0.9373	0.9720	0.9888	0.9952
Shapiro Francia	0.0533	0.0827	0.1675	0.3115	0.4736	0.6353	0.7714	0.8661	0.9292	0.9637
Jarque Bera	0.0023	0.0009	0.0002	0.0001	0.0000	0.0031	0.0369	0.1554	0.3446	0.5595
JB Robusto	0.0165	0.0039	0.0006	0.0005	0.0001	0.0002	0.0000	0.0002	0.0050	0.0400
Adjusted JB	0.0187	0.0032	0.0005	0.0001	0.0023	0.0245	0.1110	0.2645	0.4526	0.6439
Agostino-Pearson	-	0.1545	0.3898	0.6281	0.8057	0.8979	0.9541	0.9816	0.9931	0.9971
Chi Pearson	0.0905	0.0798	0.1138	0.1618	0.2061	0.2626	0.2826	0.3544	0.4338	0.4569
Filliben	0.0441	0.0630	0.1305	0.2553	0.4137	0.5654	0.7135	0.8221	0.8993	0.9477

Tabla 5.3: Potencias de las pruebas de normalidad bajo distribución alterna-tiva  $U(0,10)$  con distintos tamaños de muestra.

Para la potencia bajo la hipótesis alternativa de que los datos provie-nen de una distribución uniforme (Tabla 5.3, Figura 5.3), el test de mayor potencia en este caso sería Agostino-Pearson que alcanza una potencia lo suficientemente buena para  $n = 50$ , seguido de otros como Shapiro-Wilk. El test con menor potencia es JB Robusto y otro de los que tienen la potencia más baja es el chi-cuadrado. Podemos ver cómo a medida que aumenta el tamaño de la muestra la potencia aumenta en la mayoría de los tests.



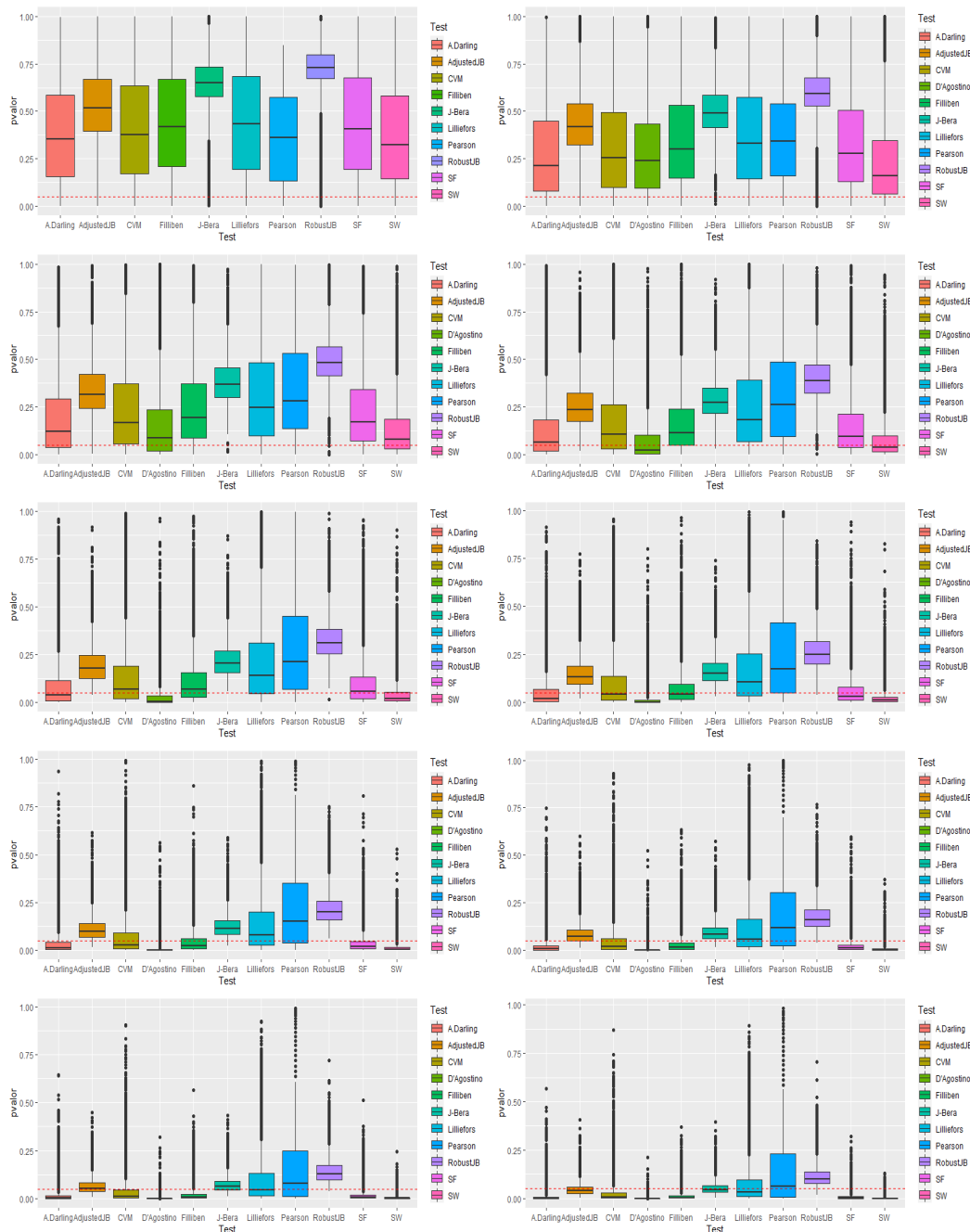


Figura 5.3: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa  $U(0,10)$  con distintos tamaños de muestra.

### 5.2.3. Distribución doble exponencial

Test de normalidad	10	20	30	40	50	60	70	80	90	100
Lilliefors	0.1433	0.2196	0.2927	0.3603	0.4263	0.4996	0.5551	0.6136	0.6625	0.7044
Cramer Von Mises	0.1597	0.2630	0.3663	0.4538	0.5334	0.6102	0.6764	0.7314	0.7819	0.8209
Anderson Darling	0.1617	0.2700	0.3726	0.4627	0.5375	0.6148	0.6802	0.7360	0.7913	0.8208
Shapiro-Wilk	0.1578	0.2607	0.3594	0.4465	0.5181	0.5904	0.6505	0.7033	0.7561	0.7872
Shapiro Francia	0.1911	0.3222	0.4293	0.5235	0.5904	0.6627	0.7192	0.7649	0.8124	0.8386
Jarque Bera	0.0650	0.2176	0.3332	0.4390	0.5072	0.5801	0.6465	0.6886	0.7431	0.7749
JB Robusto	0.2114	0.3863	0.5085	0.6098	0.6782	0.7440	0.7952	0.8345	0.8743	0.8945
Adjusted JB	0.2235	0.3452	0.4421	0.5329	0.5889	0.6565	0.7056	0.7415	0.7871	0.8147
Agostino-Pearson	-	0.3035	0.3819	0.4597	0.5071	0.5632	0.6164	0.6500	0.6990	0.7290
Chi Pearson	0.1358	0.1427	0.1968	0.2392	0.2747	0.3259	0.3655	0.3985	0.4393	0.4711
Filliben	0.1851	0.3203	0.4293	0.5243	0.5931	0.6640	0.7214	0.7671	0.8144	0.8398

Tabla 5.4: Potencias de las pruebas de normalidad bajo distribución Laplace(3,5) con distintos tamaños de muestra.

Para la potencia bajo la hipótesis alternativa de que los datos provienen de una distribución doble exponencial (Tabla 5.4, Figura 5.4) vemos como la potencia aumenta a medida que aumenta el tamaño de la muestra. Tests como Jarque Bera Robusto alcanzan una potencia aceptable a partir de  $n = 80$ , siendo el test de mayor potencia seguido de otros como Filliben o Shapiro-Francia. En este caso el test chi-cuadrado es el test de menor potencia con diferencia alcanzando en  $n = 100$  únicamente una potencia de 0.4711.

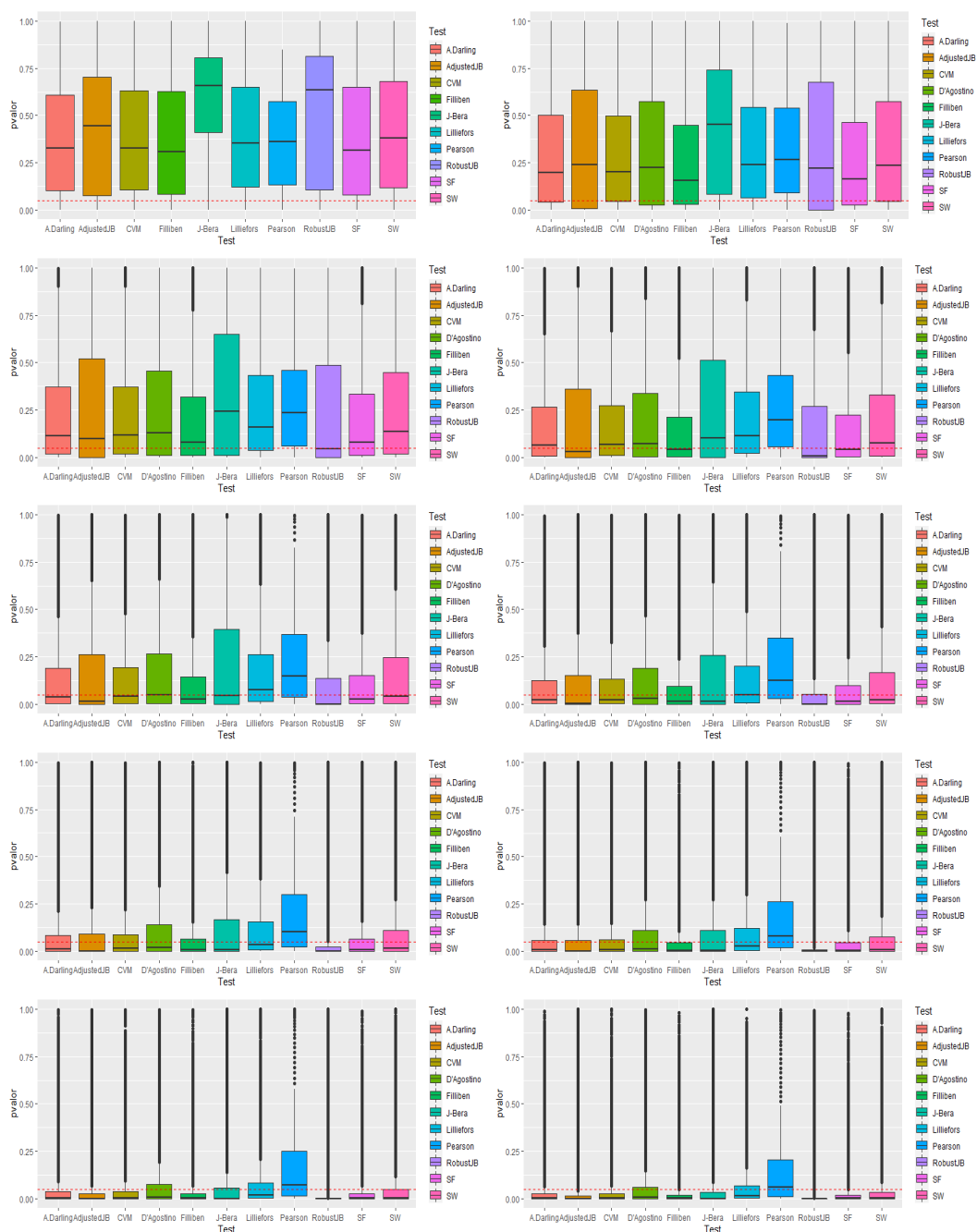


Figura 5.4: Distribución de los p-valores de las pruebas de normalidad bajo distribución Laplace(3,5) con distintos tamaños de muestra.

### 5.2.4. Distribución mixtura gaussiana

Ahora vamos a estudiar la potencia de dos tipos de distribuciones mixtura gaussiana, para distintas probabilidades (en concreto  $p \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$ ). Realizaremos este estudio para tres tamaños de muestra distintos ( $n = 20$ ,  $n = 50$  y  $n = 100$ ). En concreto las mixturas sobre las que estudiaremos la potencia de los distintos test son:

- Mixturas gaussianas centradas:  $N(0, 1)$  con probabilidad  $1-p$  y  $N(0, 1000)$  con probabilidad  $p$ .
- Mixturas gaussianas descentradas:  $N(0, 1)$  con probabilidad  $1 - p$  y  $N(1000, 1)$  con probabilidad  $1 - p$ .

Ambas distribuciones representan de alguna forma la introducción de ruido en una normal estándar, repartiéndose este ruido en ambas colas en el primero de los casos y solamente en la cola derecha en el segundo de los casos.

#### Mixtura gaussiana centrada y tamaño muestral $n=20$

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.2146	0.3636	0.6533	0.8834	0.9884	0.9988	0.9964	0.5774
Cramer Von Mises	0.2169	0.3655	0.6546	0.8837	0.9886	0.9988	0.9955	0.5791
Anderson Darling	0.2175	0.3666	0.6547	0.8838	0.9887	0.9986	0.9903	0.5793
Shapiro Wilk	0.2159	0.3647	0.6533	0.8830	0.9883	0.9948	0.9652	0.5781
Shapiro Francia	0.2180	0.3672	0.6557	0.8840	0.9888	0.9958	0.9751	0.5809
Jarque Bera	0.1949	0.3478	0.6443	0.8749	0.9198	0.7806	0.5713	0.5678
JB Robusto	0.2260	0.3723	0.6588	0.8849	0.9888	0.9914	0.9568	0.5848
Adjusted JB	0.2232	0.3706	0.6573	0.8834	0.9664	0.9038	0.7591	0.5835
Agostino Pearson	0.2192	0.3675	0.6560	0.8811	0.9434	0.8408	0.6659	0.5817
Chi Pearson	0.2143	0.3636	0.6514	0.8825	0.9886	0.9986	0.9955	0.5762
Filliben	0.2155	0.3649	0.6548	0.8838	0.9887	0.9958	0.9748	0.5799

Tabla 5.5: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 20 y distintas probabilidades.

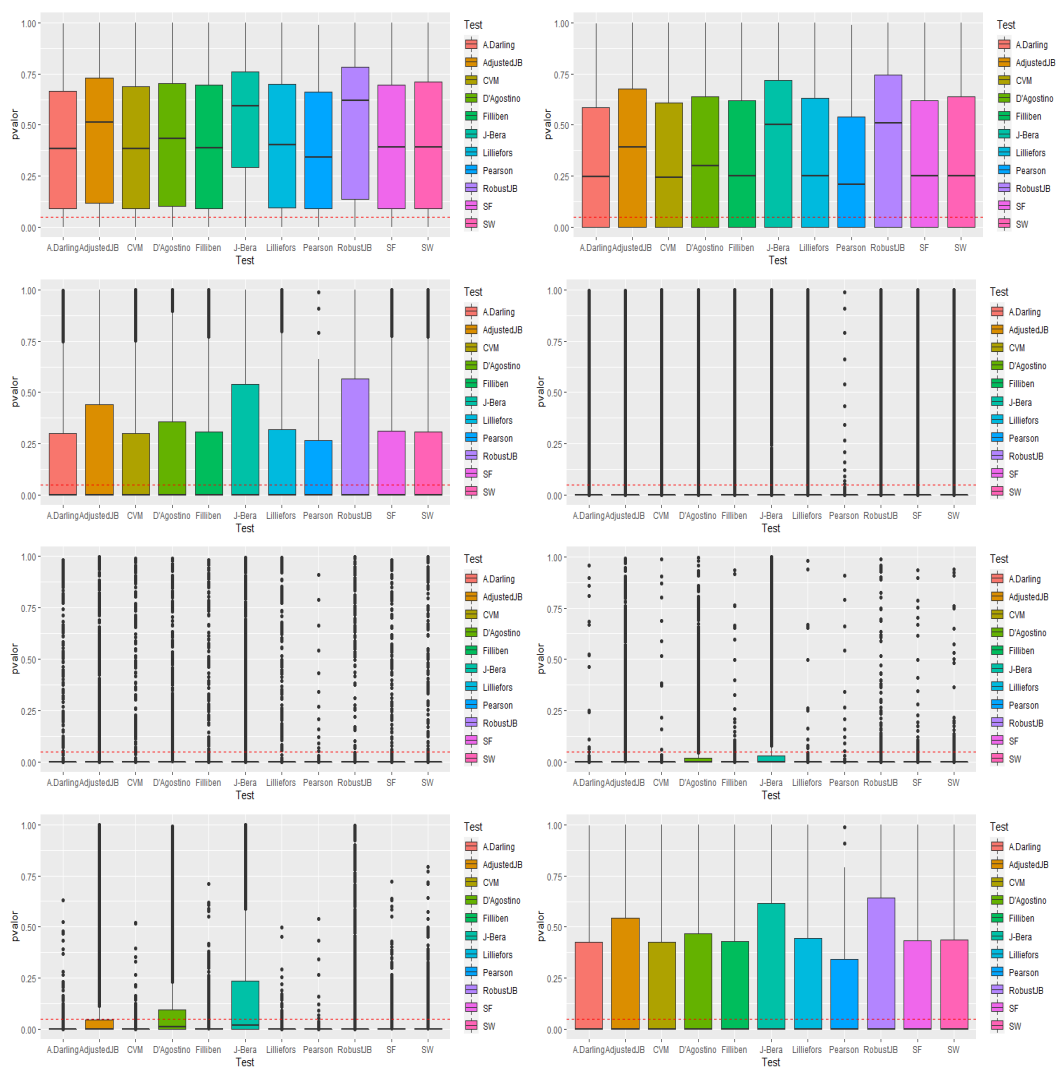


Figura 5.5: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 20 y distintas probabilidades.

### Mixtura gaussiana centrada y tamaño muestral n=50

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.4270	0.6539	0.9267	0.9952	1.0000	1.0000	1.0000	1.0000
Cramer Von Mises	0.4253	0.6530	0.9269	0.9951	1.0000	1.0000	1.0000	1.0000
Anderson Darling	0.4265	0.6536	0.9271	0.9951	1.0000	1.0000	1.0000	1.0000
Shapiro Wilk	0.4259	0.6556	0.9273	0.9951	1.0000	1.0000	1.0000	0.9984
Shapiro Francia	0.4308	0.6570	0.9272	0.9951	1.0000	1.0000	1.0000	0.9984
Jarque Bera	0.4207	0.6507	0.9251	0.9949	0.9995	0.9883	0.9160	0.7377
JB Robusto	0.4372	0.6610	0.9274	0.9951	1.0000	1.0000	0.9998	0.9943
Adjusted JB	0.4352	0.6597	0.9266	0.9950	0.9999	0.9958	0.9548	0.8268
Agostino Pearson	0.4341	0.6581	0.9263	0.9950	0.9993	0.9803	0.8957	0.7122
Chi Pearson	0.4296	0.6550	0.9274	0.9950	1.0000	1.0000	1.0000	1.0000
Filliben	0.4304	0.6568	0.9271	0.9951	1.0000	1.0000	1.0000	0.9989

Tabla 5.6: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 50 y distintas probabilidades.

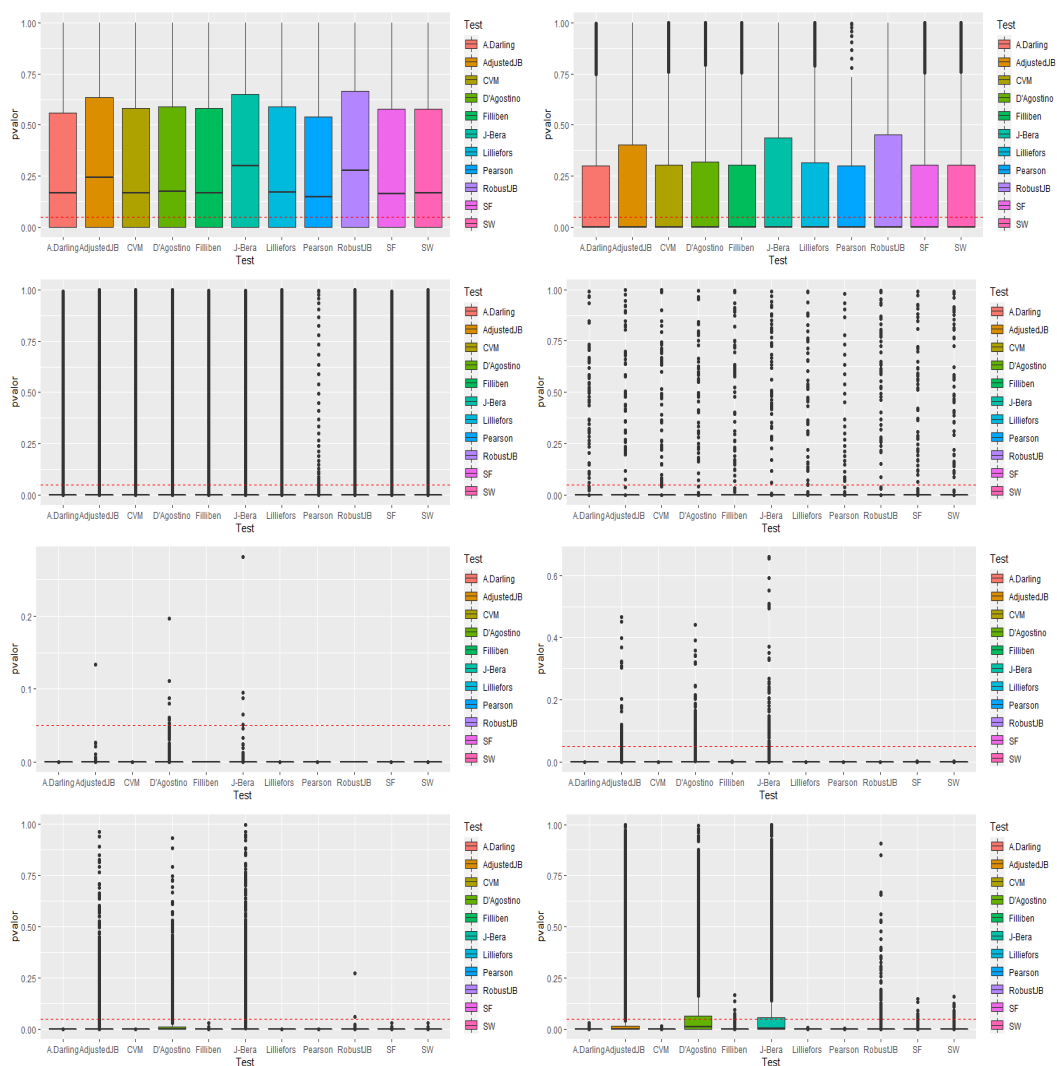


Figura 5.6: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 50 y distintas probabilidades.

## Mixtura gaussiana centrada y tamaño muestral n=100

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.6495	0.8708	0.9954	1.0000	1.0000	1.0000	1.0000	1.0000
Cramer Von Mises	0.6498	0.8710	0.9954	1.0000	1.0000	1.0000	1.0000	1.0000
Anderson Darling	0.6506	0.8721	0.9954	1.0000	1.0000	1.0000	1.0000	1.0000
Shapiro Wilk	0.6512	0.8718	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000
Shapiro Francia	0.6524	0.8725	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000
Jarque Bera	0.6496	0.8710	0.9951	1.0000	1.0000	1.0000	0.9971	0.9552
JB Robusto	0.6552	0.8724	0.9951	1.0000	1.0000	1.0000	1.0000	1.0000
Adjusted JB	0.6542	0.8726	0.9951	1.0000	1.0000	1.0000	0.9986	0.9712
Agostino Pearson	0.6527	0.8720	0.9951	1.0000	1.0000	1.0000	0.9922	0.9210
Chi Pearson	0.6518	0.8714	0.9951	1.0000	1.0000	1.0000	1.0000	1.0000
Filliben	0.6519	0.8719	0.9950	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla 5.7: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 100 y distintas probabilidades.

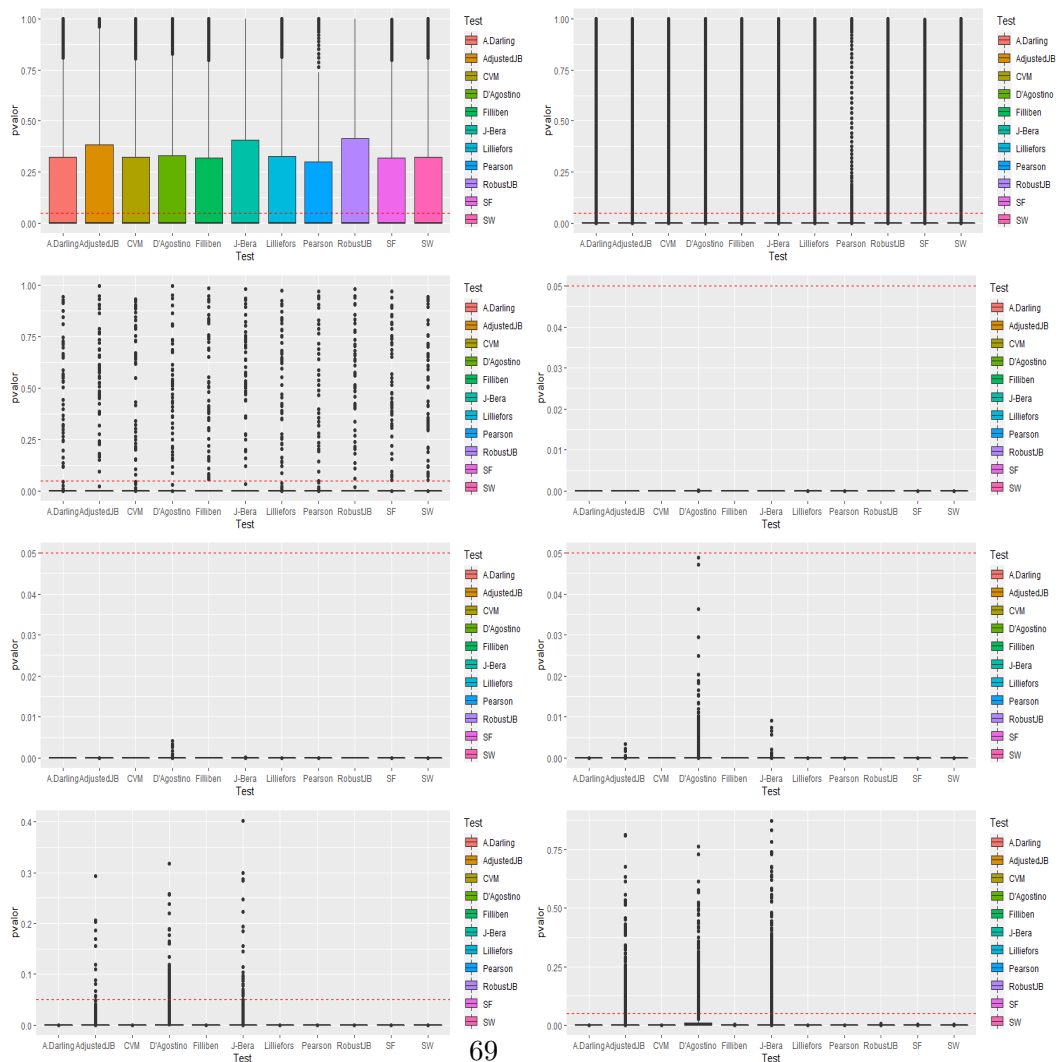


Figura 5.7: Distribución de los p-valoros de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana centrada con tamaño de muestra 100 y distintas probabilidades.



### Mixtura gaussiana descentrada y tamaño muestral n=20

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.2150	0.3644	0.6548	0.8841	0.9887	0.9992	1.0000	1.0000
Cramer Von Mises	0.2172	0.3664	0.6560	0.8842	0.9889	0.9992	1.0000	1.0000
Anderson Darling	0.2178	0.3674	0.6561	0.8843	0.9890	0.9992	1.0000	1.0000
Shapiro Wilk	0.2162	0.3654	0.6546	0.8843	0.9885	0.9992	1.0000	1.0000
Shapiro Francia	0.2183	0.3679	0.6569	0.8843	0.9890	0.9992	1.0000	1.0000
Jarque Bera	0.1952	0.3485	0.6433	0.8381	0.6176	0.2345	0.0493	0.0112
JB Robusto	0.2263	0.3730	0.6598	0.8837	0.9036	0.6086	0.2523	0.1128
Adjusted JB	0.2235	0.3713	0.6580	0.8751	0.7938	0.4126	0.1242	0.0406
Agostino Pearson	0.2195	0.3682	0.6551	0.8513	0.8130	0.8211	0.9251	0.9706
Chi Pearson	0.2149	0.3648	0.6533	0.8834	0.9889	0.9993	1.0000	1.0000
Filliben	0.2158	0.3656	0.6560	0.8841	0.9889	0.9992	1.0000	1.0000

Tabla 5.8: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 20 y distintas probabilidades.

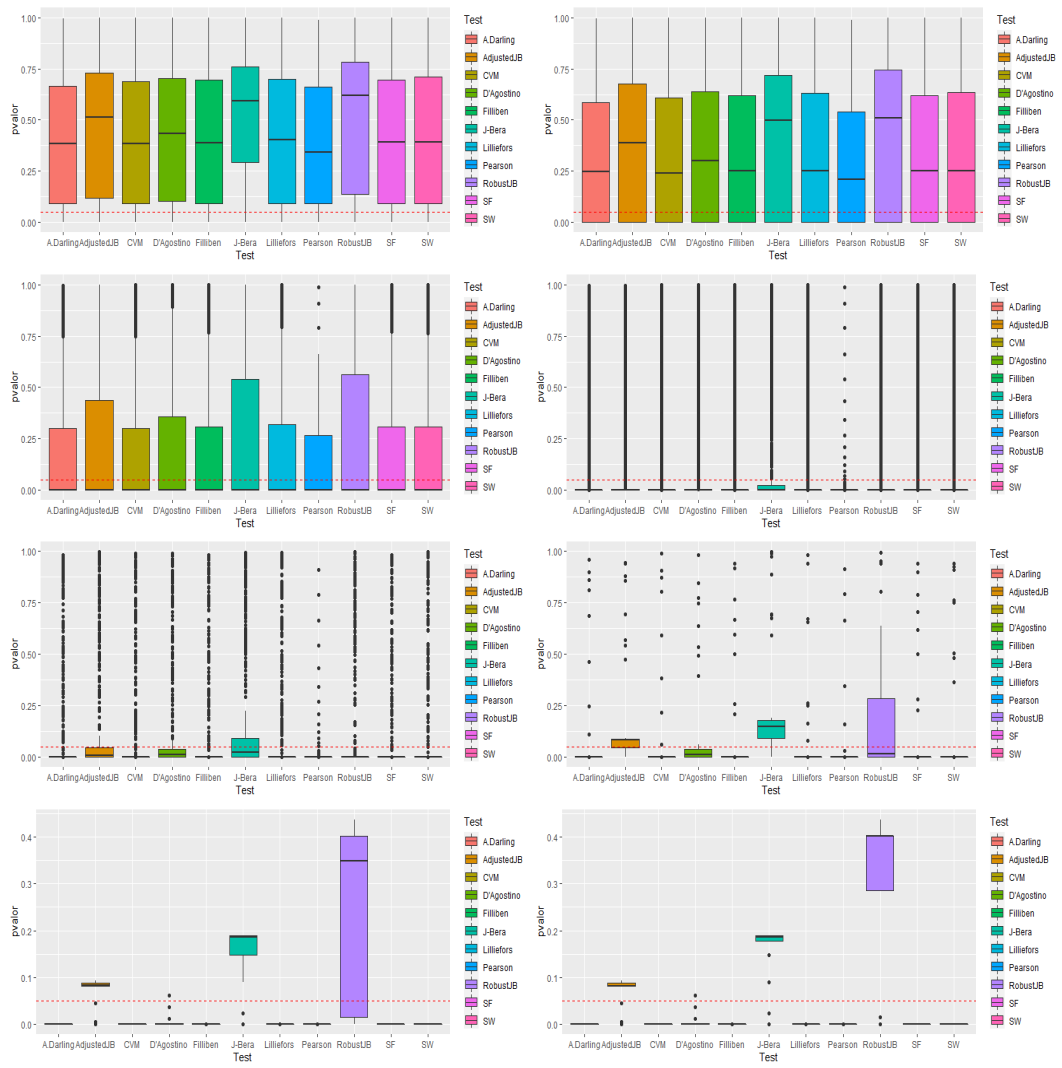


Figura 5.8: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 20 y distintas probabilidades.

### Mixtura gaussiana descentrada y tamaño muestral n=50

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.4289	0.6558	0.9274	0.9953	1.0000	1.0000	1.0000	1.000
Cramer Von Mises	0.4274	0.6549	0.9276	0.9952	1.0000	1.0000	1.0000	1.000
Anderson Darling	0.4283	0.6552	0.9277	0.9952	1.0000	1.0000	1.0000	1.000
Shapiro Wilk	0.4308	0.6568	0.9278	0.9952	1.0000	1.0000	1.0000	1.000
Shapiro Francia	0.4320	0.6580	0.9277	0.9952	1.0000	1.0000	1.0000	1.000
Jarque Bera	0.4219	0.6517	0.9256	0.9950	1.0000	1.0000	1.0000	1.000
JB Robusto	0.4383	0.6620	0.9279	0.9952	0.9945	0.7822	0.2404	0.034
Adjusted JB	0.4363	0.6607	0.9271	0.9951	1.0000	1.0000	1.0000	1.000
Agostino Pearson	0.4325	0.6591	0.9268	0.9951	NA	NA	NA	NA
Chi Pearson	0.4321	0.6573	0.9282	0.9950	1.0000	1.0000	1.0000	1.000
Filliben	0.4316	0.6578	0.9276	0.9952	1.0000	1.0000	1.0000	1.000

Tabla 5.9: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 50 y distintas probabilidades.

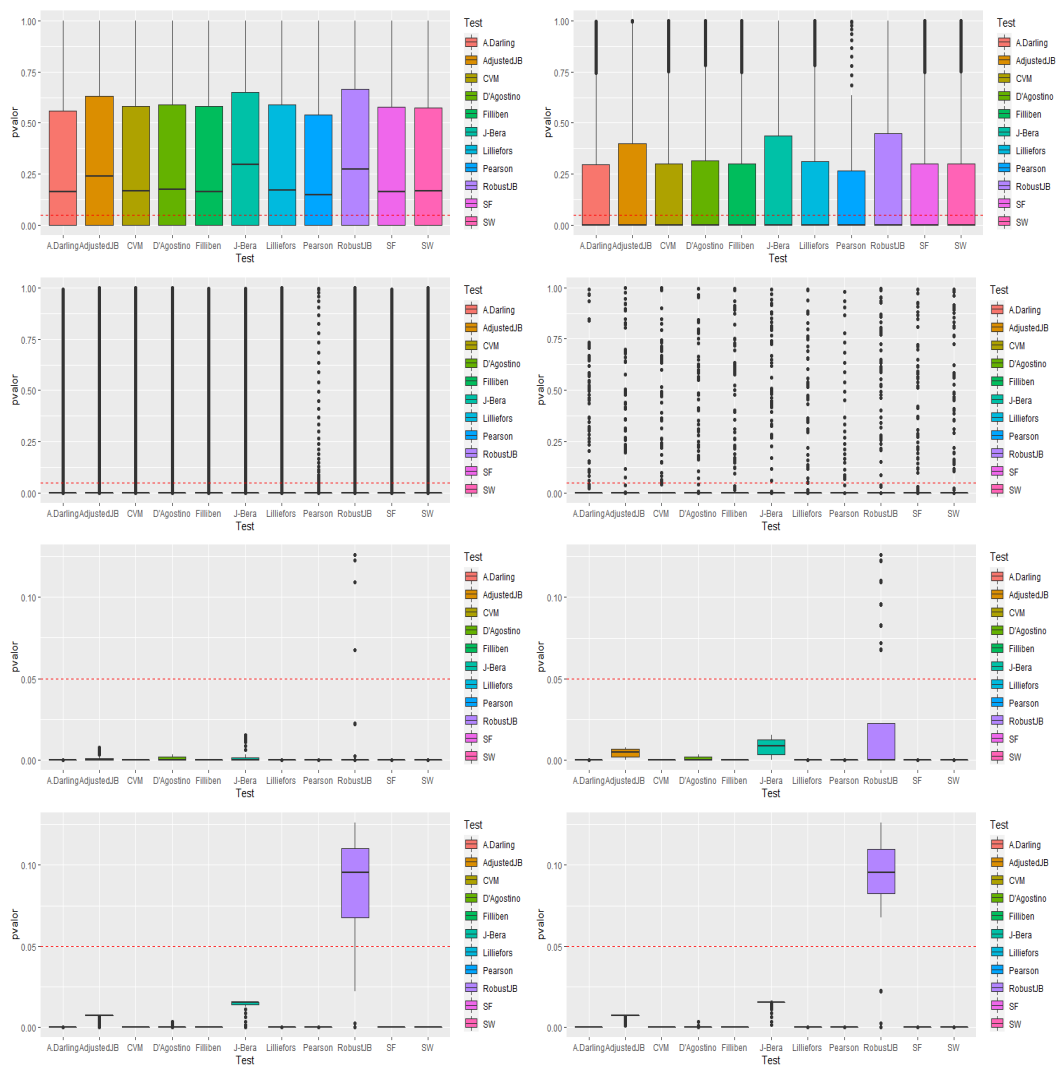


Figura 5.9: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 50 y distintas probabilidades.

### Mixtura gaussiana descentrada y tamaño muestral n=100

	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5
Lilliefors	0.6507	0.8718	0.9956	1.0000	1.0000	1.0000	1.0000	1.0000
Cramer Von Mises	0.6511	0.8720	0.9956	1.0000	1.0000	1.0000	1.0000	1.0000
Anderson Darling	0.6516	0.8729	0.9956	1.0000	1.0000	1.0000	1.0000	1.0000
Shapiro Wilk	0.6519	0.8724	0.9952	1.0000	1.0000	1.0000	1.0000	1.0000
Shapiro Francia	0.6531	0.8731	0.9952	1.0000	1.0000	1.0000	1.0000	1.0000
Jarque Bera	0.6502	0.8716	0.9953	1.0000	1.0000	1.0000	1.0000	1.0000
JB Robusto	0.6558	0.8730	0.9953	1.0000	1.0000	1.0000	1.0000	1.0000
Adjusted JB	0.6548	0.8732	0.9953	1.0000	1.0000	1.0000	1.0000	1.0000
Agostino Pearson	0.6534	0.8726	0.9953	1.0000	1.0000	NA	NA	NA
Chi Pearson	0.6532	0.8723	0.9954	1.0000	1.0000	1.0000	1.0000	1.0000
Filliben	0.6526	0.8725	0.9952	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla 5.10: Potencias de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 100 y distintas probabilidades.

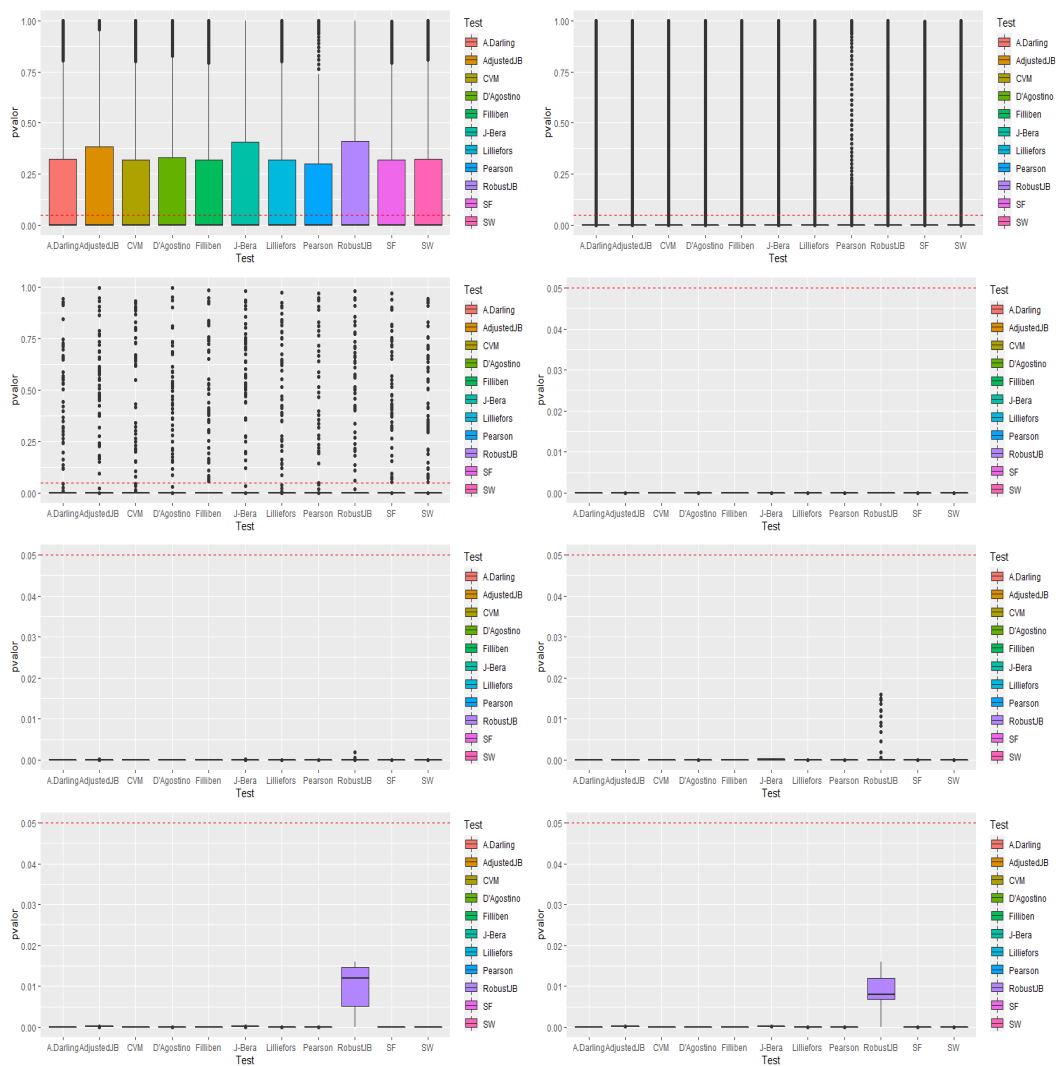


Figura 5.10: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa mixtura gaussiana descentrada con tamaño de muestra 100 y distintas probabilidades.

En resumen, trabajando bajo la distribución alternativa mixtura gaussiana, se observa un incremento de la potencia al aumentar el tamaño muestral, hasta llegar al máximo en algunos casos. Además, la potencia es mayor para aquellas simulaciones en las que  $p$  toma valores cercanos a 0.5. En particular, en los casos de  $p = 0.01$  y  $p = 0.03$  podemos ver potencias bajas, mientras que en los casos de probabilidades como pueden ser  $p = 0.3$  o  $p = 0.4$  la potencia es alta, llegando a los valores que se consideran aceptables (en torno a 0.8 y superiores). Nótese el comportamiento extraño para el caso  $p = 0.5$  y tamaño muestral  $n = 20$ , donde las potencias se hacen más pequeñas que

en el caso  $p = 0.4$ .

En el caso de la mixtura descentrada las potencias son muy similares, pero podríamos destacar el test de Agostino-Pearson como aquel de menor potencia que los otros, alcanzando aún así valores de potencia aceptables. En el caso de la mixtura descentrada se destaca que el test de JB y sus modificaciones tienen potencias muy bajas para las simulaciones con tamaños pequeños y probabilidades como  $p = 0.4$  o  $p = 0.5$ . Para el test de Agostino-Pearson hay veces en las que la salida es NA, probablemente por la aparición de valores cercanos a cero en algún denominador.

### 5.2.5. Distribución t de Student

A continuación se estudiará la potencia de los distintos test de normalidad para la distribución t de Student, con la cuál trabajaremos variando los grados de libertad como podremos ver en las tablas, así como los tamaños de las muestras que vamos a manejar (en concreto nos centramos en los tamaños muestrales  $n \in \{20, 50, 100\}$ ).

#### Tamaño de muestra $n=20$ y distintos grados de libertad

	1	2	3	4	5	6	7	8
Lilliefors	0.8460	0.4470	0.2555	0.1813	0.1316	0.1050	0.0861	0.0808
Cramer Von Mises	0.8816	0.5118	0.3029	0.2218	0.1591	0.1257	0.1057	0.0951
Anderson Darling	0.8816	0.5308	0.3206	0.2347	0.1744	0.1389	0.1158	0.1043
Shapiro Wilk	0.8717	0.5312	0.3396	0.2511	0.1891	0.1497	0.1315	0.1132
Shapiro Francia	0.8947	0.5846	0.3860	0.2931	0.2247	0.1814	0.1557	0.1351
Jarque Bera	0.8234	0.4927	0.3152	0.2252	0.1729	0.1326	0.1097	0.0943
JB Robusto	0.9113	0.6319	0.4348	0.3339	0.2667	0.2196	0.1899	0.1654
Adjusted JB	0.8853	0.6040	0.4166	0.3275	0.2591	0.2148	0.1873	0.1654
Agostino-Pearson	0.8614	0.5647	0.3854	0.2966	0.2343	0.1917	0.1666	0.1488
Chi Pearson	0.7797	0.3390	0.1804	0.1219	0.0936	0.0815	0.0691	0.0645
Filliben	0.8945	0.5831	0.3849	0.2917	0.2233	0.1798	0.1536	0.1318

Tabla 5.11: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 20 y diferentes grados de libertad.

	9	10	15	20	25	30	35	40
Lilliefors	0.0739	0.0720	0.0615	0.0568	0.0509	0.0515	0.0534	0.0518
Cramer Von Mises	0.0833	0.0826	0.0700	0.0611	0.0598	0.0551	0.0558	0.0522
Anderson Darling	0.0913	0.0902	0.0720	0.0606	0.0590	0.0551	0.0550	0.0553
Shapiro Wilk	0.1046	0.0966	0.0806	0.0677	0.0635	0.0589	0.0562	0.0587
Shapiro Francia	0.1243	0.1158	0.0944	0.0754	0.0715	0.0667	0.0631	0.0664
Jarque Bera	0.0836	0.0746	0.0553	0.0434	0.0413	0.0347	0.0330	0.0335
JB Robusto	0.1562	0.1408	0.1149	0.0941	0.0842	0.0826	0.0782	0.0792
Adjusted JB	0.1553	0.1375	0.1174	0.0944	0.0834	0.0809	0.0798	0.0805
Agostino-Pearson	0.1362	0.1247	0.1068	0.0858	0.0766	0.0737	0.0722	0.0725
Chi Pearson	0.0623	0.0617	0.0592	0.0529	0.0493	0.0474	0.0463	0.0518
Filliben	0.1221	0.1134	0.0917	0.0716	0.0694	0.0643	0.0608	0.0626

Tabla 5.12: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 20 y diferentes grados de libertad (continuación).



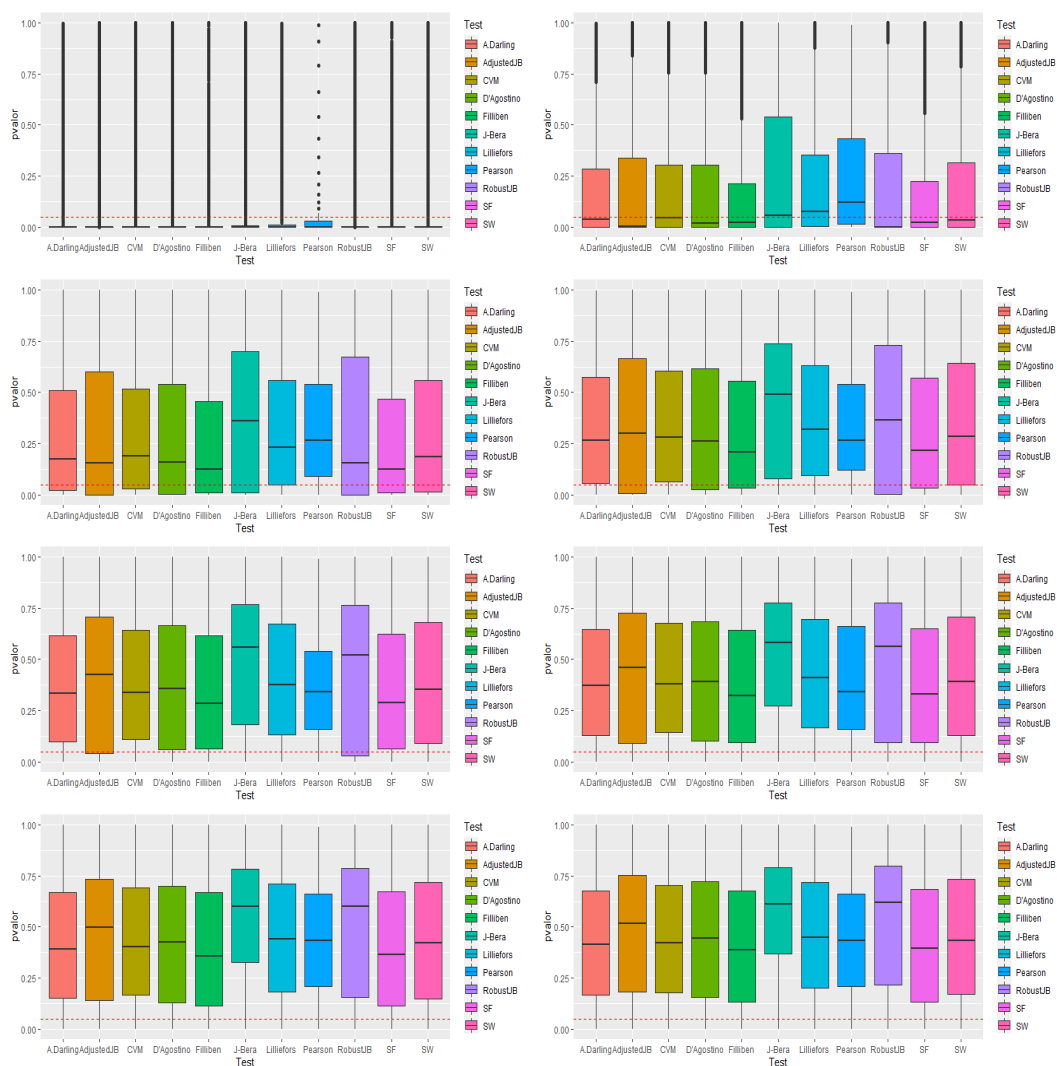


Figura 5.11: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 20 y diferentes grados de libertad.

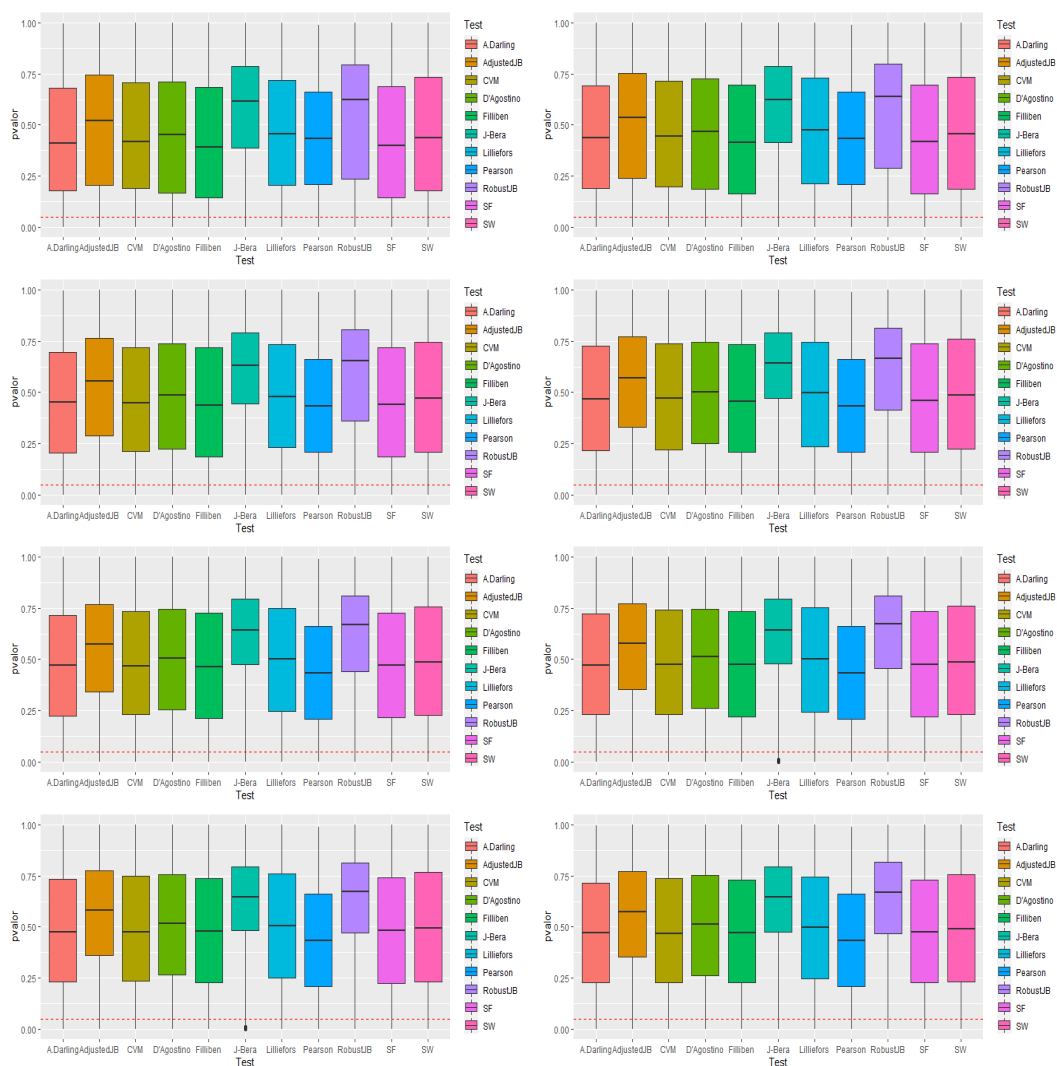


Figura 5.12: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 20 y diferentes grados de libertad(continuación).

**Tamaño de muestra n=50 y distintos grados de libertad**

	1	2	3	4	5	6	7	8
Lilliefors	0.9943	0.7736	0.4793	0.2973	0.2067	0.1551	0.1249	0.1036
Cramer Von Mises	0.9975	0.8404	0.5713	0.3752	0.2676	0.2020	0.1567	0.1370
Anderson Darling	0.9974	0.8557	0.6082	0.4091	0.2977	0.2320	0.1805	0.1547
Shapiro Wilk	0.9969	0.8586	0.6378	0.4583	0.3542	0.2862	0.2345	0.1994
Shapiro Francia	0.9976	0.8905	0.6998	0.5279	0.4134	0.3431	0.2822	0.2456
Jarque Bera	0.9945	0.8593	0.6610	0.4968	0.3921	0.3175	0.2632	0.2281
JB Robusto	0.9986	0.9153	0.7510	0.5866	0.4786	0.4018	0.3322	0.2932
Adjusted JB	0.9969	0.8901	0.7208	0.5617	0.4586	0.3861	0.3174	0.2797
Agostino-Pearson	0.9935	0.8561	0.6639	0.5039	0.4013	0.3387	0.2757	0.2422
Chi Pearson	0.9835	0.6418	0.3314	0.1927	0.1289	0.0987	0.0828	0.0761
Filliben	0.9977	0.8902	0.7012	0.5291	0.4145	0.3429	0.2831	0.2446

Tabla 5.13: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 50 y diferentes grados de libertad.

	9	10	15	20	25	30	35	40
Lilliefors	0.0975	0.0885	0.0670	0.0578	0.0570	0.0543	0.0526	0.0550
Cramer Von Mises	0.1179	0.1087	0.0797	0.0595	0.0641	0.0583	0.0552	0.0576
Anderson Darling	0.1324	0.1235	0.0850	0.0672	0.0664	0.0597	0.0571	0.0572
Shapiro Wilk	0.1726	0.1544	0.1119	0.0820	0.0795	0.0689	0.0642	0.0616
Shapiro Francia	0.2166	0.1931	0.1375	0.1029	0.0926	0.0826	0.0768	0.0739
Jarque Bera	0.2004	0.1752	0.1227	0.0891	0.0806	0.0665	0.0641	0.0601
JB Robusto	0.2599	0.2355	0.1674	0.1312	0.1181	0.1007	0.0955	0.0908
Adjusted JB	0.2501	0.2262	0.1625	0.1259	0.1122	0.0970	0.0937	0.0903
Agostino-Pearson	0.2144	0.1911	0.1401	0.1103	0.0997	0.0873	0.0852	0.0795
Chi Pearson	0.0679	0.0633	0.0608	0.0571	0.0532	0.0548	0.0533	0.0505
Filliben	0.2159	0.1914	0.1369	0.1003	0.0926	0.0802	0.0749	0.0721

Tabla 5.14: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 50 y diferentes grados de libertad (continuación).

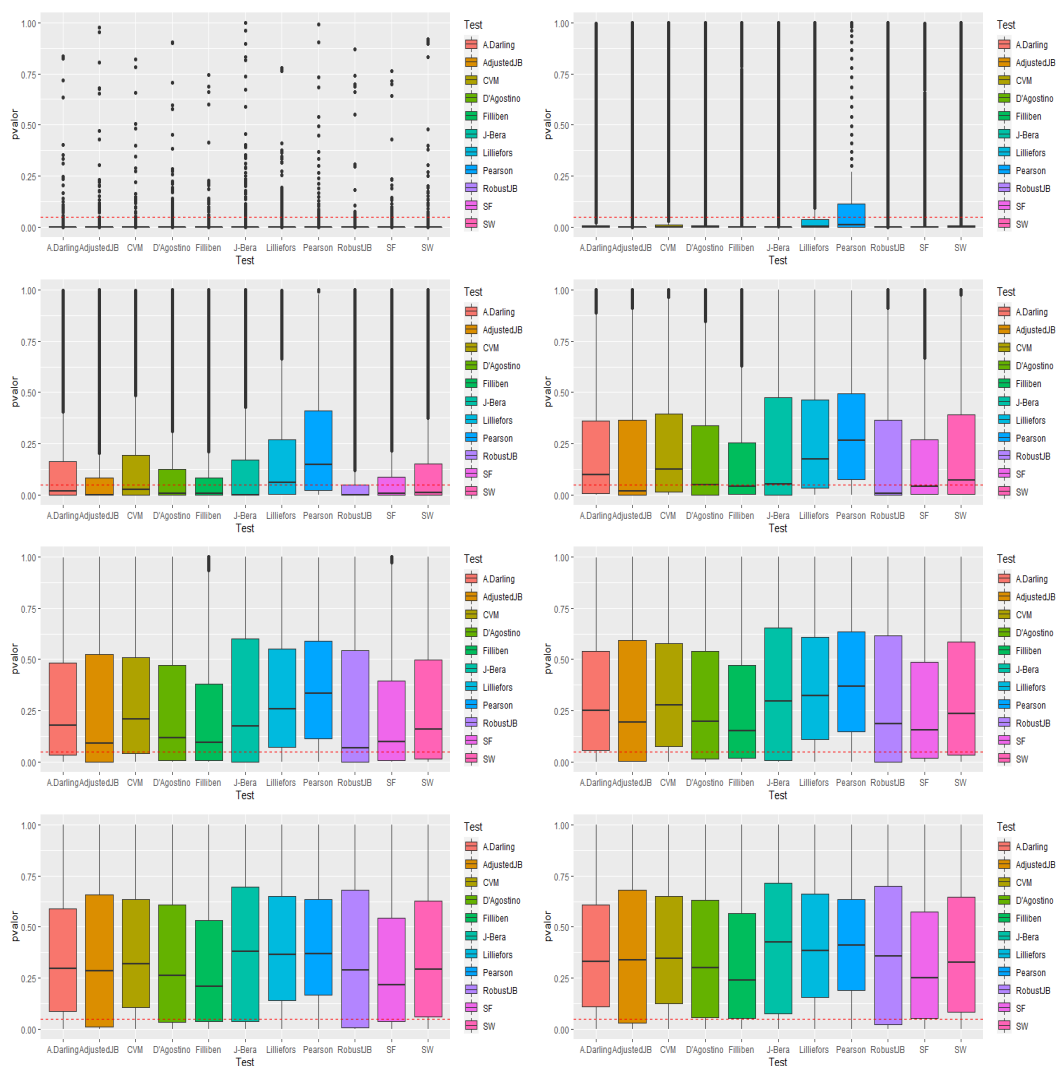


Figura 5.13: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 50 y diferentes grados de libertad.

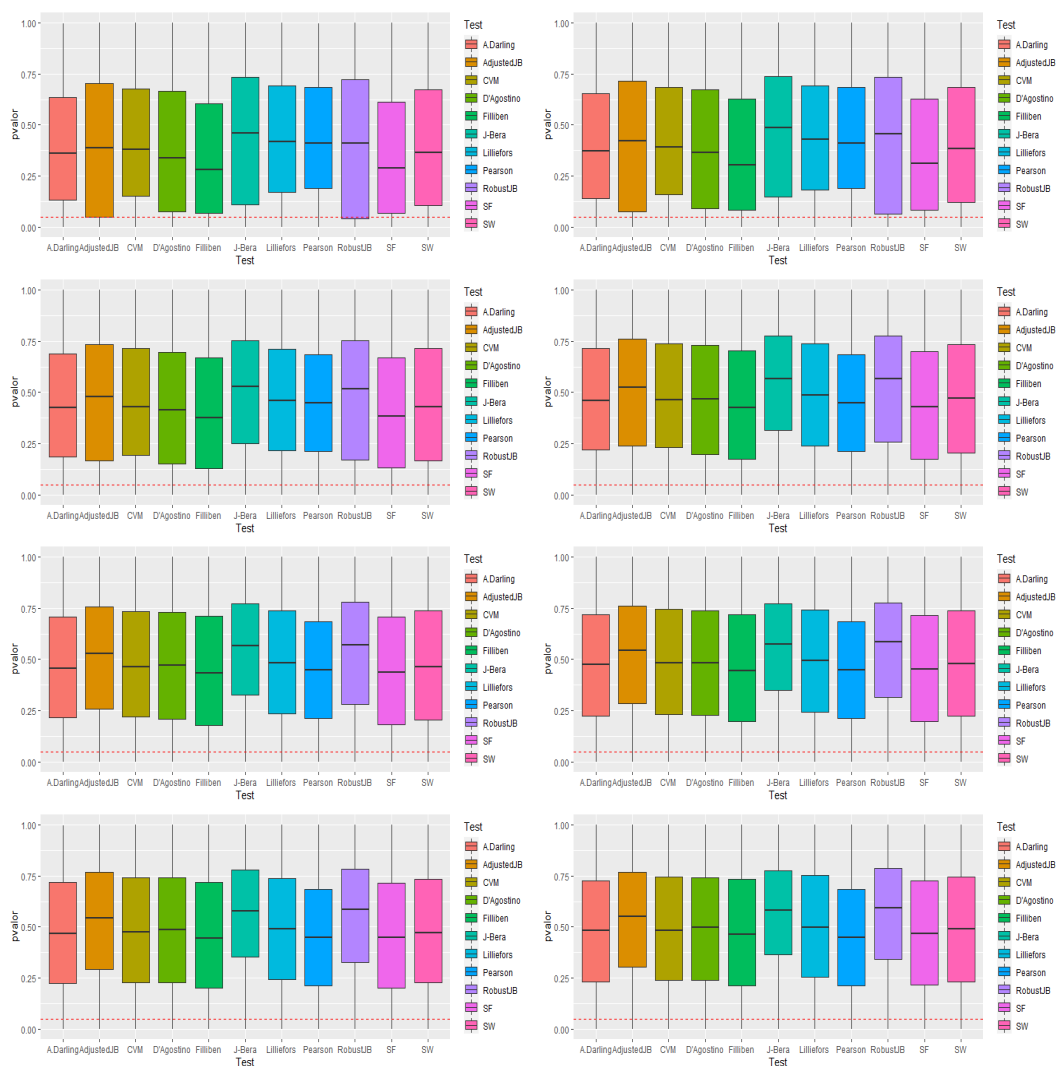


Figura 5.14: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 50 y diferentes grados de libertad (continuación).

### Tamaño de muestra n=100 y distintos grados de libertad

	1	2	3	4	5	6	7	8
Lilliefors	1.0000	0.9624	0.7295	0.4870	0.3375	0.2387	0.1827	0.1510
Cramer Von Mises	1.0000	0.9819	0.8249	0.5995	0.4347	0.3169	0.2504	0.1952
Anderson Darling	1.0000	0.9867	0.8514	0.6467	0.4827	0.3570	0.2902	0.2248
Shapiro Wilk	1.0000	0.9866	0.8763	0.7106	0.5614	0.4398	0.3743	0.3088
Shapiro Francia	1.0000	0.9903	0.9063	0.7689	0.6278	0.5141	0.4458	0.3737
Jarque Bera	1.0000	0.9862	0.8966	0.7566	0.6246	0.5141	0.4500	0.3751
JB Robusto	1.0000	0.9941	0.9313	0.8173	0.6897	0.5786	0.5082	0.4312
Adjusted JB	1.0000	0.9903	0.9145	0.7891	0.6649	0.5575	0.4945	0.4157
Agostino-Pearson	1.0000	0.9820	0.8772	0.7295	0.5982	0.4868	0.4246	0.3533
Chi Pearson	0.9999	0.8843	0.5238	0.2912	0.1716	0.1339	0.1091	0.0912
Filliben	1.0000	0.9911	0.9083	0.7715	0.6310	0.5171	0.4485	0.3763

Tabla 5.15: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 100 y diferentes grados de libertad.

	9	10	15	20	25	30	35	40
Lilliefors	0.1257	0.1084	0.0751	0.0692	0.0588	0.0606	0.0573	0.0558
Cramer Von Mises	0.1619	0.1412	0.0895	0.0760	0.0647	0.0638	0.0589	0.0558
Anderson Darling	0.1885	0.1640	0.0992	0.0803	0.0679	0.0657	0.0610	0.0587
Shapiro Wilk	0.2677	0.2347	0.1427	0.1116	0.0946	0.0859	0.0735	0.0709
Shapiro Francia	0.3275	0.2875	0.1823	0.1391	0.1121	0.1037	0.0886	0.0879
Jarque Bera	0.3296	0.2948	0.1837	0.1377	0.1176	0.0968	0.0841	0.0820
JB Robusto	0.3804	0.3386	0.2189	0.1711	0.1433	0.1192	0.1072	0.1047
Adjusted JB	0.3693	0.3307	0.2130	0.1664	0.1414	0.1182	0.1038	0.1001
Agostino-Pearson	0.3108	0.2789	0.1784	0.1365	0.1146	0.1032	0.0913	0.0868
Chi Pearson	0.0776	0.0732	0.0630	0.0584	0.0526	0.0527	0.0526	0.0466
Filliben	0.3299	0.2903	0.1834	0.1397	0.1204	0.1038	0.0871	0.0867

Tabla 5.16: Potencias de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 100 y diferentes grados de libertad (continuación).

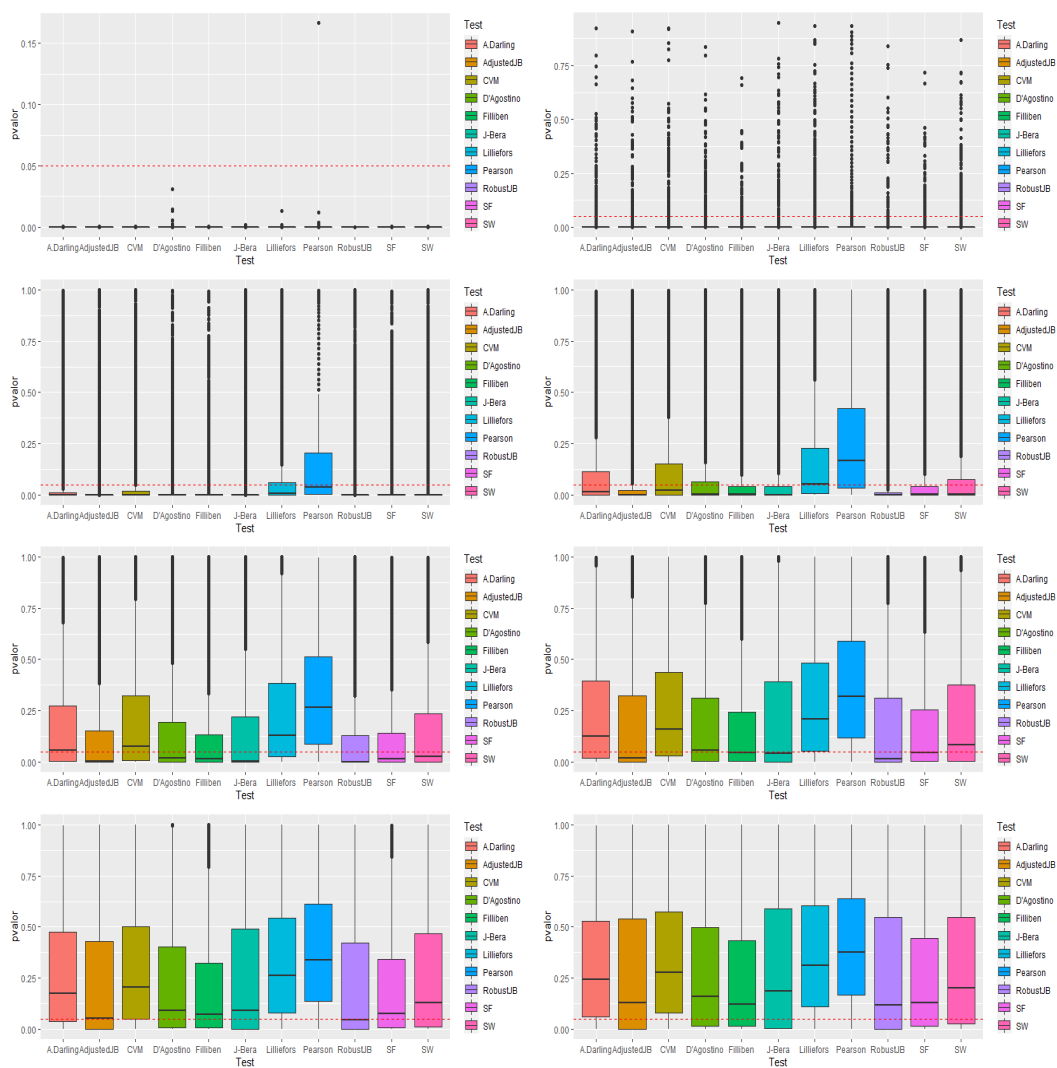


Figura 5.15: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 100 y diferentes grados de libertad.

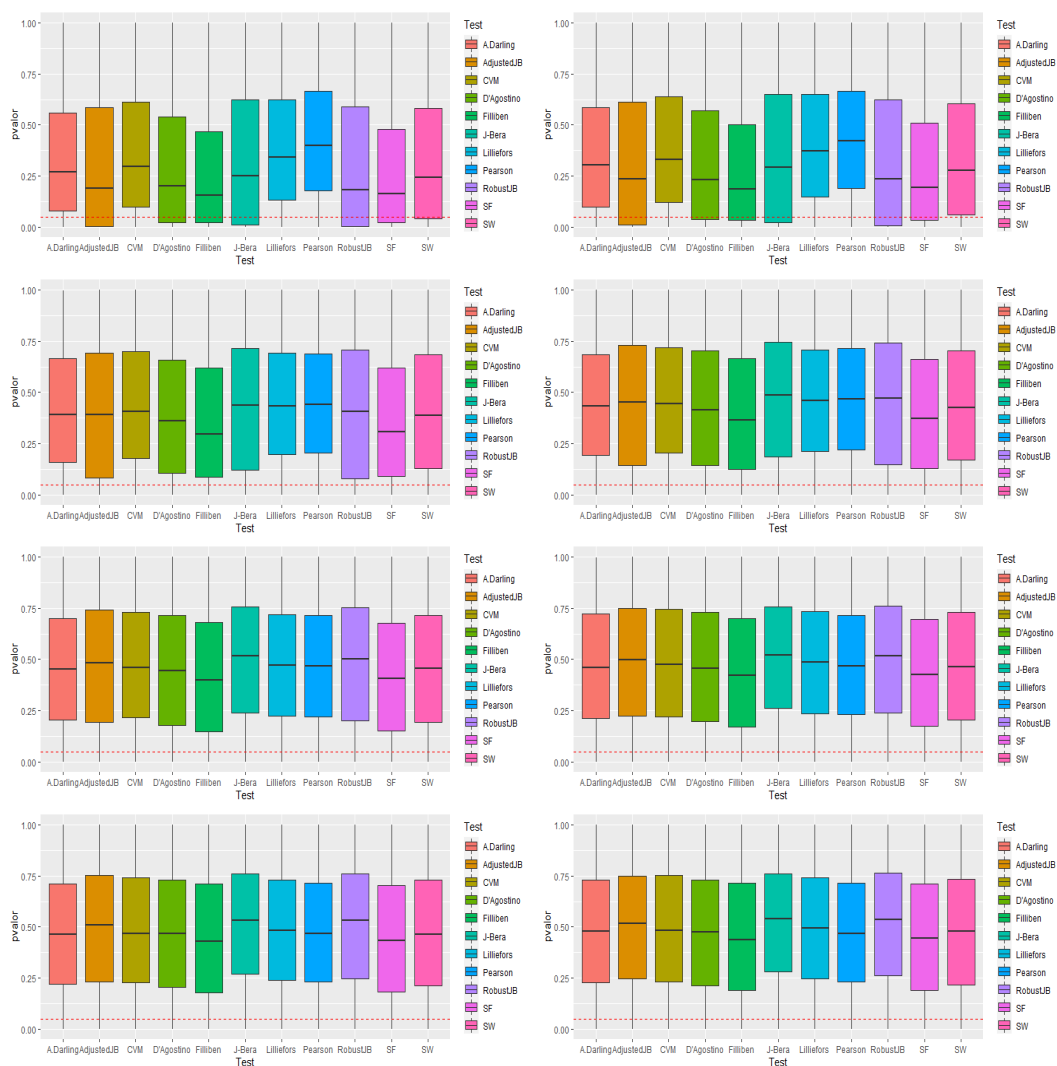


Figura 5.16: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa t de Student con tamaño de muestra 100 y diferentes grados de libertad (continuación).

Podemos observar claramente como para los tamaños de muestra  $n = 20$  y  $n = 50$  a medida que aumentamos los grados de libertad de la distribución, el comportamiento de las potencias de los test se va asemejando cada vez más al que tienen bajo la hipótesis nula de normalidad (va disminuyendo la potencia). Esto es intuitivo por la convergencia de la distribución t de Student a una distribución normal. Nótese que Jarque Bera Robusto y Adjusted JB son los test de mayor potencia, y chi-cuadrado y Lilliefors los de menor. Cabe destacar que para tamaño de muestra  $n = 100$  podemos ver cómo, a pesar de aumentar los grados de libertad, la potencia de los



test se asemeja menos o necesita más grados de libertad para asemejarse al comportamiento bajo la hipótesis nula de normalidad.

### 5.2.6. Distribución chi-cuadrado

Al igual que con la anterior distribución, también estudiaremos la potencia de los distintos test de normalidad para la  $\chi^2$ , variando los grados de libertad y con los anteriores tres tamaños muestrales.

#### Tamaño de muestra $n = 20$ y distintos grados de libertad

	1	2	3	4	5	6	7	8
Lilliefors	0.8807	0.5774	0.4105	0.3221	0.2612	0.2259	0.2034	0.1773
Cramer Von Mises	0.9552	0.7248	0.5406	0.4161	0.3381	0.2940	0.2564	0.2265
Anderson Darling	0.9705	0.7725	0.5928	0.4602	0.3748	0.3254	0.2830	0.2508
Shapiro Wilk	0.9855	0.8346	0.6627	0.5304	0.4358	0.3756	0.3312	0.2903
Shapiro Francia	0.9756	0.7975	0.6227	0.5005	0.4161	0.3572	0.3220	0.2816
Jarque Bera	0.7138	0.4799	0.3526	0.2815	0.2380	0.2084	0.1861	0.1595
JB Robusto	0.8171	0.6009	0.4676	0.3895	0.3334	0.2943	0.2752	0.2486
Adjusted JB	0.8136	0.6093	0.4728	0.4012	0.3414	0.3005	0.2836	0.2537
Agostino-Pearson	0.8090	0.6079	0.4737	0.4012	0.3437	0.3032	0.2839	0.2533
Chi Pearson	0.9540	0.6543	0.4124	0.2805	0.2155	0.1733	0.1485	0.1251
Filliben	0.9715	0.7832	0.6035	0.4820	0.4017	0.3436	0.3091	0.2712

Tabla 5.17: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 20 y diferentes grados de libertad.

	9	10	15	20	25	30	35	40
Lilliefors	0.1659	0.1529	0.1202	0.0960	0.0879	0.0792	0.0739	0.0690
Cramer Von Mises	0.2035	0.1889	0.1383	0.1144	0.1009	0.0909	0.0850	0.0813
Anderson Darling	0.2234	0.2127	0.1537	0.1237	0.1103	0.0963	0.0906	0.0866
Shapiro Wilk	0.2597	0.2445	0.1815	0.1387	0.1264	0.1090	0.1003	0.0907
Shapiro Francia	0.2541	0.2380	0.1785	0.1416	0.1288	0.1126	0.1038	0.0941
Jarque Bera	0.1419	0.1335	0.1082	0.0826	0.0738	0.0640	0.0602	0.0519
JB Robusto	0.2233	0.2111	0.1741	0.1403	0.1310	0.1164	0.1087	0.0995
Adjusted JB	0.2311	0.2159	0.1772	0.1450	0.1334	0.1174	0.1110	0.1036
Agostino-Pearson	0.2303	0.2155	0.1747	0.1429	0.1296	0.1154	0.1092	0.0975
Chi Pearson	0.1196	0.1158	0.0833	0.0714	0.0667	0.0621	0.0622	0.0643
Filliben	0.2460	0.2269	0.1733	0.1355	0.1227	0.1070	0.0993	0.0900

Tabla 5.18: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 20 y diferentes grados de libertad (continuación).

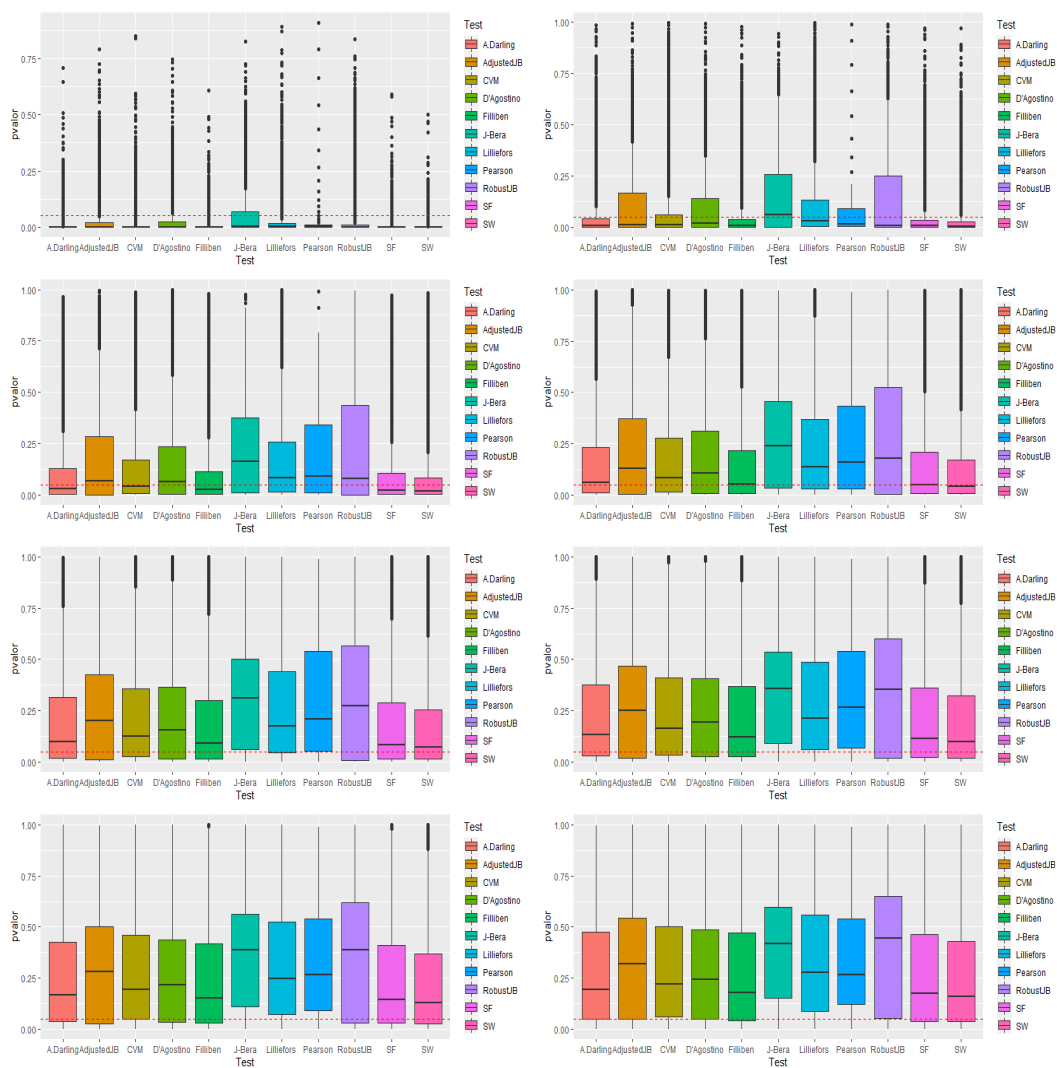


Figura 5.17: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 20 y diferentes grados de libertad.

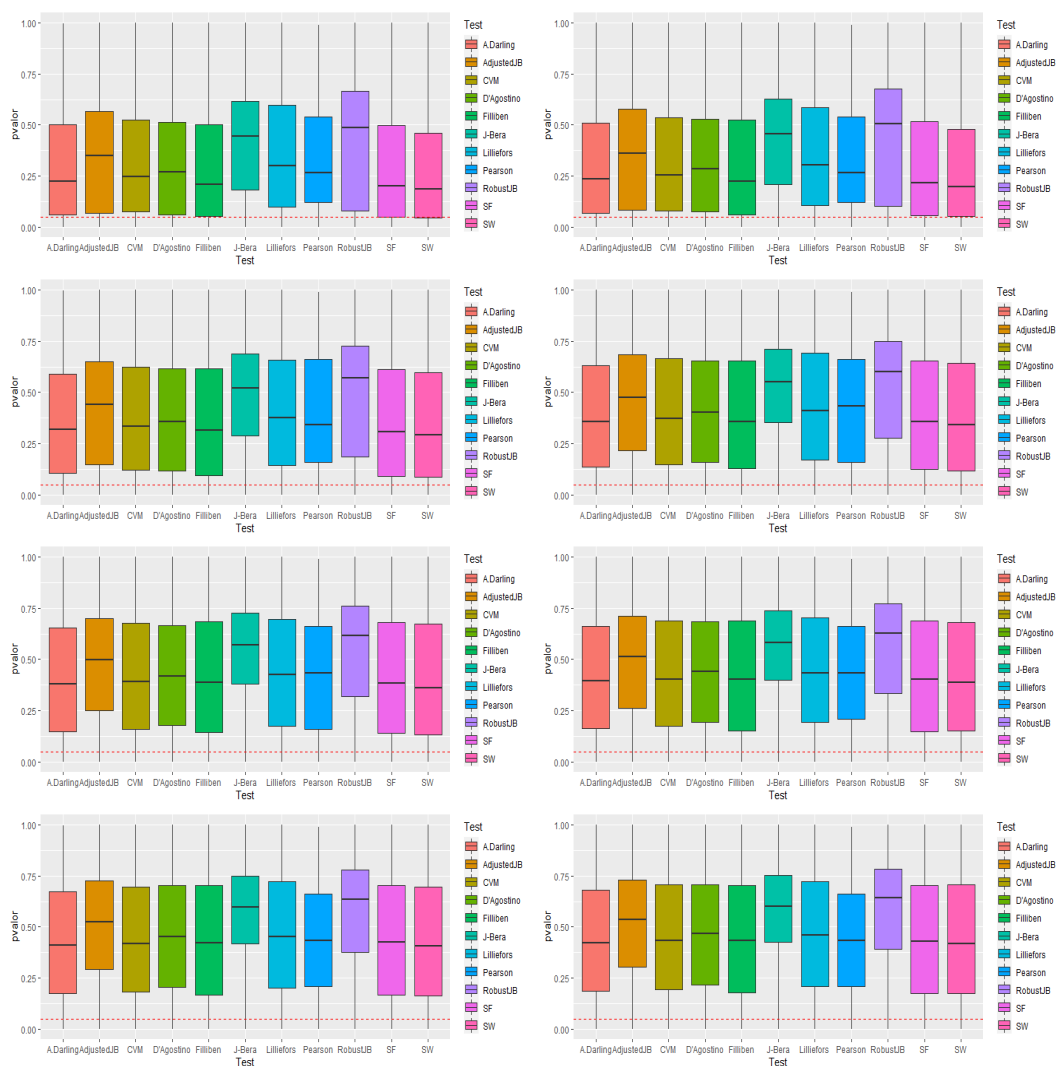


Figura 5.18: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 20 y diferentes grados de libertad (continuación).

### Tamaño de muestra $n = 50$ y distintos grados de libertad

	1	2	3	4	5	6	7	8	9
Lilliefors	0.9999	0.9644	0.8274	0.6827	0.5905	0.5061	0.4052	0.4041	0.3544
Cramer Von Mises	1.0000	0.9908	0.9337	0.8330	0.7397	0.6444	0.5799	0.5184	0.4593
Anderson Darling	1.0000	0.9966	0.9670	0.8877	0.8053	0.7156	0.6443	0.5807	0.5201
Shapiro Wilk	1.0000	0.9993	0.9893	0.9480	0.8904	0.8201	0.7557	0.6947	0.6346
Shapiro Francia	1.0000	0.9983	0.9813	0.9242	0.8564	0.7837	0.7155	0.6582	0.5998
Jarque Bera	0.9980	0.9335	0.8628	0.7590	0.6721	0.5997	0.5420	0.4926	0.4496
JB Robusto	0.9957	0.9465	0.8612	0.7690	0.6927	0.6262	0.5759	0.5329	0.4916
Adjusted JB	0.9990	0.9673	0.8912	0.8022	0.7215	0.6536	0.5069	0.5493	0.5080
Agostino-Pearson	0.9985	0.9627	0.8837	0.7944	0.7159	0.6480	0.5903	0.5446	0.5038
Chi Pearson	1.0000	0.9851	0.8546	0.6471	0.5014	0.3842	0.3123	0.2676	0.2341
Filliben	1.0000	0.9980	0.9769	0.9166	0.8435	0.7690	0.7007	0.6417	0.5851

Tabla 5.19: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 50 y diferentes grados de libertad.

	10	15	20	25	30	35	40	50	60
Lilliefors	0.3387	0.2350	0.1911	0.1567	0.1351	0.1188	0.1159	0.1014	0.0909
Cramer Von Mises	0.4331	0.3013	0.2360	0.1877	0.1609	0.1407	0.1328	0.1185	0.1047
Anderson Darling	0.4908	0.3409	0.2648	0.2095	0.1795	0.1586	0.1489	0.1307	0.1135
Shapiro Wilk	0.5950	0.4298	0.3345	0.2681	0.2318	0.2019	0.1847	0.1626	0.1409
Shapiro Francia	0.5620	0.4064	0.3201	0.2601	0.2293	0.1942	0.1804	0.1602	0.1417
Jarque Bera	0.4231	0.3002	0.2436	0.1990	0.1770	0.1525	0.1400	0.1202	0.1102
JB Robusto	0.4655	0.3474	0.2894	0.2399	0.2174	0.1897	0.1739	0.1549	0.1401
Adjusted JB	0.4791	0.3576	0.2901	0.2424	0.2173	0.1899	0.1753	0.1541	0.1382
Agostino-Pearson	0.4789	0.3590	0.2935	0.2447	0.2210	0.1927	0.1777	0.1540	0.1406
Chi Pearson	0.2099	0.1379	0.1179	0.0970	0.0880	0.0850	0.0763	0.0734	0.0689
Filliben	0.5457	0.3929	0.3101	0.2499	0.2218	0.1888	0.1736	0.1548	0.1364

Tabla 5.20: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 50 y diferentes grados de libertad (continuación).

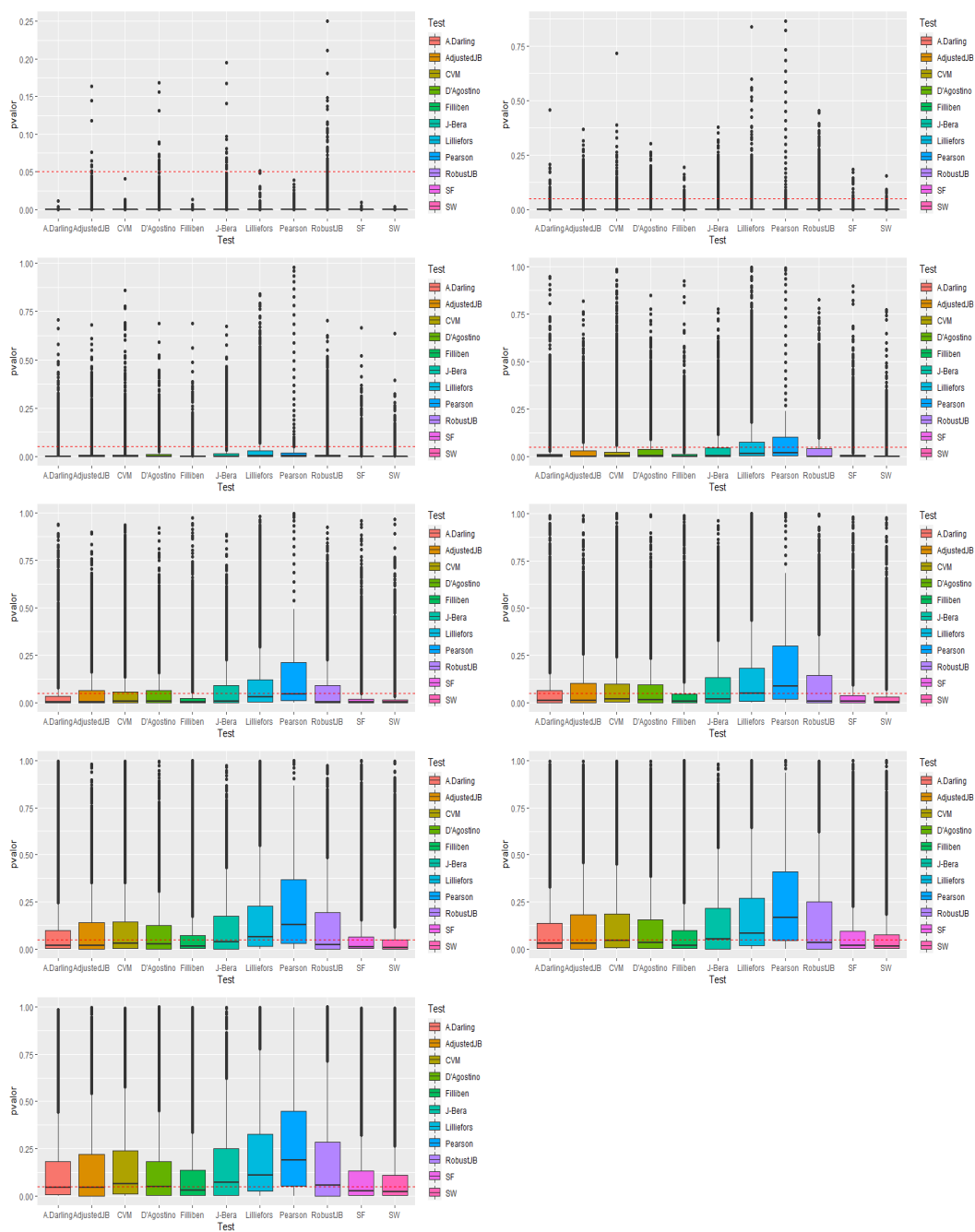


Figura 5.19: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 50 y diferentes grados de libertad.

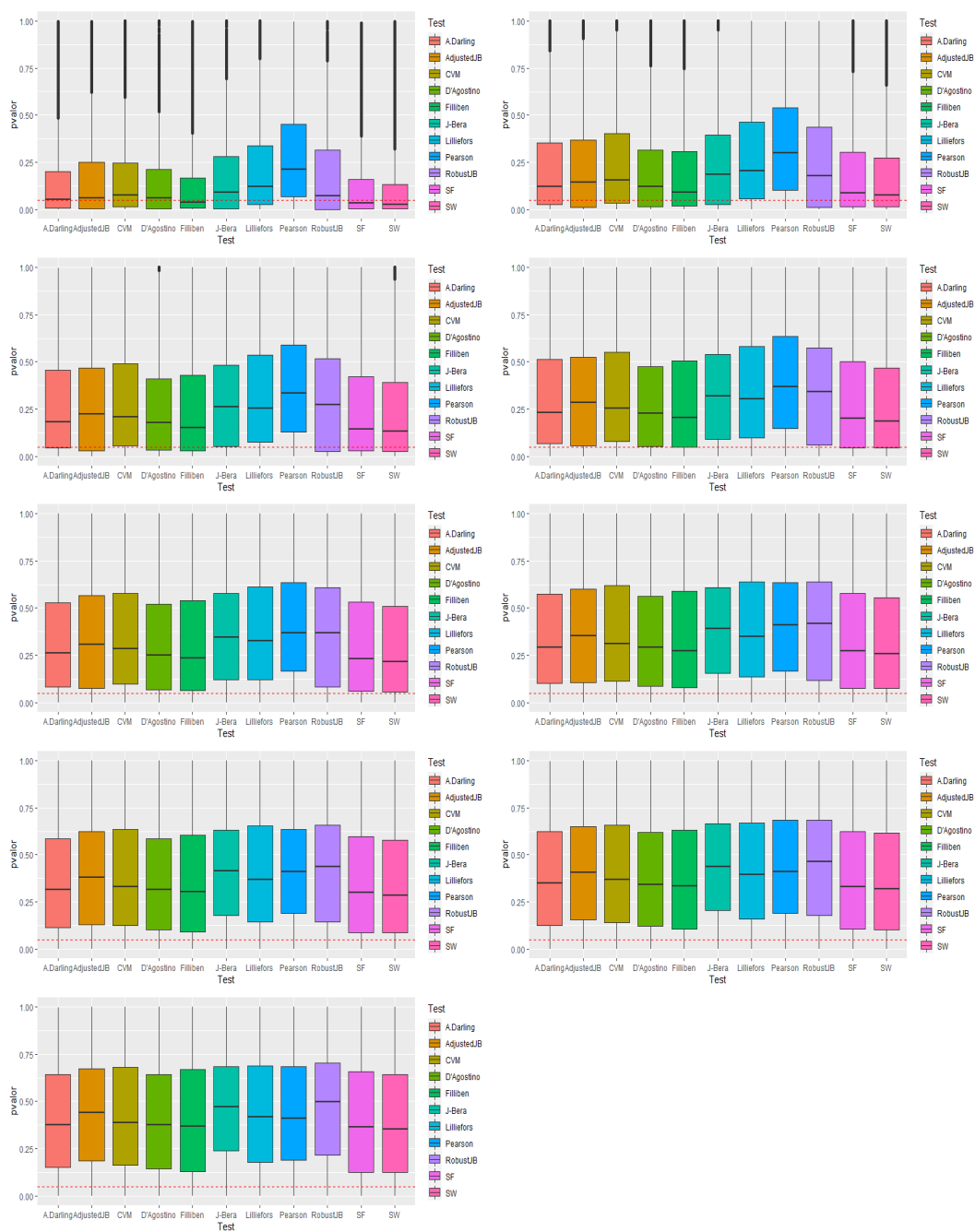


Figura 5.20: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 50 y diferentes grados de libertad (continuación).

**Tamaño de muestra  $n = 100$  y distintos grados de libertad**

	1	2	3	4	5	6	7	8	9
Lilliefors	1.0000	0.9996	0.9912	0.9503	0.8926	0.8239	0.7633	0.7068	0.6484
Cramer Von Mises	1.0000	1.0000	0.9986	0.9920	0.9674	0.9287	0.8907	0.8335	0.7819
Anderson Darling	1.0000	1.0000	0.9994	0.9974	0.9859	0.9643	0.9349	0.8920	0.8511
Shapiro Wilk	1.0000	1.0000	1.0000	0.9998	0.9980	0.9923	0.9775	0.9580	0.9360
Shapiro Francia	1.0000	1.0000	1.0000	0.9994	0.9959	0.9862	0.9671	0.9405	0.9139
Jarque Bera	1.0000	0.9999	0.9980	0.9912	0.9735	0.9433	0.9103	0.8612	0.8245
JB Robusto	1.0000	0.9998	0.9962	0.9835	0.9609	0.9262	0.8940	0.8460	0.8093
Adjusted JB	1.0000	1.0000	0.9986	0.9932	0.9773	0.9517	0.9218	0.8771	0.8432
Agostino-Pearson	1.0000	0.9999	0.9989	0.9929	0.9775	0.9523	0.9204	0.8751	0.8421
Chi Pearson	1.0000	1.0000	0.9954	0.9482	0.8348	0.7148	0.6075	0.5045	0.4403
Filliben	1.0000	1.0000	0.9999	0.9994	0.9951	0.9843	0.9631	0.9324	0.9045

Tabla 5.21: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 100 y diferentes grados de libertad.

	10	15	20	25	30	35	40	50	60
Lilliefors	0.6010	0.4359	0.3358	0.2783	0.2488	0.2120	0.1930	0.1577	0.1388
Cramer Von Mises	0.7403	0.5489	0.4212	0.3508	0.3000	0.2547	0.2291	0.1825	0.1643
Anderson Darling	0.8062	0.6166	0.4770	0.3980	0.3407	0.2921	0.2603	0.2063	0.1839
Shapiro Wilk	0.9014	0.7397	0.5980	0.5115	0.4444	0.3835	0.3386	0.2858	0.2446
Shapiro Francia	0.8771	0.7074	0.5698	0.4858	0.4251	0.3701	0.3230	0.2730	0.2348
Jarque Bera	0.7821	0.6032	0.4823	0.4087	0.3577	0.3094	0.2671	0.2311	0.1997
JB Robusto	0.7717	0.6073	0.4974	0.4253	0.3808	0.3333	0.2904	0.2536	0.2224
Adjusted JB	0.8051	0.6321	0.5140	0.4382	0.3888	0.3408	0.2965	0.2564	0.2242
Agostino-Pearson	0.7989	0.6328	0.5123	0.4407	0.3897	0.3426	0.2990	0.2569	0.2263
Chi Pearson	0.3795	0.2327	0.1719	0.1435	0.1232	0.1100	0.1040	0.0833	0.0859
Filliben	0.8671	0.6929	0.5553	0.4737	0.4131	0.3592	0.3121	0.2621	0.2286

Tabla 5.22: Potencias de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 100 y diferentes grados de libertad (continuación).

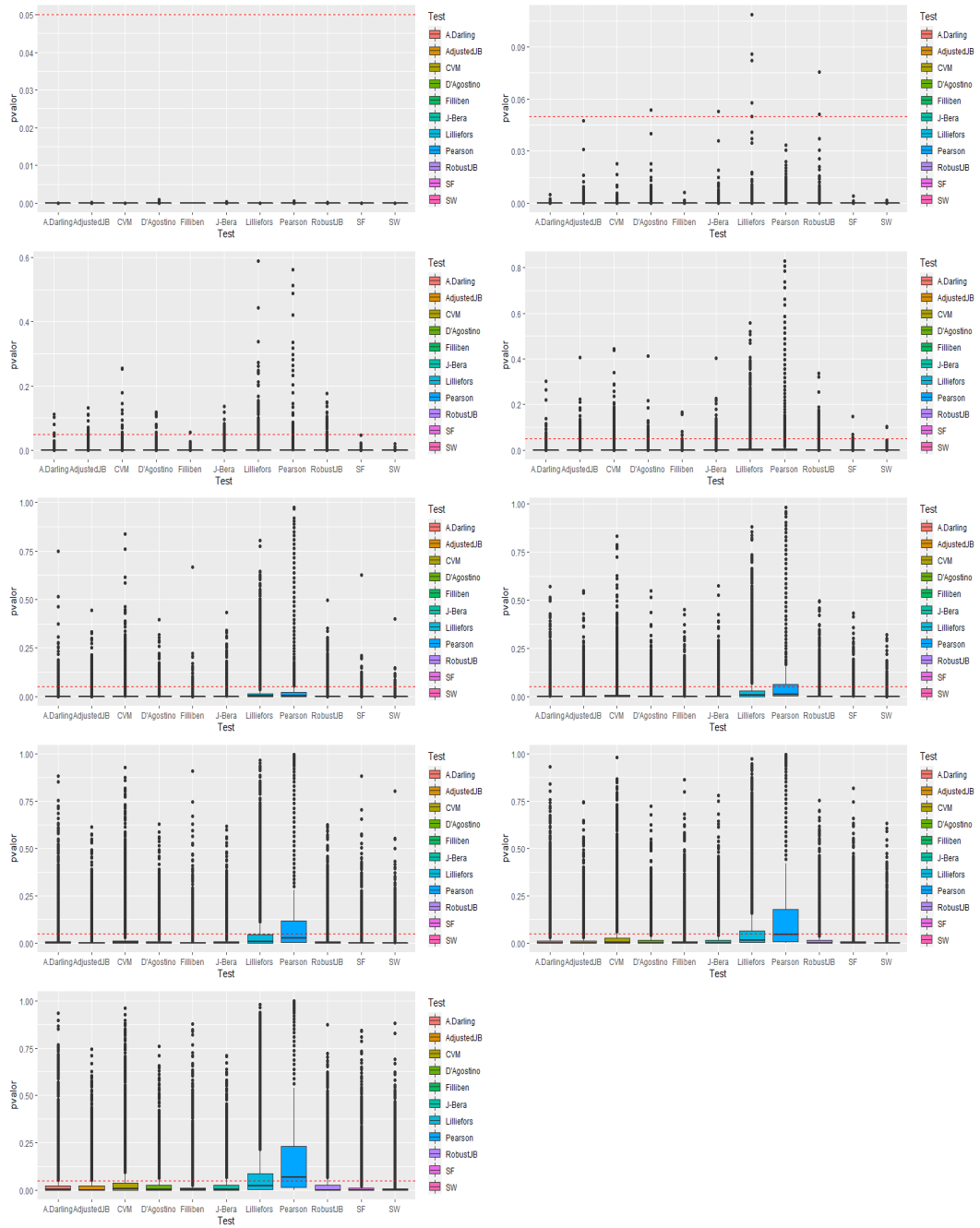


Figura 5.21: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 100 y diferentes grados de libertad.



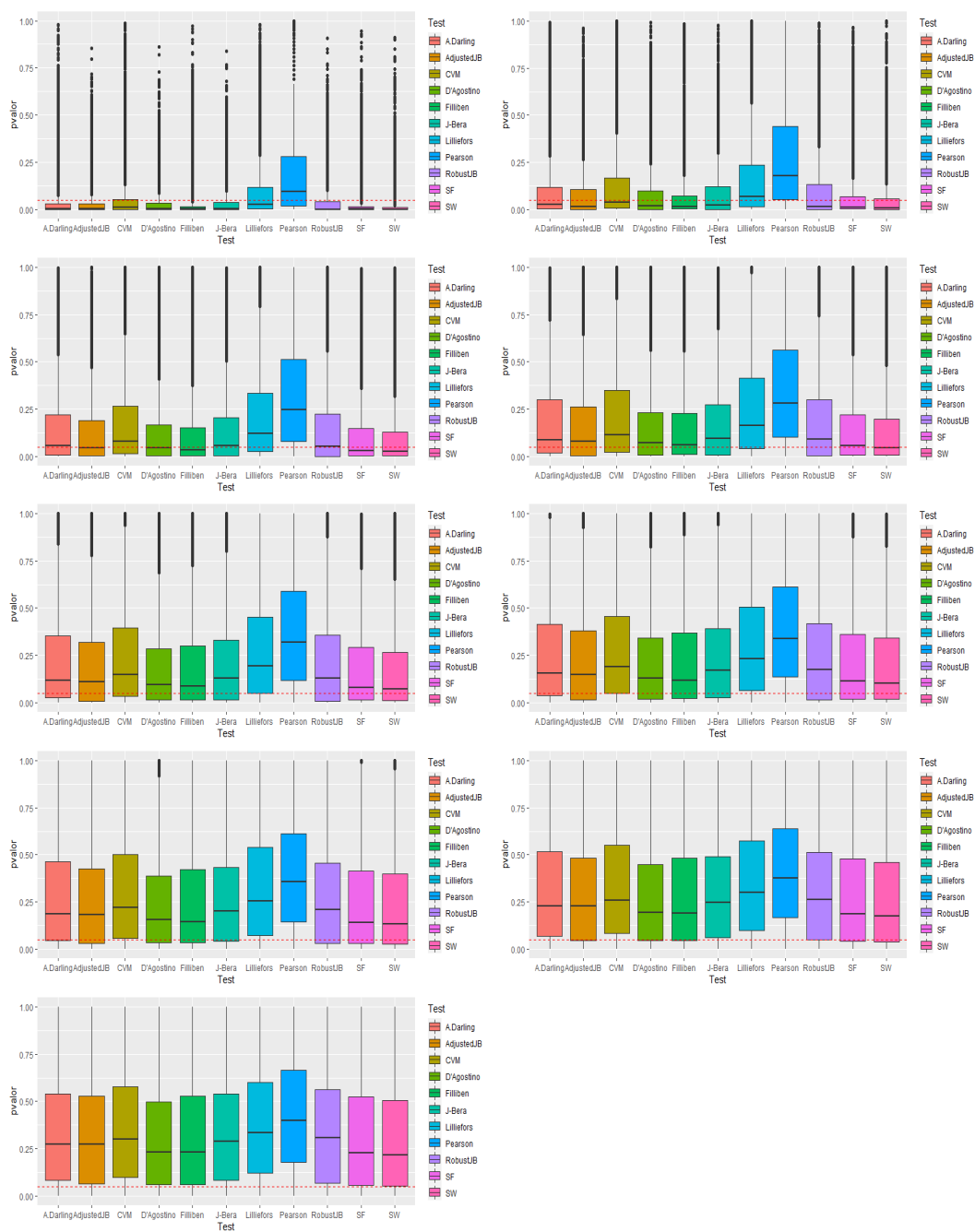


Figura 5.22: Distribución de los p-valores de las pruebas de normalidad bajo distribución alternativa chi-cuadrado con tamaño de muestra 100 y diferentes grados de libertad (continuación).

Al igual que ocurría en el caso anterior, la potencia de los test de normalidad bajo distribución alternativa chi-cuadrado se va asemejando cada

vez más a la potencia bajo la hipótesis nula de normalidad a medida que vamos aumentando los grados de libertad de la distribución, aunque podemos ver que esta vez es de forma más lenta que en el caso de la  $t$  de Student. De nuevo, esto se debe a la convergencia asintótica de la distribución chi-cuadrado a la distribución normal. Nótese que el test de mayor potencia para  $n = 20$  es Adjusted JB, seguido de JB Robusto y Agostino-Pearson, mientras que los de menor potencia son JB y chi-cuadrado. En el caso de  $n = 50$  Agostino-Pearson y JB Robusto son los de menor potencia, manteniéndose chi-cuadrado como el de menor potencia seguido de Lilliefors. Para el tamaño muestral  $n = 100$ , cabe destacar que la similitud de la potencia con la de la normal se hace notar de forma mucho más leve con respecto a lo que ocurría con muestras más pequeñas a pesar de aumentar los grados de libertad. El test chi-cuadrado sigue siendo de forma clara el de menor potencia, y destaca Shapiro Wilk como el de mayor potencia.

## Capítulo 6

# Conclusiones

Este Trabajo de Fin de Grado utiliza elementos teóricos de Estadística, junto con otros de corte más práctico como, por ejemplo, simulaciones realizadas con el programa RStudio. A partir de los conceptos y distribuciones que se explican en el Capítulo 2, se asienta la base de la experimentación que se realizó en los siguientes capítulos, aplicando los métodos gráficos y pruebas de normalidad.

Con respecto a los métodos gráficos, se concluye que el diagrama de violín es el más completo, ya que recoge la distribución del estadístico y además permite incluir en su interior un diagrama de cajas. Estos diagramas de violín son muy fáciles de manejar debido a que a simple vista nos permiten decidir si la distribución de los datos se asemeja lo suficiente a la de una normal. Otros métodos gráficos que también permiten la interpretación de una manera sencilla son los QQ-plot, los PP-plot, así como las curvas de Lilliefors.

Como conclusión principal con respecto a los test de normalidad, se tiene que por regla general los test de mayor potencia dentro de los estudiados (y por ello los más recomendables) son Jarque Bera Robusto y Adjusted Jarque Bera, seguidos de Agostino-Pearson, Shapiro-Wilk y Shapiro-Francia. Por otro lado, el test que tiene menor potencia tiende a ser el test chi-cuadrado de Pearson. Se destaca que en el caso de trabajar con la distribución exponencial, la potencia de todos los test se incrementa de forma significativa, debido a la clara asimetría de la misma. Tal y como era de esperar, bajo la hipótesis de normalidad, resulta indiferente el hecho de utilizar un test u otro para el contraste, puesto que todos poseen potencias muy similares, cercanas al nivel de significación fijado. Se ha observado también que para las distribuciones t de Student y  $\chi^2$ , a medida que los grados de libertad de las mismas incrementan, el comportamiento de la potencia tiende cada vez más a parecerse al que se tiene bajo la hipótesis nula de normalidad. Esto es de esperar debido a la convergencia de ambas distribuciones a una distribución normal. En el caso de la mixtura gaussiana, se comprueba que la poten-

cia de los test es más alta para las simulaciones que usan probabilidades en torno al 0.5 y más baja para las que usan probabilidades extremas como 0.01 y 0.02. Se llegan a alcanzar potencias máximas de 1, especialmente para muestras grandes. Esto es intuitivo, puesto que probabilidad pequeñas indican que la mayoría de los datos provienen de una única distribución normal. Otra observación a tener en cuenta viene a la hora de trabajar con muestras grandes. Las pruebas de contraste son muy sensibles y no toleran bien las desviaciones en la normalidad, por ende si tenemos una muestra grande que presente esta característica es muy posible que se rechace la hipótesis nula de normalidad, debido simplemente al ruido inherente al proceso de medición de la variable. Asimismo, si la muestra es muy pequeña, el test puede fallar a la hora de detectar la normalidad. Se concluye a raíz de esto que es aconsejable complementar el uso de tests de normalidad con alguno de los métodos gráficos.

# Bibliografía

- [1] Blom, G., *Statistical Estimates and Transformed Beta-Variables*. John Wiley and Sons. New York (1958).
- [2] Brzezinski, M., The Chen–Shapiro test for normality, *The Stata Journal: Promoting communications on statistics and Stata*, 12(3), (2012) pp.368-374.
- [3] Castillo Gutiérrez, S., Lozano Aguilera, E.D., Q-Q Plot normal. Los puntos de posición gráfica. *Iniciación a la investigación* (2010).
- [4] Chen, L., Shapiro, S., An alternative test for normality based on normalized spacings. *Journal of Statistical Computation and Simulation*, 53(3-4), (1995), pp.269-287.
- [5] Cotterill, D., Csorgo, M., On the limiting distribution of and critical values for the multivariate Cramer-Von Mises statistic. *The Annals of Statistics*, 10(1), (1982).
- [6] Cramér, H., On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1), (1928), pp.141-180.
- [7] Elmore, K., Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and Forecasting*, 20(5), (2005), pp.789-795.
- [8] Filliben, J., The probability plot correlation coefficient test for normality. *Technometrics*, 17(1), (1975), pp.111-117.
- [9] Frozini, B.V., “On the distribution and power of a goodness-of-fit statistic with parametric and nonparametric applications”. *Goodness-of-fit*, (Ed. by Revesz P., Sarkadi K., Sen P.K.) Amsterdam-Oxford-New York: North-Holland, pp.133-154, (1987).
- [10] Geary, R., The ratio of the mean deviation to the standard deviation as a test of normality. *Biometrika*, 27(3-4), (1935), pp.310-332.
- [11] Gel, Y., Gastwirth, J., A robust modification of the Jarque–Bera test of normality. *Economics Letters*, 99(1), (2008), pp.30-32.

- [12] Hintze, J., Nelson, R., Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2), (1998), pp.181-184.
- [13] Hogg, R., McKean, J., Craig, A., Introduction to Mathematical Statistics. 8th ed. Pearson Education. New York (2018).
- [14] Hubert, M., Vandervieren, E., An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12),(2008), pp.5186-5201.
- [15] Iman, R., Graphs for use with the Lilliefors test for normal and exponential distributions. *The American Statistician*, 36(2), (1982), pp.109-112.
- [16] Jarque, C., Bera, A., A Test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2), (1987), pp.163-172
- [17] Kolmogorov, A.N., Sulla Determinazione Empirica di Una Legge di Distribuzione, *Giornale dell'Istituto Italiano degli Attuari*, 4. 83-91 (1933).
- [18] Lewis, P., Distribution of the Anderson-Darling statistic. *The Annals of Mathematical Statistics*, 32(4), (1961), pp.1118-1124.
- [19] Lilliefors, H., On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318),(1967), pp.399-402.
- [20] Parzen, E., On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), (1962), pp.1065-1076.
- [21] Pearson, K., On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is such that It Can Be Reasonably Supposed to Have Arisen from Random Sampling. *Philosophical Magazine, Series 5*, 50(302), (1900), pp.157 – 175.
- [22] Ralph, B., D'agostino, R., Belanger, A., D'agostino, Jr., A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4), (1990), pp.316-321.
- [23] Reynolds, D.A., Gaussian mixture models. *Encyclopedia of biometrics*, 741, (2009), pp.659-663.
- [24] Shapiro, S., Francia, R., An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67(337), ( 1972), pp.215-216.

- [25] Shapiro, S., Wilk, M., Chen, H., A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63(324), (1968), pp.1343-1372.
- [26] Sheskin, D.J., *Handbook of Parametric and Nonparametric Statistical Procedures*. 3rd edition. Chapman and Hall/CRC, New York (2003).
- [27] Smirnov, N., Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2),(1948), pp.279-281.
- [28] Stephens, M.A., “Tests Based on EDF Statistics.” In RB D’Agostino, MA Stephens (eds.), *Goodness-of-Fit Techniques*, volume 68 of *Statistics: a Series of Textbooks and Monographs*, 1 edition, chapter 3, pp. 97–185. Marcel Dekker (1986).
- [29] Torabi, H., Montazeri, N.H., Grané, A., A test for normality based on the empirical distribution function. *SORT-Statistics and Operations Research Transactions*, 1, (2016), pp.55-88.
- [30] Tukey, J.W., The future of data analysis. *The Annals of Mathematical Statistics*, 33(1), (1962), pp.1-67.
- [31] Tukey, J.W., *Exploratory Data Analysis*, 1st ed. Addison-Wesley Publishing Co, Reading (1977).
- [32] Tukey, J.W., *Exploratory Data Analysis (Limited Preliminary Edition)*, Addison-Wesley Publishing Co, Reading. (1970).
- [33] Uyanto, S.S., An extensive comparisons of 50 univariate goodness-of-fit tests for normality. *Austrian Journal of Statistics*, 51(3), pp.45–97, (2022).
- [34] von Mises, R., *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik [Probability Calculus and Its Application in Statistics and Theoretical Physics]*. F. Deuticke, Leipzig (1931).
- [35] Wilk, M.B., Gnanadesikan, R., Probability plotting methods for the analysis of data. *Biometrika*, 55(1), (1968), pp 1-17.
- [36] Yu, B., Lemeshko, S., About the effect of rounding on the properties of tests for testing statistical hypotheses. *Journal of Physics: Conference Series*, 1715(1), (2021), p.012063.

# Apéndice A

## Apéndice

Las Tablas A.1, A.2 y A.3 recogen los valores que conforman cada una de las muestras introducidas al inicio del Capítulo 3.

Índice	Normal	Uniforme	Exponencial	Doble exponencial	Mixtura de normales
1	0.20	2.88	4.22	7.22	-9.06
2	1.85	7.88	2.88	5.88	6.28
3	10.79	4.09	6.65	-3.65	3.10
4	3.35	8.83	0.16	2.84	-2.04
5	3.65	9.40	0.28	2.72	-10.95
6	11.58	0.46	1.58	4.58	6.61
7	5.30	5.28	1.57	4.57	-7.62
8	-3.33	8.92	0.73	3.73	-25.18
9	-0.43	5.51	13.63	16.63	-4.42
10	0.77	4.57	0.15	2.85	-15.67
11	9.12	9.57	5.02	-2.02	-2.57
12	4.80	4.53	2.40	0.60	-13.02
13	5.00	6.78	1.41	1.49	-20.27
14	3.55	5.73	1.89	4.89	2.44
15	0.22	1.03	0.94	3.94	9.40
16	11.93	9.00	4.25	-1.25	11.98
17	5.49	2.46	7.82	-4.82	5.83
18	-6.83	0.42	2.39	5.39	-11.53
19	6.51	3.28	2.95	5.95	-7.99
20	0.64	9.55	20.21	-17.21	-13.54

Tabla A.1: Tabla que recoge los datos de las muestras.



Índice	Normal	Uniforme	Exponencial	Doble exponencial	Mixtura de normales
21	-2.34	8.90	4.22	-1.22	2.10
22	1.91	6.93	4.83	7.83	-11.76
23	-2.13	6.41	7.43	10.43	-6.92
24	-0.64	9.94	6.74	-3.74	-2.66
25	-0.13	6.56	5.84	-2.84	29.10
26	-5.43	7.09	8.03	11.03	-11.00
27	7.19	5.44	7.48	-4.48	8.63
28	3.77	5.94	7.85	10.85	4.55
29	-2.69	2.89	0.16	3.16	-14.27
30	9.27	1.47	2.99	5.99	5.06
31	5.13	9.63	10.84	13.84	25.73
32	1.52	9.02	2.53	0.47	9.03
33	7.48	6.91	1.30	4.30	5.86
34	7.39	7.95	12.98	15.98	-1.12
35	7.11	0.25	6.15	9.15	-25.74
36	6.44	4.78	3.95	-0.95	21.26
37	5.77	7.58	3.25	6.15	-17.52
38	2.69	2.16	6.27	9.27	13.94
39	1.47	3.18	2.94	5.94	30.87
40	1.10	2.32	5.65	-2.65	-19.61
41	-0.47	1.43	2.10	0.90	11.79
42	1.96	4.15	36.06	39.06	-1.45
43	-3.33	4.14	4.23	7.23	-23.75
44	13.84	3.69	1.13	4.23	-14.30
45	9.04	1.52	5.50	8.50	-18.00
46	-2.62	1.39	11.24	-8.24	-7.77
47	0.99	2.33	6.82	9.82	-19.93
48	0.67	4.66	2.88	0.12	12.15
49	6.90	2.66	13.63	16.63	36.45
50	2.58	8.58	6.56	-3.56	-16.51
51	4.27	0.46	0.45	2.55	15.45
52	2.86	4.42	1.53	1.47	14.46
53	2.79	7.99	5.34	8.34	7.88
54	9.84	1.22	1.57	4.57	-8.70
55	1.87	5.61	4.87	-1.87	0.64
56	10.58	2.07	9.44	-6.44	2.59
57	-4.74	1.28	2.82	5.82	7.57
58	5.92	7.53	12.88	-9.88	-1.13
59	3.62	8.95	5.24	-2.24	17.96
60	4.08	3.74	5.12	8.12	-2.08
61	4.90	6.65	5.14	8.14	19.74
62	0.49	0.95	1.42	1.58	-13.99
63	1.33	3.84	7.82	10.82	-16.74
64	-2.09	2.74	0.21	3.21	49.07

Tabla A.2: Tabla que recoge los datos de las muestras (continuación).

Índice	Normal	Uniforme	Exponencial	Doble exponencial	Mixtura de normales
65	-2.36	8.15	0.49	3.49	-5.93
66	4.52	4.49	0.49	3.49	8.23
67	5.24	8.10	1.40	4.40	13.67
68	3.27	8.12	1.48	1.52	-4.12
69	7.61	7.94	4.86	-1.86	13.06
70	13.25	4.40	4.62	7.62	13.66
71	0.54	7.54	8.21	-5.21	-1.46
72	-8.55	6.29	8.10	11.10	-1.79
73	8.03	7.10	12.68	-9.68	5.00
74	-0.55	0.01	7.61	-4.61	33.15
75	-0.44	4.75	1.90	1.10	-9.84
76	8.13	2.20	1.91	4.19	-10.88
77	1.58	3.80	2.33	5.33	2.85
78	-3.10	6.13	0.21	3.21	4.61
79	3.91	3.52	1.60	1.40	10.00
80	2.31	1.11	3.22	6.22	-4.22
81	3.03	2.44	2.86	5.86	-12.94
82	4.93	6.68	1.08	1.92	22.91
83	1.15	4.18	22.49	-19.49	-3.17
84	6.22	7.88	9.29	-6.29	-8.37
85	1.90	1.03	3.43	-0.43	-1.10
86	4.66	4.35	7.20	-4.20	0.87
87	8.48	9.85	8.66	11.66	22.39
88	5.18	8.93	6.22	-3.22	5.36
89	1.37	8.86	7.32	10.32	13.50
90	8.74	1.75	7.69	10.69	-1.62
91	7.97	1.31	0.02	3.02	8.70
92	5.74	6.53	5.54	-2.54	-0.50
93	4.19	3.44	1.50	4.50	5.02
94	-0.14	6.57	5.96	-2.96	-12.00
95	9.80	3.20	5.57	8.57	-13.26
96	-0.00	1.88	0.34	2.66	31.46
97	13.94	7.82	2.40	0.60	17.48
98	10.66	0.94	7.85	-4.85	-11.94
99	1.82	4.67	1.30	1.70	-6.76
100	-2.13	5.12	9.28	-6.28	-17.35

Tabla A.3: Tabla que recoge los datos de las muestras (continuación).