UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Bondad de ajuste: Una revisión de algunos métodos existentes

Claudia Alejandra Huaylla

Directora: Dra. Mariela Sued.
Codirectora: Dra. Gilda Garibotti.

Julio 2015

Agradecimientos

Primero quiero agradecerle a mi directora Mariela Sued, que sin ella esto no hubiese sido posible, por su tiempo, por ayudarme a entender los temas y por guiarme.

A la Dra Gilda Garibotti que fue mi codirectora en Bariloche, quien me brindó toda su ayuda y tiempo.

A Juan Carlos Pardo-Fernandez por toda la información que me brindó sobre el tema.

Al Jurado, muchas gracias por tomarse el tiempo para leer la tesis.

A mi papá, gracias por confiar en mi, por darme fuerzas para seguir cada vez que estaba desanimada y por asegurarte que pueda terminar de estudiar a pesar de que ya no estés entre nosotros. Sé que estarías orgulloso de ver cómo tu esfuerzo no fue en vano.

A mi mamá, gracias por estar siempre, por soportar mi silencio y nervios los días previos a los parciales y finales, por ayudarme en todo.

A mi querido esposo, que aceptó tener un matrimonio a distancia para que yo pueda terminar de cursar y terminar el esfuerzo de muchos años. Gracias por ser incondicional.

A toda mi familia, por estar siempre.

A mis amigos, Carla, Tatiana, Mariela, Pauli, Alan, Bar, Anita, Miriam, y a mis compañeros de cursada y de estudio.

A todos los profesores que padecieron todas mis consultas y dudas durante cada materia que fui cursando.

Quiero agradecerle particularmente al Dr Carlos Sánchez, quien fue muchas veces mi consejero académico.

Índice

Capítulo 1

Introducción

En las ciencias experimentales, muchas veces es necesario cotejar los datos obtenidos con un modelo. Un modelo suele estar caracterizado por una función de distribución o una familia de distribuciones. Si el modelo está caracterizado por una función de distribución, sus parámetros estarán completamente especificados. Pero si tenemos una familia de distribuciones, los parámetros no estarán especificados y en este caso deberemos estimarlos. Para decidir si el modelo propuesto *ajusta bien los datos* se realizará un test de hipótesis.

La hipótesis nula puede ser simple o compuesta. En el primer caso la función de distribución está completamente especificada, mientras que en el segundo tendremos una familia de distribuciones. A este tipo de test se lo suele llamar **Test de bondad de ajuste**. En esta tesis presentaremos una revisión de algunos métodos existentes para bondad de ajuste y acompañaremos la presentación de cada método con un pequeño estudio de simulación. Gran parte del material presentado en los primeros capítulos fue estudiado del libro *Comparing Distributions*, Thas (2010).

En el **Capítulo 2**, mencionaremos dos herramientas gráficas (qq-plot y pp-plot) que ayudan a determinar si un modelo propuesto es adecuado para describir los datos observados. Aplicaremos estas herramientas a diferentes muestras generadas a partir de diferentes distribuciones.

En el **Capítulo 3** presentaremos el test χ^2. Este test fue introducido por **Pearson** en 1900. Originalmente fue construido para testear una hipótesis nula multinomial con los parámetros especificados. Luego se adaptó para ser aplicado en el caso de variables continuas. El test consiste, esencialmente, en comparar valores observados con los esperados bajo la hipótesis nula.

Luego presentaremos el test de Kolmogorov-Smirnov (K-S), que se basa en considerar una distancia entre la distribución empírica y la distribución de la hipótesis nula, cuando esta es simple.

Para ambos procedimientos (χ^2 y K-S) adaptamos el estadístico propuesto para poder contemplar el caso de hipótesis nula compuesta, y presentamos una propuesta de remuestreo para el cómputo de p-valores.

En el **Capítulo 4**, se presentan los estimadores no paramétricos de la densidad basados en núcleos, introducidos por Rosenblatt [2] y Parzen [1].

En el **Capítulo 5**, utilizaremos los estimadores no paramétricos de la función de densidad para abordar problemas de bondad de ajuste. Presentaremos dos estadísticos propuestos inicialmente por Bickel y Rosenblatt [1] para hipótesis nula simple, extendidos por Fan [4] al caso de hipótesis nula compuesta. Por último, realizamos un estudio de simulación calculando p-valores con técnicas de remuestreo.

Capítulo 2

Herramientas gráficas: PP Plot y QQ Plot

Tanto los histogramas como los boxplot suelen ser útiles a la hora de vizualizar datos. En esta sección presentaremos algunas otras herramientas gráficas utilizadas para determinar si ciertas observaciones provienen de un modelo propuesto. Estos procedimientos pueden considerarse descriptivos, pero resultan muy informativos. Por lo general, se realiza un gráfico donde cada punto representa una observación y el modelo propuesto es *bueno* en la medida que los puntos pueden ser aproximados por una recta. Apartamientos de esta estructura sugieren considerar otros modelos.

Comenzaremos considerando el caso simple, donde se desea determinar si la distribución de los datos coincide con cierta F_0. Es decir, dada una muestra $X_1, ..., X_n$ de una variable aleatoria X con función de distribución F, queremos ver si F_0 es un buen ajuste para nuestros datos proponiendo $H_0 : F = F_0$ como hipótesis nula. El pp-plot (probability vs. probability plot) se realiza utilizando la siguiente idea: bajo H_0 resulta que $F(X_i) = F_0(X_i)$ para $i = 1, ..., n,$. En este caso los puntos $(F(X_i), F_0(X_i))$ se encuentran sobre la recta identidad. Como no conocemos F podemos estimarla. En esta instancia, utilizaremos una versión corregida del estimador empírico de F, definiendo F_p siendo

$$F_p(X_i) = \frac{\#\{X_j : X_j \leq X_i\} - 0.5}{n}$$

El pp-plot se consigue graficando los siguientes puntos:

$$\{(F_p(X_i), F_0(X_i)), i = 1, ..., n\}$$

Bajo H_0, los puntos deberían estar cerca de la recta identidad.

El qq-plot compara los percentiles empíricos con los correspondientes a F_0, graficando

$$\{(F_0^{-1}F_p(X_i), X_i), i = 1, ..., n\}$$

Nuevamente, cuando el modelo es correcto los puntos están próximos de la recta identidad.

En los siguientes gráficos generamos datos con distintas distribuciones y en cada caso se quiere determinar si la distribución normal estandar es un buen modelo para los valores observados. Es decir, en cada caso consideramos $H_0 : F = F_0$, siendo F_0 la distribución normal estandar.

En la Figura 2.1 graficamos el qq-plot y el pp-plot de una muestra de tamaño 100 generada a partir de una distribución $N(0, 1)$.

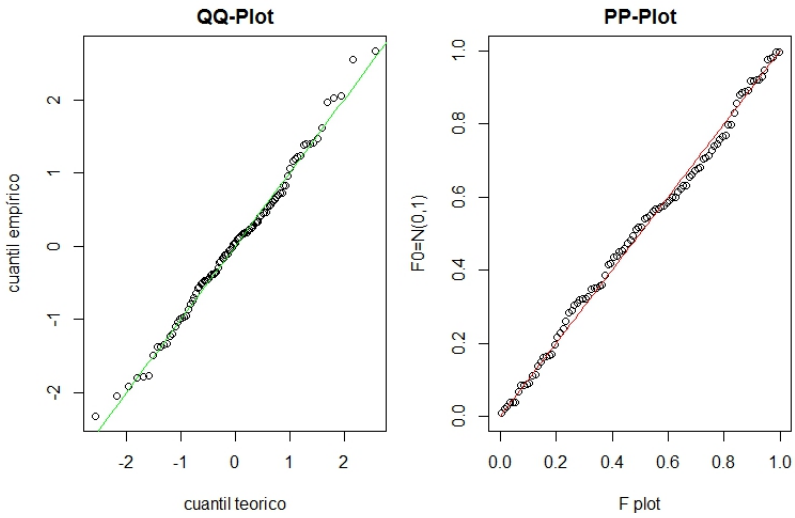

Figure 2.1: $X \sim N(0,1), \quad n = 100$

Podemos observar que los puntos se encuentran sobre una recta, por lo tanto visualmente no rechazo H_0.

En la Figura 2.2 graficamos el qq-plot y el pp-plot de una muestra de tamaño 100 generada a partir de una distribución de Student con cinco grados de libertad: $X \sim t_5$.

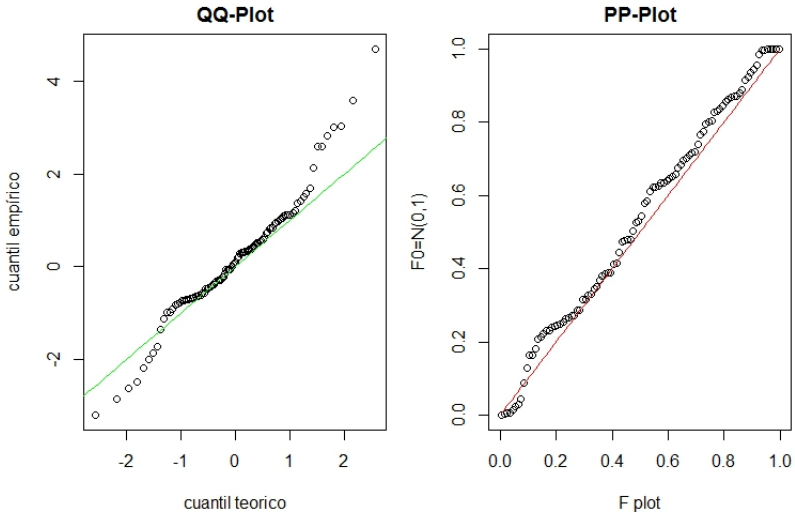

Figure 2.2: $X \sim t_5, \quad n = 100$

Si bien t_5 es una distribución simétrica, podemos ver en el qq-plot que los puntos alrededor del origen se encuentran sobre la recta. Sin embargo, los puntos en los extremos se alejan de la recta, sugiriendo que los datos no provienen de la distribución normal estandar. En el pp-plot observamos oscilaciones en torno a la recta identidad que no están presentes cuando los datos se generan bajo el modelo normal (Figura 2.1).

En la Figura 2.3 graficamos el qq-plot y el pp-plot de una muestra de tamaño 100 generada a partir de una distribución exponencial con $\lambda = 1$: $X \sim \varepsilon(1)$.

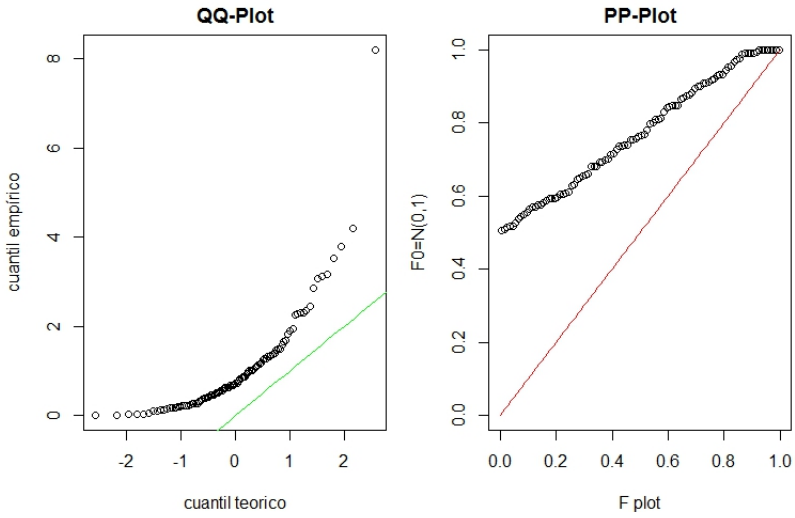

Figure 2.3: $X \sim \varepsilon(1), \quad n = 100$

La distribución no es simétrica y podemos ver que los puntos se alejan de la recta, en el caso del qq-plot y del pp-plot.

En la Figura 2.4 graficamos el qq-plot y el pp-plot de una muestra de tamaño 100 generada a partir de una distribución chi cuadrado con 5 grados de libertad, centrada: $X \sim \chi^2_5 - 5$.

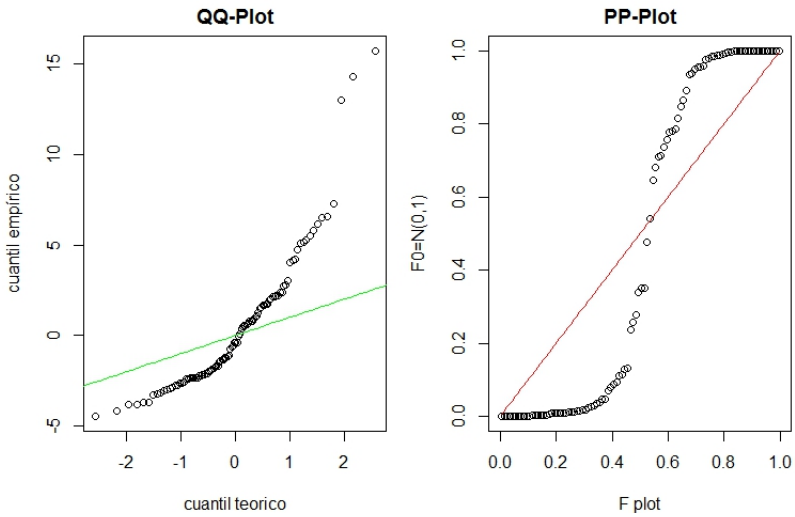

Figure 2.4: $X \sim \chi^2_5 - 5, \quad n = 100$

En este caso también podemos ver que los puntos no se encuentran sobre una linea recta.

El qq-plot y el pp-plot nos brindan información acerca de nuestro supuesto, pero no es un método exacto.

Cuando la hipótesis nula es compuesta, podemos reemplazar F_0 por $F_{\hat{\theta}}$, siendo $\hat{\theta}$ un estimador consistente del parámetro bajo el modelo.

Si se trata de un modelo de posición y escala, los graficos presentados siguen siendo de utilidad, salvo por el hecho de que ahora los puntos deben estar cerca de una recta, y no necesariamente de la recta identidad. Es decir, supongamos que $X \sim \mu + \sigma Z$, para $Z \sim F_0$, para $\sigma > 0$. En tal caso tenemos que $F(x) = F_0((x-\mu)/\sigma)$ y por consiguiente $F^{-1}(u) = \sigma F_0^{-1}(u) + \mu$. De esta forma tenemos que los percentiles teóricos y los empíricos se relacionan linealmente. Esperamos entonces que los puntos

$$\{(F_0^{-1}(F_p(X_i)), X_i), i = 1, ..., n\}$$

sigan una relación lineal, con ordenada al origen μ y pendiente σ. El comando **qqnorm** realiza este último gráfico, cuando F_0 es la función de distribución normal estandar.

En la Figura 2.5 graficamos el qq-plot de una muestra de tamaño 100 generada a partir de una distribución $N(1, 1)$.

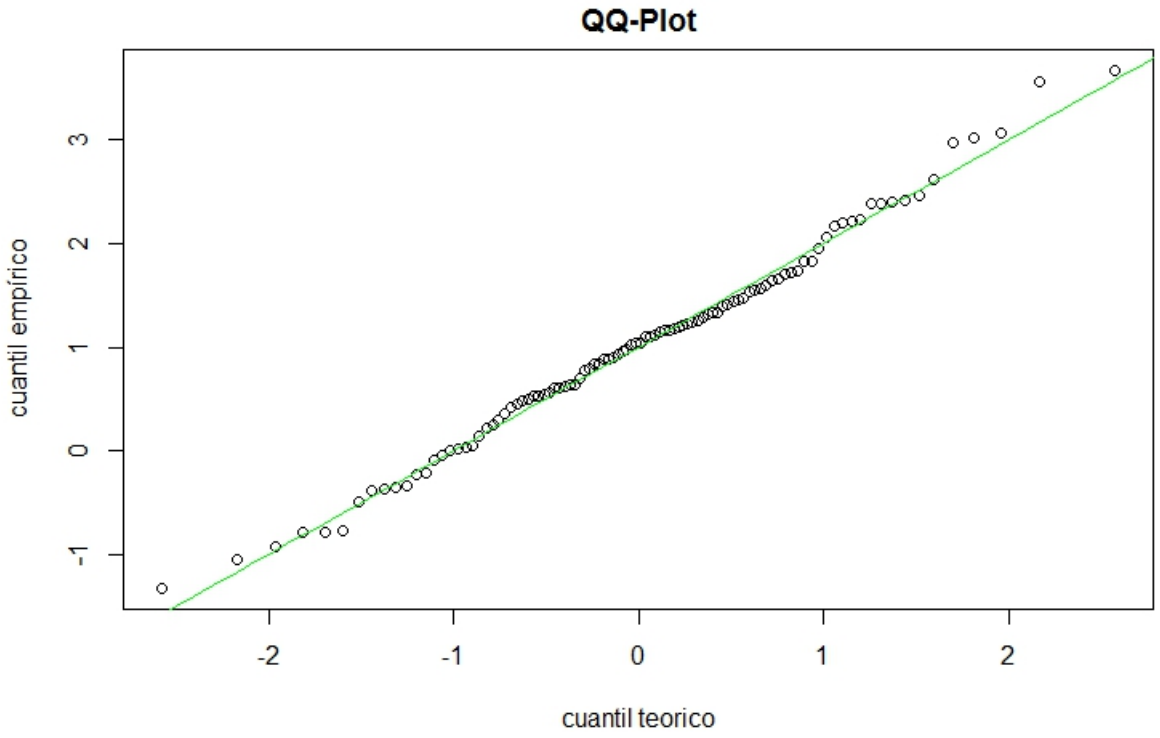

Figure 2.5: $X \sim N(1, 1), \quad n = 100$

En la Figura 2.6 graficamos el qq-plot de una muestra de tamaño 100 generada a partir de una distribución $N(0, 2)$.

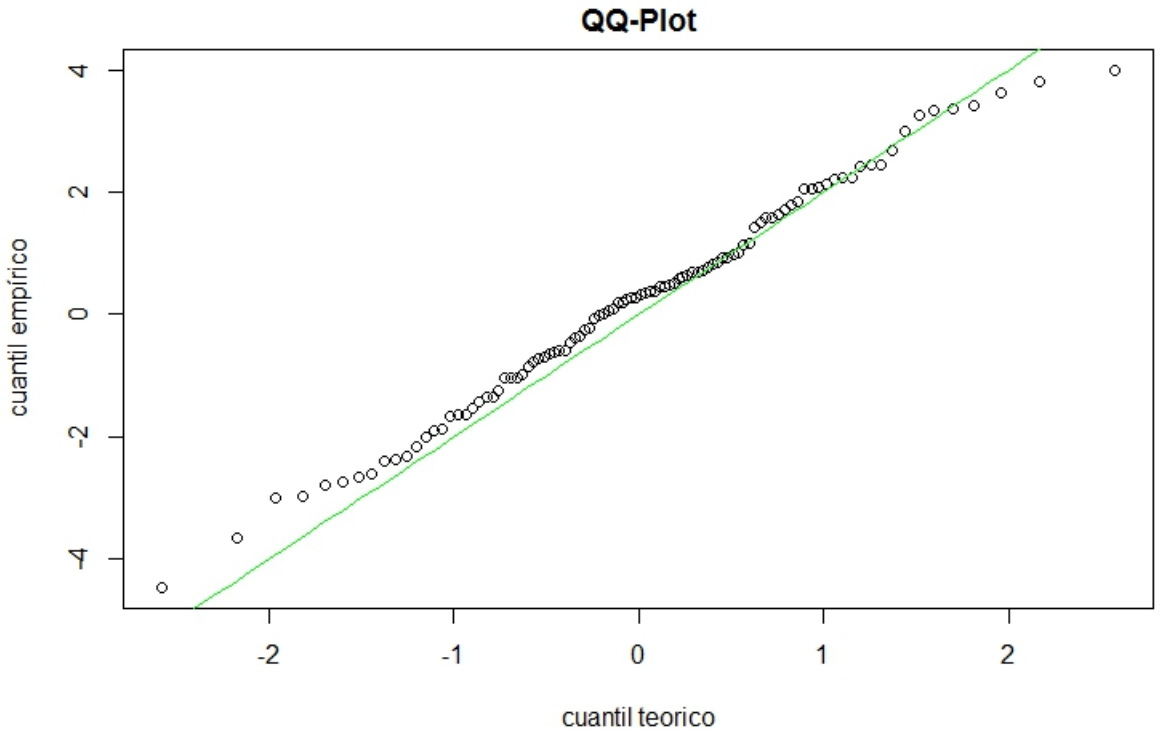

Figure 2.6: $X \sim N(0, 2), \quad n = 100$

Podemos ver en ambos gráficos que los puntos se encuentran sobre una recta, indicando que la familia normal ajusta bien los datos. Además encontrando la recta que ajusta mejor a los puntos uno puede calcular la ordenada al origen y la pendiente, conociendo así los parámetros de la normal a partir de la cual se obtuvieron los datos originales.

Capítulo 3

Bondad de ajuste

En general tendremos observaciones $X_1, ..., X_n$ i.i.d de una variable aleatoria X con distribución F desconocida y se desea testear si F coincide con cierta distribución F_0 (hipótesis simple) o pertenece a cierta familia paramétrica $\mathcal{F}$ (hipótesis compuesta). Más específicamente, si la hipótesis nula es simple estamos bajo el siguiente escenario:

$$H_0 : F = F_0 \quad vs \quad H_1 : F \neq F_0, \tag{3.1}$$

siendo F_0 una función de distribución conocida. Es decir, con parámetros especificados. A modo de ejemplo, podemos estar interesados en saber si $X \sim N(0, 1)$. Si H_0 es una hipótesis compuesta, sólo se conoce la familia a la que pertenece, pero no sus parámetros. En tal caso tenemos

$$H_0 : F \in \mathcal{F} = \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\} \quad \text{vs} \quad H_1 : F \notin \mathcal{F}\,, \tag{3.2}$$

siendo F_θ una función de distribución con parámetro θ. Por ejemplo, podemos estar interesados en saber si los datos provienen de una distribución normal, sin especificar cuáles deberían ser sus parámetros.

Nuestro objetivo será encontrar un modelo que *ajuste bien a los datos*. Cuando no rechazamos H_0 el modelo propuesto es compatible con los datos observados y, en tal caso, decimos que proporciona un buen ajuste para los mismos.

3.1 Test Chi cuadrado

El test de Pearson χ^2 es el más conocido y el más antiguo. Este test originalmente fue construido para testear H_0 simple, siendo F_0 una distribución multinomial. Durante muchos años este fue el único test de bondad de ajuste utilizado.

Si bien el test de Pearson fue diseñado para datos multinomiales, se lo puede usar con datos continuos pero éstos deben ser agrupados o categorizados.

Este hecho nos lleva a perder información y el método puede resultar menos potente que otros. Actualmente disponemos de otras técnicas para datos continuos, sin embargo muchos siguen eligiendo el test de Pearson agrupando los datos.

3.1.1 Hipótesis nula simple

El test χ^2 propone comparar la frecuencia observada sobre un conjunto de intervalos con las esperadas bajo H_0. Está inspirado en el siguiente resultado, cuya demostración puede encontrarse en el Capítulo 17 del libro de Van der Vaart [10]

Teorema 3.1.1. *Sea* $\mathbf{N} = (N_1, .., N_k)$ *un vector aleatorio con distribución multinomial de parámetros* n *y* $\pi = (\pi_1, ..., \pi_k) > 0$ *:* $\mathbf{N} \sim \mathcal{M}(n, \pi)$*, entonces*

$$\sum_{j=1}^{k} \frac{(N_j - n\pi_j)^2}{n\pi_j} \longrightarrow \chi^2_{k-1} \quad \textit{en distribución.} \tag{3.3}$$

Vamos a describir ahora cómo este resultado puede ser utilizado en el contexto de bondad de ajuste. Sea X una variable aleatoria con distribución F. Se trata de testear $H_0 : F = F_0$. Para ello, consideremos una partición finita de $\mathbb{R}$, $\mathcal{P} : I_1, ..., I_k$ donde cada I_j es un intervalo. Sea $X_1, ..., X_n$ una muestra aleatoria i.i.d con distribución F. Vamos a contar cuántas observaciones pertenecen a cada intervalo I_j, introduciendo las variables

$$N_j = \sum_{i=1}^{n} \mathbf{I}_{I_j}(X_i),$$

siendo

$$\mathbf{I}_B(x) = \begin{cases} 1 & \text{si } x \in B, \\ 0 & \text{En caso contrario.} \end{cases}$$

El vector $\mathbf{N} = (N_1, .., N_k)$ tiene distribución multinomial de parámetros $(\pi_1, \ldots, \pi_k)$, siendo $\pi_j = P(X \in I_j)$, para $1 \leq j \leq k$. Bajo H_0, utilizamos π_{0j} para denotar la probabilidad de que X pertenezca al intervalo I_j. Es decir, $\pi_{0j} = P(X \in I_j)$ cuando $X \sim F_0$. El estadístico de Pearson, se define como

$$X^2_n = \sum_{j=1}^{k} \frac{(N_j - n\pi_{0j})^2}{n\pi_{0j}}. \tag{3.4}$$

Notemos que N_j es la frecuencia observada del intervalo I_j, mientras que $n\pi_{0j}$ es la esperada bajo H_0. Si la diferencia entre lo observado y lo esperado es muy pequeña, podemos decir que F_0 ajusta bien nuestros datos. Valores grandes del estadístico sugieren que el modelo no es adecuado, siendo que lo observado difiere de lo esperado bajo H_0. El Teorema 3.1.1 sugiere considerar el siguiente criterio para tener una región de rechazo de nivel α:

- Si $X^2_n \geq \chi^2_{k-1,\alpha}$, rechazamos H_0
- Si $X^2_n < \chi^2_{k-1,\alpha}$, no rechazamos H_0,

donde $\chi^2_{l,\alpha}$ denota al percentil α de la distribución chi-cuadrado con l grados de libertad:
$P(T_l > \chi^2_{l,\alpha}) = \alpha$, cuando $T_l \sim \chi^2_l$
Los p-valores asociados al procedimiento descripto se obtienen según la ecuación
p-valor $= P(T_{k-1} \geq X^2_n)$, con $T_{k-1} \sim \chi^2_{k-1}$

3.1.2 Hipótesis nula compuesta

En esta sección no tenemos un F_0 específico, sino una familia de distribuciones, por ejemplo la familia normal. Nuestro escenario será:

$$H_0 : F \in \mathcal{F} = \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\} \quad \text{vs} \quad H_1 : F \notin \mathcal{F}\,. \tag{3.5}$$

Dado que no tenemos una distribución con parámetros conocidos, un procedimiento natural es estimar los parámetros bajo H_0, y luego usar el estadístico del test χ^2. Si consideramos $\tilde{p}_i$

utilizando $F_{\tilde{\theta}}$, con $\tilde{\theta}$ estimador de máxima verosimilitud utilizando $\mathbf{N}_i$, el número de observaciones que se encuentran en el i-ésimo intervalo, Fisher demostró que el estadístico

$$Q_n = \sum_{i=1}^{k} \frac{(N_i - n\widetilde{p}_i)^2}{n\widetilde{p}_i}$$

converge en distribución a una variable χ^2_{k-1-s}, siendo s el número de parámetros estimados. Sin embargo, si se calcula $\hat{\theta}$, el estimador de máxima verosimilitud basado en las observaciones originales $X_1, ..., X_n$, y se calcula $\hat{p}_i$ utilizando $F_{\hat{\theta}}$, se obtiene el estadístico

$$\hat{Q}_n = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}. \tag{3.6}$$

Notemos que $\hat{Q}_n$ coincide con el estadístico del test chi cuadrado definido en (3.4) pensando en una hipótesis nula simple con $F_{\hat{\theta}}$. La distribución asintótica de $\hat{Q}_n$ no es χ^2. Chernoff y Lehmann [3] demostraron el siguiente resultado.

Teorema 3.1.2. *Bajo H_0 la distribución asintótica de $\hat{Q}_n$ es*

$$W = \sum_{i=1}^{k-1-s} Z_i^2 + \sum_{i=k-s}^{k-1} \lambda_i Z_i^2 \tag{3.7}$$

siendo $Z_i \sim N(0,1)$ independientes y $\lambda_i = \lambda_i(\theta)$ valores entre 0 y 1.

Notemos que el primer término en (3.7) corresponde a una variable aleatoria con distribución χ^2_{k-1-s} y por consiguiente, $P(\chi^2_{k-1-s} \geq \hat{Q}_n) \leq P(W \geq \hat{Q}_n)$, siendo $\lambda_i > 0$. Es decir, si calculamos p-valores de manera ingenua, haciendo

$$\text{p-naive} = P(\chi^2_{k-s-1} \geq \hat{Q}_n)\ , \tag{3.8}$$

ignorando el segundo término en la expresión (3.7), obtenemos una cota inferior para verdaderos p-valores asintóticos. Tenemos entonces que si p-naive es grande, podemos considerar que el modelo proporciona un buen ajuste para los datos observados. Sin embargo, valores pequeños de p-naive no deberían ser considerados para rechazar H_0.

3.1.3 Remuestreo para el test χ^2

Vamos ahora a considerar un procedimiento de tipo Bootstrap para el cálculo de p-valores correspondientes al estadístico $\hat{Q}_n$, definido en (3.6). Es decir, bajo el problema planteado en (3.5), queremos utilizar el estadístico $\hat{Q}_n$ y aproximar su distribución bajo H_0 mediante técnicas paramétricas de remuestreo, como se describe a continuación:

1) Dada una muestra aleatoria $X_1, ..., X_n$, denotamos con $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a un estimador de θ bajo el modelo $\mathcal{F} = \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\}$.

2) Calculamos el estadístico $\hat{Q}_n = \hat{Q}_n(X_1, \ldots, X_n)$, como en (3.6).

3) Fijado $Nboot$, para $t \in \{1, \ldots, Nboot\}$, repetimos los siguientes pasos:

   3.1) Generamos una muestra $X_1^*, ..., X_n^*$, con distribución $F_{\hat{\theta}}$.

3.2) Calculamos el estadístico de la muestra bootstrap: $\hat{Q}^*_{t,n} = \hat{Q}_n(X^*_1, \ldots, X^*_n)$, utilizando $\hat{\theta}^* = \hat{\theta}(X^*_1, \ldots, X^*_n)$ en la construcción de $\hat{Q}^*_{t,n}$.

4) Finalmente calculamos la proporción de valores $\hat{Q}^*_{t,n}$ mayores que $\hat{Q}_n$:

$$\text{p-boot} = \frac{1}{Nboot} \sum_{t=1}^{Nboot} I_{\hat{Q}^*_{t,n} \geq \hat{Q}_n}. \tag{3.9}$$

Notemos que el $\hat{Q}^*_n$ es un estadístico de tipo χ^2, pensando en una hipótesis nula simple dada por $F_{\hat{\theta}^*}$.

Para terminar la sección queremos mencionar que la elección de la cantidad de intervalos es muy importante y la decisión puede ser diferente cuando tomamos intervalos distintos. Como regla general, para los procedimientos que utilizan aproximaciones asintóticas se recomienda elegir los intervalos de forma tal que el número esperado de observaciones en cada uno de ellos sea mayor o igual a cinco.

3.1.4 Simulación

En esta sección calculamos el nivel y la potencia empírica de los procedimientos descriptos en las secciones anteriores, mediante un estudio de simulación. En todos los casos, consideramos la hipótesis nula compuesta que asume que los datos provienen de una distribución normal: $H_0 : F \in \mathcal{F} = \{F_\theta, \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\}$, donde F_θ denota la distribución normal $N(\mu, \sigma^2)$. En cada caso, se realizaron $Nrep = 1000$ replicaciones con tamaños muestrales $n = 15, 50, 100$, generando datos normales (H_0), y tres escenarios alternativos, distribución uniforme, t-student y exponencial. Para cada conjunto de datos, calculamos el p-naive y el p-boot con $Nboot = 1000$, definidos en (3.8) y (3.9), respectivamente.

La elección de la partición con la que se construyen los estadísticos χ^2 juegan un rol importante a la hora de evaluar la bondad de un ajuste. En la práctica, el usuario necesita información sobre la variable en cuestión para saber (al menos) que rango considerar a la hora de hacer la partición. En muchos casos, se dispone de información previa sobre los datos a estudiar y se la utiliza en la confección de los intervalos. Cuando esto no sucede, técnicas gráficas ayudan a determinar la partición. En nuestra simulación consideramos los intervalos provistos (por default) por el comando hist, tomando "hist(muestra)$breaks", cada vez que calculamos el estadístico.

Las siguientes tablas muestran la proporción de veces (en las $Nrep = 1000$ replicaciones) en que los p-valores resultaron menores que α, tomando $\alpha = 0.05$ en la Tabla 1 y $\alpha = 0.20$ en la Tabla 2. Los datos generados bajo H_0 permiten estudiar el nivel empírico, mientras que en los restantes escenarios se manifiesta la potencia de los procedimientos.

Tabla 1. Proporción de rechazos de H_0 a nivel $\alpha = 0.05$ en $Nrep = 1000$ replicaciones para cada procedimiento, con diferentes tamaños muestrales.

	n=15		n=50		n=100	
	p-boot	p-naive	p-boot	p-naive	p-boot	p-naive
$N(0,1)$	0.048	0.070	0.061	0.113	0.060	0.118
$U(-1,1)$	0.008	0.140	0.206	0.291	0.350	0.409
t_5	0.122	0.159	0.357	0.453	0.555	0.662
$\varepsilon(1)$	0.284	0.429	0.781	0.913	0.977	0.9980

Tabla 2. Proporción de rechazos de H_0 a nivel $\alpha = 0.20$ en $Nrep = 1000$ replicaciones para cada procedimiento, con diferentes tamaños muestrales.

	n=15		n=50		n=100	
	p-boot	p-naive	p-boot	p-naive	p-boot	p-naive
$N(0,1)$	0.20	0.288	0.214	0.278	0.218	0.294
$U(-1,1)$	0.098	0.473	0.620	0.611	0.913	0.734
t_5	0.284	0.380	0.535	0.616	0.710	0.787
$\varepsilon(1)$	0.517	0.705	0.931	0.983	0.9983	1

En ambas tablas podemos observar que cuando generamos muestras con distribución $N(0,1)$ y utilizamos el valor de p-boot, el nivel empírico está muy cerca del nominal, como es de esperarse ya que estamos generando muestras bajo H_0.

Sin embargo podemos ver que usando el p-naive la proporción de rechazos es mayor que el nivel nominial del procedimiento. Entendemos que este hecho se debe a que, como hemos mencionado, p-naive proporciona una cota inferior para el verdadero p-valor asociado al test chi cuadrado. Es decir, valores pequeños de p-naive pueden dar lugar a rechazos indebidos de la hipótesis nula. Así, el nivel empírico del test basado en el valor p-naive , resulta superior al nivel nominal, como puede apreciarse en las primeras filas de las Tabla 1 y Tabla 2, en las columnas correspondientes a p-naive.

Practicamente en todos los casos, los valores observados con p-naive son superiores a los obtenidos con p-boot.

Cuando la generación de datos se hace según una distribución diferente a la contemplada por H_0 (últimas tres filas en cada una de las tablas), observamos que la potencia empírica crece cuando aumenta el tamaño muestral.

3.2 Test de Kolmogorov Smirnov

En esta sección introducimos un test basado en la función de distribución empírica. El estadístico del test calcula cierta distancia entre la función de distribución empírica de los datos y la función F_0 propuesta bajo H_0, cuando la hipótesis nula es simple. Comenzaremos revisando algunos resultados sobre la función de distribución empírica.

3.2.1 La función de distribución empírica

Sea X una variable aleatoria con función de distribución F. Dada una muestra $X_1, ..., X_n$ con distribución F, definimos la función de distribución empírica siendo

$$F_n(x) = \frac{1}{n}\#\{X_i : X_i \leq x : i = 1...n\} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,x]}(X_i) \tag{3.10}$$

Cada término $\mathbf{I}_{(-\infty,x]}(X_i)$ es una variable aleatoria de Bernoulli con probabilidad de éxito

$$p = P(\mathbf{I}_{(-\infty,x]}(X_i) = 1) = P(X_i \leq x) = F(x).$$

Además, la independencia de las variables $X_1, \ldots, X_n$ garantiza la independencia de las variables $\mathbf{I}_{(-\infty,x]}(X_i)$, y por consiguiente, tenemos el siguiente resultado:

Lema 3.2.1. $nF_n(x) \sim B(n, p = F(x))$

Notemos que F_n es la función de distribución de una variable aleatoria discreta que a cada punto X_i observado le asigna peso $\frac{1}{n}$. Por consiguiente, para cada $x \in \mathbb{R}$, esta función le asigna la proporción de los valores observados menores o iguales a x. Tenemos entonces que F_n es una función de distribución acumulada:

1. $F_n \in [0, 1] \ \ \forall x \in \mathbb{R}$
2. F_n es continua por la derecha.
3. F_n es no decreciente.
4. $\lim_{x\to-\infty} F_n(x) = 0$
5. $\lim_{x\to\infty} F_n(x) = 1$

En el siguiente resultado se presentan algunas propiedades de la función de distribución empírica.

Proposición 3.2.2. *Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Se denota por F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, ..., X_n$. Sea $x \in \mathbb{R}$. Se verifica lo siguiente:*

a) $P(F_n(x) = \frac{j}{n}) = \binom{n}{j} F(x)^j (1 - F(x))^{n-j} \quad , j = 0, ..., n.$

b) $\mathbb{E}(F_n(x)) = F(x), \mathbb{V}ar(F_n(x)) = \frac{1}{n}F(x)(1 - F(x))$

c) $F_n(x) \xrightarrow{c.s} F(x)$

d) $\frac{\sqrt{n}(F_n(x)-F(x))}{\sqrt{F(x)(1-F(x))}} \xrightarrow{D} Z$, *siendo* $Z \sim N(0, 1)$

Demostración 1. *a) y b) se deducen del Lema 3.2.1. Sea* $Y_i = \mathbf{I}_{(-\infty,x]}(X_i)$*, se tiene que* $F_n(X) = \bar{Y}_n$*, y por la Ley fuerte de los grandes números, sabemos que* $\bar{Y}_n \longrightarrow \mathbb{E}(Y_1) = F(X)$*, por lo tanto se cumple c). d) es consecuencia de aplicar el teorema central del límite.*

Los siguientes gráficos representan la función de distribución empírica con muestras generadas con distribución $N(0,1)$ de tamaño 20, 50 y 100. En cada gráfico también está la verdadera función de distribución.

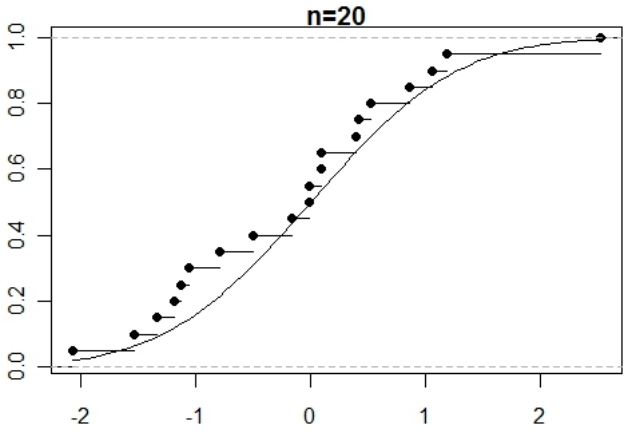

Figure 3.1: Función de distribución empírica con $n = 20$

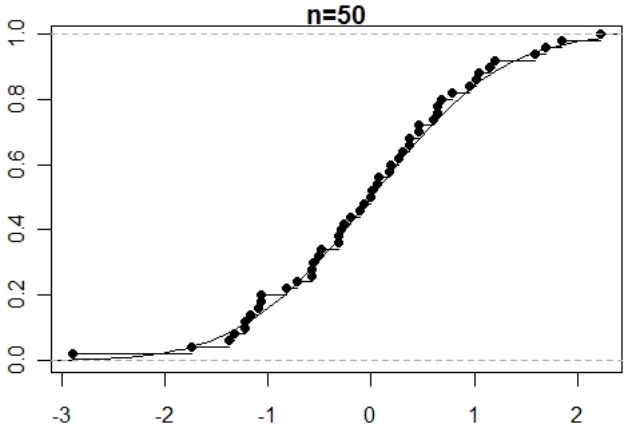

Figure 3.2: Función de distribución empírica con $n = 50$

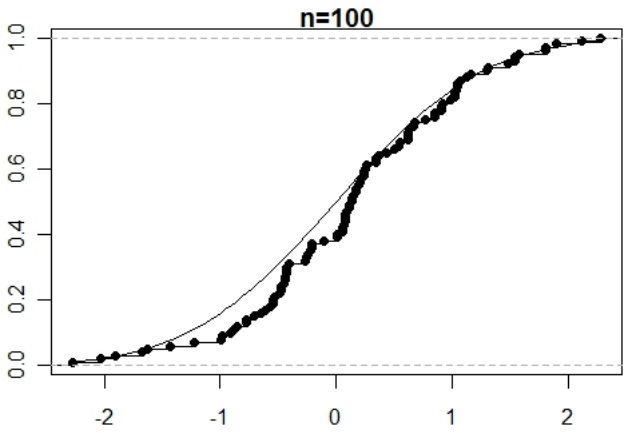

Figure 3.3: Función de distribución empírica con $n = 100$

Si bien la Proposición 3.2.2 establece la convergencia puntual (en cada x) de F_n a F, el siguiente resultado garantiza que en realidad la función de distribución empírica converge uniformemente a la distribución que genera los datos.

Teorema 3.2.3 (Teorema de Glivenko-Cantelli)**.** *Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas definidas en el espacio de probabilidad $(\Omega, \mathcal{A}, P)$ con función de distribución común F. Sea F_n la función de distribución empírica obtenida de las n primeras variables aleatorias $X_1, ..., X_n$. Entonces*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{c.s} 0 \tag{3.11}$$

3.2.2 El estadístico del Test de Kolmogorov-Smirnov

El teorema de Glivenko-Cantelli sugiere comparar la distribución empírica de los datos con la propuesta bajo H_0. Es decir, si $X_i \sim F$ y queremos determinar si $H_0 : F = F_0$, podemos considerar el siguiente estadístico

$$D_n = \sup_x |F_n(x) - F_0(x)|, \tag{3.12}$$

popularizado como el estadístico de **Kolmogorov-Smirnov**. Si H_0 es verdadera, se espera que la función empírica se aproxime a F_0; si esto no sucede, resulta una evidencia para rechazar H_0. Resta conocer la distribución del estadístico bajo H_0, para poder determinar qué valores de D_n resultan grandes y permiten rechazar la hipótesis nula. El siguiente resultado muestra que D_n es un estadístico de distribución libre.

Proposición 3.2.4. *Sean $X_1, \ldots, X_n$ i.i.d. con distribución F_0 absolutamente continua, entonces la distribución de D_n no depende de F_0.*

Demostración 2. *Recordemos que dada una función de distribución F_0, su inversa generalizada está definida por*

$$F_0^{-1}(y) = \inf\{x : F_0(x) \geq y\}$$

y satisface las siguientes propiedades:

- $F_0^{-1}(y) \leq x \Leftrightarrow F_0(x) \geq y$

- $U \sim U(0,1)$ *entonces* $F_0^{-1}(U) \sim F_0$

Dado que el resultado trata apenas de la distribución de D_n*, construiremos ad-hoc variables con distribución* F_0 *que facilitarán la demostración del resultado y permitirán caracterizar la distribución de* D_n*. Para ello, consideremos* $U_1, \ldots, U_n$ *i.i.d.,* $U_i \sim U(0,1)$*, y tomemos* $X_i := F_0^{-1}(U_i)$*, de forma tal que* $X_i \sim F_0$*. En tal caso, tenemos que*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,x]}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,x]}(F_0^{-1}(U_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{(-\infty,F_0(x)]}(U_i) = F_n^U(F_0(x)),$$

donde F_n^U *denota la función de distribución empírica de* $U_1, \ldots, U_n$*. Haciendo entonces un cambio de variable, tenemos que* D_n *puede ser escrito como*

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| = \sup_{x\in\mathbb{R}} |F_n^U(F_0(x)) - F_0(x)| = \sup_{u\in(0,1)} |F_n^U(u) - u|,$$

de donde concluímos que la distribución del estadístico D_n *no depende de* F_0*.*

Podemos usar entonces el estadístico D_n para testear $H_0 : F = F_0$ dado que su distribución bajo H_0 es *conocida*. Este procedimiento se conoce como el test de **Kolmogorov-Smirnov**. En la tabla [B1] incluída en el apendice se muestran los valores critícos para diferentes niveles, variando el tamaño muestral. Sin embargo, a medida que el tamaño muestral crece, se puede demostrar que

$$\sqrt{n} D_n \xrightarrow{d} D,$$

siendo D una variable aleatoria con distribución

$$F_D(s) = 1 - \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 s^2}.$$

Para valores de $n < 35$ la distribución de D_n bajo H_0 se puede ver tabulada en una tabla [B1]. Para $n \geq 35$ podemos calcular el valor crítico a nivel α usando la distribución asintótica de $\sqrt{n}D_n$.

Cabe mencionar que si bien la definición de D_n involucra el cálculo de un supremo en toda la recta, es fácil ver que D_n puede obtenerse mediante la siguiente fórmula,

$$D_n = \max\{D_n^+, D_n^-\}$$

siendo

$$D_n^+ = \max_{1\leq i\leq n} |F_n(X_i) - F_0(X_i)| \quad y \quad D_n^- = \max_{1\leq i\leq n} |F_n(X_i^-) - F_0(X_i)|,$$
$$F(x^-) = \lim_{h \longrightarrow 0^+} F(x-h)$$

3.2.3 Kolmogorov-Smirnov para hipótesis nula compuesta

Hasta el momento vimos que con el estadístico D_n podemos testear el caso de hipótesis nula simple $H_0 : F = F_0 \;\; vs \;\; H_1 : F \neq F_0$. Para adaptar este procedimiento al caso en que la hipótesis nula es compuesta, resulta natural reemplazar F_0 por $F_{\hat{\theta}}$, y definir

$$\hat{D}_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_{\hat{\theta}}(x)|. \tag{3.13}$$

Sin embargo éste deja de ser un estadístico de distribución libre.

Cuando $\mathcal{F}$ es la familia de distribuciones normales, Lilliefors [5] realizó un estudio de Monte Carlo para aproximar la distribución del estadístico $\hat{D}_n$, utilizando diferentes tamaños muestrales. Pocos años después consideró también la familia exponencial [6].

En el caso normal, el estadístico se construye comparando la distribución empírica de los datos con la distribución normal $F_{\hat{\theta}}$, donde $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, siendo $\hat{\mu}$ el promedio muestral y $\hat{\sigma}^2$ el estimador insesgado de la varianza (dividiendo por $n-1$). En el apéndice incluímos las tablas [B2] y [B3], donde figuran los valores critícos para diferentes niveles, variando el tamaño muestral, según el caso normal y exponencial, respectivamente. También propuso un factor de corrección para aproximar la distribución asintótica de $\hat{D}_n$.

Para cualquier otra familia, como no existen tablas habrá que usar un procedimiento alternativo, por ejemplo Bootstrap, para decidir la veracidad de H_0.

Si comparamos los valores críticos que obtuvo Lilliefors [B2] con los valores de la tabla [B1], asumiendo la hipótesis nula simple, resultan ser aproximadamente $\frac{2}{3}$ de esos valores. Por lo tanto si usamos los valores críticos usuales, sin tener en cuenta que estimamos los parámetros, tendremos un procedimiento conservativo, en el sentido que el nivel de significación resulta ser inferior al nominal. En la librería nortest, se puede realizar el test de Lilliefors para el caso normal, con el comando **lillie.test**.

Por último, cabe mencionar que técnicas de procesos empíricos permiten obtener la distribución asintótica de $\sqrt{n}\hat{D}_n$.

3.2.4 Remuestreo para el test Kolmogorov-Smirnov

Vamos ahora a presentar un procedimiento de tipo Bootstrap para el cálculo de p-valores asociados al estadístico $\hat{D}_n$.

Bajo el problema planteado en (3.5), queremos utilizar el estadístico $\hat{D}_n$ y aproximar su distribución bajo H_0 mediante técnicas paramétricas de remuestreo, como se describe a continuación:

1) Dada una muestra aleatoria $X_1, ..., X_n$, denotamos con $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a un estimador de θ bajo el modelo $\mathcal{F} = \{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\}$.

2) Calculamos el estadístico $\hat{D}_n = \hat{D}_n(X_1, \ldots, X_n)$, como en (3.13).

3) Fijado $Nboot$, para $t \in \{1, \ldots, Nboot\}$, repetimos los siguientes pasos:

   3.1) Generamos una muestra $X_1^*, ..., X_n^*$, con distribución $F_{\hat{\theta}}$.

   3.2) Calculamos el estadístico de la muestra bootstrap:

$$\hat{D}_{t,n}^* = \sup_x |F_{t,n}^*(x) - F_{\hat{\theta}^*}(x)|$$

   siendo $F_{t,n}^*(x)$ la función de distribución empírica de la muestra bootstrap y utilizando $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$ en la construcción de $\hat{D}_{t,n}^*$.

4) Finalmente calculamos la proporción de valores $\hat{D}_{t,n}^*$ mayores que $\hat{D}_n$:

$$\text{p-ks-boot} = \frac{1}{Nboot} \sum_{t=1}^{Nboot} I_{\hat{D}_{t,n}^* \geq \hat{D}_n}. \tag{3.14}$$

3.2.5 Simulación K-S

Para terminar esta sección, presentamos los resultados de una simulación que permite comparar el comportamiento de los p-valores obtenidos mediante procedimiento naive, según la distribución de Lilliefors, y mediante la propuesta de remuestreo.

- p-Lilliefors
- p-ks-naive es el p-valor que se obtiene al proceder como si tuviéramos H_0 simple con los parámetros estimados.
- p-ks-boot

Además calculamos el nivel y la potencia empírica de los procedimientos descriptos en las secciones anteriores, mediante un estudio de simulación. En todos los casos, consideramos la hipótesis nula compuesta que asume que los datos provienen de una distribución normal:

$H_0 : F \in \mathcal{F} = \{F_\theta, \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\}$, donde F_θ denota la distribución normal $N(\mu, \sigma^2)$.

En cada caso, se realizaron $Nrep = 1000$ replicaciones con tamaños muestrales $n = 15, 50, 100$, generando datos normales (H_0), y tres escenarios alternativos. Para cada conjunto de datos, se calculó el p-Lillie, el p-ks-naive y el p-ks-boot con $Nboot = 1000$, definido en (3.14).

Las siguientes tablas muestran la proporción de veces en las $Nrep = 1000$ replicaciones en que los p-valores resultaron menores que α, tomando $\alpha = 0.05$ en la Tabla 3 y $\alpha = 0.20$ en la Tabla 4.

Los datos generados bajo H_0 nos permiten estudiar el nivel empírico, mientras que en los otros escenarios se manifiesta la potencia de los procedimientos.

Tabla 3. Proporción de rechazos de H_0 a nivel $\alpha = 0.05$ en $Nrep = 1000$ replicaciones para cada procedimiento, con diferentes tamaños muestrales.

	n=15			n=50			n=100		
	p-Lillie	p-ks-naive	p-ks-boot	p-Lillie	p-ks-naive	p-ks-boot	p-Lillie	p-ks-naive	p-ks-boot
$N(0,1)$	0.050	0	0.048	0.054	0	0.056	0.043	0	0.046
$U(-1,1)$	0.080	0	0.074	0.254	0.004	0.260	0.591	0.013	0.584
t_5	0.109	0.002	0.107	0.203	0.014	0.203	0.324	0.028	0.319
$\varepsilon(1)$	0.426	0.012	0.421	0.965	0.385	0.965	1	0.9141	1

Tabla 4. Proporción de rechazos de H_0 a nivel $\alpha = 0.20$ en $Nrep = 1000$ replicaciones para cada procedimiento, con diferentes tamaños muestrales.

	n=15			n=50			n=100		
	p-Lillie	p-ks-naive	p-ks-boot	p-Lillie	p-ks-naive	p-ks-boot	p-Lillie	p-ks-naive	p-ks-boot
$N(0,1)$	0.195	0.005	0.207	0.189	0.008	0.189	0.222	0.004	0.222
$U(-1,1)$	0.293	0.004	0.296	0.639	0.066	0.642	0.888	0.188	0.890
t_5	0.280	0.027	0.289	0.431	0.081	0.436	0.586	0.131	0.589
$\varepsilon(1)$	0.696	0.178	0.696	0.991	0.839	0.993	1	0.9966	1

Comencemos por notar que los valores obtenidos usando p-Lillie y p-ks-boot, son similares. En ambas tablas podemos observar que cuando generamos muestras con distribución $N(0,1)$ el nivel empírico calculado usando el p-Lillie y p-ks-boot están muy cerca del nivel nominal del test. Esto es de esperarse ya que estamos generando muestras bajo H_0. Además, tal como esperabamos a partir de los resultados presentados por Lilliefors, el procedimiento p-ks-naive resulta conservativo, rechazando H_0 más de lo indicado por el nivel nominal del procedimiento.

Se ve que cuando los datos se generan con distribuciones en la hipótesis alternativa (últimas tres filas de cada tabla), la potencia crece con el tamaño muestral. Cabe mencionar que si usamos el procedimiento deducido a partir de p-ks-naive, la potencia empírica crece más lento que si usamos los otros p-valores.

Podemos concluir que usar el p-ks-naive nos puede conducir a resultados erróneos.

3.2.6 Comparamos el test K-S con el test χ^2

1) El test K-S no necesita agrupar los datos, a diferencia del test χ^2.

2) El test χ^2 se aplica para distribuciones continuas y discretas, mientras que el test K-S sólo se aplica para distribuciones continuas.

3) El estadístico K-S es fácil de calcular, pero requiere ordenar los datos y eso puede ser más costoso que el test χ^2.

3.3 Apéndice Capítulo 3

[B1] Tabla Standard para el caso H_0 simple. Valores críticos de $d_{\alpha,n}$, siendo $\alpha = P(\sup_x |F_n(x) - F_0(x)| > d_{\alpha,n})$.

Tamaño de la muestra (n)	Nivel de significación del test (α)				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.210	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
$n \geq 35$	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

[B2] Tabla confeccionada por Lilliefors cuando H_0 es una familia Normal. Esta tabla contiene los valores críticos de $\widehat{D} = sup|F_N(x) - F_{\hat{\theta}}(x)|$ usando el Método de Monte Carlo.

Tamaño de	Nivel de significación del test (α)				
la muestra (n)	0.20	0.15	0.10	0.05	0.01
4	0.300	0.319	0.352	0.381	0.417
5	0.285	0.299	0.315	0.337	0.405
6	0.265	0.277	0.294	0.319	0.364
7	0.247	0.258	0.276	0.300	0.348
8	0.233	0.244	0.261	0.285	0.331
9	0.223	0.233	0.249	0.271	0.311
10	0.215	0.224	0.239	0.258	0.294
11	0.206	0.217	0.230	0.249	0.284
12	0.199	0.212	0.223	0.242	0.275
13	0.190	0.202	0.214	0.234	0.268
14	0.183	0.194	0.207	0.227	0.261
15	0.177	0.187	0.201	0.220	0.257
16	0.173	0.182	0.195	0.213	0.250
17	0.169	0.177	0.189	0.206	0.245
18	0.166	0.173	0.184	0.200	0.239
19	0.163	0.169	0.179	0.195	0.235
20	0.160	0.166	0.174	0.190	0.231
25	0.149	0.153	0.165	0.180	0.203
30	0.131	0.136	0.144	0.161	0.187
$n \geq 30$	$\frac{0.736}{\sqrt{N}}$	$\frac{0.768}{\sqrt{N}}$	$\frac{0.805}{\sqrt{N}}$	$\frac{0.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

[B3] Tabla confeccionada por Lilliefors cuando H_0 es una familia Exponencial. Esta tabla contiene los valores criticos de $\hat{D}$ usando el Método de Monte Carlo.

Tamaño de la muestra (n)	Nivel de significación del test (α)				
	0.20	0.15	0.10	0.05	0.01
3	0.451	0.479	0.511	0.551	0.60
4	0.396	0.422	0.449	0.487	0.548
5	0.359	0.382	0.406	0.442	0.504
6	0.331	0.351	0.375	0.408	0.470
7	0.309	0.327	0.350	0.382	0.442
8	0.291	0.308	0.329	0.360	0.419
9	0.277	0.291	0.311	0.341	0.399
10	0.263	0.277	0.295	0.325	0.380
11	0.251	0.264	0.283	0.311	0.365
12	0.241	0.254	0.271	0.298	0.351
13	0.232	0.245	0.261	0.287	0.338
14	0.224	0.237	0.252	0.277	0.326
15	0.217	0.2290	0.244	0.269	0.315
16	0.211	0.222	0.236	0.261	0.306
17	0.204	0.215	0.229	0.253	0.297
18	0.199	0.210	0.223	0.246	0.289
19	0.193	0.204	0.218	0.239	0.283
20	0.188	0.199	0.212	0.234	0.278
25	0.170	0.180	0.191	0.210	0.247
30	0.155	0.164	0.174	0.192	0.226
$n \geq 30$	$\frac{0.86}{\sqrt{n}}$	$\frac{0.91}{\sqrt{n}}$	$\frac{0.96}{\sqrt{n}}$	$\frac{1.06}{\sqrt{n}}$	$\frac{1.25}{\sqrt{n}}$

Capítulo 4

Estimación no paramétrica de la densidad

Dada una variable aleatoria X continua, su comportamiento puede ser descripto mediante su función de densidad f. En diferentes situaciones, puede ser de interés estimar dicha densidad mediante una muestra aleatoria $X_1, ..., X_n$ de la variable X. Esto puede realizarse mediante un método paramétrico o no paramétrico. Los métodos paramétricos asumen que la densidad es de la forma $f(\cdot, \theta)$, con $\theta \in \mathbb{R}^s$, es decir, consideran que la densidad está unívocamente determinada, salvo por cierto parámetro θ finito dimensional. En tal caso, basta estimar θ para disponer de un estimador de la densidad f. Sin embargo, cuando no tenemos información sobre f y queremos estimarla usando un método paramétrico, podemos llegar a resultados erróneos. Los métodos no paramétricos surgen como solución a este problema. Fueron introducidos por Rosenblatt y Parzen, y apenas asumen condiciones de regularidad sobre la función de densidad.

4.1 Estimadores de la densidad por núcleos

A continuación daremos una idea intuitiva de la construcción de los estimadores no paramétricos de densidad por núcleos. Sea X una variable aleatoria continua con densidad f. Dada una muestra aleatoria $X_1, ..., X_n$ con densidad $f(x)$, nuestro objetivo es estimar la función $f(x)$ a partir de la muestra. Para todo $h > 0$ se satisface lo siguiente:

$$P(X \in (x-h, x+h)) = \int_{x-h}^{x+h} f(t)dt$$

Podemos aproximar $P(X \in (x-h, x+h))$ por la proporción de elementos de la muestra aleatoria que se encuentran dentro de ese intervalo, y de esta manera sustituimos $P(X \in (x-h, x+h))$ por $\frac{\sharp\{X_i \in (x-h, x+h)\}}{n}$, y aproximamos $\int_{x-h}^{x+h} f(t)dt$ por $f(x)2h$, si h es suficientemente pequeño y f continua en x. Luego

$$f(x)\ 2h \approx \frac{\sharp\{X_i \in (x-h, x+h)\}}{n} \tag{4.1}$$

Podemos deducir de (4.1) un estimador reemplazando las aproximaciones por igualdades y obtenemos lo siguiente:

$$\hat{f}(x) = \frac{1}{2h} \frac{\sharp\{X_i \in (x-h, x+h)\}}{n}$$

$\hat{f}(x)$ se puede escribir como:

$$\frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbf{I}_{[-1,1]}\left(\frac{x-X_i}{h}\right).$$

Sea $w(x)=\frac{1}{2}I_{[-1,1]}(x)$, tenemos que $\hat{f}(x)=\sum_{i=1}^{n}\frac{1}{nh}w(\frac{x-Xi}{h})$.

Notemos que $w(x)$ es una función de densidad pues $w\geq 0$ y $\int w(x)=1$ entonces $\hat{f}(x)$ también lo es. Siendo $w(x)$ discontinua, lo mismo ocurre con $\hat{f}(x)$. Si sustituimos $w(x)$ por una función de densidad suave $K(x)$, obtendremos un nuevo estimador de la densidad, dado por

$$\hat{f}_K(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right). \tag{4.2}$$

A la función K se la denomina función núcleo (o Kernel) y a h se lo llama ventana o parámetro de suavizado. En general K es una función de densidad continua, unimodal y simétrica alrededor del 0. $\hat{f}_K$ es suave, pues lo hereda de K, y su comportamiento dependerá de los valores que tome h.

En este trabajo consideraremos núcleos K satisfaciendo las siguientes propiedades:

K.1 $K\geq 0$ y $\int K(u)du=1$

K.2 $K(u)=K(-u)$ y $||K||_\infty<\infty$

K.3 $\int \left|u^i\right|\left|K^j(u)\right|du<\infty$ para $i=1,2,\frac{5}{2}$, $j=1,2$

Proposición 4.1.1. *Sea f una función de densidad continua y acotada, y K un núcleo verificando **K.1** y **K.2**. Entonces, $\hat{f}_K(x)\xrightarrow{p}f(x)$, si $h\to 0$ y $nh\to\infty$.*

Demostración: Vamos a probar bajo los supuestos realizados, que el error cuadrático medio del estimador converge a cero:

$$ECM(\hat{f}_K(x))=\mathbb{E}((\hat{f}_K(x)-f(x))^2)\to 0.$$

Recordemos que

$$ECM(\hat{f}_K(x))=sesgo(\hat{f}_K(x))^2+\mathbb{V}ar(\hat{f}_K(x)),$$

donde $Sesgo(\hat{f}_K(x))=\mathbb{E}(\hat{f}_K(x))-f(x)$. Analizaremos cada término por separado. Para estudiar el sesgo, notemos que, siendo K una función simétrica,

$$\mathbb{E}(\hat{f}_K(x))=\mathbb{E}\left(\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)\right)=\frac{1}{h}\mathbb{E}\left(K\left(\frac{x-X}{h}\right)\right)=\frac{1}{h}\int K\left(\frac{u-x}{h}\right)f(u)du \tag{4.3}$$

Sea $y=\frac{u-x}{h}$ entonces $dy=\frac{1}{h}du$ siendo $u=yh+x$, y por consiguiente

$$\mathbb{E}(\hat{f}_K(x))=\int K(y)f(yh+x)dy$$

Tenemos entonces que

$$|Sesgo(\hat{f}_K(x))|=|\mathbb{E}(\hat{f}_K(x))-f(x)|\leq\int K(y)|f(yh+x)-f(x)|dy=\int G_h(y)dy,$$

con $G_h(y) = K(y)|f(yh+x) - f(x)|$. Ahora bien, siendo f continua, tenemos que $G_h(y)$ converge a cero cuando $h \downarrow 0$, para todo y. Además, si M acota a f, tenemos que $G_h(y) \leq 2MK(y)$, y por consiguiente G_h está dominada por una función integrable. Podemos entonces apelar al teorema de la convergencia dominada para concluir que

$$\lim_{h\to 0} |Sesgo(\hat{f}_K(x))| = 0.$$

Para estudiar la varianza del estimador, notemos que

$$\mathbb{V}ar(\hat{f}_K(x)) = \mathbb{V}ar\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right) =$$
$$\frac{1}{nh^2}\mathbb{V}ar\left(K\left(\frac{x-X}{h}\right)\right) \leq \frac{1}{nh^2}\mathbb{E}\left(K^2\left(\frac{x-X}{h}\right)\right)$$

Siendo K un núcleo acotado, tenemos entonces que

$$\mathbb{V}ar(\widehat{f}_K(x)) \leq ||K||_\infty \frac{1}{nh^2}\mathbb{E}\left(K\left(\frac{x-X}{h}\right)\right)$$

Vimos que $\mathbb{E}(K(\frac{x-X}{h}))/h$ converge a $f(x)$, y por consiguiente, como $nh \uparrow \infty$, concluímos que

$$\mathbb{V}ar(\widehat{f}_K(x)) \to 0. \quad \square$$

Veremos ahora cómo la condición **K.3** permite obtener una mejor caracterización del sesgo y la varianza del estimador $\widehat{f}_K(x)$.

Proposición 4.1.2. *Sea f una función de densidad $C^3(\mathbb{R})$ con $f''(\cdot)$ tal que $\|f'''\|_{L^2} < \infty$ y K un núcleo univariado que satisface **K.1**,**K.2** y **K.3**. Tenemos entonces que*

$$|\mathbb{E}(\hat{f}_K(x)) - f(x)| \leq h^2 \frac{|f''(x)|}{2}\int K(y)y^2dy + o(h^2)\,, \tag{4.4}$$

$$\mathbb{V}ar(\hat{f}_K(x)) \leq \frac{1}{nh}f(x)\int K^2(y)dy + o((nh)^{-1}). \tag{4.5}$$

Demostración: Vimos en la demostración de la Proposición 4.1.1 que

$$Sesgo(\hat{f_K}(x)) = \mathbb{E}(\hat{f_K}(x)) - f(x) = \int K(y)\left\{f(yh+x) - f(x)\right\}dy.$$

Bajo los supuestos de la proposición, haciendo un desarrollo de Taylor de orden 2 alrededor de $x_0 = x$, expresando el resto en la forma integral, obtenemos

$$f(x+yh) = f(x) + f'(x)hy + \frac{1}{2}f''(x)h^2y^2 + \int_x^{x+hy}\frac{f'''(t)}{2!}(x+hy-t)^2dt$$

de donde deducimos que

$$\mathbb{E}(\hat{f}_K(x)) - f(x) =$$
$$f'(x)h\int K(y)y\,dy + \frac{h^2f''(x)}{2}\int K(y)y^2\,dy + \int\int_x^{x+hy}\frac{f'''(t)}{2!}(x+hy-t)^2dtK(y)dy$$

y por consiguiente

$$|\mathbb{E}(\hat{f}_K(x)) - f(x)| =$$
$$\frac{|f''(x)|}{2}h^2\int y^2K(y)dy + \frac{\|f'''\|_{L^2}}{2}\int\left(\int_x^{x+hy}(x+hy-t)^4dt\right)^{\frac{1}{2}}K(y)dy| =$$
$$h^2\frac{|f''(x)|}{2}\int K(y)y^2dy + \frac{\|f'''\|_{L^2}}{2!\sqrt{5}}h^{\frac{5}{2}}\int|y|^{\frac{5}{2}}K(y)dy.$$

Tenemos entonces que

$$|\mathbb{E}(\hat{f}_K(x)) - f(x)| \leq h^2\frac{|f''(x)|}{2}\int K(y)y^2dy + o(h^2).$$

Analicemos ahora la varianza del estimador. En la demostración de la Proposición 4.1.1 vimos que

$$\mathbb{V}ar(\hat{f}_K(x)) \leq \frac{1}{nh^2}\mathbb{E}\left(K^2\left(\frac{x-X}{h}\right)\right),$$

y por consiguiente, reproduciendo los cálculos hechos para el estudio del sesgo, concluímos que

$$\mathbb{V}ar(\hat{f}_K(x)) \leq \frac{1}{nh}f(x)\int K^2(y)dy + o((nh)^{-1}). \quad \square$$

De (4.4) y (4.5) podemos observar que si h es muy grande, el sesgo aumenta, mientras que la varianza disminuye. Si nh es muy pequeño, la varianza no tiende a 0. Es decir, fijando n la elección de h resulta fundamental en la implementación de $\hat{f}(x)$ para equilibrar el sesgo y la varianza. Tenemos entonces el siguiente corolario:

Corolario 4.1.3. *Bajo los supuestos de la Proposición (4.1.2), tenemos que*

$$ECM(\widehat{f}(x)) \leq h^4\frac{(f''(x))^2}{4}\left(\int K(y)y^2dy\right)^2 + o(h^4) + \frac{1}{nh}f(x)\int K^2(y)dy + o((nh)^{-1}). \quad (4.6)$$

y por consiguiente $\hat{f}_K(x)$ converge en probabilidad a $f(x)$ si $h \to 0$ y $nh \to \infty$.

Los núcleos más conocidos son:

- Gaussiano o Normal :
$$K_G(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}.$$
- Uniforme:
$$K_U(x) = \begin{cases} \frac{1}{2} & \text{si } |x| \leq 1, \\ 0 & \text{si} |x| > 0. \end{cases}$$
- Triangular:
$$K_T(x) = \begin{cases} 1-|x| & \text{si} |x| \leq 1, \\ 0 & \text{si} |x| > 1. \end{cases}$$
- Epanechnikov:
$$K_E(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{si} |x| \leq 1, \\ 0 & \text{si } |x| > 1. \end{cases}$$

En la práctica, la elección del núcleo no parece jugar un papel muy importante en la estimación de la densidad. Sin embargo, el estimador es muy sensible al valor de la ventana utilizado, como muestra la siguiente figura.

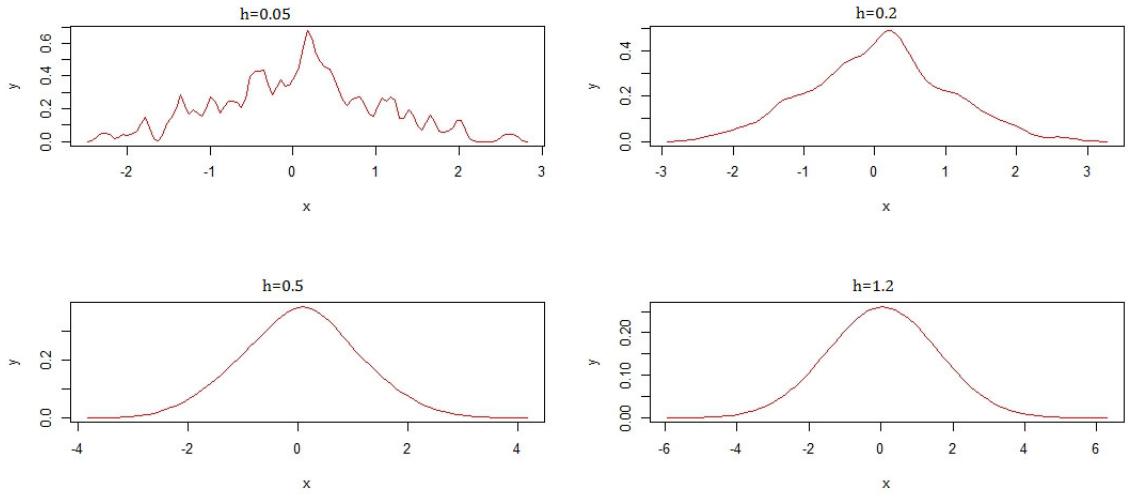

Figure 4.1: Usamos una muestra de tamaño 200, $X_i \sim N(0,1)$, el núcleo $N(0,1)$ y diferentes valores de h

En la siguiente sección resumiremos algunas de las propuestas existentes para seleccionar el parámetro de suavizado h.

4.2 Selección de la ventana

4.2.1 Método Plug-in

El error cuadrático medio asintótico (ECMA) de $\hat{f}_K(x)$, surge de despreciar en (4.6) los términos de orden pequeño, dando origen al error cuadrático medio asintótico integrado, definido por

$$ECMAI(\hat{f}_K(x)) = \frac{\|K\|_{L^2}^2}{nh} + \frac{h^4 \|f''\|_{L^2}^2}{4} \left(\int y^2 K(y) dy \right)^2 .$$

Elegimos el valor de h de forma tal de minimizar esta última expresión. Con este criterio obtenemos

$$h_{opt} = \left(\frac{\|K\|_{L^2}^2}{\|f''\|_{L^2}^2 (\int y^2 K(y) dy)} \right)^{\frac{1}{5}} n^{-\frac{1}{5}} \tag{4.7}$$

Observar que en este caso necesitamos conocer el valor de $\|f''\|_{L^2}^2$. En la práctica, podemos estimar este valor mediante un estimador no paramétrico con un valor de h fijado *a priori*.

Otros autores proponen considerar la *regla de referencia a la normal*, desarrollada por Silverman. Este método consiste en suponer que f sigue una distribución Normal $N(\mu, \sigma^2)$, en cuyo caso $\|f''\|_{L^2}^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}}$. Por otra parte, si además elegimos un núcleo Gaussiano, resulta que

$$\int y^2 K_G(y) dy = 1 \ , \quad \int K_G^2(y) dy = \frac{1}{2\sqrt{\pi}}$$

de donde concluímos que

$$h_{opt} = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \simeq \frac{1.059}{n^{\frac{1}{5}}}\hat{\sigma},$$

siendo $\hat{\sigma}$ un estimador consistente de σ. En general se propone:

$$\widehat{\sigma} = min\left\{\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \frac{\widehat{Q}}{1.34}\right\} \tag{4.8}$$

donde $\widehat{Q}$ es un estimador del rango intercuartil. Recordemos que el rango intercuartil es el percentil 75 menos el percentil 25. Dividimos a $\widehat{Q}$ por 1.34 y de esa manera conseguimos que $\frac{\widehat{Q}}{1.34}$ sea un estimador consistente de σ, bajo normalidad.

4.2.2 Elección de la ventana mediante convalidación cruzada

Los métodos de convalidación consisten en minimizar la distancia entre f y $\widehat{f}_K$, en este caso usaremos la norma en $L^2(\mathbb{R})$

$$\|\widehat{f}_K(x) - f(x)\|_{L_2} = \sqrt{\int_{-\infty}^{+\infty}(\widehat{f}_K(x) - f(x))^2 dx}.$$

Definimos

$$h_{opt} = \min_h \|\widehat{f}_K(x) - f(x)\|_{L_2}^2 \tag{4.9}$$

Reescribimos lo anterior,

$$\|\widehat{f}_K(x) - f(x)\|_{L_2}^2 = \int(\widehat{f}_K(x) - f(x))^2 dx = \int \widehat{f}_K^2(x)dx - 2\int \widehat{f}_K f(x)dx + \int f^2(x)dx$$

Pero $f(x)$ no depende de h, entonces minimizar

$$\int \widehat{f}_K^2(x)dx - 2\int \widehat{f}_K f(x)dx$$

es equivalente a minimizar $\|\hat{f}_K(x) - f(x)\|_{L_2}^2$.
Ahora estimemos cada término en función de nuestra muestra.

$$\int \hat{f}_K(x) = \int\left(\frac{1}{nh}\sum K\left(\frac{x - X_i}{h}\right)\right)^2 dx = \frac{1}{n^2h}\sum_{i,j}^{n}\int K\left(\frac{x - X_i}{h}\right)K\left(\frac{x - X_j}{h}\right)dx =$$

$$\frac{1}{n^2h}\sum_{i,j}^{n}\int K(y)K\left(y - \frac{X_j - X_i}{h}\right)dy = \frac{1}{n^2h}\sum_{i,j}^{n}K * K\left(\frac{X_i - X_j}{h}\right)$$

Recordemos que: $\int \widehat{f}_K(x)f(x)dx = \mathbb{E}_f(\hat{f}_K(x))$.
Estimamos la $\mathbb{E}_f(\widehat{f}_K(x))$ como $\frac{1}{n}\sum_i^n \hat{f}_{K,-i}(X_i)$ siendo $\hat{f}_{K,-i}(X_i)$ el estimador Leave-one-out de $f(X_i)$,

$$\hat{f}_{K,-i}(X_i) = \frac{1}{h(n-1)}\sum_{j\neq i}^{n}K\left(\frac{X_i - X_j}{h}\right) \tag{4.10}$$

Restando $\frac{1}{n^2 h}\sum_{i,j}^{n} K * K\left(\frac{X_i - X_j}{h}\right)$ y (4.10) queda definido $\hat{h}_{CV}$.

$$\hat{h}_{CV} = argmin_h CV(h) = argmin_h \frac{1}{n^2 h}\sum_{i,j}^{n} K * K\left(\frac{X_i - X_j}{h}\right) - \frac{2}{h(n-1)}\sum_{i\neq j}^{n} K\left(\frac{X_i - X_j}{h}\right)$$

Este método fue uno de los primeros intentos de buscar el parámetro h de forma automática. Una de las desventajas del método es que $CV(h)$ puede tener varios mínimos locales.

4.2.3 Validación cruzada por Máxima Verosimilitud

Vimos que para la elección del parámetro de suavizado h se puede usar el método de validación cruzada. Dado un h estimamos la verosimilitud de X_i a partir del estimador no paramétrico de la densidad calculado con el resto de la muestra y ese valor de h. Notemos con

$$\hat{f}_{h,-i}(X_i) = \frac{1}{h(n-1)}\sum_{j\neq i}^{n} K\left(\frac{X_i - X_j}{h}\right) \tag{4.11}$$

Definimos la verosimilitud de la muestra por validación cruzada para el valor h como

$$L_{CV}(h) = \prod_{i=1}^{n} \hat{f}_{h,(-i)}(X_i)$$

y tomo h_{LCV} siendo el valor que maximiza L_{CV}:

$$h_{CV} = argmax_h L_{CV}(h)$$

Sin embargo este método no es tan utilizado.

Capítulo 5

Test de bondad de ajuste basado en estimadores por núcleos

En este capítulo veremos cómo las técnicas de estimación no paramétrica de la función de densidad pueden ser utilizadas para tratar problemas de bondad de ajuste. Estos procedimientos fueron introducidos por Bickel *et al.* [1] para el caso de hipótesis nula simple, mientras que Fan [4] considera el caso compuesto. Nuestra presentación considerará variables aleatorias unidimensionales, aunque los trabajos mencionados tratan el caso multivariado. Es decir, consideran el problema para variables tomando valores en $\mathbb{R}^d$. Recientemente, Boente *et al* [2] estudiaron este tipo de técnicas para problemas de bondad de ajuste con datos direccionales.

Sean $X_1, \ldots, X_n$ observaciones independientes de una variable aleatoria X con función de densidad $f(x)$. Comenzaremos estudiando el caso de la hipótesis nula simple. Consideraremos luego el caso de hipótesis nula compuesta.

5.1 Estadístico del test de bondad de ajuste por núcleos

5.1.1 Hipótesis nula simple

Vamos a construir tests para

$$H_0 : f(x) = f_0(x) \quad vs \quad H_1 : f(x) \neq f_0(x)$$

Para determinar si $f_0(x)$ ajusta bien nuestros datos, Bickel y Rosenblatt, en [1], proponen estimar la discrepancia entre $f_0(x)$ y $f(x)$. Una posible medida de discrepancia está dada por el error cuadrático integrado, definido por $I = \int [f(x) - f_0(x)]^2 dx$. La medida I se puede usar como indicador de los errores de un modelo. Un estimador de tipo plug-in de I queda definido al sustituir $f(x)$ por $\hat{f}(x)$, un estimador no paramétrico de la densidad. En adelante, consideraremos estimadores basados en núcleos, como los estudiados en el Capítulo 4, definidos por

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

siendo K el núcleo y h el parámetro de suavizado. De esta forma, podemos estimar I con

$$I_{s,n} = \int [\hat{f}(x) - f_0(x)]^2 dx.$$

Como se discute en Fan [4], la elección de h juega un rol crucial en el comportamiento asintótico del estimador $I_{s,n}$.

Vamos ahora a presentar una modificación a esta primera propuesta, que contemple el sesgo de $\hat{f}(x)$. Es decir, hemos visto en la fórmula (4.3) del Capítulo 4 que $E[\hat{f}(x)] = K_h * f_0(x)$ bajo H_0, siendo $(K_h * f_0)(x)$ la convolución entre $f_0(x)$ y $K_h(u) := \frac{1}{h}K(\frac{u}{h})$, definida por

$$(K_h * f_0)(x) = \int K_h(x-u)f_0(u)du = \int \frac{1}{h}K\left(\frac{x-u}{h}\right)f_0(u)du.$$

Este hecho sugiere considerar

$$J_{s,n} = \int [\hat{f}(x) - (K_h * f_0)(x)]^2 dx$$

para cuantificar discrepancias entre $f(x)$ y $f_0(x)$. Bickel (Teorema 4.1) [1] estudia el comportamiento asintótico de $J_{s,n}$ y, bajo condiciones de regularidad para el núcleo K y para f_0, demuestra que bajo H_0

$$n\sqrt{h}\frac{J_{s,n} - \mu(K)}{\sigma(K, f_0)} \longrightarrow N(0,1) \tag{5.1}$$

cuando $h \to 0$ y $nh \to \infty$, siendo

$$\mu(K) = \frac{1}{nh}\int K^2(u)du,$$
$$\sigma^2(K, f_0) = 2\left(\int \left[\int K(u+v)K(v)dv\right]^2 du\right)\left(\int f_0^2(x)dx\right).$$

Notemos que $\mu(K)$ y $\sigma^2(K, f_0)$ quedan unívocamente determinados por K y f_0. Este resultado sugiere rechazar H_0 cuando observamos valores grandes del estadístico $J_{s,n}$. Más precisamente, fijado el nivel α, rechazamos H_0 si el valor observado del estadístico $J_{s,n}$ supera a $\mu(K) + z_{1-\alpha}\frac{\sigma^2(K,f_0)}{n\sqrt{h}}$.

5.1.2 Hipótesis nula compuesta

Consideremos ahora el caso en el que

$$H_0 : f \in \mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^s\} \quad vs \quad H_1 : f \notin \mathcal{F}$$

donde ahora, para cada $\theta \in \Theta$, f_θ es una función de densidad. A modo de ejemplo, podemos pensar que $\mathcal{F}$ es la familia de densidades normales, indexada por $\theta = (\mu, \sigma^2)$, la media y la varianza.

Vamos ahora a adaptar los estadísticos definidos en la sección anterior, reemplazando f_0 por $f_{\hat{\theta}}$, siendo $\hat{\theta}$ un estimador *bueno* del parámetro del modelo. Bajo condiciones de regularidad, podemos utilizar el estimador de máxima verosimilitud, siendo el argumento que maximiza la función de verosimilitud. Es decir,

$$\hat{\theta} = argmax_{\theta\in\Theta} L(\theta)$$

para

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

Siguiendo con el modelo normal, el estimador de máxima verosimilitud de θ es $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, con $\hat{\mu} = n^{-1}\sum X_i$, $\hat{\sigma}^2 = n^{-1}\sum(X_i - \hat{\mu})^2$. Fan propuso los siguientes estadísticos basándose en el trabajo de Bickel y Rosenblatt [1]:

$$I_n = \int [\hat{f}(x) - f_{\hat{\theta}}(x)]^2 dx, \tag{5.2}$$

$$J_n = \int [\hat{f}(x) - (K_h * f_{\hat{\theta}})(x)]^2 dx. \tag{5.3}$$

La distribución asintótica de I_n depende de la relación entre el parámetro de regularidad y el tamaño muestral. Por otra parte, bajo condiciones de regularidad para el estimador $\hat{\theta}$, para la familia de densidades del modelo y para el núcleo, Fan (Teorema 4.1) demuestra que bajo H_0, cuando $h \to 0$ y $nh \to \infty$, la distribución asintótica de J_n coincide con la presentada en (5.1). Es decir, tenemos que

$$n\sqrt{h}\frac{J_n - \mu(K)}{\sigma(K, f_\theta)} \longrightarrow N(0,1) \tag{5.4}$$

cuando $h \to 0$ y $nh \to \infty$, siendo

$$\mu(K) = \frac{1}{nh}\int K^2(u)du,$$
$$\sigma^2(K, f_\theta) = 2\left(\int \left[\int K(u+v)K(v)dv\right]^2 du\right)\left(\int f_\theta^2(x)dx\right).$$

Notemos que en el presente contexto, la varianza asintótica depende de la densidad f_θ. Este hecho sugiere rechazar H_0 para valores grandes de

$$n\sqrt{h}\frac{J_n - \mu(K)}{\widehat{\sigma}}$$

siendo $\hat{\sigma}$ un estimador de $\sigma(K, f_\theta)$. Podemos considerar un estimador $\hat{\sigma}$ de tipo plug-in, reemplazando en $\sigma(K, f_\theta)$ la función de densidad f_θ por $\hat{f}$, o bien por $f_{\hat{\theta}}$.

En síntesis, la hipótesis nula es rechazada a nivel α para valores de J_n que superan el valor crítico $\mu(K) + z_{1-\alpha}\frac{\hat{\sigma}}{n\sqrt{h}}$.

Cabe mencionar que la mejor elección del parámetro de suavizado es un problema abierto.

5.1.3 Procedimientos de remuestreo para la distribución de I_n y J_n

Vamos ahora a considerar un procedimiento de tipo Bootstrap para el cálculo de p-valores correspondientes a los estadísticos I_n y J_n, definidos en (5.2) y (5.3), respectivamente, cuando la hipótesis nula es compuesta. Recordemos que ambos estadísticos dependen de h, el parámetro de suavizado que se utiliza en la estimación no paramétrica de la función de densidad. Siguiendo las ideas presentadas en el Capítulo 3, vamos a aproximar la distribución de los estadísticos mencionados bajo H_0 mediante técnicas paramétricas de remuestreo, como se describe a continuación:

1) Dada una muestra aleatoria $X_1, ..., X_n$ denotamos con $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a un estimador de θ bajo el modelo $\mathcal{F} = \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^s\}$.

2) Calculamos el estadístico $I_n = I_n(X_1, \ldots, X_n)$, como en (5.2).

3) Fijado $Nboot$, para $t \in \{1, \ldots, Nboot\}$, repetimos los siguientes pasos:

   3.1) Generamos una muestra $X_1^*, ..., X_n^*$, con densidad $f_{\hat{\theta}}$.

   3.2) Calculamos el estadístico de la muestra bootstrap:

$$I_{t,n}^* = \int [\hat{f}^*(x) - f_{\hat{\theta}^*}(x)]^2 dx$$

   utilizando $\hat{\theta}^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$ y $\hat{f}^*(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i^*}{h}\right)$ en la construcción de $I_{t,n}^*$.

4) Finalmente calculamos la proporción de valores $I^*_{t,n}$ mayores que I_n:

$$\text{p-Iboot} = \frac{1}{Nboot} \sum_{t=1}^{Nboot} I_{\{I^*_{t,n} \geq I_n\}}. \tag{5.5}$$

De la misma manera, procedemos con el estadístico J_n:

1) Dada una muestra aleatoria $X_1, ..., X_n$ denotamos con $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ a un estimador de θ bajo el modelo $\mathcal{F} = \{f_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\}$.

2) Calculamos el estadístico $J_n = J_n(X_1, \ldots, X_n)$, como en (5.3).

3) Fijado $Nboot$, para $t \in \{1, \ldots, Nboot\}$, repetimos los siguientes pasos:

   3.1) Generamos una muestra $X^*_1, ..., X^*_n$, con densidad $f_{\hat{\theta}}$.

   3.2) Calculamos el estadístico de la muestra bootstrap:

$$J^*_{t,n} = \int [\hat{f}^*(x) - (K_h * f_{\hat{\theta}^*})(x)]^2 dx.$$

   utilizando $\hat{\theta}^* = \hat{\theta}(X^*_1, \ldots, X^*_n)$ y $\hat{f}^*(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X^*_i}{h}\right)$ en la construcción de $J^*_{t,n}$.

4) Finalmente calculamos la proporción de valores $J^*_{t,n}$ mayores que J_n:

$$\text{p-Jboot} = \frac{1}{Nboot} \sum_{t=1}^{Nboot} I_{\{J^*_{t,n} \geq J_n\}}. \tag{5.6}$$

5.2 Estudio de simulación

En esta sección calculamos el nivel y la potencia empírica de los procedimientos descriptos en la sección anterior, mediante un estudio de simulación. En todos los casos, consideramos la hipótesis nula compuesta que asume que los datos provienen de una distribución normal:

$$H_0 : F \in \mathcal{F} = \{f_\theta : \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}\},$$

donde f_θ denota la función de densidad normal con media μ y varianza σ^2. Se realizaron $Nrep = 1000$ replicaciones con tamaño muestral $n = 100$, generando datos normales con media cero y varianza uno, (H_0), y tres escenarios alternativos. En todos los casos consideramos el núcleo normal. Consideramos cuatro diferentes valores de h a la hora de definir los estadísticos I_n y J_n. Elegimos h en $\{0.05, 0.2, 0.5, 1\}$. En cada caso, se estimaron los parámetros del modelo con $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$, siendo $\hat{\mu}$ el promedio muestral y $\hat{\sigma}^2$ el estimador insesgado de la varianza (dividiendo por $n - 1$). Para cada conjunto de datos, calculamos p-Iboot y p-Jboot con $Nboot = 200$.

En el Capítulo 3 mencionamos que, en la práctica, la selección de los intervalos para aplicar el test χ^2 se realiza en base a los propios datos. En el presente contexto, con el propósito de proponer un procedimiento *data driven*, podemos optar por elegir el parámetro de suavizado de forma tal que dependa de la muestra. Incluímos en la tabla los resultados correspondientes a los estadísticos construídos utilizando h_{cv}, un procedimiento de validación cruzada para la selección de ventana. En nuestra simulación, dada una muestra, usamos el comando density(muestra)$bw para hallar

este valor. En tal caso, el procedimiento Bootstrap calcula el estadístico utilizando el parámetro de suavizado asociado a la muestra $X_1^*, \ldots, X_n^*$.

Las siguientes tablas muestran la proporción de veces (en las $Nrep = 1000$ replicaciones) en que los p-valores calculados mediante los procedimientos Bootstrap descriptos en la sección anterior resultaron menores que α, tomando $\alpha = 0.05$ en las Tablas 5 y 6 y $\alpha = 0.2$ en las Tablas 7 y 8. Los resultados correspondientes al estadístico I_n, definido en (5.2), se encuentran en las Tablas 5 y 7, mientras que en las Tablas 6 y 8 trabajamos con el estadístico J_n, definido en (5.3). Los datos generados bajo H_0 permiten estudiar el nivel empírico, mientras que en los restantes escenarios se manifiesta la potencia de los procedimientos.

Tabla 5 -Usamos I_n. Proporción de rechazos de H_0 a nivel $\alpha = 0.05$ en $Nrep = 1000$ con muestras de tamaño 100, considerando diferentes valores de h.

h	h_{cv}	0.05	0.2	0.5	1
$N(0,1)$	0.049	0.059	0.060	0.041	0
$\mathcal{U}(-1,1)$	0.677	0.802	0.987	0.244	0
t_5	0.329	0.1537	0.2505	0.1767	0.0045
$\varepsilon(1)$	1	0.999	1	0.960	0.005

Tabla 6 -Usamos J_n. Proporción de rechazos de H_0 a nivel $\alpha = 0.05$ en $Nrep = 1000$ con muestras de tamaño 100, considerando diferentes valores de h.

h	h_{cv}	0.05	0.2	0.5	1
$N(0,1)$	0.041	0.064	0.062	0.052	0.052
$\mathcal{U}(-1,1)$	0.989	0.795	0.934	0.839	0.018
t_5	0.445	0.1577	0.2805	0.4112	0.4960
$\varepsilon(1)$	1	1	1	1	1

Tabla 7-Usamos I_n. Proporción de rechazos de H_0 a nivel $\alpha = 0.2$ en $Nrep = 1000$ con muestras de tamaño 100, considerando diferentes valores de h.

h	h_{cv}	0.05	0.2	0.5	1
$N(0,1)$	0.199	0.218	0.212	0.169	0.002
$\mathcal{U}(-1,1)$	1	0.959	1	0.939	0
t_5	0.515	0.3743	0.49108	0.3184	0.0060
$\varepsilon(1)$	1	0.999	1	0.993	0.014

Tabla 8 -Usamos J_n. Proporción de rechazos de H_0 a nivel $\alpha = 0.2$ en $Nrep = 1000$ con muestras de tamaño 100, considerando diferentes valores de h.

h	h_{cv}	0.05	0.2	0.5	1
$N(0,1)$	0.199	0.210	0.219	0.216	0.184
$\mathcal{U}(-1,1)$	1	0.953	0.994	0.993	0.510
t_5	0.672	0.3723	0.5260	0.6460	0.7136
$\varepsilon(1)$	1	1	1	1	1

En las Tablas 5, 7 y 8 podemos observar que cuando generamos muestras con distribución $N(0,1)$ y elegimos el parámetro de suavizado mediante el criterio de convalidación cruzada ($h = h_{cv}$), el nivel empírico se condice con el nivel nominal del test.

El nivel utilizando el estadístico J_n resulta menos sensible a la elección del parámetro de suavizado. En la primera fila de las Tablas 6 y 8 obtenemos valores más estables a lo largo de las diferentes columnas, respecto de lo que observamos en las Tablas 5 y 7.

De hecho el procedimiento basado en I_n se vuelve sumamente conservativo a medida que h aumenta (Tablas 5 y 7).

En lo que respecta a la potencia de los procedimientos, observamos que cuando generamos datos con distribución uniforme la potencia disminuye con $h = 1$, tanto para I_n como para J_n. Entendemos que esta elección de h ejerce un efecto de sobresuavizado, impidiendo identificar diferencias respecto del modelo nulo.

Cuando generamos bajo los modelos t_5 y $\varepsilon(1)$, la potencia basada en el estadístico I_n disminuye al aumentar el valor de h. Sin embargo, la potencia empírica basada en el estadístico J_n no disminuye cuando h aumenta, como se puede observar en las últimas filas de las Tablas 6 y 8.

En términos generales, valores grandes de h no permiten que $\hat{f}(x)$ estime bien a la función de densidad de los datos.

5.2.1 Gráficos relacionados con los estadísticos del test

Con el propósito de analizar los resultados presentados en las tablas de la sección anterior, presentaremos algunos gráficos que permitirán mostrar cómo el parámetro de suavizado incide en el resultado del procedimiento propuesto para medir bondad de ajuste. Recordemos que en todos los casos se trata de determinar si la distribución normal resulta un buen modelo, siendo que los datos se generan bajo diferentes distribuciones. Tenemos entonces que $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ y $f_{\hat{\theta}}$ representa la densidad normal $N(\hat{\mu}, \hat{\sigma}^2)$. Para cada una de las distribuciones utilizadas a la hora de generar los datos, representamos en un mismo gráfico las siguientes funciones:

- $f(x)$: función de densidad con la que se generan los datos.
- $\hat{f}(x)$: estimador no paramétrico de la densidad, utilizando un núcleo Gaussiano, y parámetro de suavizado h, especificado en cada caso.
- $f_{\hat{\theta}}(x)$: función de densidad con los parámetros estimados bajo el modelo contemplado en H_0.
- $(K_h * f_{\hat{\theta}})(x)$: suavizado de la función de densidad estimada bajo el modelo.

En todos los casos observamos que el estimador no paramétrico de la densidad se torna más suave en la medida que aumentamos el valor del parámetro de suavizado. De hecho, cuando $h = 1$ el efecto de sobresuavizado es tal que el procedimiento propuesto pierde potencia. Es decir, perdemos capacidad para detectar que el modelo normal no es adecuado, siendo que el estimador no paramétrico resulta muy parecido a una densidad normal. Probablemente, el hecho de estar trabajando con un núcleo normal colabore en este sentido. En la figura 5.1, 5.2, 5.3, 5.4, 5.5 generamos una muestra de tamaño $n = 100$ con distribución normal de media cero y varianza uno.

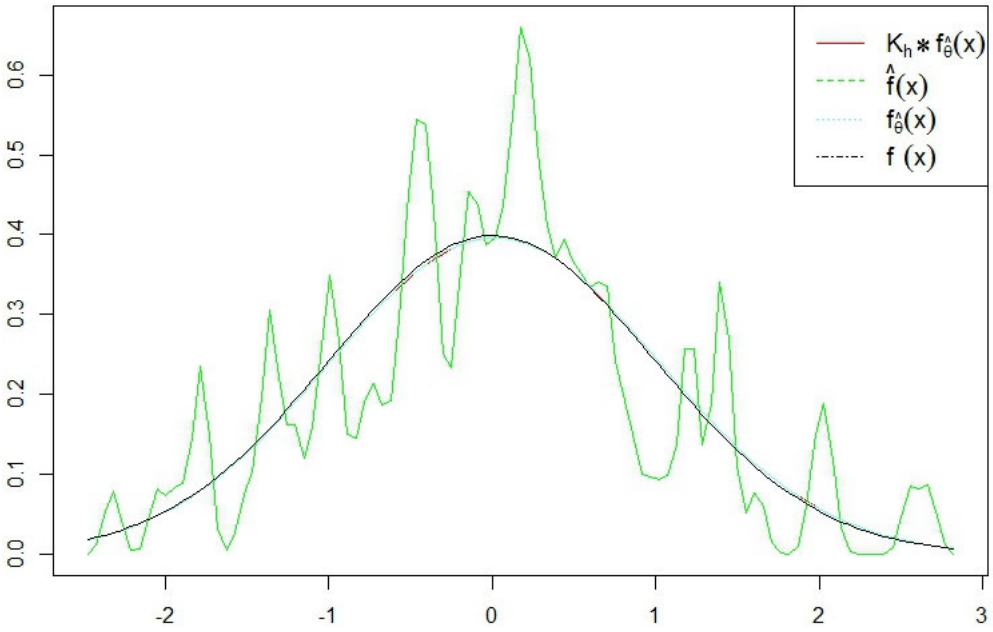

Figure 5.1: $X \sim N(0,1)$, $h = 0.05$

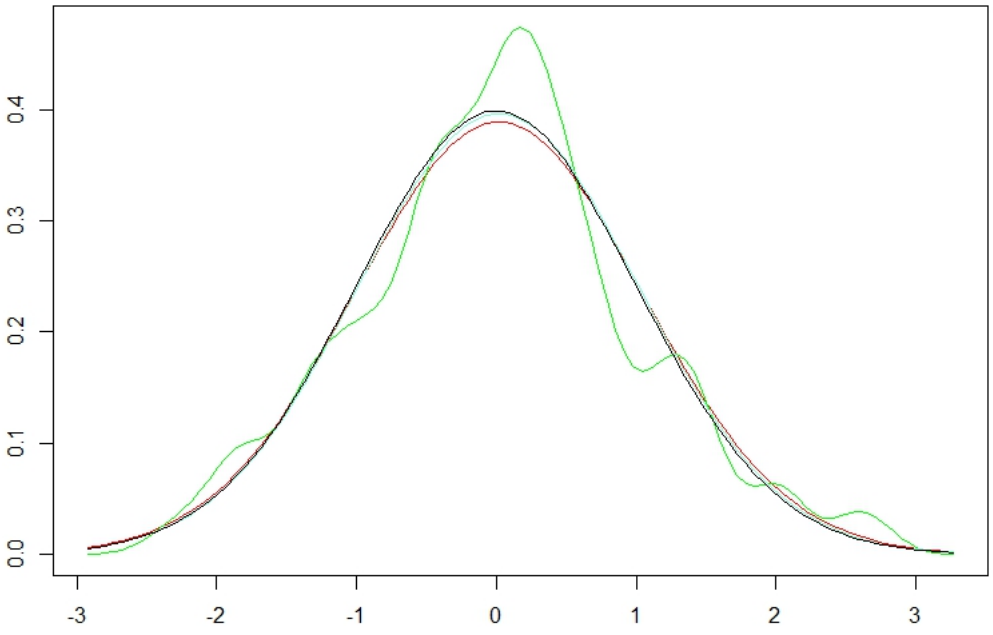

Figure 5.2: $X \sim N(0,1)$, $h = 0.2$

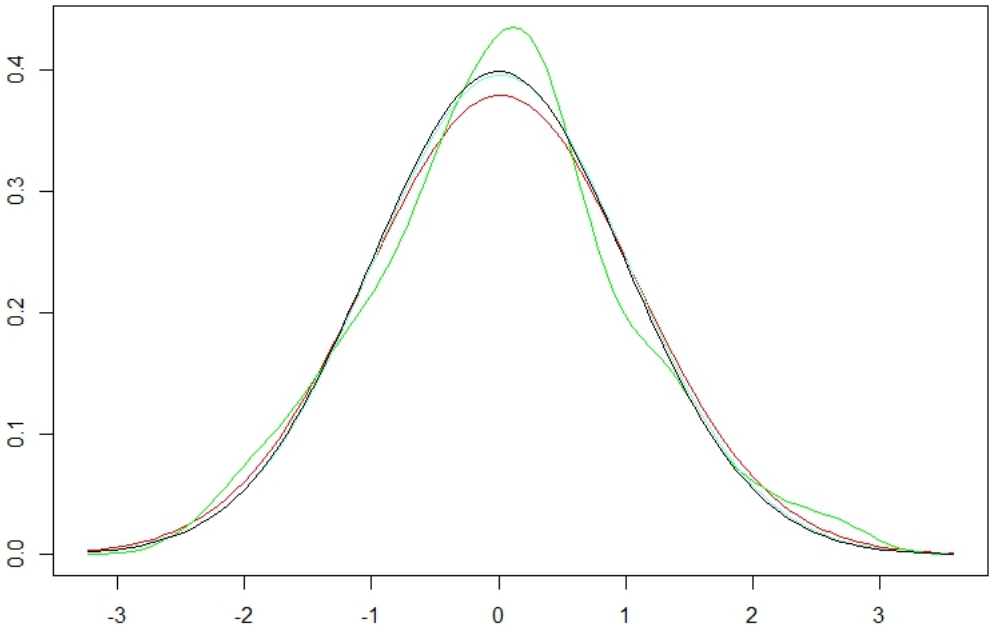

Figure 5.3: $X \sim N(0,1)$, $h = h_{cv}$

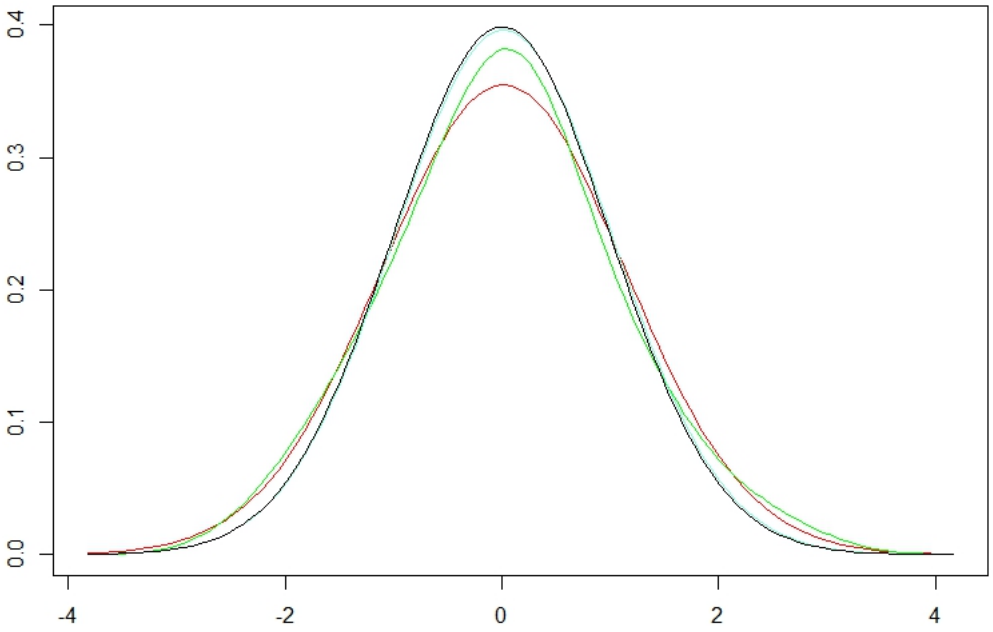

Figure 5.4: $X \sim N(0,1)$, $h = 0.5$

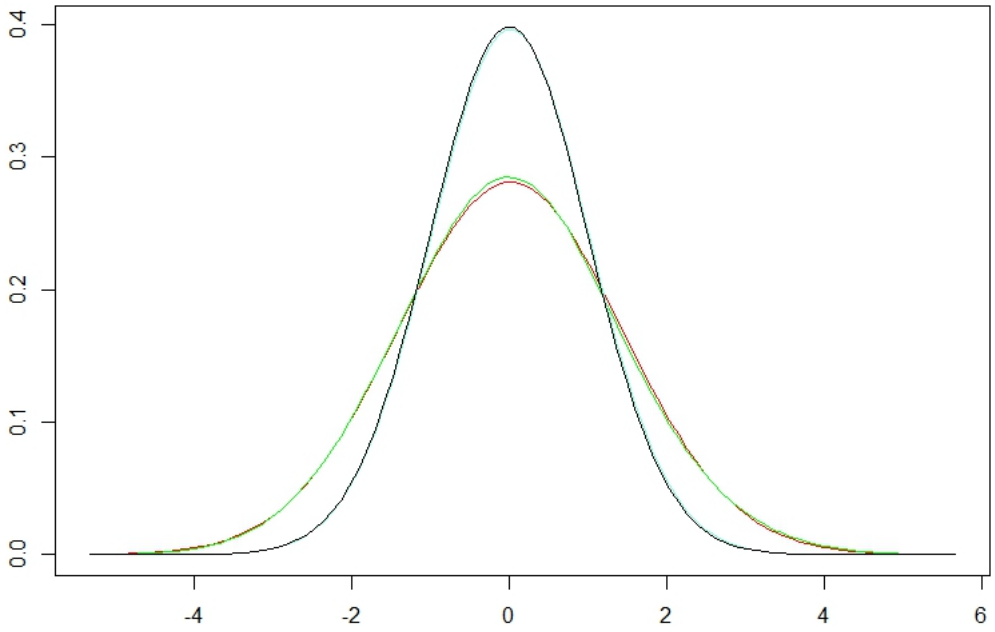

Figure 5.5: $X \sim N(0,1)$, $h = 1$

En las figuras 5.6, 5.7, 5.8, 5.9 y 5.10 usamos una muestra de tamaño $n = 100$ con distribución $\varepsilon(1)$.

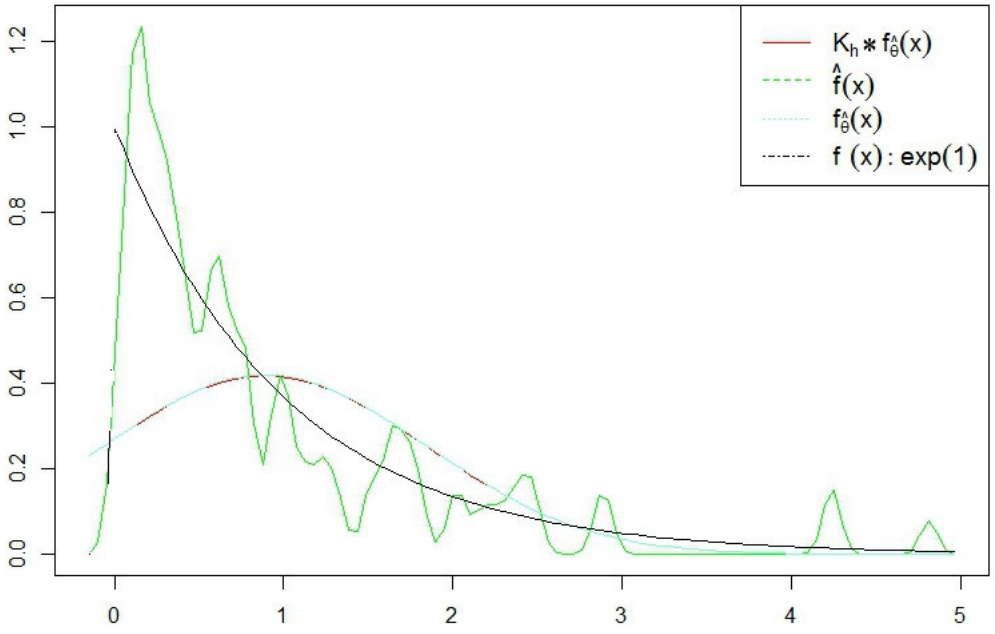

Figure 5.6: $X \sim \varepsilon(1)$, $h = 0.05$

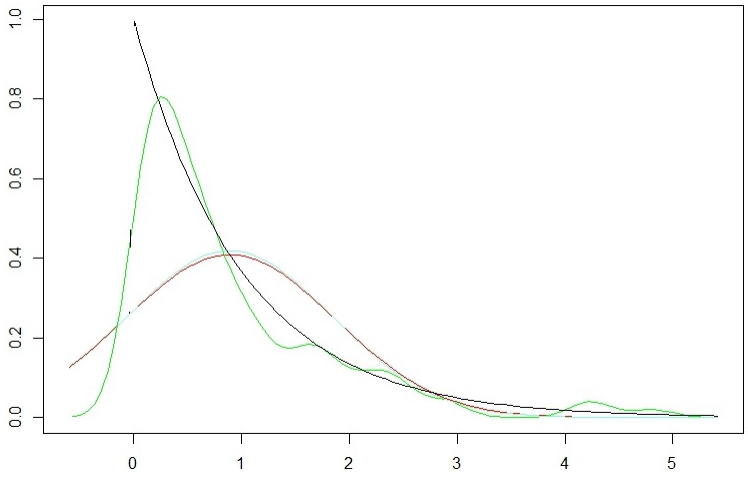

Figure 5.7: $X \sim \varepsilon(1)$, $h = 0.2$

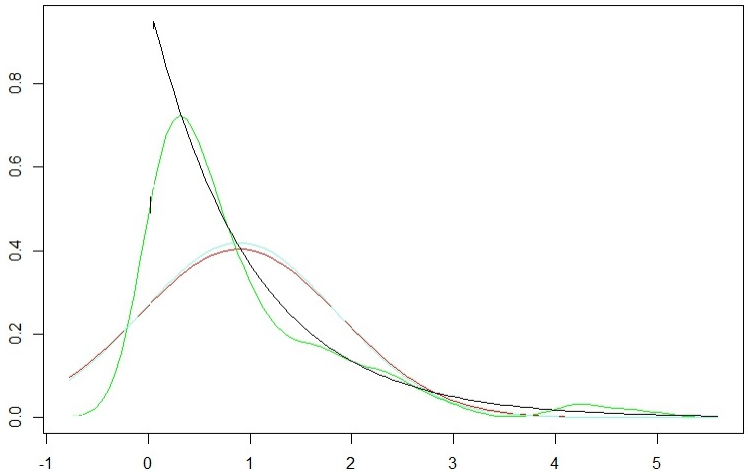

Figure 5.8: $X \sim \varepsilon(1)$, $h = h_{cv}$

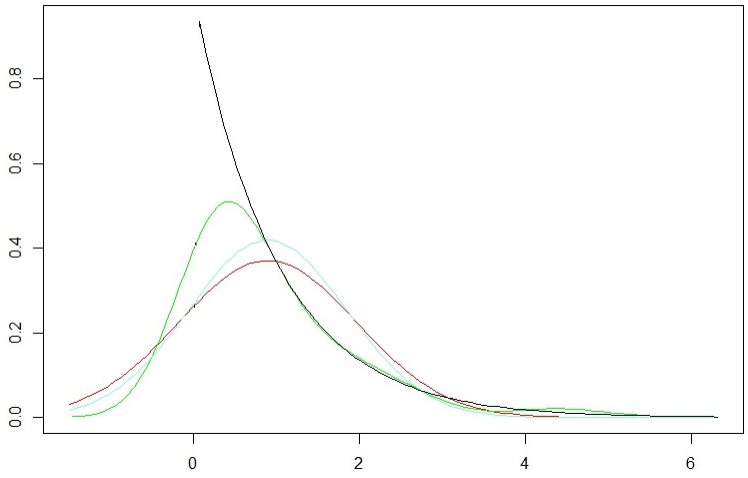

Figure 5.9: $X \sim \varepsilon(1)$, $h = 0.5$

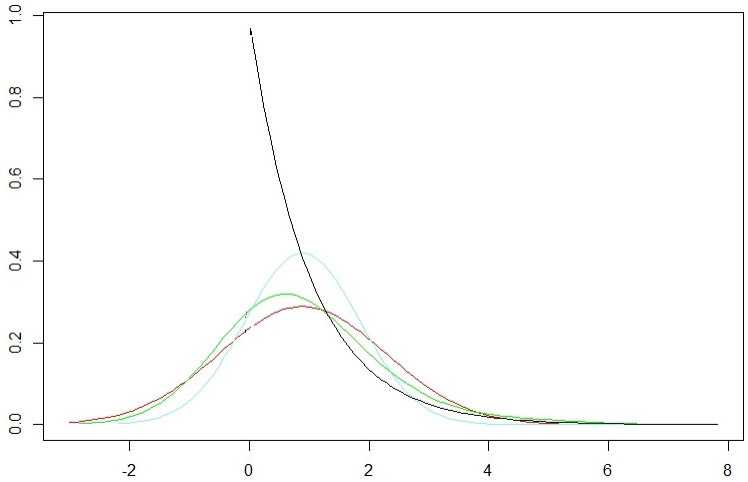

Figure 5.10: $X \sim \varepsilon(1)$, $h = 1$

En las figuras 5.11, 5.12, 5.13, 5.14 y 5.15 nuestra muestra ha sido generada con distribución $\mathcal{U}(-1, 1)$, y tamaño $n = 100$.

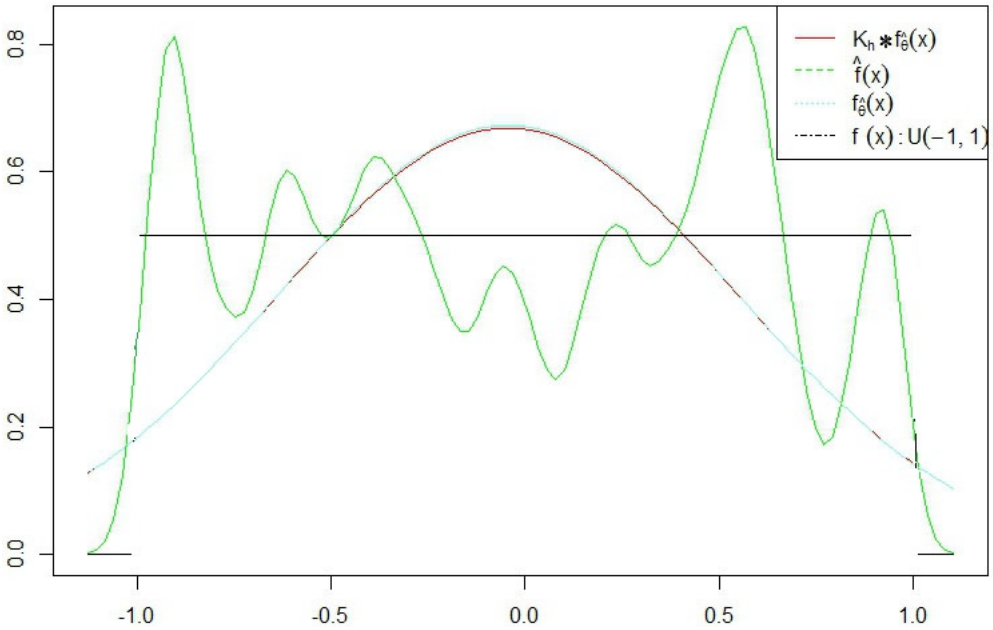

Figure 5.11: $X \sim U(-1,1)$, $h = 0.05$

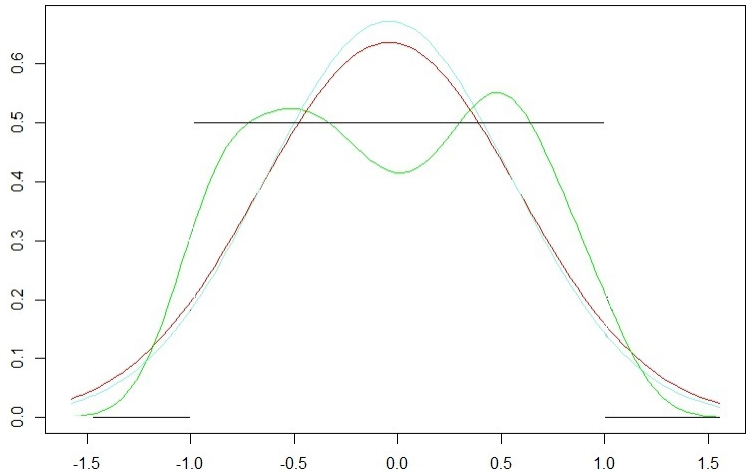

Figure 5.12: $X \sim U(-1,1)$, $h = 0.2$

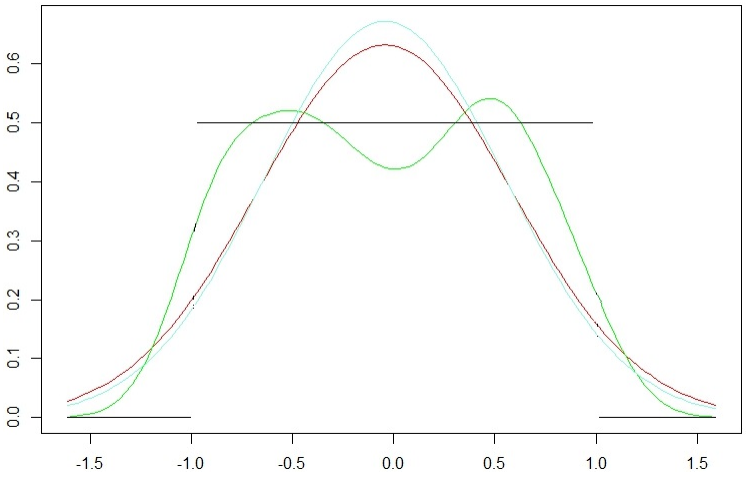

Figure 5.13: $X \sim U(-1,1)$, $h = h_{cv}$

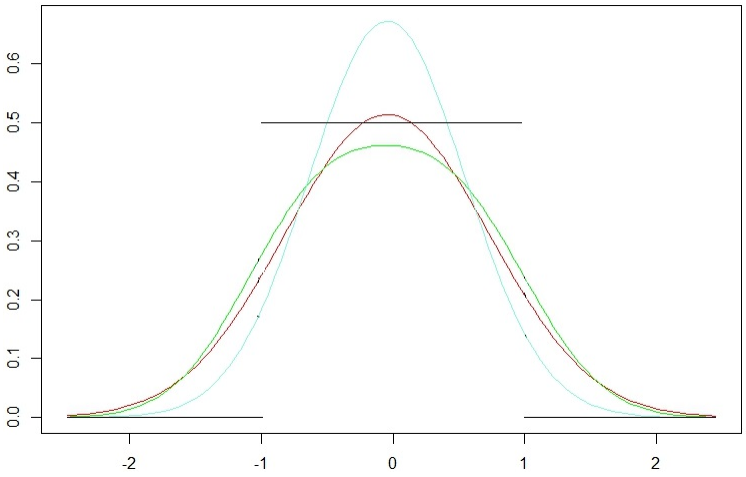

Figure 5.14: $X \sim U(-1,1)$, $h = 0.5$

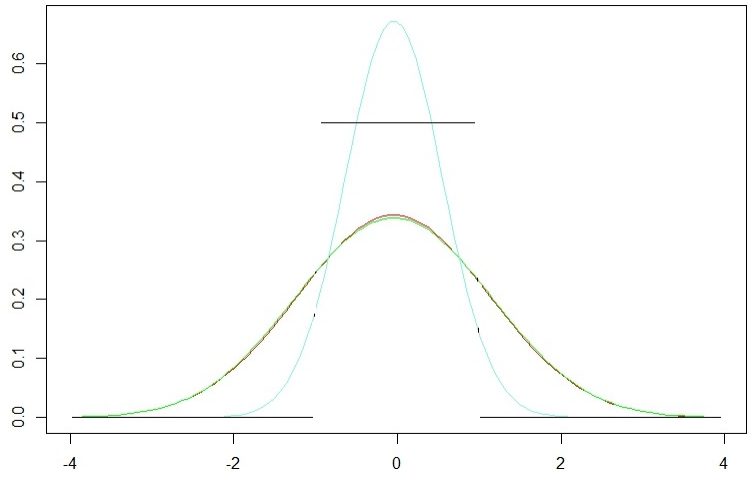

Figure 5.15: $X \sim U(-1,1)$, $h = 1$

Como mencionamos en la sección anterior la elección de h juega un rol crucial en la decisión que tomemos.

5.3 Comparación Final

Para terminar esta presentación, incluímos una tabla resumiendo los resultados obtenidos con los diferentes métodos, en lo que respecta a nivel y potencia empírica, utilizando $n = 100$ y los p-valores obtenidos con el método bootstrap, a los cuales llamamos p-boot.

Tabla con el nivel y la potencia empírica usando p-boot :

Proporción de rechazos de H_0 a nivel $\alpha = 0.20$ en $Nrep = 1000$ replicaciones para $n = 100$, en cada caso usamos el p-boot.

	$N(0,1)$	$U(-1,1)$	t_5	$\varepsilon(1)$
χ^2	0.218	0.913	0.710	0.983
$K-S$	0.222	0.890	0.589	1
I_n	0.199	1	0.515	1
J_n	0.199	1	0.672	1

En todos los casos, el nivel empírico se condice con el nominal y ningún método resulta uniformemente mas potente. De todas formas, cabe destacarse el buen desempeño del test χ^2, entendiendo que este puede justificar su popularidad.

Bibliografía

[1] P.J.Bickel and Rosenblatt. *On some global measures of the deviations of density function estimates*. The Annals of statistics. 1973,Vol 1, 1071-1096.

[2] G.Boente, D.Rodriguez and W.Gonzalez Manteiga. *Goodness of fit Test for Directional Data*.Scandinavian Journal of Statistics.2013.

[3] H.Chernoff and E.L. Lehmann. *The use of maximum likelihood estimates in χ^2 tests for goodness of fit*. Annals of Mathematical Statistics,1954.

[4] Y.Fan.*Testing the googness of Fit of a Parametric Density Function by Kernel Method*. Cambridge University Press. 1994, Vol 10, 316-356.

[5] H.W.Lilliefors. *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown*. Journal of the American Statistical Association, 1967 Vol 62,399-402.

[6] H.W.Lilliefors. *On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown*. Journal of the American Statistical Association,1969 Vol 64,387-389.

[7] H.B.Mann and Wald. *On the choice of the number of class intervals in the application of the chi square test.*Annals of Mathematical Statistics,1942.

[8] F.J.Massey.Jr *The kolmogorov-Smirnov Test for Goodness of Fit*. Journal of the American Statistical Association, vol. 46 68-78.

[1] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Statist.,* 33, 1065-1076.

[2] Rosenblatt, M. (1955). Remarks on some nonparametric estimates of a density function. *Ann. Statist.,* 27, 832-837.

[9] O.Thas *Comparing Distributions*. Springer.2010.

[10] A.W.Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press,1998.