Facultad
de
Ciencias

TEST DE BONDAD DE AJUSTE MULTIVARIANTE
(Multivariate goodness of fit test)

Trabajo de Fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autora: Andrea Diego Gutiérrez

Directora: Alicia Nieto Reyes

Octubre - 2018

Resumen

En este trabajo abordamos el concepto de función de profundidad de Tukey, que permite establecer un orden de datos procedentes de una distribución multivariante mediante la asignación de un número real a cada dato. Se define posteriormente la profundidad de Tukey aleatoria, basada en un conjunto finito de vectores, con la que trabajamos computacionalmente. Sabemos que si P y Q son distribuciones discretas en $\mathbb{R}^p$, entonces sus funciones de profundidad de Tukey, o de Tukey aleatoria, coinciden si y solo si P = Q. Tratamos de comprobar empíricamente la primera implicación para distribuciones continuas, concretamente para normales multivariantes. Para ello, usaremos el contraste de hipótesis basados en los estadísticos de Kolmogorov-Smirnov y Chi-Cuadrado.

Palabras clave: Chi-Cuadrado, distribución multivariante, Kolmogorov-Smirnov, profundidad estadística de Tukey aleatoria, p-valor, test de hipótesis.

Abstract

In this report we study the Tukey depth, which enables us to establish an order of data drawn from a multivariate distribution, assigning a real number to each multivariate datum. We also study the random Tukey depth, based on a finite set of vectors, with which we work computationally. It is proved that if P and Q are discrete distributions in $\mathbb{R}^p$, then their Tukey depth, or their random Tukey depth, are equal if and only if P = Q. We want to empirically prove that the first implication is true for continue distributions, in particular for multivariate normal distributions. To do so, we use hypothesis tests based on the Kolmogorov-Smirnov and the Chi-Cuadrado statistics.

Key words: Chi-Squared, hypothesis testing, Kolmogorov-Smirnov, multivariate distribution, p-value, random Tukey depth.

Índice general

Capítulo 1

Introducción: Función de profundidad estadística de datos

Establecer un orden para un conjunto de datos puede resultar sencillo en el caso unidimensional, en el que está definido un orden de menor a mayor en la recta real. No obstante, si trabajamos en dimensiones mayores, esta tarea no es tan sencilla. Para ello se han definido diferentes conceptos que permiten ordenar datos multivariantes. Es el caso de la función de profundidad estadística, que establece un orden de un conjunto de datos con respecto a una distribución de probabilidad. Procedemos a enunciar formalmente dicho concepto, introducido en Zuo y Serfling [13]:

Definición 1.1 Sea $\mathcal{P}$ la clase de distribuciones sobre los conjuntos de Borel de $\mathbb{R}^p$ y P_X la distribución de probabilidad de un vector aleatorio X. Una función de profundidad estadística de datos es una función $D(\cdot,\cdot)$: $\mathbb{R}^p \times \mathcal{P} \longrightarrow \mathbb{R}$ acotada y no negativa satisfaciendo lo siguiente:

1. Invarianza afín: $D(Ax + b, P_{AX+b}) = D(x, P_X)$, para cualquier vector aleatorio X definido en $\mathbb{R}^p$, cualquier matriz A no singular $p \times p$ y cualquier $b \in \mathbb{R}^p$.

2. Maximalidad en el centro de simetría. Para cualquier $P \in \mathcal{P}$ que tenga centro de simetría θ, se verifica $D(\theta, P) = sup_{x\in\mathbb{R}^p} D(x, P)$, .

3. Monotonicidad con respecto al punto más profundo. Para cualquier $P \in \mathcal{P}$ con punto más profundo θ y cualquier $\alpha \in [0, 1]$, se tiene que $D(x, P) \leq D(\theta + \alpha(x - \theta), P)$.

4. Desvanecimiento en el infinito: cuando $\|x\| \longrightarrow \infty$, $D(x, P) \longrightarrow 0$ para cada $P \in \mathcal{P}$.

Analizando la definición, tenemos lo siguiente:

En dimensión 1, el punto más profundo de un conjunto de datos es la mediana y la profundidad disminuye cuanto más lejos se encuentre un punto de dicho valor.

En dimensiones mayores, imaginemos que tenemos una nube de puntos, los que se encuentran en el centro de la misma son los más profundos, y a medida que nos alejamos en

cualquier dirección la profundidad disminuye.

1.1. Funciones de profundidad de Tukey y de Tukey aleatoria

Dentro de las funciones de profundidad nos centraremos en la función de profundidad de Tukey y en la función de profundidad de Tukey aleatoria.
Con este propósito definimos en primer lugar la función de profundidad unidimensional para un punto $x \in \mathbb{R}$ con respecto a una probabilidad P con función de distribución F:

$$D_1(x, P) := min\{P(-\infty, x], P[x, \infty)\}$$

En el caso de que la función de distribución F sea continua se tiene que $P\{x\} = 0$, por lo que $P[x, \infty) = P\{x\} + P(x, \infty) = P(x, \infty)$. En consecuencia,

$$D_1(x, P) := min\{P(-\infty, x], P[x, \infty)\} = min\{F(x), 1 - F(x)\}$$

Abordamos ahora el concepto de función de profundidad de Tukey para un punto $x \in \mathbb{R}^p$ con respecto a una distribución de probabilidad P definida en $\mathbb{R}^p$. Según Tukey [12]:

Definición 1.2 Sea $v \in \mathbb{R}^p$ y Π_v la proyección de $\mathbb{R}^p$ sobre el subespacio unidimensional generado por v. Sea P_v la marginal de P relativa al subespacio anterior. Se define la profundidad de Tukey de un punto $x \in \mathbb{R}^p$ con respecto a una distribución P como

$$D_T(x, P) := inf\{D_1(\Pi_v(x), P_v) : v \in \mathbb{R}^p\}$$

Es decir, se calculan las profundidades unidimensionales de todas las proyecciones unidimensionales de x, con respecto a las correspondientes marginales de P, y se toma el ínfimo.

Para una distribución continua se puede simplificar la definición anterior si se trabaja con vectores definidos sobre la esfera unidad, $\mathbb{S}^{p-1}$. Al considerar todos los vectores de $\mathbb{S}^{p-1}$, estamos cogiendo un vector y su opuesto y se tiene que $F_{-v}(x) = 1 - F_v(x)$.

Definición 1.3 Sea $v \in \mathbb{S}^{p-1}$ y Π_v la proyección de $\mathbb{R}^p$ sobre el subespacio unidimensional generado por v. Sea P una distribución continua, P_v la marginal de P relativa al subespacio anterior y F_v la correspondiente función de distribución. Se define la profundidad de Tukey de un punto $x \in \mathbb{R}^p$ con respecto a P como

$$D_T(x, P) := inf\{D_1(\Pi_v(x), P_v) : v \in \mathbb{S}^{p-1}\} = inf\{F_v(\Pi_v(x)) : v \in \mathbb{S}^{p-1}\}$$

Las definiciones anteriores presentan un problema si queremos trabajar computacionalmente con ellas ya que requieren calcular todas las proyecciones unidimensionales, lo cual puede ser muy costoso. Por ello se ha definido el concepto de función de profundidad de Tukey aleatoria, en inglés "Random Tukey depth", en la que se considera el mínimo de un número finito de proyecciones unidimensionales elegidas aleatoriamente. Basándonos en [4], de Cuesta Albertos y Nieto Reyes:

Definición 1.4 Sean $P \in \mathcal{P}$, $\nu \in \mathcal{P}$ una distribución absolutamente continua y $v_1, ..., v_k$ vectores aleatorios, independientes e idénticamente distribuidos con distribución ν. Sea $x \in \mathbb{R}^p$ con $p > 1$, se define la profundidad de Tukey aleatoria de x con respecto a P basado en los k vectores anteriores como

$$D_{RT}(x, P) := min\{D_1(\Pi_{v_i}(x), P_{v_i}) : i = 1, ..., k\}$$

Cuesta Albertos y Nieto Reyes identifican en [5] el conjunto de vectores aleatorios, independientes e idénticamente distribuidos, al que nos referimos en la definición 1.4, con un conjunto de vectores aleatorios e idénticamente distribuidos definidos sobre la esfera unidad.
Así, si P es una distribución continua y tomamos un conjunto de vectores V sobre la esfera unidad y el conjunto de sus opuestos, $-V$, tenemos lo siguiente:

Definición 1.5 Sea $P \in \mathcal{P}$ una distribución continua. Sean $\nu \in \mathcal{P}$ una distribución absolutamente continua y $v_1, ..., v_k$ vectores aleatorios definidos sobre la esfera unidad, independientes e idénticamente distribuidos con distribución ν. Sean $v_{k+1}, ..., v_{2k}$ los vectores opuestos a los anteriores, es decir cumpliendo $v_{k+1} = -v_1$, $v_{2k} = -v_k$. Sea $x \in \mathbb{R}^p$ con $p > 1$, se define la profundidad de Tukey aleatoria de x con respecto a P basado en los $2k$ vectores anteriores como

$$D_{RT}(x, P) := min\{F_{v_i}(\Pi_{v_i}(x)) : i = 1, ..., 2k\}$$

1.2. Caracterización de distribuciones mediante las profundidades de Tukey y de Tukey aleatoria

Acercándonos al propósito de nuestro trabajo, presentaremos ahora varios resultados que muestran la caracterización de distribuciones por medio de la profundidad de Tukey y de Tukey aleatoria.
En primer lugar tratamos las distribuciones discretas, posteriormente las absolutamente continuas para llegar por último a los estudios existentes respecto a las distribuciones continuas en relación a los conceptos de profundidad de Tukey y de Tukey aleatoria.

Comencemos por aclarar el significado de distribución discreta, distribución continua y distribución absolutamente continua.

Se dice que una variable aleatoria sigue una distribución discreta si su soporte es finito o infinito numerable. Por otro lado, una variable aleatoria es continua si su función de distribución es continua. Un caso particular de variable aleatoria continua es la absolutamente continua, la cual se define como variable continua que presenta función de densidad.

En [5], de Cuesta Albertos y Nieto Reyes, se propone que si P y Q son distribuciones discretas en $\mathbb{R}^p$, entonces sus funciones de profundidad de Tukey aleatoria, basadas en un conjunto de vectores que basta que sea finito, coinciden si y solo si $P = Q$. Enunciemos detalladamente dicho resultado. Adoptaremos la notación $D_V(x, P)$ para la profundidad de Tukey aleatoria, donde V representa un conjunto finito de vectores aleatorios, independientes e idénticamente distribuidos.

Teorema 1.1 Sea $x \in \mathbb{R}^p$. Sean P y Q dos medidas de probabilidad y supongamos que P es discreta. Consideremos el espacio de probabilidad (Ω, κ) en el que definimos un conjunto V a lo sumo numerable de vectores aleatorios, independientes e idénticamente distribuidos.
Sea $\Omega_0 := \{\omega \in \Omega : D_{V(\omega)}(x, P) = D_{V(\omega)}(x, Q)$, para cada $x \in \mathbb{R}^p\}$.
Entonces $\kappa(\Omega_0) \in \{0, 1\}$ y $\kappa(\Omega_0) = 1$ si y solo si $P = Q$.

Respecto a la caracterización de distribuciones absolutamente continuas por medio de la función de profundidad de Tukey, destacamos en primer lugar el estudio de Koshevoy. Dicho autor establece, bajo ciertas consideraciones, la caracterización mediante la profundidad de Tukey de distribuciones absolutamente continuas con soporte compacto. Posteriormente, Hassari y Regaieg enuncian este resultado pero para distribuciones absolutamente continuas con soporte conexo.

Resulta lógico preguntarnos si ocurre lo mismo para las distribuciones continuas. En [7], de González Ruiz, se aborda este problema. Aunque no se llega a demostrar que si P y Q son distribuciones continuas en $\mathbb{R}^p$, entonces sus funciones de profundidad de Tukey aleatoria basadas en un conjunto finito de vectores coinciden si y solo si $P = Q$, sí que se prueban resultados relacionados asumiendo una serie de restricciones.

Nuestro propósito es comprobar empíricamente, mediante un test de hipótesis, que la primera implicación se cumple para distribuciones normales. Tras un análisis de las pruebas de bondad de ajuste existentes, escogeremos la prueba de Kolmogorov-Smirnov, con una serie de modificaciones que nos permitirán trabajar con la profundidad de Tukey aleatoria, y la prueba de Chi-Cuadrado. El interés radica no solo en dicha comprobación sino en la idea de que a partir de datos multidimensionales podemos, gracias al concepto de profundidad estadística, transformar dichos datos en datos unidimensionales.

Capítulo 2

Test de hipótesis

Un test de hipótesis tiene como objetivo rechazar una propiedad que se supone cierta para una población estadística basándose en observaciones sobre una muestra de la población. En él se contrasta una hipótesis nula, que denotaremos como H_0, frente a una hipótesis alternativa, H_1.
La hipótesis nula es la afirmación de que el valor de un parámetro poblacional θ es igual a un valor establecido (por ejemplo, $\theta = \theta_0$), mientras que la hipótesis alternativa es la afirmación de que el parámetro es $<$, $>$ o bien $\neq$.

Para realizar el contraste, a partir de los datos muestrales se calcula un estadístico, llamado estadístico de prueba. Por otro lado, el experimentador establece un nivel de significación, α, el cual se corresponde con la probabilidad de rechazar H_0 cuando esta es cierta.
Un estadístico T es una variable aleatoria que sigue cierta distribución. Con α se define una región crítica (o región de rechazo) de forma que si el estadístico calculado pertenece a dicha región decidiremos rechazar H_0. Denominamos valor crítico a cualquier valor que separa la región crítica de los valores del estadístico de prueba que no conducen al rechazo de la hipótesis nula.
Si la hipótesis alternativa es de la forma $\theta < \theta_0$, la región crítica se encuentra en el extremo izquierdo bajo la curva definida por la función de densidad del estadístico, y hablamos de una prueba de cola izquierda. Si H_1 es de la forma $\theta > \theta_0$, la región crítica está en el extremo derecho bajo la curva y se dice que es una prueba de cola derecha. Por último, si H_1 es de la forma $\theta \neq \theta_0$, la región crítica se localiza en las dos regiones extremas bajo la curva y hablamos de prueba de dos colas.

Además, a partir de un estadístico de prueba podemos calcular un p-valor. El p-valor (p) es una probabilidad, $0 \leq p \leq 1$, e indica lo verosímil que resulta obtener una muestra como la actual si es cierta H_0. Valores altos señalan que lo anterior es bastante probable, mientras que valores cercanos a 0 indican lo contrario.
Basándonos en dicho valor decidiremos rechazar H_0 (y aceptar H_1) o bien que no podemos rechazarla. Pero, ¿qué criterio seguiremos para ello? Una vez el experimentador establece

el nivel de significación se aplica lo siguiente:

- Si $p \leq \alpha$, rechazar H_0.
- Si $p > \alpha$, no tenemos evidencia suficiente para rechazar H_0.

El nivel de significación que suele emplearse es 0.05 o 0.01. Denominamos a α la probabilidad de cometer un error de tipo I (es decir, la probabilidad de rechazar H_0 cuando esta es cierta). Por otro lado, se dice que se comete un error de tipo II cuando no se rechaza H_0 siendo H_1 la hipótesis verdadera. Denotamos la probabilidad de cometer un error de tipo II como $1 - \beta$. β es la potencia del contraste, esto es, la probabilidad de aceptar H_1 cuando esta es cierta.

Disminuir α implica reducir el tamaño de la región crítica a la que nos hemos referido anteriormente y, en consecuencia, lleva a un aumento del valor $1 - \beta$. Es decir, si disminuimos la probabilidad de cometer un error de tipo I también disminuye la potencia del test.

Por otro lado, aumentar la potencia conlleva un aumento de α. Una solución a obtener la potencia deseada sin un aumento excesivo de la probabilidad de error de tipo I es aumentar el tamaño muestral.

Una propiedad importante de los p-valores, a la que se hace referencia en [10], es que, bajo la hipótesis nula, se distribuyen como una uniforme en el intervalo [0, 1]. Nos basaremos en esto para desarrollar nuestro trabajo.

Pero, antes, vamos a probar dicha propiedad. Concretamente lo probaremos para el caso de que tengamos una muestra con una distribución continua, basándonos en [8]. Sea T una variable aleatoria que representa todos los valores posibles del estadístico bajo la hipótesis nula. Sea P la variable aleatoria que representa los p-valores correspondientes a T. Denotemos por t el valor del estadístico observado actualmente. Se define el p-valor correspondiente a t, p, como la probabilidad de obtener un valor del estadístico T que sea al menos tan extremo como el que representa a los datos muestrales (esto es, t), asumiendo que H_0 es cierta.

Basándonos en [1], para una prueba de cola izquierda el p-valor es $\mathbb{P}(T \leq t|H_0)$, para una prueba de cola derecha, $\mathbb{P}(T \geq t|H_0)$, y para una prueba de dos colas se define como $2min\{\mathbb{P}(T \leq t|H_0), \mathbb{P}(T \geq t|H_0)\}$.

Supongamos por ejemplo que p es de la forma, $\mathbb{P}(T \geq t|H_0)$ y que T sigue una distribución continua que denotaremos por F_T. Entonces, $p = 1 - \mathbb{P}(T < t|H_0) = 1 - F_T(t)$.

De esta forma, podemos escribir la variable aleatoria P como:

$$P = 1 - F_T(T) \quad \textbf{(1)}$$

Definamos $F_T(T)$ como la variable aleatoria Y, $Y = F_T(T)$. Sea F_Y la función de distribución de Y, entonces

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F_T(T) \leq y) \quad \textbf{(2)}$$

Una función de distribución es monótona creciente y continua por la derecha. En el caso de ser continua y estrictamente monótona, es biyectiva y en consecuencia está definida su función inversa, a la que denominamos función cuantil. Así, si F es una función de distribución cumpliendo lo anterior, su inversa $F^{-1}(z)$ para una probabilidad $z\epsilon[0,1]$, es el único número real x tal que $F(x) = z$.

En el caso de que la función de distribución sea discreta o monótona no estricta, se define la función cuantil como $F^{-1}(z) = inf\{x\epsilon\mathbb{R} : z \leq F(x)\}$, con $z\epsilon[0,1]$.

Haciendo uso de la función cuantil en el paso (2), tenemos que

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F_T(T) \leq y) = \mathbb{P}(T \leq F_T^{-1}(y)) = F_T(F_T^{-1}(y)) = y \quad \textbf{(3)}$$

Ahora sea X una variable aleatoria que sigue una distribución uniforme continua con soporte [a, b], a, b $\epsilon\mathbb{R}$. Entonces, su función de distribución es de la siguiente forma:

$$F_X(x) = \begin{cases} 0 \text{ si } x < a \\ \dfrac{x-a}{b-a} \text{ si } a \leq x < b \\ 1 \text{ si } x \geq b \end{cases}$$

Basándonos en esto, la función de distribución de una variable aleatoria con distribución uniforme continua y soporte [0, 1] está definida por $F_X(x) = x \ \forall x \in [0, 1]$.

En el paso (3) vimos que $F_Y(y) = y$ si $y \in [0, 1]$, por lo que podemos decir que Y se trata de una variable aleatoria continua con distribución uniforme y soporte [0, 1]. Por (1), el p-valor es $P = 1 - F_T(T) = 1 - Y$, de donde concluimos que P también es una variable aleatoria con distribución uniforme en [0, 1].

2.1. Interpretación de un histograma de p-valores

Cuando se tienen miles de p-valores de un test estadístico, representarlos en un histograma nos da mucha información. En primer lugar, nos permite saber si el test que estamos usando es adecuado al experimento o si, por las características de la distribución de la que tomamos la muestra, deberíamos cambiar de test. En el caso de que sea adecuado, nos da una idea del peso que tiene la hipótesis alternativa frente a la nula.

Como sabemos que bajo la hipótesis nula los p-valores se distribuyen como una U[0, 1], el histograma de un conjunto de p-valores podría asemejarse a alguno de la figura 2.1.

Los p-valores que no apoyan la hipótesis alternativa se distribuyen uniformemente en el intervalo [0, 1], mientras que los que la apoyan se pueden condensar en el intervalo [0, 0.05]. Además, hay que tener en cuenta que a la hora de realizar un test se puede obtener algún valor que sea un falso positivo (esto es, obtener un p-valor menor que 0.05 cuando en realidad la hipótesis alternativa no es cierta) así como algún falso negativo (es decir, un p-valor mayor que 0.05 cuando es cierta la hipótesis alternativa).

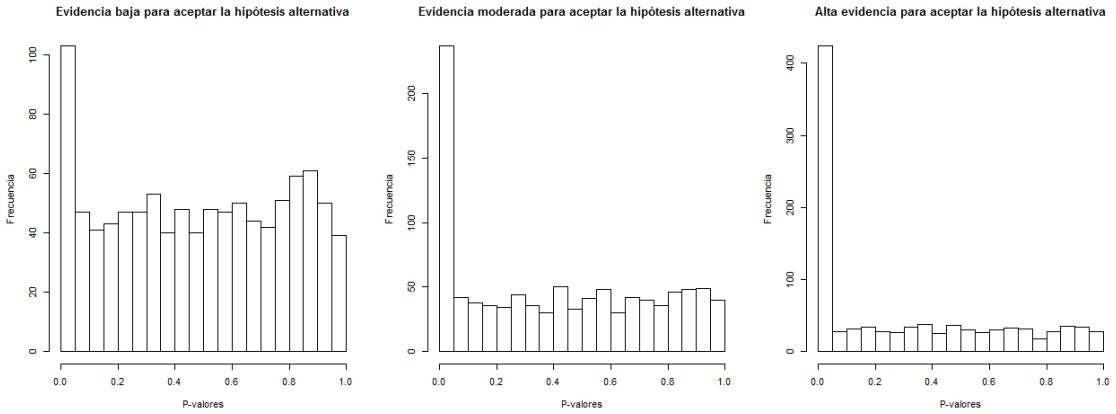


Figura 2.1: De izquierda a derecha, en el primer histograma se representan 950 valores procedentes de una U[0,1] y 50 valores pertenecientes al intervalo [0, 0.05]. En el segundo, 800 valores procedentes de una U[0,1] y 200 en el intervalo [0, 0.05]. En el tercero, 600 valores procedentes de una U[0,1] y 400 valores en [0, 0.05].

2.2. Pruebas de bondad de ajuste

2.2.1. Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov o prueba K-S permite comparar dos distribuciones de probabilidad entre sí. Existen dos versiones de esta prueba según se comparen dos distribuciones empíricas o una distribución empírica con una teórica. A la primera la denominamos prueba de Kolmogorov-Smirnov para dos muestras y a la segunda prueba de Kolmogorov-Smirnov para una muestra.

En ambos casos hablaremos de muestra aleatoria simple, es decir un conjunto de variables aleatorias independientes e idénticamente distribuidas.

Kolmogorov-Smirnov para dos muestras

Sean $X_1, ..., X_m$ e $Y_1, ..., Y_n$ dos muestras aleatorias simples procedentes de poblaciones continuas con funciones de distribución F y G respectivamente.

Asumimos que las $X's$ son independientes entre sí y que las $Y's$ también son independientes entre sí. Además de esto, supondremos independencia entre las dos muestras, es decir que las $X's$ son independientes de las $Y's$.

El propósito es determinar si existen diferencias entre las funciones de distribución de X e Y.

$$H_0 : F(t) = G(t) \ \forall \text{ t.}$$

$$H_1 : F(t) \neq G(t) \text{ para algún t.}$$

Para calcular el estadístico de prueba de Kolmogorov-Smirnov se usa la información disponible, es decir las funciones de distribución empíricas. Denotemos por F_m y G_n las funciones de distribución empíricas de la muestra X y la muestra Y respectivamente. Dado $t\epsilon\mathbb{R}$

$$F_m(t) = \frac{\sharp X's \leq t}{m}$$
$$G_n(t) = \frac{\sharp Y's \leq t}{n} \ .$$

Escrito de otra forma:

$$F_m(t) = \begin{cases} 0 \text{ si } t < X_{(1)} \\ \dfrac{j}{m} \text{ si } X_{(j)} \leq t < X_{(j+1)} \\ 1 \text{ si } t \geq X_{(m)} \end{cases} \quad G_n(t) = \begin{cases} 0 \text{ si } t < Y_{(1)} \\ \dfrac{k}{n} \text{ si } Y_{(k)} \leq t < Y_{(k+1)} \\ 1 \text{ si } t \geq Y_{(n)} \end{cases}$$

con $j\epsilon\{1, ..., m-1\}$ y $k\epsilon\{1, ..., n-1\}$.

Observar que $F_m(t)$ y $G_n(t)$ toman un número finito de valores diferentes y estos valores diferentes se dan en el primer caso para t igual a $X_{(j)}$ con $j\epsilon\{1, ..., m\}$ y en el segundo caso para t igual a $Y_{(k)}$ con $k\epsilon\{1, ..., n\}$. Por ello en el estadístico de prueba únicamente consideraremos $t = U_{(i)}$ con $i\epsilon\{1, ..., N\}$, representando así los $N = m + n$ valores ordenados de menor a mayor de la muestra combinada $X_{(1)}, ..., X_{(m)}, Y_{(1)}, ..., Y_{(n)}$.

Sea $d = mcd(m, n)$, definimos el estadístico de prueba como

$$S_2 = \frac{mn}{d} max_{i=1,...,N}\{|F_m(U_{(i)}) - G_n(U_{(i)})|\}.$$

El criterio que se adopta para rechazar H_0 es que S_2 sea mayor o igual que s_α, valor al que denominamos valor crítico. Este se toma de forma que la probabilidad de cometer un error de tipo I sea igual a α, es decir $P_0(S_2 \geq s_\alpha) = \alpha$.

Cuando tratamos con tamaños muestrales grandes ($m, n \geq 100$) definimos el estadístico de prueba de la siguiente forma.

$$S_2^* = \sqrt{\frac{mn}{N}} max_{i=1,...,N}\{|F_m(U_{(i)}) - G_n(U_{(i)})|\} = \frac{d}{\sqrt{mnN}} S_2.$$

¿Qué criterio usaremos para rechazar la hipótesis nula? Analicemos en primer lugar lo que ocurre para el estadístico que hemos llamado anteriormente S_2^*.
Basándonos en [3], de Chicken y Hollander, sea $a > 0$, cuando $min(m, n) \to \infty$ entonces

$$P_0(S_2^* < a) \to \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2a^2}.$$

Definamos la función $Q(a)$ como

$$Q(a) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2a^2}.$$

Sea q_α el valor que tomamos de forma que $Q(q_\alpha) = \alpha$, donde α representa el nivel de significación. El criterio para rechazar la hipótesis nula es:

Rechazar H_0 si $S_2^* \geq q_\alpha$.
En otro caso no tenemos evidencia para rechazar H_0.

En nuestro caso, decidimos trabajar con un nivel de significación $\alpha = 0.05$, para el cual $q_\alpha = 1.358$.

Kolmogorov-Smirnov para una muestra

Supongamos ahora que disponemos únicamente de una muestra aleatoria simple, $X_1, ..., X_m$, con función de distribución continua F. Deseamos conocer si dicha distribución coincide con la de una distribución concreta que denotaremos como F_0. Por ejemplo, F_0 puede tratarse de una distribución normal de media 0 y desviación estándar 1.

Las hipótesis a contrastar son las siguientes:

$$H_0 : F(t) = F_0(t) \ \forall \text{ t.}$$
$$H_1 : F(t) \neq F_0(t) \text{ para algún t.}$$

El estadístico de prueba en este caso es $S = sup_{-\infty<t<\infty}\{|F_m(t) - F_0(t)|\}$, donde F_m representa la función de distribución empírica de la muestra X.
Si pensamos en la representación gráfica de las funciones $F_m(t)$ y $F_0(t)$, con t en el eje de abscisas, el estadístico calcula la mayor distancia vertical entre estas. El supremo se encuentra o bien para un valor de t igual a X_i, con $i\epsilon\{1, ..., m\}$, o bien justo a la izquierda de un X_i. Esto se puede observar en la figura 2.2. Como F_0 es una función continua creciente y F_m es una función creciente a saltos, cuando $F_0(t) < F_m(t)$, el supremo se

encuentra para t igual a algún X_i, que es donde se produce el salto. En cambio, cuando $F_0(t) > F_m(t)$, el supremo se localiza justo a la izquierda de un X_i.

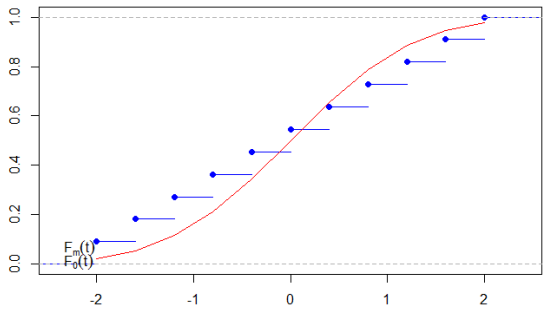


Figura 2.2: En azul, la función de distribución empírica de la muetra X = {-2, -1.6, -1.2, -0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6, 2}, frente a la función de distribución teórica de una N(0,1) en rojo.

Computacionalmente, S se calcula como

$$S = max_{i=1,...,m}\ R_i,$$

$$\text{donde } R_i = max\{|F_m(X_{(i)}) - F_0(X_{(i)})|, |F_m(X_{(i-1)}) - F_0(X_{(i)})|\} = max\{|\frac{i}{m} - F_0(X_{(i)})|, |\frac{i-1}{m} - F_0(X_{(i)})|\}$$

En el caso de que existan valores repetidos, se ordenan los valores diferentes, $Y_{(1)} < ... < Y_{(k)}$, y se calcula S como

$$S = max_{i=1,...,k}\ \{|F_m(Y_{(i)}) - F_0(Y_{(i)})|, |F_m(Y_{(i-1)}) - F_0(Y_{(i)})|\}.$$

El criterio para rechazar la hipótesis nula es que S sea mayor o igual que s_α, valor que se toma de forma que $P_{F_0}(S \geq s_\alpha) = \alpha$.

2.2.2. Prueba Chi-Cuadrado

La prueba χ^2 de Pearson tiene varias utilidades, por un lado permite comparar una distribución observada con una teórica y, por otro lado, se emplea para averiguar si dos variables son independientes a partir de datos en tablas de contingencia. Nos centraremos en lo primero ya que es lo realmente interesante en el curso de este trabajo.

Disponemos de una muestra aleatoria simple de tamaño m y deseamos conocer si los datos se corresponden con cierta distribución teórica. Tomamos el recorrido de la distribución teórica y lo descomponemos en un número finito k de clases, $C_1, ..., C_k$. A continuación clasificamos las observaciones muestrales según dichas clases y comparamos

las frecuencias observadas de cada C_i con las probabilidades que les corresponderían con la distribución teórica.

Sea O_i, con $i = 1, ..., k$, la frecuencia observada correspondiente a la clase C_i y E_i la frecuencia esperada. E_i se calcula como mp_i, donde m es el tamaño muestral y p_i es la probabilidad de concurrencia asociada a la clase C_i. Para conocer si las frecuencias observadas concuerdan con las esperadas se realiza un contraste de hipótesis con el siguiente estadístico:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Las hipótesis a contrastar son:

$$H_0 : E_i = O_i \ \forall \text{ i, con } i = 1, ..., k.$$

$$H_1 : \exists \text{ i tal que } E_i \neq O_i.$$

Pearson probó que, cuando H_0 es cierta y $m \longrightarrow \infty$, la distribución del estadístico χ^2 es la de una Chi-cuadrado con k-1 grados de libertad. A la hora de realizar el test con nivel de significación α, rechazaremos la hipótesis nula si $\chi^2 \geq \chi^2_{\alpha,k-1}$. El valor $\chi^2_{\alpha,k-1}$ se encuentra tabulado y corresponde al valor de una Chi-cuadrado de k-1 grados de libertad cumpliendo que la probabilidad de encontrar un valor mayor o igual que este sea α.

Por último, notar que para realizar el test requeriremos que todas las frecuencias esperadas sean mayores o iguales a 5, como aparece en [3], de forma que la aproximación de χ^2 sea buena. Puede ocurrir que debamos agrupar varias clases a fin de que se cumpla este requisito.

Capítulo 3

Trabajando con R

En este capítulo presentamos cómo se realiza el cálculo computacional de la función de profundidad de Tukey aleatoria para una distribución continua, la cual usaremos posteriormente en los test de hipótesis.

El cálculo de la función de profundidad de Tukey aleatoria empírica de cada uno de los puntos de la muestra $datau = \{U_i, \text{ con } i = 1, ..., N\}$, respecto a la probabilidad empírica de la muestra $datax = \{X_i, \text{ con } i = 1, ..., m\}$, lo realizo de la siguiente forma:

```
depthu.RT=function(datax,datau,maproj){
  m<-nrow(datax)
  p<-ncol(datax)
  N<-nrow(datau)
  nproj<-ncol(maproj)
  Prodx=datax%*%maproj
  Produ=datau%*%maproj
  d=numeric(m+1)
  du=numeric(N)
  for(i in 1:N){
    Prod<-matrix(rbind(Prodx,Produ[i,]),m+1,nproj)
  # apply(Prod,2,rank,ties="max") cuenta, para cada elemento
  # de cada columna de Prod, el número de valores de dicha
  # columna menores o iguales que dicho elemento.
  # apply(-Prod,2,rank,ties="min")-1 cuenta, para cada elemento
  # de cada columna de Prod, el número de valores de dicha
  # columna mayores que dicho elemento.
    d1<-apply(Prod,2,rank,ties="max")
    d2<-c(d1[1:m,],d1[m+1,]-1)
    d<-apply(cbind(d2,(apply(-Prod,2,rank,ties="min")-1)),1,min)
    if(d[m+1]==0){
```

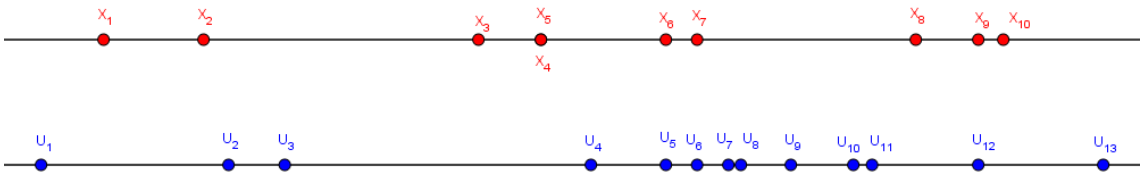
```
      du[i]=0
    } else {
      du[i]<-(d[m+1])/m
    }
  }
  return(du)
}
```

Observar que *datax* es una matriz $m \times p$, donde m indica el tamaño muestral y p la dimensión, y *datau* es una matriz $N \times p$. Por otro lado, *maproj* hace referencia a la matriz de proyección, que será de dimensión $p \times nproj$, donde *nproj* indica el número de proyecciones que deseamos realizar. En dimensión 1 no hay que proyectar, luego damos a maproj el valor 1. En dimensión $p > 1$ construiremos *maproj* de forma que cada columna sea un vector aleatorio unitario con distribución $N(0, 1)$. Así conseguimos un conjunto de vectores aleatorios e idénticamente distribuidos definidos sobre la esfera unidad $\mathbb{S}^{p-1}$.

A continuación se presenta un ejemplo con datos unidimensionales para mostrar cómo funciona *depthu.RT*. La profundidad de Tukey aleatoria de un punto U_i respecto a un conjunto *datax* procedente de una distribución continua viene dada por la siguiente definición:

$$\frac{min\{\sharp(datax \cap (-\infty, U_i]), \sharp(datax \cap (U_i, \infty))\}}{\sharp datax}$$

Así, por ejemplo, sea *datax* = { -9, -7.4, -3, -2, -2, 0, 0.5, 4, 5, 5.4 } y sea *datau* = { -10, -7, -6.1, -1.2, 0, 0.5, 1, 1.2, 2, 3, 3.3, 5, 7 }.



El resultado devuelto por la función depthu.RT es el siguiente:

0 0.2 0.2 0.5 0.4 0.3 0.3 0.3 0.3 0.3 0.3 0.1 0

Donde cada elemento es la profundidad de Tukey aleatoria de cada uno de los valores de *datau* respecto al conjunto *datax*.

¿Cómo funciona *depthu.RT* para dimensión mayor que 1? Se comienza calculando el producto de matrices $Prodx = datax\,\% * \%maproj$, que da lugar a una matriz $m \times nproj$. Cada elemento $Prodx[i, j]$ es la proyección del dato $datax[i,]$ sobre el subespacio unidimensional generado por el vector $maproj[, j]$.

$$Prodx = \begin{pmatrix} X_{11} & X_{12} & ... & X_{1p} \\ X_{21} & X_{22} & ... & X_{2p} \\ ... & ... & ... & ... \\ X_{m1} & X_{m2} & ... & X_{mp} \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & ... & v_{1nproj} \\ v_{21} & v_{22} & ... & v_{2nproj} \\ ... & ... & ... & ... \\ v_{p1} & v_{p2} & ... & v_{pnproj} \end{pmatrix} =$$

$$= \begin{pmatrix} X_{11}v_{11} + X_{12}v_{21} + ... + X_{1p}v_{p1} & ... & X_{11}v_{1nproj} + X_{12}v_{2nproj} + ... + X_{1p}v_{pnproj} \\ X_{21}v_{11} + X_{22}v_{21} + ... + X_{2p}v_{p1} & ... & X_{21}v_{1nproj} + X_{22}v_{2nproj} + ... + X_{2p}v_{pnproj} \\ ... & ... & ... \\ X_{m1}v_{11} + X_{m2}v_{21} + ... + X_{mp}v_{p1} & ... & X_{m1}v_{1nproj} + X_{m2}v_{2nproj} + ... + X_{mp}v_{pnproj} \end{pmatrix}$$

Así, $Prodx[i,j] = \sum_{r=1}^{p} X_{ir}v_{rj}$

De forma similar se construye la proyección $Produ = datau\,\%*\%maproj$, de dimensión $N \times nproj$. A continuación se define para cada $i = 1, ..., N$ la matriz $Prod$ como

$$Prod = \begin{bmatrix} Prodx \\ Produ[i,] \end{bmatrix}$$

de dimensión $(m+1)$ x $nproj$.
Para cada $Prod$ se hace lo siguiente:

Para cada columna $Prod[,j]$ (para cada proyección) se cuenta el número de elementos de $Prodx[,j]$ que son menores o iguales que $Produ[i,j]$ y, por otro lado, se cuenta cuántos son mayores. De estos dos números se toma el menor.

Una vez se realiza esto para todos los j, $j = 1, ..., nproj$, se obtienen $nproj$ números de los cuales se escoge el mínimo y se divide entre el tamaño muestral de $datax$. Dicho valor corresponde a la profundidad de Tukey aleatoria del dato $datau[i,]$ respecto al conjunto $datax$.

A la hora de pasar el test de Kolmogorov-Smirnov usaremos la definición 1.5 de profundidad de Tukey aleatoria. La razón es que el cálculo del estadístico de prueba se basa en la diferencia de funciones de distribución y queremos mantener esto. Computacionalmente la denominaremos $depthu2.RT$ y es como $depthu.RT$ cambiando el bucle "for"por

```
for(i in 1:N){
  Prod<-matrix(rbind(Prodx,Produ[i,]),m+1,nproj)
  d<-apply(apply(Prod,2,rank,ties="max"),1,min)
  du[i]<-(d[m+1]-1)/m}
```

Además, en este caso consideraremos maproj como una matriz formada por un conjunto de vectores V y sus opuestos, el conjunto $-V$. Por ello, en dimensión 1 tomaremos maproj

como una matriz de dos columnas constituidas por 1 y -1. En dimensión $p > 1$ construiremos *maproj* de forma que cada columna de 1 a nproj/2 sea un vector aleatorio unitario con distribución $N(0, 1)$. Las columnas de nproj/2 + 1 a nproj serán los vectores opuestos a los anteriores.

Observar que cuando $datau = datax$ las funciones anteriores se pueden simplificar. En esta situación, emplearemos $depth.RT$ en lugar de $depthu.RT$ y $depth2.RT$ en lugar de $depthu2.RT$.

```
depth.RT=function(data,maproj){
  m<-nrow(data)
  Prod=data%*%maproj
  apply(cbind(apply(Prod,2,rank,ties="max"),
              (apply(-Prod,2,rank,ties="min")-1)),1,min)/m}
```

De manera similar, definimos $depth2.RT$ cambiando la última línea de $depth.RT$ por

```
apply(apply(Prod,2,rank,ties="max"),1,min)/m
```

Por último también nos interesa el cálculo de la profundidad de Tukey aleatoria teórica, concretamente para un distribución normal. Presentamos en primer lugar la función para una muestra con distribución normal de media "med" y desviación estándar "desv" y, en segundo lugar, para una distribución normal multivariante de vector de media "med" y matriz de covarianza "mcov".

```
depth.real.RT=function(datau,med,desv){
  N<-nrow(datau)
  dur=numeric(N)
  for(i in 1:N){
    dur[i]=min(pnorm(datau[i],med,desv),
               pnorm(datau[i],med,desv,lower.tail=FALSE))}
  return(dur)}
```

Ahora para una distribución normal multivariante. Calculamos las profundidades unidimensionales de las proyecciones del punto estudiado y nos quedamos con la mínima. La media y la desviación estándar de la marginal relativa a un vector de proyección las obtenemos de la siguiente manera.

Denotemos la proyección del vector aleatorio $X_i = (X_{i1}, X_{i2}, ..., X_{ip})$ sobre el vector $v_j = (v_{1j}, v_{2j}, ..., v_{pj})^T$ como $X_i v_j$.

$E[X_i v_j] = E[\sum_{r=1}^{p} X_{ir} v_{rj}] = \sum_{r=1}^{p} E[X_{ir} v_{rj}] = \sum_{r=1}^{p} v_{rj} E[X_{ir}] = \mu\ v_j.$

Por otro lado, sea la matriz de covarianza $Cov(X_i)$, a la que también denotaremos como $Var(X_i)$. Como $Var(X_i) = E[(X_i - \mu)(X_i - \mu)^T] \in \mathbb{R}^{p\times p}$, tenemos que

$$Var(v_j^T X_i^T) = E[(v_j^T(X_i^T - \mu))(v_j^T(X_i^T - \mu))^T] = E[v_j^T(X_i^T - \mu)(X_i^T - \mu)^T v_j]$$
$$= v_j^T E[(X_i^T - \mu)(X_i^T - \mu)^T]v_j = v_j^T Var(X_i^T)v_j$$

Para el cálculo computacional de la profundidad de Tukey aleatoria teórica de una distribución normal multivariante usaremos la siguiente función.

```
depth.real.Rp.RT=function(data,maproj,med,mcov){
  N<-nrow(data)
  p<-ncol(data)
  nproj<-ncol(maproj)
  Prod=data%*%maproj
  dur1<-matrix(0,N,nproj)
  dur=numeric(N)
  med2=med%*%maproj
  for(i in 1:N){
    for(j in 1:nproj){
      medj=med2[,j]
      varj=t(maproj[,j])%*%mcov%*%maproj[,j]
      desvj=sqrt(varj)
      dur1[i,j]=min(pnorm(Prod[i,j],medj,desvj),
                    pnorm(Prod[i,j],medj,desvj,lower.tail=FALSE))}
    dur[i]=min(dur1[i,])}
  return(dur)}
```

Los cálculos de acuerdo a la definición 1.5 los llamaremos $depth2.real.RT$ y $depth2.real.Rp.RT$. Para el primer caso cambiaríamos

```
for(i in 1:N){
  dur[i]=min(pnorm(datau[i],med,desv),
             pnorm(datau[i],med,desv,lower.tail=FALSE))}
```

por

```
for(i in 1:N){
  dur[i]=min(pnorm(datau[i],med,desv),pnorm(-datau[i],med,desv))}
```

Para el caso de la función $depth2.real.Rp.RT$ lo que haremos es eliminar la parte de $pnorm$ con $lower.tail = FALSE$ y tomar maproj de forma que contenga un conjunto de vectores cumpliendo las condiciones anteriormente mencionadas así como sus opuestos.

3.1. Test computacional de Kolmogorov-Smirnov para dos muestras

Retomando lo visto en el capítulo 2, queremos comprobar computacionalmente la caracterización de distribuciones normales y normales multivariantes por medio de la profundidad de Tukey aleatoria.

Sean $X_1, ..., X_m$ e $Y_1, ..., Y_n$ dos muestras aleatorias simples procedentes de poblaciones con funciones de distribución F y G respectivamente. Denotemos por F_m y G_n las funciones de distribución empíricas. Sea $N = m+n$ y $U_{(1)}, ..., U_{(N)}$ los N valores ordenados de menor a mayor de las muestras combinadas $X_1, ..., X_m$ e $Y_1, ..., Y_n$.

Las hipótesis que deseamos contrastar son:

$$H_0 : D_{RT}(U_i, F_m) = D_{RT}(U_i, G_n) \text{ para todo } i \text{ con } i = 1, ..., N.$$
$$H_1 : D_{RT}(U_i, F_m) \neq D_{RT}(U_i, G_n) \text{ para algún } i \text{ con } i = 1, ..., N.$$

Donde $D_{RT}(U_i, F_m)$ y $D_{RT}(U_i, G_n)$ representan las funciones de profundidad de Tukey aleatoria empíricas.

Vamos a hacerlo basándonos en el estadístico de Kolmogorov-Smirnov. Para obtener mejores resultados trabajaremos con tamaños muestrales grandes (m, $n \geq 100$). Modificaremos la expresión de $S_2^* = \sqrt{\frac{mn}{N}} max_{i=1,...,N}\{|F_m(U_{(i)}) - G_n(U_{(i)})|\}$ por

$$J_2 = \sqrt{\frac{mn}{N}} max_{i=1,...,N}\{|D_{RT}(U_i, F_m) - D_{RT}(U_i, G_n)|\}.$$

Ahora implementamos la función para calcular el estadístico J_2 y el p-valor. Para el cálculo de este último nos hemos basado en el código implementado en Matlab. Como datos de entrada hay que introducir *depthux*, que se trata de un vector con las profundidades de Tukey aleatoria del conjunto *datau* respecto al conjunto *datax*, y *depthuy*, vector con las profundidades de *datau* respecto a *datay*.

```
kstest2pval=function(m,n,depthux,depthuy){
  N=m+n
  dif=numeric(N)
  for(i in 1:N){
    dif[i]=abs(depthux[i]-depthuy[i])}
  difmax=max(dif)
  k=m*n/N
  lambda=max((sqrt(k)+0.12+0.11/sqrt(k))*difmax,0)
  j=c(1:101)
  pvalor=2*sum((-1)^(j-1)*exp(-2*lambda*lambda*j^2))
  pvalor=min(max(pvalor,0),1)
  return(pvalor)
}
```

3.1.1. Resultados

Vamos a analizar conjuntos de datos provenientes de una N(0,1) o de una normal multivariante con media dada por el vector nulo y matriz de covarianza igual a la matriz identidad.

Para dimensión 1 hacemos lo siguiente:

```
p=1; m=100
r=1000; g=numeric(r)
for(i in 1:r){
  datax<-matrix(rnorm(m),m,1)
  datay<-matrix(rnorm(m),m,1)
  datau<-rbind(datax,datay)
  maproj<-matrix(c(1,-1),1,2)
  dux=depthu2.RT(datax,datau,maproj)
  duy=depthu2.RT(datay,datau,maproj)
  g[i]=kstest2pval(m,m,dux,duy)
}
k1=sort(g)
sum(k1<0.05)
hist(g,breaks=20)
```

Otro ejemplo, para dimensión 2:

```
p=2; m=200
nproj=20; nproj1=10
r=1000; g=numeric(r)
med=c(0,0); mcov<-diag(2)
for(i in 1:r){
  datax<-mvrnorm(m,med,mcov)
  datay<-mvrnorm(m,med,mcov)
  datau<-rbind(datax,datay)
  maproj1<-matrix(rnorm(p*nproj1),p,nproj1)
  for(j in 1:nproj1){
    maproj1[,j]<-maproj1[,j]/sqrt(sum(maproj1[,j]^2))}
  maproj<-cbind(maproj1,-maproj1)
  dux=depthu2.RT(datax,datau,maproj)
  duy=depthu2.RT(datay,datau,maproj)
  g[i]=kstest2pval(m,m,dux,duy)}
k1=sort(g)
sum(k1<0.05)
hist(g,breaks=20)
```

Para dimensión 1 se toma maproj como una matriz formada por dos columnas con un 1 y un -1. Para dimensiones mayores, ¿cuántas proyecciones tomar? Basándonos en [4], de Cuesta Albertos y Nieto Reyes, el número de proyecciones empleadas para que la profundidad de Tukey aleatoria sea una buena aproximación de la profundidad de Tukey dependerá de varios factores como la dimensión y el tamaño muestral. Se han llevado a cabo simulaciones que muestran que, en la mayoría de los casos, 250 proyecciones aleatorias son suficientes.

En dicho artículo se analizan casos concretos de distribuciones normales, llegando a la conclusión de que para dimensión 2, 10 proyecciones permiten obtener una buena aproximación de la profundidad de Tukey, para dimensión 8, bastarían 60 y para dimensión 50, 250.
En nuestro caso vamos a analizar muestras bidimensionales con 20 proyecciones: 10 vectores y sus opuestos, y con 100 proyecciones: 50 vectores y sus opuestos.

En cada bucle for se calculan 1000 p-valores y se analiza si se distribuyen como una uniforme en [0,1]. Para ello se representan dichos p-valores mediante un histogramas. Además se mira si el 5 por ciento de estos, en este caso 50, es menor que 0.05. Los resultados obtenidos son los siguientes.

Dimensión 1

En primer lugar trabajamos con muestras procedentes de N(0,1) de tamaño 100 y, en segundo, analizamos qué ocurre si aumentamos el tamaño a 200. Cada bucle for calcula 1000 p-valores y realizo este proceso tres veces (tres veces para m=100 y otras tres para m=200). En la siguiente tabla aparece la media de rechazos de H_0, es decir de p-valores menores que 0.05, la desviación y la media del tiempo computacional.

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
100	42	1.15	1.6
200	49	4.16	7

Tabla 3.1: Resultados del test K-S para las profundidades de Tukey aleatoria de dos muestras N(0,1).

En la figura 3.1 se presenta uno de los tres histogramas obtenidos para m=200:

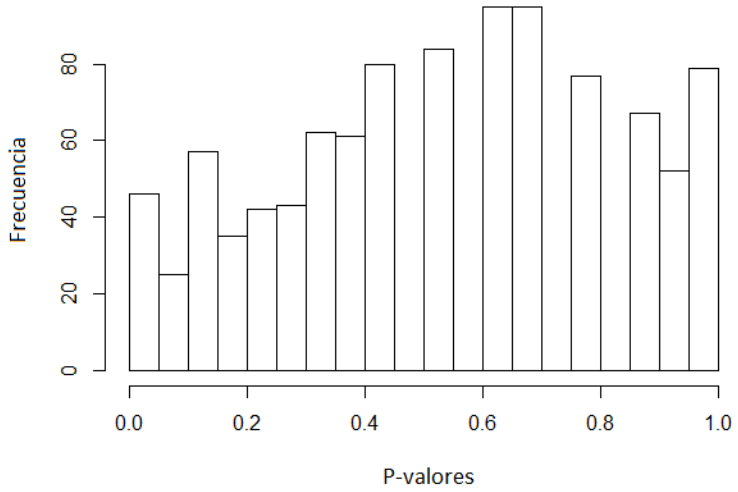


Figura 3.1: Histograma de los p-valores del test K-S para dos muestras N(0,1) de tamaño 200.

Dimensión 2

En este caso tomamos muestras de tamaño 200 procedentes de normales multivariantes de media dada por el vector nulo y matriz de covarianza la identidad y variamos el número de proyecciones: 20 y 100. En cada caso, realizamos tres veces el proceso de calcular 1000 p-valores. La media de rechazos, desviación y tiempo computacional son los siguientes:

Nº de proyecciones	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
20	44	3.05	22
100	38	1.52	100

Tabla 3.2: Resultados del test K-S para las profundidades de Tukey aleatoria de dos muestras normales bivariantes de media (0,0) y matriz de covarianza Id_2.

Algunos de los histogramas obtenidos:

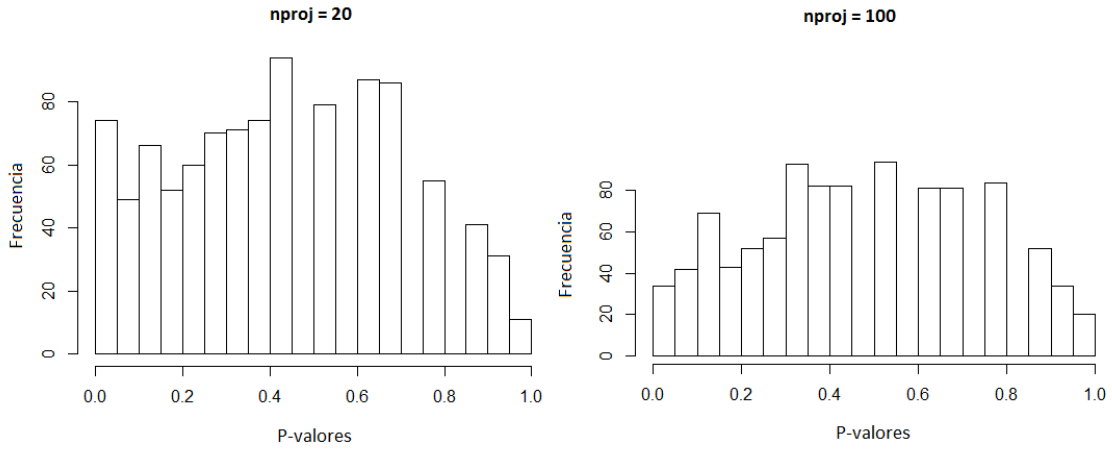


Figura 3.2: Histograma de los p-valores del test K-S para dos muestras normales bivariantes. A la izquierda, un caso en el que se han calculado las profundidades de Tukey aleatoria con 20 proyecciones, y a la derecha con 100 proyecciones.

Observamos que para dimensión 1 el número de rechazos de H_0 se acerca a 50, esto es el 5 por cierto de los p-valores se encuentra en el intervalo [0, 0.05]. No obstante, al analizar el histograma (figura 3.1) vemos que no es como el de una uniforme en [0,1], ya que por ejemplo hay unos 80 valores en el intervalo [0.4, 0.45], pero no hay valores en [0.45, 0.5]. Para dimensión 2 el número de rechazos de H_0 es menor de 50 y al representar los histogramas también se alejan del de una U[0,1]. Concluimos que este test no funciona demasiado bien para los conceptos que estamos tratando. Esto puede deberse a varias razones; en primer lugar, al proyectar estamos perdiendo la independencia de los datos; en segundo lugar, pueden existir valores repetidos. Basándonos en [3], el test está bien definido en caso de valores repetidos, pero cambia la distribución de los p-valores. Por último, se requiere un tiempo computacional elevado: para dimensión 1 y tamaño muestral 200, 7 minutos, incrementándose a 22 minutos para dimensión 2 y 20 proyecciones, y a 1 hora 40 minutos si empleamos 100 proyecciones.

3.2. Test computacional de Kolmogorov-Smirnov para una muestra

Sea $X_1, ..., X_m$ una muestra aleatoria simple procedente de una población con distribución continua F. Sea $D_{RT}(X_i, F)$ la profundidad de Tukey aleatoria teórica del punto X_i respecto a dicha distribución y sea $D_{RT}(X_i, F_m)$ la profundidad de Tukey aleatoria empírica.

Las hipótesis que deseamos contrastar son:

$$H_0 : D_{RT}(X_i, F_m) = D_{RT}(X_i, F) \text{ para todo } i \text{ con } i = 1, ..., m.$$
$$H_1 : D_{RT}(X_i, F_m) \neq D_{RT}(X_i, F) \text{ para algún } i \text{ con } i = 1, ..., m.$$

Vamos a hacerlo basándonos en el estadístico de Kolmogorov-Smirnov y empleando las funciones depth2.RT, depth2.real.RT (para dimensión 1) y depth2.real.Rk2.RT (para dimensiones mayores). Para el cálculo del estadístico de prueba computacionalmente, el cual incluimos en el anexo, modificamos la expresión de S de la sección 2.2.1 por

$$J = max_{i=1,\dots,m}\{max\{|D_{RT}(X_i, F_m) - D_{RT}(X_i, F)|, |D_{RT}(X_i, F_m) - \frac{1}{m} - D_{RT}(X_i, F)|\}\}$$

Por cómo está definido el estadístico de prueba, nos interesa que al comparar la profundidad empírica y la profundidad teórica de un punto X_i, se trate de la misma proyección. De esta forma se comparan las mismas marginales (las mismas funciones de distribución).

Aquí surge la pregunta de cuántas proyecciones emplear en el cálculo de la profundidad. Ya abordamos este tema anteriormente, ahora bien, teniendo en cuenta que a la hora de pasar el test interesa que las profundidades empírica y teórica de un punto se obtengan para la misma proyección, parece lógico pensar que entre el número de proyecciones valoradas funcione mejor para un número pequeño. De esta forma, la probabilidad de comparar las mismas marginales será mayor.

Para obtener datos del número de veces de las que pasamos el test que se comparan las mismas marginales, construimos la función NumberEqualProy. Esta requiere introducir los siguientes valores de entrada: $dP1 = depth2.RT.NL(datau, maproj)\$proj$ y $dP2 = depth2.real.Rp.RT.NL(datau, maproj, med, mcov)\$RealProj$, es decir la proyección a la que corresponde la profundidad calculada para cierto punto. Las funciones $depth2.RT.NL$ y $depth2.real.Rp.RT.NL$ aparecen en el anexo.

```
NumberEqualProj=function(dP1,dP2){
  projDif=dP1-dP2
  NEqualProj=length(which(projDif==0))
  return(NEqualProj)}
```

Por ejemplo, para dimensión 2 tomamos una muestra de tamaño 200 de una normal multivariante con media dada por el vector nulo y matriz de covarianza igual a la identidad y calculamos la profundidad de Tukey aleatoria (empírica) y la profundidad de Tukey aleatoria teórica. Además, comparamos para cada uno de los 200 puntos a qué proyección corresponde cada profundidad calculada. De esta forma, cuanto más cercano sea el valor de NumberEqualProy a 200, mejor para pasar el test. Empleamos 20 y 100 proyecciones y en cada caso calculamos NumberEqualProy 1000 veces, obteniendo las medias y desviaciones que aparecen en la siguiente tabla.
Observamos que NumberEqualProj es mucho mayor para 20 proyecciones que para 100, por lo que trabajaremos con 20 proyecciones.

Nº de proyecciones	Media NumberEqualProj	Desviación
20	102	11.42
100	25	5.59

3.2.1. Resultados

Como en el test de Kolmogorov-Smirnov para dos muestras, el procedimiento que realizamos es calcular 1000 p-valores y analizar si se distribuyen como una uniforme en [0,1].

Un ejemplo para dimensión 2:

```
p=2; m=200
nproj=20; nproj1=10
med=c(0,0); mcov<-diag(2)
r=1000; g=numeric(r)
for(i in 1:r){
  datau<-mvrnorm(m,med,mcov)
  maproj1<-matrix(rnorm(p*nproj1),p,nproj1)
  for(j in 1:nproj1){
    maproj1[,j]<-maproj1[,j]/sqrt(sum(maproj1[,j]^2))}
  maproj<-cbind(maproj1,-maproj1)
  depthu=depth2.RT(datau,maproj)
  depthureal=depth2.real.Rk2.RT(datau,maproj,med,mcov)
  g[i]=kstest1pval(depthu,depthureal,alternative="two.sided")}
k1=sort(g)
sum(k1<0.05)
hist(g,breaks=20)
```

En cada caso analizado repetiremos el proceso de calcular 1000 p-valores cinco veces. En las tablas que se presentan a continuación aparecen las medias del número de rechazos de H_0 y su desviación, así como alguno de los histogramas obtenidos.

Para dimensión 1 tomaremos muestras procedentes de N(0,1) de tamaño 200. Para dimensiones superiores, aumentaremos el tamaño muestral y trabajaremos con distribuciones normales multivariantes N(μ,Σ), de media $\mu = (0_1, 0_2, ..., 0_p)$ y matriz de covarianza

$$\Sigma = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1p} \\ a_{21} & a_{22} & ... & a_{2p} \\ ... & ... & ... & ... \\ a_{p1} & a_{p2} & ... & a_{pp} \end{pmatrix}$$

con $a_{ii} = 1 \ \forall$ i, i=1, ..., p, y a_{ij} con $i \neq j$ con valor 0, 0.2, 0.5 o 0.9 según el caso.

Dimensión 1, tamaño muestral 200

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
200	46	3.2	0.05

Tabla 3.3: Resultados del test K-S para las profundidades de Tukey aleatoria de una muestra N(0,1).

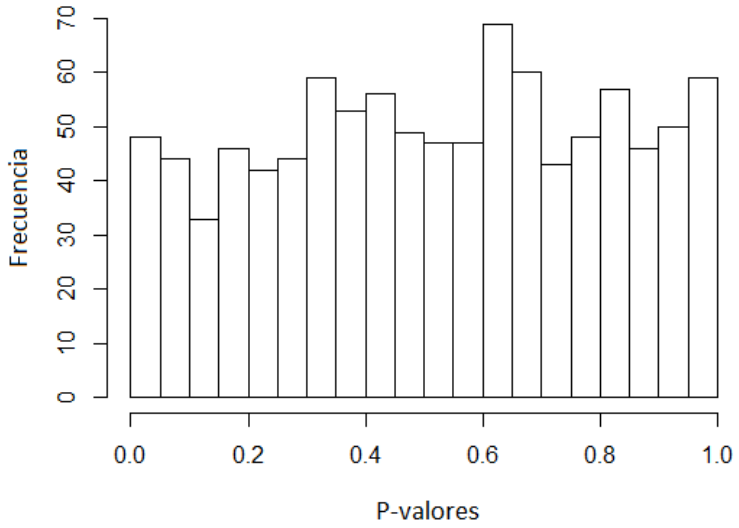


Figura 3.3: Histograma de los p-valores obtenidos con el test K-S para una muestra N(0,1) de tamaño 200.

Dimensión 2, tamaño muestral 200

Matriz de covarianza (valor de los a_{ij} con $i \neq j$)	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
0	62	6.18	1
0.2	57	6.63	1
0.5	59	6.28	1
0.9	59	5.39	1

Tabla 3.4: Resultados del test K-S para las profundidades de Tukey aleatoria de una muestra normal bivariante de media vector nulo y matriz de covarianza Σ.

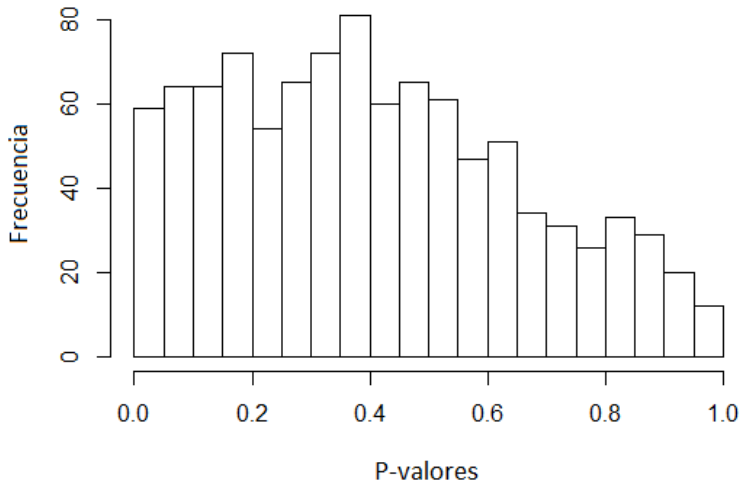


Figura 3.4: Histograma de los p-valores obtenidos con el test K-S para una muestra normal bivariante de tamaño 200.

Dimensión 3, tamaño muestral 300

Matriz de covarianza (valor de los a_{ij} con $i \neq j$)	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
0	51	10.82	1.5
0.2	51	3.74	1.8
0.5	53	9.36	1.5
0.9	58	7.56	1.7

Tabla 3.5: Resultados del test K-S para las profundidades de Tukey aleatoria de una muestra normal multivariante de media μ=(0,0,0) y matriz de covarianza Σ.

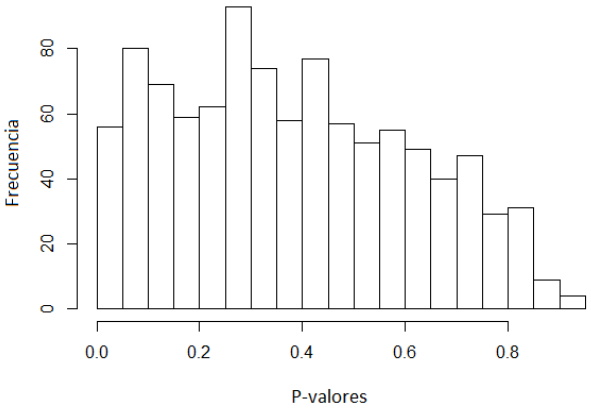


Figura 3.5: Histograma de los p-valores obtenidos con el test K-S para una muestra normal trivariante de tamaño 300.

Dimensión 4, tamaño muestral 500

Matriz de covarianza (valor de los a_{ij} con $i \neq j$)	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
0	48	9.42	2.4
0.2	50	8.33	2.4
0.5	55	5.26	2.4
0.9	60	3.46	2.4

Tabla 3.6: Resultados del test K-S para las profundidades de Tukey aleatoria de una muestra normal multivariante de media μ=(0,0,0,0) y matriz de covarianza Σ.

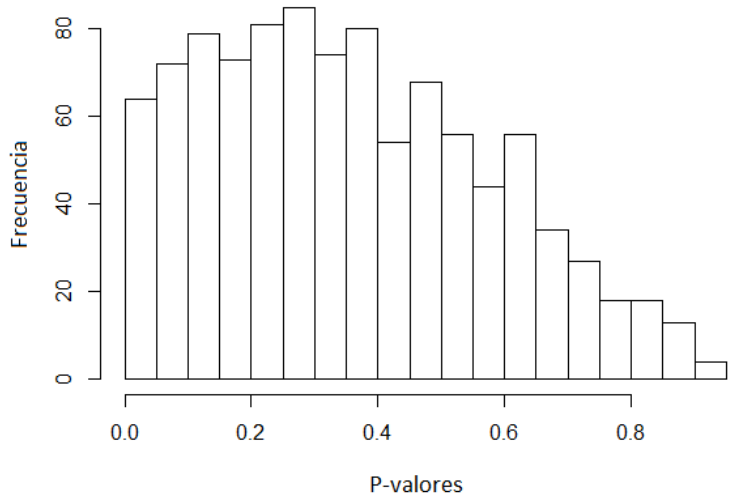


Figura 3.6: Histograma de los p-valores obtenidos con el test K-S para una muestra normal tetravariante de tamaño 500.

Dimensión 5, tamaño muestral 800

Matriz de covarianza (valor de los a_{ij} con $i \neq j$)	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
0	49	6.63	4.05
0.2	54	5.06	4.07
0.5	59	5.5	4.08
0.9	62	3.83	4.11

Tabla 3.7: Resultados del test K-S para para las profundidades de Tukey aleatoria de una muestra normal multivariante de media μ=(0,0,0,0,0) y matriz de covarianza Σ.

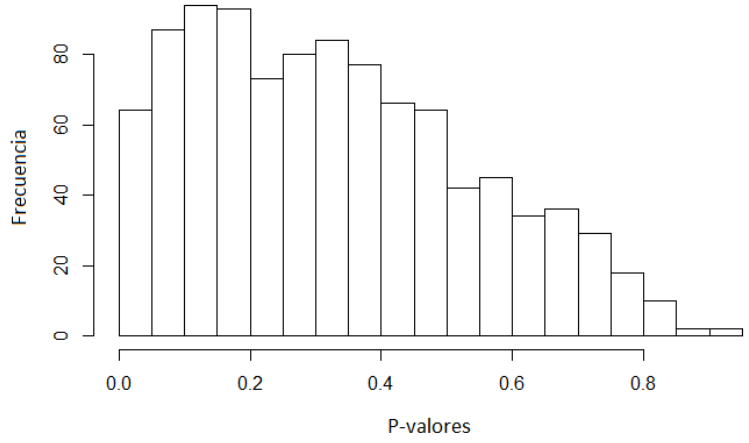


Figura 3.7: Histograma de los p-valores obtenidos con el test K-S para una muestra normal pentavariante de tamaño 800.

En primer lugar, observar que con el test de Kolmogorov-Smirnov para una muestra el tiempo computacional es mucho menor que con el test K-S para dos muestras. Por ejemplo, para muestras normales bivariantes de tamaño 200 y empleando 20 proyecciones, se registra un tiempo computacional de 1 minuto con el primer test frente a 22 minutos con el segundo.

Por otro lado, el test K-S de una muestra realizado para dimensión 1, presenta un número de rechazos de la hipótesis nula cercano al 5 % y, además, el histograma se parece al de una U[0, 1]. Al aumentar la dimensión, el número de rechazos de H_0 se incrementa ligeramente y los histogramas se van distanciando del de una U[0,1]. La razón de que a medida que aumentamos la dimensión el test no funcione tan bien, puede deberse a varios motivos:

- Al realizar proyecciones se pierde la independencia de los datos.
- En el cálculo del estadístico de prueba no se comparan siempre las mismas proyecciones.

3.3. Test computacional Chi-Cuadrado

Abordamos ahora el problema de otra forma: en lugar de comparar la profundidad empírica y la profundidad teórica para cada punto, establecemos unas clases (intervalos contenidos en el intervalo [0, 0.5]), calculamos las profundidades empírica y teórica de los puntos de la muestra y miramos cuántos valores hay en cada clase.

Retomando lo visto en la sección 2.2.2, descomponemos el recorrido de la profundidad teórica, es decir [0, 0.5], en un número finito k de clases, $C_1, ..., C_k$. Sea O_i, con $i = 1, ..., k$, la frecuencia observada correspondiente a la clase C_i y E_i la frecuencia esperada.

Contrastaremos las siguientes hipótesis:

$$H_0 : E_i = O_i \ \forall \text{ i, con } i = 1, ..., k.$$

$$H_1 : \exists \text{ i tal que } E_i \neq O_i.$$

Para obtener las frecuencias observadas y esperadas, empleamos la siguiente función, a la que pasamos como entrada depth.RT o depth.real.RT (depth.real.Rk2.RT para dimensión mayor que 1), respectivamente. Definimos los extremos de los intervalos (clases) mediante el vector "vect", si establecemos 10 clases de longitud 0.05 cada una, la función sería así:

```
contdux=function(dux){
  vect=c(0,0.05,0.1,0.15,0.2,0.25,0.3,0.35,0.4,0.45,0.5)
  n=length(vect)
  num=numeric(n-1)
  for(i in 1:n-2){
    num[i]=sum(dux>=vect[i] & dux<vect[i+1])
  }
  num[n-1]=sum(dux>=vect[n-1]&dux<=vect[n])
  return(num)}
```

Por ejemplo, tomamos una muestra de una N(0,1) de tamaño 50 y calculamos las profundidades empíricas mediante la función depth.RT. Ordenamos dichas profundidades obteniendo los siguientes valores:

0 0.02 0.02 0.04 0.04 0.06 0.06 0.08 0.08 0.10 0.10 0.12 0.12 0.14 0.14 0.16 0.16 0.18 0.18
0.20 0.20 0.22 0.22 0.24 0.24 0.26 0.26 0.28 0.28 0.30 0.30 0.32 0.32 0.34 0.34 0.36 0.36
0.38 0.38 0.40 0.40 0.42 0.42 0.44 0.44 0.46 0.46 0.48 0.48 0.50

Pasando el vector formado por dichos valores a la función contdux, obtenemos el número de valores en cada clase:

[0,0.05)	[0.05,0.1)	[0.1,0.15)	[0.15,0.2)	[0.2,0.25)
5	4	6	4	6

Observar que a medida que aumentamos la dimensión, hay más valores de profundidad cercanos a 0 y disminuyen los valores próximos a 0.5. Esto se trata de una propiedad de la

[0.25,0.3)	[0.3,0.35)	[0.35,0.4)	[0.4,0.45)	[0.45,0.5]
4	6	4	6	5

profundidad de Tukey, a la que se hace referencia en [2]. Representamos gráficamente la profundidad de Tukey aleatoria teórica de muestras normales de tamaño 50 para dimensión 1, 2, 4 y 5:

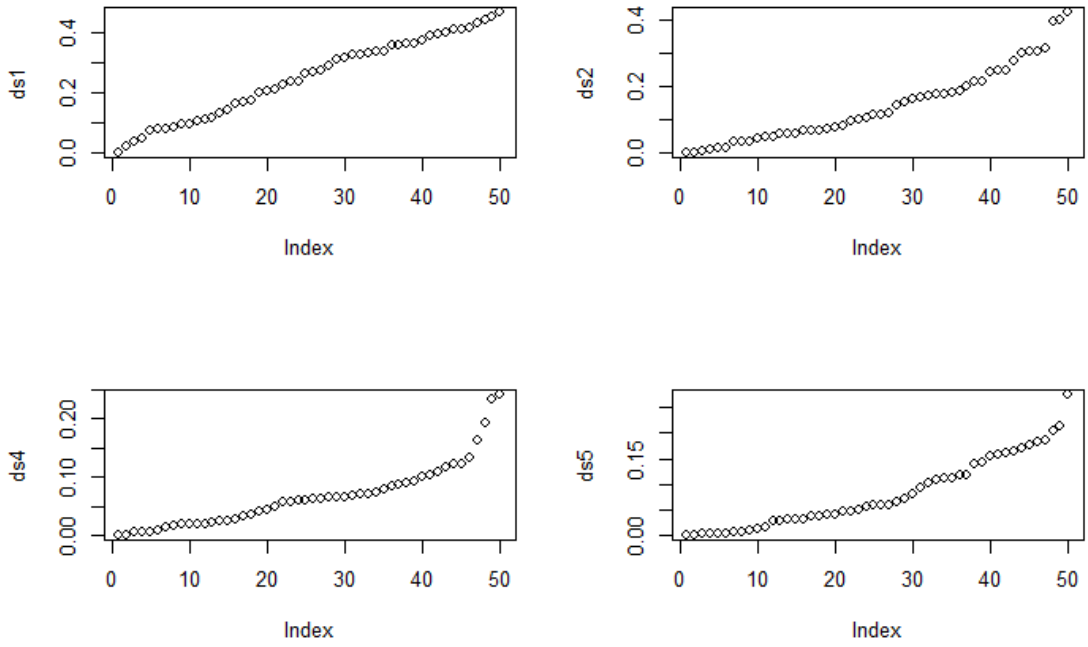


Figura 3.8: Se representan los valores ordenados de la profundidad de Tukey aleatoria teórica para dimensiones 1 (ds1), 2 (ds2), 4 (ds4) y 5 (ds5).

Dado que el test de Chi-Cuadrado requiere que las frecuencias esperadas de cada clase sean mayores o iguales a 5, a medida que aumentemos la dimensión, variaremos las clases uniendo intervalos del extremo próximo a 0.5.

3.3.1. Resultados

Dimensión 1

En primer lugar, establecemos las clases en las que vamos a dividir el recorrido de la función de profundidad de Tukey. En este caso, lo dividimos en diez intervalos de longitud 0.05 como indicamos arriba: [0, 0.05), [0.05, 0.1), [0.1, 0.15), ..., [0.45, 0.5]. Ahora, tomamos una muestra de una N(0,1) de tamaño 400, calculamos las profundidades empírica y teórica de cada punto, realizamos el recuento de valores en cada clase establecida y pasamos el test de Chi-Cuadrado. Hacemos esto 1000 veces y representamos los p-valores obtenidos.

```
p=1; m=400
nproj=1
med=0; desv=1
r=1000; g=numeric(r)
for(i in 1:r){
  datax<-matrix(rnorm(m),m,p)
  maproj<-matrix(1)
  dx=depth.RT(datax,maproj)
  dr=depth.real.RT(datax,med,desv)
  contdx=contdux(dx)
  contdr=contdux(dr)
  pcontdr=contdr/m
  g[i]=chisq.test(x=contdx,p=pcontdr,correct=FALSE)$p.value
}
k1=sort(g)
sum(k1<0.05)
hist(g,breaks=20)
```

Repetimos el proceso anterior cinco veces, calculamos la media del número de rechazos de H_0 así como la desviación, obteniendo los siguientes resultados:

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
400	65	2.41	0.07

Tabla 3.8: Resultados del test Chi-Cuadrado para las profundidades de Tukey aleatoria de una muestra procedente de una N(0,1).

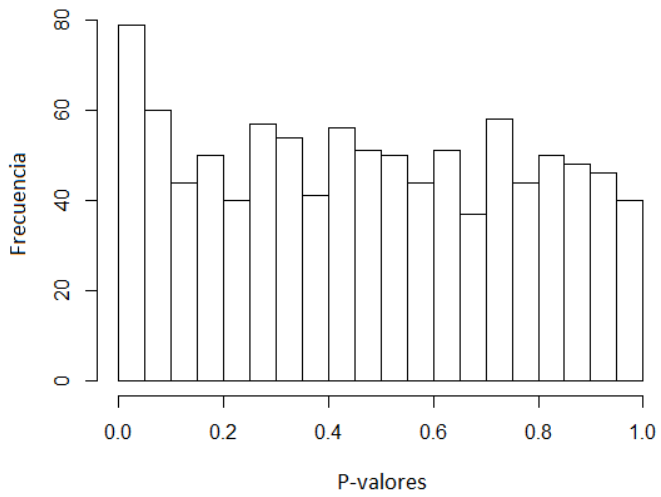


Figura 3.9: Histograma de los p-valores obtenidos con el test Chi-Cuadrado para una muestra N(0,1) de tamaño 400.

Dimensión 2

En este caso tomamos muestras de tamaño 300 procedentes de normales multivariantes de media dada por el vector nulo y matriz de covarianza la identidad y empleamos 20 proyecciones para el cálculo de la profundidad de Tukey aleatoria. Para el test de Chi-Cuadrado establecemos nueve clases: unimos las clases [0.40, 0.45) y [0.45, 0.5] definidas en el caso anterior, de forma que la frecuencia esperada de la novena clase sea al menos cinco.

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
300	51	5.95	2.43

Tabla 3.9: Resultados del test Chi-Cuadrado para las profundidades de Tukey aleatoria de una muestra procedente de una normal bivariante de media (0,0) y matriz de covarianza Id_2.

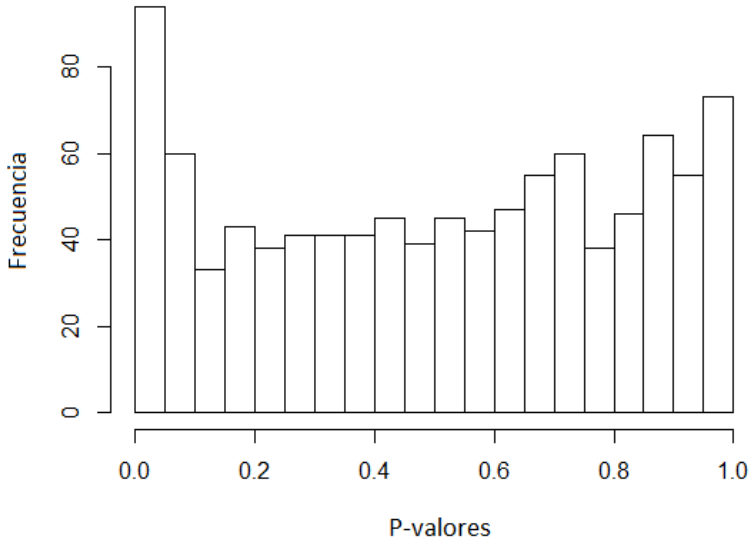


Figura 3.10: Histograma de los p-valores obtenidos con el test Chi-Cuadrado para una muestra normal bivariante de tamaño 200.

Dimensión 3

Mantenemos el número de proyecciones para el cálculo de la profundidad (20), pero aumentamos el tamaño muestral a 400. Además, como aumentan los valores de profundidad próximos a 0 y disminuyen los valores próximos a 0.5, establecemos las siguientes clases: ocho clases de longitud 0.025 ([0, 0.025), [0.025, 0.05),..., [0.175, 0.2)), dos de longitud 0.05 ([0.2, 0.25), [0.25, 0.3)) y la undécima clase de longitud 0.2 ([0.3,0.5]).

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
400	63	7.16	2.5

Tabla 3.10: Resultados del test Chi-Cuadrado para las profundidades de Tukey aleatoria de una muestra procedente de una normal multivariante de media (0,0,0) y matriz de covarianza Id_3.

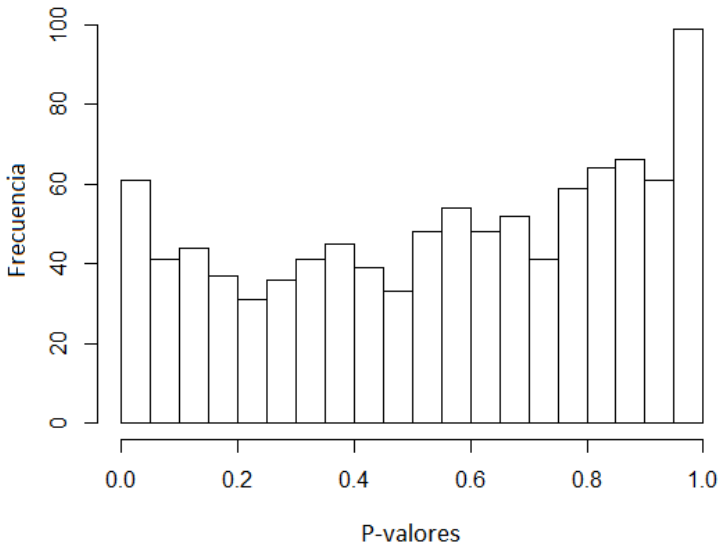


Figura 3.11: Histograma de los p-valores obtenidos con el test Chi-Cuadrado para una muestra normal trivariante de tamaño 400.

Dimensión 4

Tomamos 20 proyecciones, tamaño muestral 400 y once clases como en el caso anterior, obteniendo los siguientes resultados:

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
400	66	5.5	2.33

Tabla 3.11: Resultados del test Chi-Cuadrado para las profundidades de Tukey aleatoria de una muestra procedente de una normal multivariante de media vector nulo y matriz de covarianza Id_4.

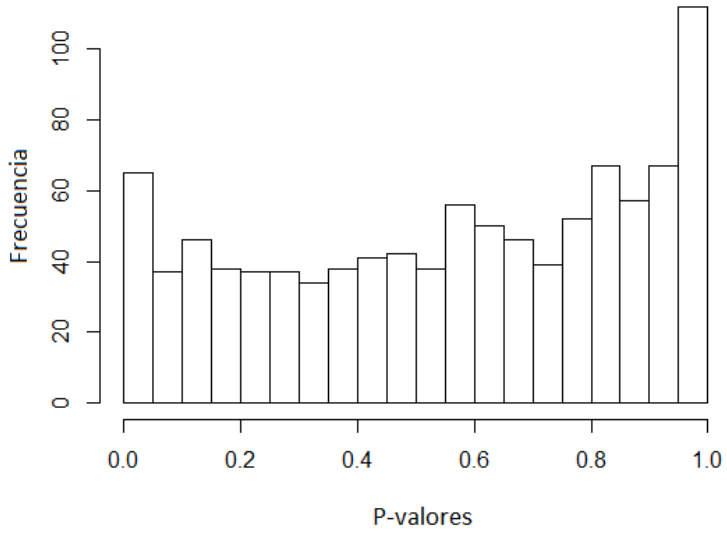


Figura 3.12: Histograma de los p-valores obtenidos con el test Chi-Cuadrado para una muestra normal tetravariante de tamaño 400.

Dimensión 5

Para dimensión cinco, en lugar de seguir aumentando el tamaño muestral, tomamos únicamente cinco clases de la siguiente forma: [0, 0.05), [0.05, 0.1), [0.1, 0.15), [0.15, 0.2), [0.2, 0.25), [0.25, 0.5]. Obtenemos los siguientes resultados:

Tamaño muestral	Nº de rechazos de H_0	Desviación	Tiempo computacional (min)
200	74	6.61	1.33

Tabla 3.12: Resultados del test Chi-Cuadrado para las profundidades de Tukey aleatoria de una muestra procedente de una normal multivariante de media vector nulo y matriz de covarianza Id_5.

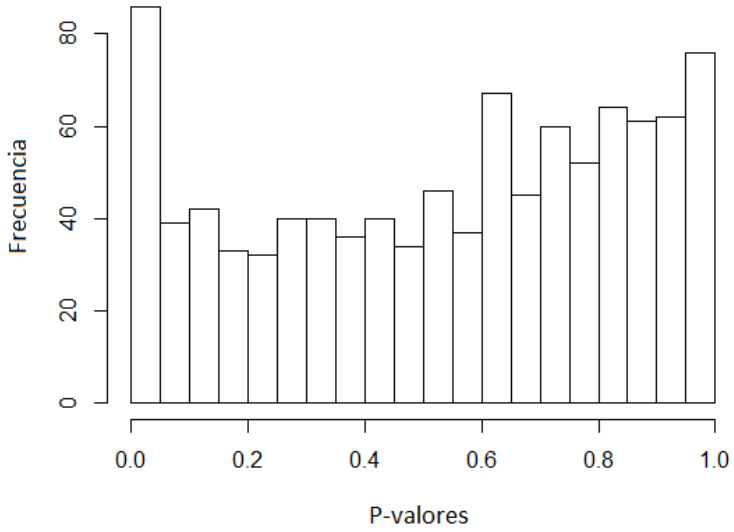


Figura 3.13: Histograma de los p-valores obtenidos con el test Chi-Cuadrado para una muestra normal pentavariante de tamaño 200.

En general, el número de rechazos de la hipótesis nula es algo superior al 5 % de las veces que pasamos el test. Al representar los histogramas de los p-valores, los valores centrales se distribuyen uniformemente, pero los valores de los extremos presentan frecuencias un poco mayores. Respecto al tiempo computacional, es similar al del test K-S para una muestra.

Capítulo 4

Conclusiones

Para finalizar, haremos un repaso del propósito del trabajo así como de los resultados obtenidos.
En primer lugar, destacar el interés de la función de profundidad estadística de Tukey para, partiendo de datos en $\mathbb{R}^p$, $p > 1$, obtener información en $\mathbb{R}$. De esta forma, se puede establecer un orden para datos multidimensionales.

Además, presenta interesantes propiedades como la que hemos tratado de comprobar: si partimos de dos muestras con la misma distribución en $\mathbb{R}^p$ y tomamos un conjunto de puntos de $\mathbb{R}^p$, la profundidad de Tukey de cada punto respecto a la primera muestra, coincide con la calculada respecto a la segunda muestra.

En el terreno computacional, resulta muy útil la definición de profundidad de Tukey aleatoria. Hemos analizado el número de proyecciones a emplear, eligiendo, en base a las dimensiones tratadas y a las características de los estadísticos de test, 20 proyecciones.

Respecto a los test de hipótesis, queríamos ver que la distribución de los p-valores seguía una U[0,1], lo que indica que no podemos rechazar la hipótesis nula (si partimos de dos distribuciones normales iguales, las profundidades coinciden).
A pesar de que los histogramas obtenidos no siguen fielmente una U[0,1], sacamos las siguientes conclusiones:

- Para las hipótesis contrastadas, funciona mejor el test de Kolmogorov-Smirnov para una muestra, en el que se realiza un contraste con la función de profundidad teórica, que el test K-S para dos muestras. Esto era de esperar debido a que hay más incertidumbre en el test asociado a dos muestras.
- Una forma diferente de abordar el problema es con el test de Chi-Cuadrado, en el que se realiza un recuento de los valores de profundidad que hay en cada clase en lugar de comparar la profundidad para cada punto. A la vista de los histogramas de los p-valores, los resultados son mejores que en los casos anteriores, pues al aumentar la dimensión no distan tanto del de una U[0,1].

- A medida que aumentamos la dimensión, la forma de los histogramas indican que el test funciona peor, por lo que incrementamos el tamaño muestral y/o, en el caso del test Chi-Cuadrado, variamos el número de clases.

El hecho de que los test no funcionen de forma "ideal" puede deberse a varias razones. Para Kolmogorov-Smirnov hemos tenido que modificar la expresión del estadístico y ya vimos que, aunque debería compararse la misma proyección, no siempre ocurre esto. Por otro lado, al proyectar los datos se está perdiendo independencia.

Al abordar el test computacional de Chi-Cuadrado, representamos la profundidad teórica y vimos que al aumentar la dimensión se incrementan los valores de profundidad cercanos a 0. En el cálculo de la profundidad empírica muchos valores resultan ser 0 y dicho cálculo podría suavizarse tal y como proponen Einmahl, Li y Liu en [6]. De esta forma, el número de valores en cada clase para la profundidad teórica y la empírica se acercaría más. No obstante, esta mejora se aleja de los objetivos de nuestro trabajo.

Por último, de cara al futuro sería interesante abordar la demostración teórica tanto de esta propiedad como de la recíproca.

Anexo

En la sección 3.2, aparece la función kstest1pval, mediante la que se calcula el p-valor asociado al test de Kolmogorov-Smirnov para una muestra.

```
kstest1pval=function(depthu,depthureal,alternative="two.sided"){
  s<-length(depthu)
  dif1=numeric(s)
  dif2=numeric(s)
  for(i in 1:s){
    dif1[i]=abs(depthu[i]-depthureal[i])
    dif2[i]=abs(depthureal[i]-depthu[i]+1/s)}
  difmax=max(c(dif1,dif2))
  pval=ifelse(alternative=="two.sided",
              1-pks(sqrt(s)*difmax),exp(-2*s*difmax^2))
  pval<-min(1,max(0,pval))
  return(pval)}
```

Observar que para el cálculo del p-valor nos hemos basado en el código del ks.test de R.

Respecto a las funciones depth2.RT.NL y depth2.real.Rp.RT.NL, que nos devuelven la proyección a la que corresponde la profundidad calculada para cierto punto, son las siguientes:

```
depth2.RT.NL=function(data,maproj){
  m<-nrow(data)
  Prod=data*maproj
  a=apply(cbind(apply(Prod,2,rank,ties="max")))
  wmin=numeric(m)
  depth=numeric(m)
  for(i in 1:m){
    wmin[i]=which.min(a[i,])
    depth[i]=a[i,wmin[i]]/m}
  newList<-list("proj"=wmin,"depth"=depth)
  return(newList)
}

depth2.real.Rp.RT.NL=function(data,maproj,med,mcov){
```

```
  m<-nrow(data)
  p<-ncol(data)
  nproj<-ncol(maproj)
  Prod=data%*%maproj
  dur1<-matrix(0,m,nproj)
  wdur=numeric(m)
  dur=numeric(m)
  med2=med%*%maproj
```

Calculamos las profundidades 1-dimensionales de las proyecciones
unidimensionales de X_i y nos quedamos con la mínima:

```
  for(i in 1:m){
    for(j in 1:nproj){
      medj=med2[,j]
      varj=t(maproj[,j])%*%mcov%*%maproj[,j]
      desvj=sqrt(varj)
      dur1[i,j]=pnorm(Prod[i,j],medj,desvj)}
    wdur[i]=which.min(dur1[i,])
    dur[i]=dur[i,wdur[i]]}
  newList<-list("RealProj"=wdur,"RealDepth"=dur)
  return(newList)
}
```

Bibliografía

[1] ARNHOLT, ALAN T.; MILITINO, ANA F.; UGARTE, M. DOLORES. 2008. *Probability and statistics with R.* 1ª ed. Boca Raton: CRC Press.

[2] CHAUDHURI, PROBAL; DUTTA, SUBHAJIT; GHOSH, ANIL K. 2011. *Some intriguing properties of Tukey's halfspace depth.* Bernoulli, 17. 1420–1434.

[3] CHICKEN, ERIC; HOLLANDER, MYLES; WOLFE, DOUGLAS A. 2014. *Nonparametric statistical methods.* 3ª ed. Hoboken, New Jersey: Wiley.

[4] CUESTA ALBERTOS, JUAN A.; NIETO REYES, ALICIA. 2008. *The random Tukey depth.* Computational Statistics and Data Analysis, 52. 4979-4988.

[5] CUESTA ALBERTOS, JUAN A.; NIETO REYES, ALICIA. 2008. *The Tukey and the random Tukey depths characterize discrete distributions.* Journal of Multivariate Analysis, 99. 2304-2311.

[6] EINMAHL, J.H.J.; LI, JUN; LIU, REGINA Y. 2015. *Bridging centrality and extremity: refining empirical data depth using extreme value statistics.* The Annals of Statistics. Vol. 43, Nº 6. 2738–2765.

[7] GONZÁLEZ RUIZ, IGNACIO. 2012. *Caracterización de distribuciones de probabilidad mediante las profundidades de Tukey y Tukey aleatoria.* Trabajo dirigido en estadística y computación por Nieto Reyes, Alicia. Universidad de Cantabria.

[8] *P-values uniformly distributed under null hypothesis* [Online, accedido 21 mayo 2018]
« http://yaikhom.com/2016/04/14/p-values-uniformly-distributed-under-null-hypothesis.html»

[9] ROBINSON, DAVID. *How to interpret a p-value histogram* [Online, accedido 21 mayo 2018]
«http://varianceexplained.org/statistics/interpreting-pvalue-histogram/»

[10] STOREY, JOHN D.; TIBSHIRANI, ROBERT. 2003. *Statistical significance for genomewide studies.* PNAS, 100 . 9440-9445.

[11] TRIOLA, MARIO F. 2013. *Estadística.* 11ª ed. Naucalpan de Juárez: Pearson Educación de México.

[12] TUKEY, J.W. 1975. *Mathematics and picturing of data.* Proc. of ICM. Vancouver, 2. 523-531.

[13] ZUO, Y.; SERFLING, R. 2000. *General notions of statistical depth function.* The Annals of Statistics, 28. 461-482.