



## TRABAJO FIN DE GRADO

---

### Elementos de Inferencia no Paramétrica

---

Realizado por: Brenda María Duarte Reyes

Supervisado por: Eduardo Conde Sánchez

FACULTAD DE MATEMÁTICAS  
DPTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA



## Agradecimientos

*A mis padres. Por su infinita paciencia. Por darme la oportunidad de recorrer este camino y ayudarme a no perderme en él.*

*A mi familia y a todas esas personas a las que quiero llamar familia. Gracias por esos pequeños impulsos que me hacían avanzar cuando perdía velocidad.*

*A mi tutor, por su dedicación y sus largas correcciones. Gracias por haberme dado la oportunidad de realizar este trabajo.*

*Por último, gracias a todas aquellas personas que a lo largo de estos años me han hecho crecer.*

*Gracias a todos por confiar en mi cuando ni yo misma lo hacía.*



# Índice general

Abstract . . . . .	6
<b>1. Contrastes de localización, simetría y bondad de ajuste</b>	<b>8</b>
1.1. Introducción . . . . .	8
1.1.1. Contrastes basados en rangos . . . . .	10
1.1.2. Contrastes de permutación y contrastes combinados . . . . .	10
1.2. Contrastes univariantes para una muestra . . . . .	14
1.2.1. Contraste de bondad de ajuste de Kolmogorov . . . . .	15
1.2.2. Contrastes de permutación univariante para simetría . . . . .	22
1.3. Contrastes multivariantes de una sola muestra . . . . .	29
1.3.1. Contrastes basados en rangos multivariantes para ten- dencia central . . . . .	30
1.3.2. Contrastes de permutación multivariante para simetría . . . . .	34
1.4. Contrastes univariantes de dos muestras . . . . .	40
1.4.1. Contraste de Wilcoxon (Mann-Whitney) . . . . .	41
1.4.2. Contrastes de permutación para tendencia central . . . . .	51
1.5. Contrastes multivariantes para dos muestras . . . . .	55
1.5.1. Contrastes multivariantes basados en rangos . . . . .	56
1.5.2. Contrastes de permutación para la tendencia central en muestras multivariantes . . . . .	64

<b>2. Contrastes de dispersión</b>	<b>66</b>
2.1. Introducción . . . . .	66
2.2. Comparación de la dispersión . . . . .	68
2.2.1. Contraste de Ansari-Bradley . . . . .	70
2.2.2. Contraste de permutaciones de Pan . . . . .	75
2.2.3. Contraste de permutaciones O'Brien . . . . .	80
2.3. Comparación conjunta de tendencia central y dispersión . . .	83
2.3.1. El contraste de Lepage . . . . .	85
2.3.2. El contraste de Cucconi . . . . .	87
2.4. Comparación de distribuciones . . . . .	91
2.4.1. Contraste de Kolmogorov-Smirnov . . . . .	92
2.4.2. Contraste de Cramér-von Mises . . . . .	98

# Abstract

This project deals with nonparametric statistical solutions for hypotheses testing problems and codes for the software environment *R* for the application of these solutions. In particular, the methods presented here consist of rank or permutation procedures which have been applied to real data related to the level of knowledge of students of the faculty of Mathematics in the University of Seville in some statistics subjects.

Chapter 1 deals with one-sample and two-sample location problems, tests for symmetry and tests to compare two independent populations in terms of central tendency. According to the number of response variables and to the number of samples, we distinguish four kinds of methods: univariate one-sample tests, multivariate one-sample tests, univariate two-sample tests and multivariate two-sample tests. In this first category are the Kolmogorov goodness-of-fit test and the permutation test for symmetry; in the second group of procedures are the multivariate rank test for central tendency and the multivariate extension of the permutation test on symmetry; among the procedures included in the third family of solutions are the Wilcoxon test and the permutation test on central tendency; finally some other general procedures in the basis of rank and permutation approaches are discussed.

Chapter 2 deals with two independent random samples, one sample drawn from each of two underlying populations. This chapter presents some

tests for comparing variabilities and distributions. For problems of variability comparisons the Ansary-Bradley test, the permutation Pan test and the permutation O'Brien test are considered. For jointly comparing central tendency and variability the Lepage test and the Cucconi test are presented. While for problems of comparing the statistical distributions of both population the Kolmogorov-Smirnov and the Cramer-von Mises are analyzed.

# Capítulo 1

## Contrastes de localización, simetría y bondad de ajuste

### 1.1. Introducción

El modelado estadístico de una situación real que involucra incertidumbre se realiza formalmente a través de una estructura compleja. El denominado *espacio de probabilidad* requiere definir un conjunto de escenarios posibles que pueden darse, una familia de subconjuntos de estos escenarios cuya verosimilitud pueda ser medida y la propia medida de dicha verosimilitud. Los términos que se usan habitualmente para designar estos conceptos son *experimento aleatorio*, *espacio de resultados*, *sigma álgebra de sucesos* o *medida de probabilidad*.

En la práctica es muy difícil modelar un experimento aleatorio mediante un espacio de probabilidad, requiere mucha información. Por esta razón se recurre a las *variables aleatorias* y a su *distribución de probabilidad*. Estas permiten, un acercamiento práctico al modelado de determinadas características del fenómeno aleatorio que se consideran de interés. No obstante,

en ocasiones esta reducción en el ámbito de la modelización es excesiva, llegando a imponer familias paramétricas de distribuciones de probabilidad que no se ajustan adecuadamente al experimento o condicionando fuertemente la validez de los contrastes que se puedan realizar al respecto. En esta situación entra en juego la *Inferencia no Paramétrica* que relaja algunas de las hipótesis permitiendo aun así construir contrastes útiles sobre las distribuciones de las variables estudiadas.

En este capítulo se presentan algunos de los elementos de Inferencia no Paramétrica desarrollados para el estudio de propiedades de simetría y localización, así como métodos orientados al ajuste de las distribuciones de las variables y a la comparación entre variables desde el punto de vista distribucional. Estos métodos están siendo adaptados en la actualidad al análisis de grandes cantidades de datos, el denominado *Big Data*, lo que refuerza el interés de estas herramientas por su proyección de futuro. La disponibilidad de interminables bases de datos juega, en cierto modo, a favor de las técnicas que se verán en lo sucesivo debido a que en muchos casos las distribuciones de los estadísticos usados en los contrastes, son desconocidas por lo que será necesario estimar los valores críticos y para ello es necesario utilizar información muestral en abundancia.

A continuación se detalla cómo se utilizan dos ideas elementales en Inferencia no Paramétrica en la construcción de contrastes de hipótesis. Una de ellas se basa en el empleo de los denominados rangos de las observaciones, es decir, las posiciones que ocupan en una ordenación de la muestra considerada. Por ejemplo, si comparamos las distribuciones de probabilidad de dos variables, bajo la hipótesis nula de igualdad distribucional, los rangos de los valores observados deberían alternar entre una y otra variable sin que predominase ningún patrón reconocible. Esta idea se puede usar para

estimar valores críticos para el contraste. La otra idea utiliza permutaciones en la asignación de los datos de la muestra a un conjunto de categorías relacionadas con la propiedad que se pretende contrastar.

### 1.1.1. Contrastes basados en rangos

Esta metodología consiste en sustituir los datos muestrales por los rangos que ocupan en la ordenación de menor a mayor y usar dicha información con arreglo al contraste que se quiera construir. Formalmente, el rango de la  $i$ -ésima observación con respecto a un conjunto de  $n$  datos viene dado por

$$R_i = \mathcal{R}(X_i) = \sum_{j=1}^n \mathbb{I}(X_j \leq X_i)$$

donde  $\mathcal{R}$  es el operador rango,  $X_i$  es la observación muestral,  $\mathbb{I}(A)$  es la función indicador del evento  $A$ , la cual vale 1 si  $A$  es verdad y 0 en otro caso. Por lo tanto, el rango de  $X_i$  dentro de  $\{X_1, \dots, X_n\}$  es igual a 1 si  $X_i$  es el mínimo valor, 2 si  $X_i$  es el segundo valor más pequeño, hasta  $n$  si  $X_i$  es el máximo valor. En caso de empate, se aplica el *rango medio*, es decir, si el rango  $r$  se asigna a  $t$  observaciones iguales a un cierto valor  $x$  ( $t \leq r$ ) esto es,  $r$  observaciones en el conjunto  $\{X_1, \dots, X_n\}$  son menores o iguales a  $x$ , entonces el rango de esas  $t$  observaciones según el criterio adoptado se ajusta al valor medio de los  $t$  rangos  $(r - t + 1), (r - t + 2), \dots, r$ .

### 1.1.2. Contrastes de permutación y contrastes combinados

Muchos de los contrastes que estudiaremos se basan en un conjunto de datos que pueden ser clasificados de acuerdo a un conjunto finito de categorías.

Por ejemplo, si se estudia el comportamiento de una enfermedad tratada con dos medicamentos diferentes, ambos experimentos comparten espacio de resultados y cada conjunto de datos establece una partición de los resultados de acuerdo a las dos categorías, una asociada a cada medicamento. Si suponemos que la hipótesis nula establece que ambos tratamientos tienen la misma efectividad, el comportamiento de la variable estudiada no debería depender de la asignación arbitraria a una u otra categoría.

Por tanto, una vez establecido un estadístico, podríamos estimar su función de distribución o calcular  $p$ -valores mediante el siguiente procedimiento,

- en la muestra de referencia se asignan arbitrariamente categorías,
- se calcula el valor del estadístico en la muestra resultante
- y se utilizan los valores obtenidos para calcular la función de distribución del estadístico bajo la hipótesis nula.

El procedimiento anteriormente descrito es útil en aquellas situaciones en las que las hipótesis en las que se basa el muestreo paramétrico son cuestionables o difíciles de verificar.

### **Metodología de combinación no paramétrica**

Un método adecuado para realizar contrastes de permutación multivariante es la denominada combinación no paramétrica. Supongamos que la hipótesis nula  $H_0$  de un problema puede descomponerse en  $k$  sub-hipótesis  $H_{01}, \dots, H_{0k}$  de manera que  $H_0$  es verdadero si y solo si todas las sub-hipótesis son verdaderas, formalmente

$$H_0 = \bigcap_{i=1}^k H_{0i} \tag{1.1}$$

De manera similar, la hipótesis alternativa  $H_1$  es verdadera si y solo si al menos una de las sub-hipótesis nulas es falsa y en consecuencia, al menos una de las sub-hipótesis alternativas es verdadera, es decir, la hipótesis alternativa se puede escribir como una unión de conjuntos

$$H_1 = \bigcup_{i=1}^k H_{1i}. \quad (1.2)$$

Supongamos que  $T = T(X)$  es un vector de  $k$  estadísticos  $T_i(X)$ . Cada  $T_i(X)$  se utiliza para contrastar  $H_{0i}$  frente a  $H_{1i}$  y sin pérdida de generalidad supondremos que  $H_{0i}$  se rechaza para valores grandes de  $T_i(X)$ .

Asumiendo, como es habitual, que cada fila de la tabla de datos corresponde a una unidad estadística y considerando que se trata de un contraste para muestras independientes de observaciones, veamos los pasos esenciales del método de combinación no paramétrica.

### Método de combinación no paramétrica

1. Calcular el vector de valores observados de  $T$  :  $T_{obs} = [T_1(X), \dots, T_k(X)]'$   
 $= [T_{1(0)}, \dots, T_{k(0)}]'$ .
2. Considerar una permutación de las filas del conjunto de datos y calcular los valores correspondientes de los estadísticos:  $T_{(1)}^* = [T_1(X_{(1)}^*), \dots, T_k(X_{(1)}^*)]'$ .
3. Llevar a cabo  $B$  repeticiones independientes del paso (2) y obtener  $T_{(b)}^* = [T_{1(b)}^*, \dots, T_{k(b)}^*]'$ ,  $b = 1, \dots, B$ .
4. Para cada  $i$  llevar a cabo una estimación de la función de nivel de significación  $Pr\{T_i^* \geq z\}$  :

$$\hat{L}_i(z) = \frac{1}{B+1} \sum_r \mathbb{I}[T_{i(r)}^* \geq z], i = 1, \dots, k$$

- .
5. Para cada  $b$  calcular  $\lambda_{i(b)}^* = \hat{L}_i(T_{i(b)}^*)$ ,  $b = 1, \dots, B$  y  $\lambda_{i(0)} = \hat{L}_i(T_{i(0)})$ ,  $i = 1, \dots, k$ .
  6. Para cada  $b$  calcular los valores  $T_{(b)}^{||*} = \psi(\lambda_{1(b)}^*, \dots, \lambda_{k(b)}^*)$  y  $T_{(0)}^{||} = \psi(\lambda_{1(0)}, \dots, \lambda_{k(0)})$  usando una función de agregación adecuada  $\psi$ . Las funciones de agregación que usaremos pueden variar pero todas son no crecientes componente a componente.
  7. Calcular el estimador del  $p$ -valor del contraste como

$$\lambda^{||} = \frac{1}{B} \sum_b \mathbb{I}[T_{(b)}^{||*} \geq T_{(0)}^{||}].$$

La decisión final debe basarse en  $\lambda^{||}$  en el sentido de que  $H_0$  debería rechazarse a favor de  $H_1$ , si  $\lambda^{||} \leq \alpha$ . El método de combinación no paramétrica es muy útil para resolver problemas complejos, en particular problemas multivariantes o problemas donde es adecuado un estadístico multivariante. La principal ventaja con respecto a otros métodos paramétricos estándares es que la distribución multivariante del estadístico no necesita ser conocida o estimada, y en particular la estructura de dependencia entre las variables no necesita ser conocida. La dependencia se tiene en cuenta implícitamente a través de la metodología de permutaciones y la aplicación de la función de agregación  $\psi$ . Esta función debe satisfacer las siguientes propiedades:

1. Debe ser no creciente en cada argumento.
2. Debe alcanzar el supremo incluso cuando solo uno de los argumentos tiende a cero.

3. Para cada nivel  $\alpha$  el valor crítico  $T_\alpha^||$  se supone que es finito y estrictamente más pequeño que el valor supremo.

Algunas funciones de agregación adecuadas son:

- La función de agregación de Fisher ómnibus:  $T_F^|| = -2 \sum_i \log(\lambda_i)$ ;
- La función de agregación de Liptak:  $T_L^|| = \sum_i \Phi^{-1}(1 - \lambda_i)$ ;
- La función de agregación de Tippett:  $T_T^|| = \max_i(1 - \lambda_i)$ .

La agregación de Tippett proporciona contrastes útiles cuando una o varias, pero no todas, las sub-hipótesis alternativas son verdaderas; la función de Liptak es apropiada cuando todas las sub-hipótesis alternativas son conjuntamente verdaderas; la solución de Fisher es intermedia entre los dos.

## 1.2. Contrastes univariantes para una muestra

La hipótesis básica que se supone en cualquier contexto de la Inferencia Estadística es que los fenómenos observados pueden estar representados por variables aleatorias con distribuciones desconocidas. Supongamos que la muestra aleatoria observada se ha extraído de una población con función de distribución  $F(x)$  desconocida. Para probar si  $F(x)$  es igual a una función especificada (sin ningún parámetro desconocido) es habitual usar el procedimiento propuesto por Kolmogorov en 1933. Dicho procedimiento se basa en la comparación entre la función de distribución empírica y la distribución especificada. Como se trata de ver si la distribución de probabilidad se adapta adecuadamente a los datos, estos contrastes son conocidos como *contrastos de bondad de ajuste*.

El contraste de Kolmogorov puede considerarse una alternativa al contraste de bondad de ajuste chi-cuadrado, válido para variables categóricas

nominales. Una diferencia importante entre los dos procedimientos es que, para variables continuas, el contraste de Kolmogorov es exacto incluso para tamaños de muestra pequeños, mientras que el contraste chi-cuadrado requiere que  $n$  sea lo suficientemente grande como para que el estadístico bajo la hipótesis nula siga aproximadamente una distribución chi-cuadrado. Esto fue propuesto por Conover en 1999 (véase referencias en [1]).

En algunas aplicaciones el contraste realizado involucra solo unos pocos aspectos de la forma funcional de  $F(x)$ , por lo tanto solo una propiedad será especificada para  $F(x)$  bajo  $H_0$ . Este es el caso de los contrastes de simetría, muy útiles en control estadístico de la calidad en procesos industriales. Para variables continuas, la propiedad de simetría en la distribución respecto al origen equivale a la propiedad:  $F(x) = 1 - F(-x)$ ,  $\forall x$ .

### 1.2.1. Contraste de bondad de ajuste de Kolmogorov

Sea  $X = X_1, \dots, X_n$  una muestra aleatoria de una población con función de distribución desconocida  $F(x)$ . La hipótesis nula es que  $F(x)$  coincide con una distribución conocida  $F_0(x)$  contra la alternativa de que esto no es cierto. El estadístico para esta prueba, propuesto por Kolmogorov en 1933, se basa en la distancia en términos de la norma del supremo entre  $F_0(x)$  y la función de distribución empírica basada en la muestra observada  $X$ . Smirnov en 1939 propuso una extensión de la prueba de Kolmogorov para comparar las distribuciones de dos poblaciones independientes.

Las estadísticas basadas en la distancia en norma del supremo entre  $F_0(x)$  y función de distribución empírica son llamadas estadísticas de tipo Kolmogorov, mientras que las estadísticas similares basadas en la distancia en norma del supremo entre dos funciones de distribución empíricas se llaman estadísticas de tipo Smirnov, (Conover, véase [1]). Formalmente, el

problema consiste en probar la hipótesis nula

$$H_0 : F(x) = F_0(x)$$

contra la alternativa

$$H_1 = F(x) \neq F_0(x).$$

La función de distribución empírica de  $X$  es

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$$

donde  $\mathbb{I}(X_i \leq x)$  tiene el valor 1 si  $X_i \leq x$  y 0 en otro caso. El contraste de Kolmogorov utiliza el siguiente estadístico

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|.$$

Considerando la hipótesis nula  $H_0 : F(x) = F_0(x)$ , podemos hacer uso de la hipótesis alternativa bilateral, que viene dada por  $H_1 : F(x) \neq F_0(x)$  o bien de la hipótesis alternativa unilateral. Suele ser de utilidad una de las siguientes formas de hipótesis alternativa unilateral:  $H_1 : F(x) < F_0(x)$  o  $H_1 : F(x) > F_0(x)$ . Si no existe evidencia para rechazar la hipótesis nula, no existen diferencias en cuanto a la elección de una u otra, mientras que si rechazamos la hipótesis nula a favor de la alternativa debemos elegir una de ellas siempre que sea útil en nuestro estudio, es decir, que nos devuelva resultados de los cuales podamos extraer información y que esta información tenga sentido.

Formalmente, la alternativa unilateral podría ser

$$H_1 : F(x) \leq F_0(x), \forall x \in \mathbb{R}$$

y

$$F(x) < F_0(x), \text{ para algún } x,$$

o bien

$$H_1 : F(x) \geq F_0(x), \forall x \in \mathbb{R}$$

y

$$F(x) > F_0(x), \text{ para algún } x.$$

Por lo tanto, el estadístico adecuado es  $D_n^+ = \max_{x \in \mathbb{R}} [F_0(x) - \hat{F}_n(x)]$  y  $D_n^- = \max_{x \in \mathbb{R}} [\hat{F}_n(x) - F_0(x)]$ , respectivamente. Los contrastes rechazan la hipótesis nula para valores grandes de los estadísticos.

Se puede demostrar que si  $X$  es una muestra aleatoria de una población absolutamente continua con función de distribución  $F_0$ , entonces la distribución del estadístico  $D_n$  no depende de  $F_0$ , sino solo del tamaño muestral  $n$  (Bagdonavicius *et al.* 2011, [2]).

A partir de la definición del estadístico, en el contraste bilateral, la hipótesis  $H_0$  se rechaza con un nivel de significación  $\alpha$  cuando  $D_n > D_\alpha(n)$ , donde  $D_\alpha(n)$  es el valor crítico del estadístico  $D_n$  que es el cuantil  $(1 - \alpha)$  de la distribución bajo la hipótesis nula de  $D_n$ . De manera equivalente, la hipótesis nula se rechaza cuando el  $p$ -valor del contraste es menor que  $\alpha$ . Para los contrastes unilaterales se aplica un procedimiento similar.

El paquete básico de  $R$  incluye la función `ks.test`, la cual lleva a cabo el contraste de Kolmogorov-Smirnov para el caso de una o dos muestras. La presencia de empates en el caso de variables no continuas puede generar advertencias en la salida del programa  $R$ .

## Análisis de las calificaciones de una asignatura del grado en Matemáticas

A lo largo de este trabajo vamos a hacer uso de un conjunto de datos reales correspondientes a las calificaciones parciales o finales en 2 grupos de una asignatura impartida en la facultad de Matemáticas de la Universidad de Sevilla.

En este conjunto de datos se consideran 2 cursos académicos, 2015 – 16 y 2016 – 17 y las calificaciones se encuentran desagregadas por sexo, grupo y curso académico.

A continuación utilizaremos el subconjunto de datos que aparece en la Tabla 1.1 que corresponden a las calificaciones de 58 estudiantes clasificados por grupo y por sexo.

	Grupo A															
<i>Sexo</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>
<i>Nota</i>	7	5	5	5.6	5.3	4	7.3	5	5	6.1	5.5	5.1	4	8.4	7	5
<i>Sexo</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>
<i>Nota</i>	8	5.8	6.4	6.3	5	6	6.4	5	5.5	7.1	2.5	5	5.4	4	5.7	6.1
	Grupo B															
<i>Sexo</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>							
<i>Nota</i>	6.8	8.7	6.5	6.4	7.6	7.5	7.7	6.3	5.6	7.1	8.4	7.8	8.1	8	7.3	7.3
<i>Sexo</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>						
<i>Nota</i>	6.1	8	7.2	9	7.3	6.3	5.1	5.8	6.1	1						

Tabla 1.1: Calificaciones finales por grupo y sexo de alumnos durante el curso académico 2016 – 17.

Calificaciones										
0.5	-0.8	-0.8	-0.4	-0.6	-1.5	0.7	-0.8	-0.8	-0.1	
-0.5	-0.7	-1.5	1.5	0.5	-0.8	1.2	-0.3	0.1	0.1	
-0.8	-0.1	0.1	-0.8	-0.5	0.6	-2.5	-0.8	-0.5	-1.5	
-0.3	-0.1	0.4	1.7	0.2	0.1	0.9	0.9	1.0	0.1	
-0.4	0.6	1.5	1.1	1.3	1.2	0.7	0.7	-0.1	1.2	
0.7	1.9	0.7	0.1	-0.7	-0.3	-0.1	-3.5			

Tabla 1.2: Calificaciones tipificadas.

El objetivo de este experimento es analizar, a un nivel de significación de  $\alpha = 0,01$ , si los resultados obtenidos se pueden considerar normales. Para empezar tipificaremos los datos de la muestra (tabla 1.2) de forma que podamos compararlos con una distribución normal estándar. La hipótesis nula del problema es

$$H_0 : F(x) = \Phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

donde  $F(x)$  es la función de distribución de las calificaciones, representadas mediante una variable aleatoria continua. La función  $\Phi(x)$  especificada en la hipótesis nula se corresponde con la función de distribución de una distribución normal estándar. La hipótesis alternativa es  $H_1 = F(x) \neq \Phi(x)$ .

Como suele ser habitual las calificaciones de la muestra inicial se encuentran en un rango de 0 a 10, debemos tipificar los datos y llevar a cabo el análisis con los nuevos datos obtenidos. Este procedimiento lo haremos mediante  $R$  de la siguiente forma:

```
datos <- read.table("2017AB.txt",header=TRUE)

exnotas=c(7,5,5,5.6,5.3,4,7.3,5,5,6.1,5.5,5.1,4,
          8.4,7,5,8,5.8,6.4,6.3,5,6,6.4,5,5.5,7.1,
          2.5,5,5.4,4,5.7,6.1,6.8,8.7,6.5,6.4,7.6,
          7.5,7.7,6.3,5.6,7.1,8.4,7.8,8.1,8,7.3,7.3,
          6.1,8,7.2,9,7.3,6.3,5.1,5.8,6.1,1)

xbarra = mean(exnotas)
sc = sd(exnotas)
media <- c(rep(xbarra,58))
sc1 <- 1/sc
```

```

cuasivar<-c(rep(1/sc))
diferencia <- exnotas-media
tipif<-diferencia*cuasivar

```

Con los nuevos datos el código *R* para el análisis es el siguiente:

```

plot(ecdf(tipif),xlim=c(-10,10),verticals=TRUE,
     xlab="Calificaciones",main="")
curve(pnorm,from=-10, to=10,add=TRUE,lty="dashed",lwd=2)
ks.test(tipif,"pnorm",alternative="two.sided")

```

y la salida de `ks.test` es

```
One-sample Kolmogorov-Smirnov test
```

```

data:  tipif
D = 0.12566, p-value = 0.319
alternative hypothesis: two-sided

```

Warning message:

```

In ks.test(tipif, "pnorm", alternative = "two.sided") :
  ties should not be present for the Kolmogorov-Smirnov test

```

```

plot(ecdf(tipif),xlim=c(-10,10),verticals=TRUE,xlab=
  ‘‘Calificaciones’’,main=‘‘ ’’) devuelve la función de distribución empíri-
ca de la muestra, y el argumento xlim indica el intervalo de visualización
del eje x. El comando curve(pnorm,from=-10, to=10,add=TRUE,
lty="dashed",lwd=2) nos permite dibujar en el mismo gráfico anterior la
función de distribución de una distribución normal estándar. El primer argu-
mento indica el tipo de distribución de probabilidad y pnorm se corresponde

```

con la distribución  $\mathcal{N}(0,1)$ . El argumento `lty` define el tipo de línea. El gráfico se muestra en la Figura 1.1.

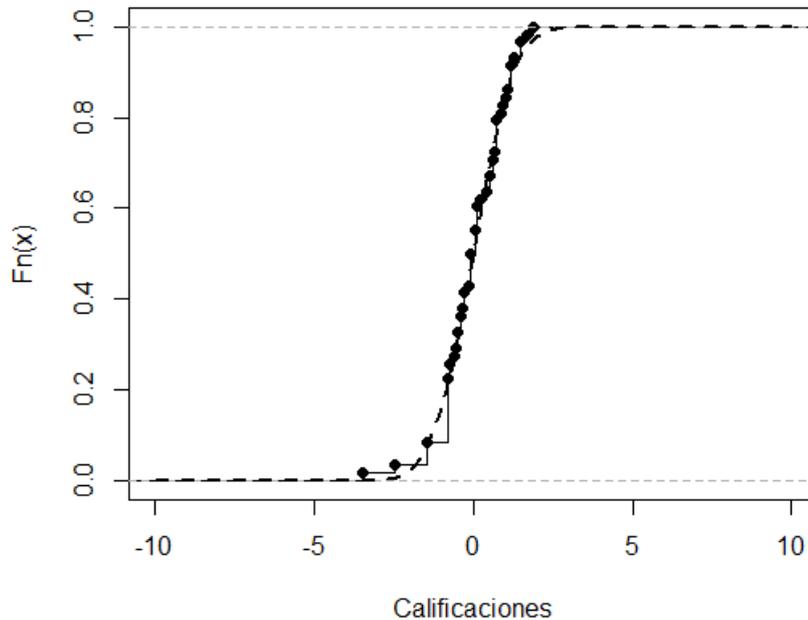


Figura 1.1: Función de distribución empírica (puntos) y función de distribución de  $\mathcal{N}(0,1)$  (línea discontinua).

Gracias a la gráfica resultante podemos observar que estamos en presencia de una nivel aceptable de bondad de ajuste.

El primer argumento de la función `ks.test` indica los datos observados (`tipif` en nuestro ejemplo), el segundo indica la distribución que hemos supuesto en la hipótesis nula y el tercero indica el tipo de alternativa. Por defecto es la alternativa bilateral; la opción "`less`" o "`greater`" se corresponden con las alternativas unilaterales.

Como se puede observar en la salida del programa aparece una adver-

tencia debido a la existencia de empates en las calificaciones. Es habitual que varias calificaciones coincidan debido a que en las actas únicamente se admiten calificaciones con un solo decimal. Por este motivo, el contraste de Kolmogorov-Smirnov pierde un poco de efectividad a la hora de llevar a cabo su análisis. El valor observado del estadístico es  $D = 0,12566$  y su correspondiente  $p$ -valor es 0,319, por lo que no hay evidencia para rechazar la hipótesis nula de que las calificaciones se distribuyen de acuerdo a una normal.

### 1.2.2. Contrastes de permutación univariante para simetría

Como se indicaba anteriormente los contrastes de simetría suelen tener un tratamiento específico, diferente al de bondad de ajuste. En esta sección vamos a ilustrar uno de estos contrastes mediante un ejemplo.

#### **Ejemplo de las calificaciones: continuación**

Para el contraste disponemos de una muestra aleatoria formada por  $n = 55$  alumnos. De estas, las 25 primeras calificaciones corresponden al grupo A, mientras que las 30 calificaciones restantes pertenecen a alumnos del grupo B (Tabla 1.3).

En este caso estudiaremos la simetría de la muestra respecto a una medida de localización. Las medidas de localización que se usan con más frecuencia en análisis estadísticos son la media y la mediana. Haremos el análisis respecto de la mediana en lugar de usar la media dado que la primera divide a la muestra en dos partes *aproximadamente* iguales, lo cual es esencial a la hora de contrastar la simetría. Para ello vamos a calcular las diferencias entre las notas finales obtenidas y la mediana (Tabla 1.4).

		Grupo A													
<i>Sexo</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>
<i>Nota</i>	2.3	7	5	1.8	1.3	5.3	6	5	4.3	5.3	5	7.7	4.3	3.5	5
<i>Sexo</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>					
<i>Nota</i>	5.3	9	5.3	4	3.5	1.5	4.3	5	5.8	5					
		Grupo B													
<i>Sexo</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>
<i>Nota</i>	6	3.3	6	7	5.5	7.3	7	2.8	7.3	5	2.5	5.2	3.5	5.3	6.2
<i>Sexo</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>
<i>Nota</i>	4.5	6.2	2	1	6	3	3.3	6.1	5.3	9	5.4	4.3	5.5	5.5	5.3

Tabla 1.3: Calificaciones finales por grupo y sexo de alumnos durante el curso académico 2015 – 16.

Nuestro interés se centra en probar si la distribución de los resultados obtenidos es simétrica o no, es decir, si existe un valor central, la mediana, respecto de la cual las calificaciones presentan un comportamiento similar a un lado y al otro.

Antes de llevar a cabo el contraste, podemos realizar un estudio cualitativo para la simetría. Para ello, basta observar las Figuras 1.2 y 1.3.

La primera de estas figuras representa un diagrama de *caja y bigotes*, muy utilizado en estadística descriptiva y que ofrece una visión de la simetría, o no, de los datos así como de las observaciones que pueden considerarse extrañas o fuera de rango (outliers). En nuestro caso podemos observar la aparente simetría de los datos bajo estudio.

La segunda figura representa un histograma de frecuencias, en nuestro caso de aspecto campaniforme, que confirmaría la impresión obtenida tras el primero de los gráficos.

El análisis cualitativo que ofrecen las representaciones gráficas permite orientarnos en cuanto al proceso de construcción de contrastes. Mediante éstos, obtendremos los datos cuantitativos necesarios para poder comparar el análisis de diferentes conjuntos de datos entre sí.

Para profundizar en el análisis, vamos a calcular el  $p$ -valor aplicando un

contraste de permutación.

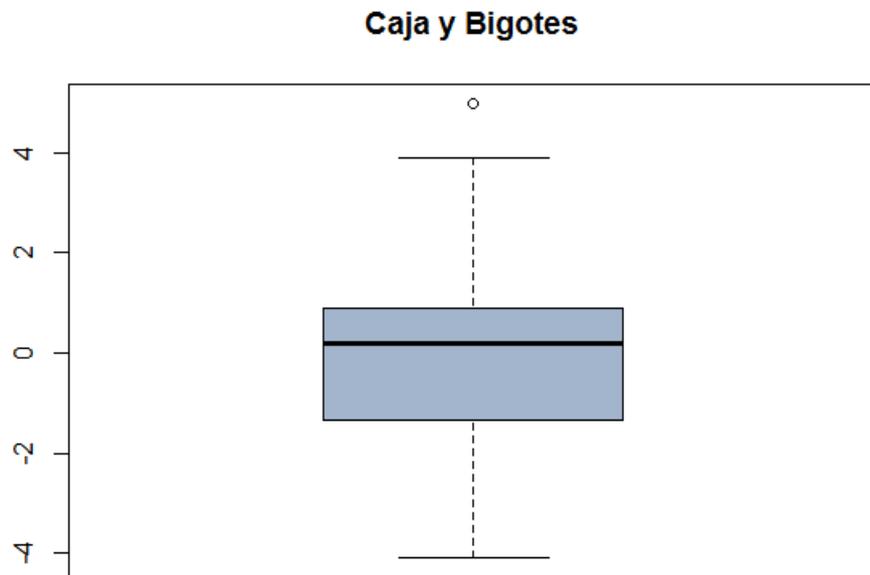


Figura 1.2: Caja y Bigotes.

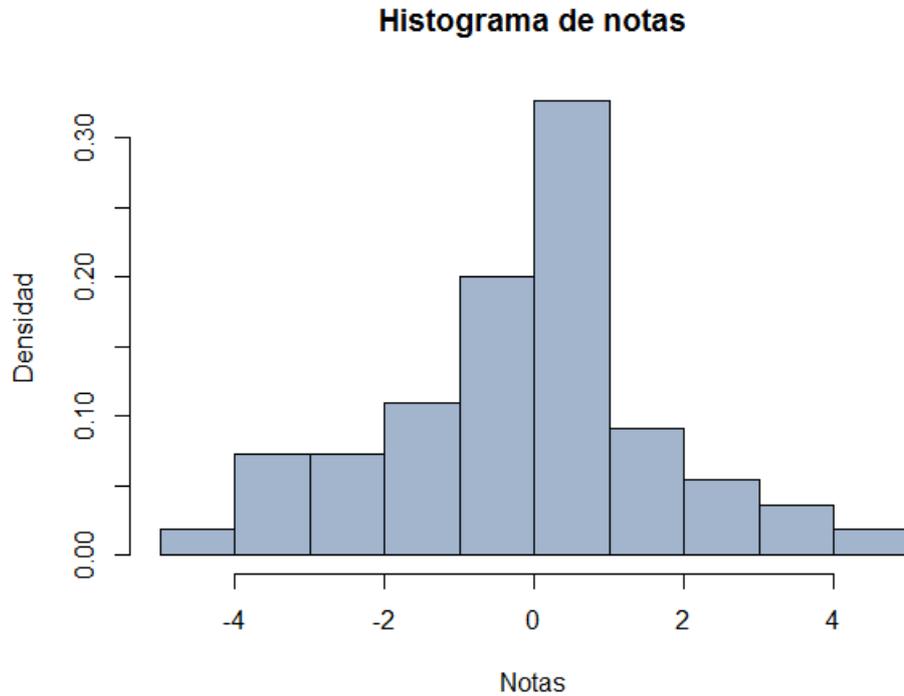


Figura 1.3: Histograma de calificaciones.

En general, dada  $F$  la función de distribución desconocida de una variable continua  $X$  y  $\{X_1, \dots, X_n\}$  una muestra aleatoria de dicha variable, el contraste de simetría de  $F$  alrededor del origen viene dado por las siguientes hipótesis

$$H_0 : F(x) = 1 - F(-x)$$

y

$$H_1 : F(x) \neq 1 - F(-x).$$

En el contraste de la hipótesis podemos considerar los datos como diferencias respecto del punto de simetría (Pesarin y Salmaso, 2010, [7]).

Calificaciones										
-2,8	1,9	-0,1	-3,3	-3,8	0,2	0,9	-0,1	-0,8	0,2	-0,1
2,6	-0,8	-1,6	-0,1	0,2	3,9	0,2	-1,1	-1,6	-3,6	-0,8
-0,1	0,7	-0,1	0,9	-1,8	0,9	1,9	0,4	2,2	1,9	-2,3
2,2	-0,1	-2,6	0,1	-1,6	0,2	1,1	-0,6	1,1	-3,1	-4,1
0,9	-2,1	-1,8	1	0,2	3,9	0,3	-0,8	0,4	0,4	0,2

Tabla 1.4: Diferencias entre la calificaciones finales y la mediana de la muestra.

En este caso, bajo la hipótesis nula, el signo de cada diferencia se considera como si se asignara aleatoriamente con probabilidad  $\frac{1}{2}$ , es decir, tenemos dos categorías, la de observaciones que están por encima de la mediana y las que están por debajo. Dichas categorías se pueden cambiar bajo la hipótesis nula obteniéndose una nueva muestra que tendría el mismo comportamiento distribucional.

Por lo tanto, una forma de construir el contraste es considerar el estadístico dado por el valor absoluto de la suma de los valores de muestra:

$$T = \left| \sum_{i=1}^n X_i \right|.$$

Su distribución,  $F_T(t|X)$ , condicionada a un conjunto dado de valores observados  $X_1, \dots, X_n$ , se obtiene bajo la suposición de que  $H_0$  es cierto al tomar aleatoriamente el signo del conjunto de datos de las diferencias. Para ello, consideramos  $B$  muestras permutadas que vienen dadas por  $X^*$  obtenidas atribuyendo aleatoriamente el signo  $+$  o  $-$  a cada  $X_i$  y calculando para cada muestra permutada el correspondiente valor  $T^*$  de  $T$ , podemos estimar la distribución nula de  $T^*$  de acuerdo con la distribución de permutación y calcular el  $p$ -valor del contraste. El código *R* para el problema del estudio de simetría de las calificaciones finales obtenidas es el siguiente:

```

> source("t2p.r")
> source("permsign.r")
> exnotas=c(-2.8,1.9,-0.1,-3.3,-3.8,0.2,0.9,-0.1,-0.8,0.2,
-0.1,2.6,-0.8,-1.6,5,0.2,3.9,0.2,-1.1,-1.6,-3.6,-0.8,-0.1,
0.7,-0.1, 0.9,-1.8,0.9,1.9,0.4,2.2,1.9,-2.3,2.2,-0.1,-2.6,
0.1,-1.6,0.2, 1.1,-0.6,1.1,-3.1,-4.1,0.9,-2.1,-1.8,1,0.2,
3.9,0.3,-0.8,0.4,0.4,0.2)
> exnotas=array(exnotas,dim=c(length(exnotas),1))
> perm.sign(x=exnotas,B=10000)

```

La salida es

```

#Permutation Test for Symmetry
$p.value
[1] 0.6988301

```

Los archivos fuentes para la ejecución del programa son "t2p.r" y "permsign.r" y se pueden cargar con el comando `source`. Estas fuentes han sido descargadas de la página web <https://www.wiley.com> correspondiente al libro [1].

El primero incluye el código para el cálculo de la función de nivel de significación y los  $p$ -valores; el último contiene el código para la computación del contraste estadístico. Esto se lleva a cabo con el comando `perm.sign` el cual calcula el estadístico  $T$  y su distribución, mediante permutaciones aleatorias de los signos de los datos observados. Para probar la simetría de la muestra, la función consta de dos argumentos: la matriz de los datos observados y el número  $B$  de permutaciones de signos aleatorios. El conjunto de datos (`exnotas`) debe ser configurado como una matriz mediante el comando `array`. Para la ejecución del programa tomaremos  $B = 10000$

permutaciones.

La función `perm.sign` permite hacer uso de `t2p` la cual calcula el  $p$ -valor como la frecuencia del conjunto de posibles repartos de signos que hacen que el estadístico  $T$  sea mayor que los valores observados.

Es interesante señalar que el  $p$ -valor resultante se obtiene a través de permutaciones aleatorias, computacionalmente, tomaremos una muestra aleatoria de  $B$  permutaciones de signo, luego realizando el contraste varias veces, especialmente para valores pequeños de  $B$ , se obtienen diferentes distribuciones empíricas del estadístico y, por lo tanto, diferentes  $p$ -valores. Esto podría llevar a que en determinadas ocasiones se aceptase la hipótesis nula, mientras que en otras se rechazase.

Para evitar este efecto aleatorizado en la decisión tomada respecto a la simetría, o no, de la variable aleatoria, deberíamos de aumentar el tamaño de  $B$ . Esto sería útil para el caso en el que el tamaño de la muestra considerada fuese pequeño, pero si nos encontramos con un tamaño de muestra grande, habría que ver si esto compensaría, pues el tiempo de ejecución del programa sería mayor. Para evitar este hecho, podemos especificar una semilla para asociarla con el conjunto de permutaciones muestreadas y usar esta semilla para identificar y recuperar el mismo conjunto de permutaciones cuando sea necesario. Para ello vamos a usar la función `set.seed(seed)`. Si volvemos a repetir el análisis sin especificar la semilla nosotros obtendremos un  $p$ -valor diferentes al anterior (aunque similar, debido a que  $B$  lo hemos elegido muy grande):

```
Permutation Test for Symmetry
```

```
$p.value
```

```
[1] 0.6984302
```

Si volvemos a hacer los cálculos, pero esta vez fijando la semilla inicial,

digamos por ejemplo que la semilla es 1234, el  $p$ -valor cambia ligeramente.

```
set.seed(1234)
perm.sign(x=exnotas,B=10000)
```

La salida es

```
Permutation Test for Symmetry
$p.value
[1] 0.6937306
```

A partir del momento en el que la semilla se ha fijado, cada vez que ejecutemos el programa, este siempre devolverá el  $p$ -valor 0,6937306. En este ejemplo, el  $p$ -valor resultante nos lleva a pensar que no hay evidencia de rechazo de  $H_0$ . Por lo tanto, podemos decir que la distribución de calificaciones finales obtenidas en los grupos A y B conjuntamente es simétrica respecto de su mediana. A partir de este resultado podemos concluir que el examen de convocatoria final de esta asignatura estuvo bien calibrado en cuanto a la medición del nivel de conocimiento de los alumnos. De forma más clara, el valor de la mediana no debería de ser un umbral a partir del cual varía el comportamiento que tienen las calificaciones de los alumnos, es decir, debe de haber la misma dispersión en ambos lados de la mediana.

### 1.3. Contrastes multivariantes de una sola muestra

Consideremos el caso de datos multivariantes. Sea  $X$  un conjunto de datos multivariantes de una muestra de tamaño  $n$  y supongamos que la variable es  $q$ -dimensional. Sea  $X = \{X_{ih}, i = 1, \dots, n, h = 1, \dots, q\}$  el conjunto de

datos, donde  $X_{ih}$  denota la  $i$ -ésima observación de la  $h$ -ésima variable. Suponemos que cada una de las  $n$  observaciones  $q$ -dimensionales  $\{X_{i1}, \dots, X_{iq}\}$  proviene de una población con función de distribución marginal  $F_i(x; \theta)$ , con  $i = 1, \dots, n$ ,  $x \in \mathcal{R}^q$  y  $\theta = (\theta_1, \dots, \theta_q)'$  es un parámetro genérico de localización. En esta sección vamos a considerar contrastes de una sola muestra con respecto al parámetro de localización  $\theta$  y una extensión multivariante del contraste de simetría presentado en la sección anterior.

### 1.3.1. Contrastes basados en rangos multivariantes para tendencia central

La variable aleatoria  $Z$  que toma valores en  $\mathcal{R}^q$  se dice que es diagonalmente simétrica respecto del 0 (vector de ceros  $q$ -dimensional) si tanto  $Z$  como  $-Z$  tienen la misma función de distribución  $F(z)$ ,  $z \in \mathcal{R}^q$ . Para funciones de distribución absolutamente continuas con función de densidad  $f(z)$  la simetría diagonal puede representarse por  $f(z) = f(-z)$ ,  $\forall z \in \mathcal{R}^q$ , esto fue propuesto por Puri y Sen en el año 1971 (véase referencias en [1]). Consideremos el problema de localización multivariante en el que la muestra de  $n$  vectores de dimensión  $q$  tiene funciones de distribución  $F_1(x; \theta), \dots, F_n(x; \theta)$  independientes y diagonalmente simétricas respecto de  $\theta$ .

Queremos probar si el parámetro de localización es nulo, es decir  $H_0 : \theta = 0$  contra  $H_1 : \theta \neq 0$ . Obsérvese que no se ha impuesto la condición habitual de idéntica distribución  $F_1(x; \theta) = \dots = F_n(x; \theta)$  en la muestra multivariante.

Para el caso general en que las componentes marginales de las variables no tienen porqué ser independientes consideremos la siguiente transformación de  $X : g(X) = \{s_i X_{ih}; i = 1, \dots, n; h = 1, \dots, q\}$  donde  $s_i = +1$  o  $s_i = -1$ . El número posible de vectores de signos  $s = (s_1, \dots, s_n)'$  es  $2^n$ . Bajo

$H_0$  la probabilidad de cada permutación es la misma, por lo tanto, podemos obtener un contraste libre de distribución para esta situación.

Consideremos ahora diferentes contrastes basados en rangos multivariantes. Sea la matriz  $n \times q$ ,  $R = [R_{ih}]$  cuyo elemento genérico  $R_{ih}$  representa el rango de  $|X_{ih}|$  en el conjunto de valores  $\{|X_{1h}|, \dots, |X_{nh}|\}$ . No se admiten empates gracias a la continuidad. Para cada variable, es decir, para cada columna de la matriz  $R$  reemplazamos los rangos por sus calibradores (*scores*)  $E^{(h)}(R_{ih})$ . Para cada variable marginal consideremos los estadísticos de orden de rango

$$T^{(h)} = \sum_{i=1}^n E^{(h)}(R_{ih})c_{ih},$$

para  $h = 1, \dots, q$ .

Los pesos  $c_{ih}$  son los signos de los valores  $X_{ih}$ , esto es,  $c_{ih} = +1$  si  $X_{ih} > 0$  y  $c_{ih} = -1$  si  $X_{ih} < 0$ . Denotemos como  $T$  el vector de estadísticos  $q$ -dimensional  $(T^{(1)}, \dots, T^{(q)})'$ . De acuerdo con la distribución de permutación  $E(T) = 0$  y  $Var(T) = E(TT') = nV$ . La matriz  $V = [v_{jk}]$  es definida positiva con elementos

$$v_{jk} = \frac{1}{n} \sum_{i=1}^n E^{(j)}(R_{ij})E^{(k)}(R_{ik})c_{ij}c_{ik}.$$

El estadístico para el problema de localización multivariante es

$$S = \frac{1}{n} T' V^{-1} T.$$

Por tanto, la hipótesis nula se rechaza para valores grandes de  $S$ . Según los calibradores podemos obtener diferentes contrastes. Algunos ejemplos son:

- $E^{(h)}(R) = 1, h = 1, \dots, q$ : contraste de signo multivariante.
- $E^{(h)}(R) = R, h = 1, \dots, q$ : generalización multivariante del contraste

basado en rango con signo de Wilcoxon de una muestra.

- $E^{(h)}(R)$  es el valor esperado de la  $R$ -ésima observación más pequeña de una muestra de tamaño  $n$  de una distribución chi-cuadrado con 1 grado de libertad  $h = 1, \dots, q$ ; contraste de calibradores normales multivariante de una muestra.

Hay que tener en cuenta que todos los contrastes no se pueden aplicar cuando la hipótesis alternativa es unilateral y requieren la suposición de continuidad de la variable de respuesta multivariante. Una solución a este problema es la extensión multivariante del contraste basado en rango con signo de Wilcoxon de una muestra, pues no requiere de la simetría, lo cual es muy útil cuando esta suposición no es realista o verosímil. Además, esta es la única solución entre estos contrastes basados en rangos cuando solo se observan los signos de las diferencias.

Por otro lado, bajo la hipótesis de simetría y cuando los rangos de la muestra de diferencias pueden ser determinados, el contraste de rangos con signos de Wilcoxon puede decirse que es superior puesto que usa más información que el contraste de signos lo que lo hace más potente bajo  $H_1$ , es decir, rechaza con mayor probabilidad cuando  $H_0$  no es verdadero. El contraste basado en el calibrador normal es menos flexible y es preferible solo en problemas específicos donde se reemplaza el rango ordinario por los calibradores normales relacionados.

### **Ejemplo de las calificaciones: continuación**

Vamos a considerar una muestra de  $n = 20$  alumnos correspondientes a uno de los grupos correspondientes a el curso académico 2015–16. Para llevar a cabo el análisis vamos a tomar las calificaciones obtenidas en la prueba

parcial, dividida en teoría y problemas. Las calificaciones se encuentran en una escala desde 0 a 10, donde el valor 5 se corresponde con el valor mínimo que debe alcanzarse para aprobar un examen.

El objetivo del análisis, por lo tanto, será estudiar si la media de las calificaciones en ambas partes es igual a 5 o no a un nivel de significación de  $\alpha = 0,05$ .

En lugar de tomar las calificaciones, tomaremos su valor menos 5, por lo que los nuevos valores se encuentran comprendidos entre  $-5$  y  $5$  y el objetivo será estudiar si la media de las nuevas calificaciones en teoría y problemas es igual a 5 o no. La muestra de datos viene representada en la tabla 1.5.

Teoría									
-2	0,5	-1,5	-5	0,5	1	-1,5	1,5	0,5	2,5
-1	0	0,5	-0,5	-1,5	-3	0,5	1	3	1
Problemas									
-3,5	-0,5	-5	-2,5	0	-1	0	-1	-0,5	-4
-2	0,5	0	-1,5	-1,5	-4	-2	-1	-1,5	-1

Tabla 1.5: Diferencias entre calificaciones y el valor 5.

Gracias al paquete ICSNP podemos hacer uso de una función  $R$  para el desarrollo de este contraste. Para llevar a cabo el análisis, los comandos utilizados son los siguientes:

```
library(ICSNP)
X1=c(3,5.5,3.5,0,5.5,6,3.5,6.5,5.5,7.5,4,5,5.5,4.5,3.5,2,5.5,6,8,6)
X2=c(1.5,4.5,0,2.5,5,4,5,4,4.5,1,3,5.5,5,3.5,3.5,1,3,4,3.5,4)
Media=rep(5,20)
Te=X1-Media
Pr=X2-Media
datos=cbind(Te,Pr)
rank.ctest(X=datos,Y=NULL,mu=NULL,scores="rank")
```

El comando `library(ICSNP)` es necesario para cargar el paquete ICSNP. Antes de instalar ICSNP, hay que descargar los paquetes `mvtnorm` y `ICS`. El comando `rank.ctest(X=datos,Y=NULL,mu=NULL,scores='rank')` lleva a cabo el contraste. El comando requiere de un marco de datos numéricos o matriz de datos ( $X$ ). El valor por defecto para el segundo argumento  $Y$  es igual a `NULL`, por lo tanto, se realiza un contraste para una muestra. El argumento `mu` es un vector que indica el valor del parámetro de localización bajo la hipótesis nula. El valor por defecto es `NULL`, que representa el origen, por lo que para este problema no se necesita especificarlo. El argumento `scores` indica la función de calibración que queremos aplicar, y `scores = 'rank'` se elige para realizar un contraste de rango signado.

La salida es:

```
Marginal One Sample Signed Rank Test
```

```
data: datos
```

```
T = 13.527, df = 2, p-value = 0.001155
```

```
alternative hypothesis: true location is not equal to c(0,0)
```

El valor observado del estadístico del contraste es 13,527 y el  $p$ -valor es igual a 0,001155. Dado que es menor que 0,05, esto conduce al rechazo de la hipótesis nula a favor de la alternativa de que el vector de medias es distinto de  $(5, 5)'$ . Por lo tanto, la media en teoría y en problemas es distinta del valor 5.

### 1.3.2. Contrastes de permutación multivariante para simetría

Vamos a presentar ahora la extensión multivariante del procedimiento de permutación para el contraste de simetría. En este nuevo problema,

la distribución desconocida que estudiamos es multivariante. El interés se centra en la simetría de las distribuciones marginales pero sin descuidar la naturaleza multivariante del problema y la posible dependencia entre las variables marginales. Formalmente  $f(x)$ , con  $x \in \mathcal{R}^q$  denota la función de probabilidad conjunta (para variables discretas) o función de densidad (para variables continuas), y  $f_i(x)$  la función marginal de la componente  $i$ -ésima de la variable. La hipótesis nula del problema es  $H_0 : f(x) = f(-x)$  y la alternativa es  $H_1 : f(x) \neq f(-x)$ . En la práctica, esto es equivalente a probar la simetría diagonal de la distribución multivariante.

Para resolver este problema, utilizamos la metodología de combinación no paramétrica (Pesarin y Salmaso, 2010,[7]). Consideramos las hipótesis nula y alternativa dadas por (1.1) y (1.2). Según la hipótesis nula de simetría, la muestra obtenida tras permutar los signos de los vectores tiene el mismo comportamiento distribucional que la original, es decir, para cada vector  $q$ -dimensional de observaciones  $(X_{i1}, \dots, X_{iq})', i = 1, \dots, n$ , los signos se pueden permutar porque la probabilidad (para variables discretas) o la densidad (para variables continuas) de observar  $(X_{i1}, \dots, X_{iq})'$  y  $(-X_{i1}, \dots, -X_{iq})'$  es igual. Bajo esta premisa, cada permutación de signos proporciona una nueva muestra con la que estimar la distribución del estadístico bajo la hipótesis nula.

### **Ejemplo de las calificaciones: continuación**

En este caso analizaremos los resultados obtenidos en la prueba intermedia, dividida en teoría y problemas, de alumnos separados por grupo y sexo durante el curso académico 2015 – 16 (Tabla 1.6).

	Grupo A									
<i>Sexo</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>H</i>
<i>Teoria</i>	3	5.5	3.5	0	5.5	6	3.5	6.5	5.5	7.5
<i>Problemas</i>	1.5	4.5	0	2.5	5	4	5	4	4.5	1
<i>Sexo</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>
<i>Teoria</i>	4	5	5.5	4.5	3.5	2	5.5	6	8	6
<i>Problemas</i>	3	5.5	5	3.5	3.5	1	3	4	3.5	4
	Grupo B									
<i>Sexo</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>H</i>	<i>H</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>
<i>Teoria</i>	4	7	9	8	4.5	6	4	5.5	6.5	8
<i>Problemas</i>	2.5	5	5	3	1	4	1	1.5	4	1
<i>Sexo</i>	<i>H</i>	<i>M</i>	<i>H</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>	<i>M</i>		
<i>Teoria</i>	2.5	1	3	3.5	7.5	6	7.5	6		
<i>Problemas</i>	1.5	1	3	3	3	2.5	3.5	4.5		

Tabla 1.6: Calificaciones parciales (teoría y problemas) por sexo y grupo de alumnos durante el curso académico 2015 – 16.

Para su estudio, vamos a tomar una muestra de  $n = 38$  observaciones obtenidas a partir de las diferencias entre las calificaciones en teoría y su mediana, así como las diferencias entre las calificaciones en problemas y su mediana. (Tabla 1.7)

Calificaciones en teoría										
2.5	0.0	2.0	5.5	0.0	-0.5	2.0	-1.0	0.0	-2.0	1.5
0.5	0.0	1.0	2.0	3.5	0.0	-0.5	-2.5	-0.5	1.5	-1.5
-3.5	-2.5	0.1	-0.5	1.5	0.0	-1.0	-2.5	2.0	4.5	2.5
2.0	-2.0	-0.5	-2.0	-0.5						
Calificaciones en problemas										
1.8	-1.2	3.2	0.8	-1.8	-0.8	-1.8	-0.8	-1.2	2.2	0.2
-2.2	-1.8	-0.2	-0.2	2.2	0.2	-0.8	-0.2	-0.8	0.8	-1.8
-1.8	0.2	2.2	-0.8	2.2	1.8	-0.8	2.2	1.8	2.2	0.2
0.2	0.2	0.8	-0.2	-1.2						

Tabla 1.7: Diferencias de las calificaciones en teoría y su mediana y diferencias entre las calificaciones en problemas y su mediana.

Los comandos de *R* para realizar el contraste son:

```
source("t2p.r")
```

```
source("comb.r")
```

```

source("permsign.r")
diffteoria=c(2.5,0.0,2.0,5.5,0.0,-0.5,2.0,-1.0,0.0,-2.0,1.5,
             0.5,0.0,1.0,2.0,3.5,0.0,-0.5,-2.5,-0.5,1.5,-1.5,
             -3.5,-2.5,0.1,-0.5,1.5,0.0,-1.0,-2.5,2.0,4.5,2.5,
             2.0,-2.0,-0.5,-2.0,-0.5)
diffproblem=c(1.8,-1.2,3.2,0.8,-1.8,-0.8,-1.8,-0.8,-1.2,2.2,0.2,
             -2.2,-1.8,-0.2,-0.2,2.2,0.2,-0.8,-0.2,-0.8,0.8,
             -1.8,-1.8,0.2,2.2,-0.8,2.2,1.8,-0.8,2.2,1.8,2.2,0.2,
             0.2, 0.2,0.8,-0.2,-1.2)

x = array(c(diffteoria,diffproblem),dim=c(length(diffteoria),2))
perm.sign(x,fun="F",B=10000)

```

El comando `source()` permite hacer uso de las funciones incluidas en los archivos `"t2p.r"`, `''comb.r"` y `"permisign.r"`. El código para realizar el contraste es `"permsign.r"`; el código para la computación del nivel de significación está incluido en `"t2p.2"` y el código para aplicación de la metodología de combinación no paramétrica está en `''comb.r"`. La función `"perm.sign"` requiere tres argumentos para el caso multivariante. El primer argumento es el conjunto de datos  $x$  (matriz de datos  $n \times q$ ) definida como un array. El argumento (`fun`) es la función de combinación: `"F"`, `"L"` y `"T"` que representan las combinaciones de Fisher, Liptak y Tippett, respectivamente. El tercer argumento representa el número de permutaciones para estimar la distribución de permutación nula del estadístico.

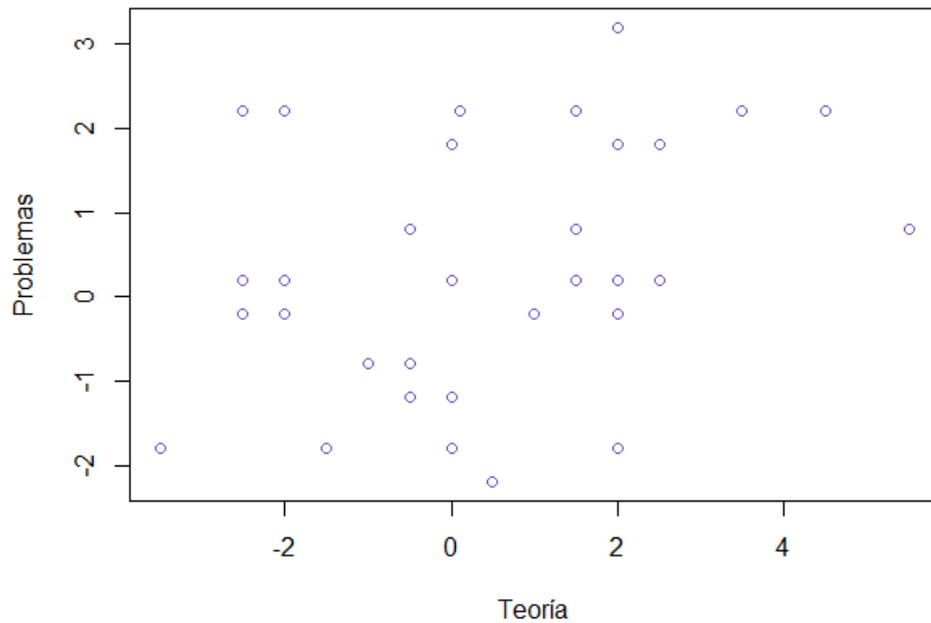


Figura 1.4: Nube de puntos para las diferencias entre las calificaciones medianas y las obtenidas en teoría y problemas.

A partir de la Figura 1.4, podemos llevar a cabo un estudio cualitativo de cual sería el resultado obtenido en el contraste estudiado. Se trata de una nube de puntos donde el eje de abscisas se corresponde con las diferencias entre las calificaciones en teoría y su mediana. Por otra parte, en el eje de ordenadas vienen representadas las diferencias entre las calificaciones en problemas y su mediana. En ella se puede observar que existe una cierta simetría con respecto a un punto de la gráfica, es decir, tomando un punto a una distancia concreta de él, podríamos encontrar otro punto simétrico al anterior con respecto al punto de simetría, todo ello de una forma aproximada. Para dar un argumento más sólido que el anterior, debemos llevar a

cabo un estudio cuantitativo de este, para obtener el  $p$ -valor del contraste.

La salida, usando la función de combinación de Fisher y 10000 permutaciones, es:

```
Multivariate Permutation Test for Symmetry
```

```
Combination Function: Fisher
```

```
$p.value
```

```
[1] 0.5340466
```

El  $p$ -valor del contraste global de simetría es 0,5340466, por lo tanto, no hay evidencia de rechazo de la hipótesis nula de simetría de la distribución multivariante. Esto nos permite deducir que por cada individuo que supera satisfactoriamente ambas partes (teoría y problemas) el valor de la nota mediana, existe otro que falla en superarlas en ambas partes. Por otro lado, por cada individuo que destaca en teoría (problemas), existe otro que destaca en problemas (teoría).

Finalmente, no hemos de olvidar la gran diferencia entre la simetría multivariante y la simetría de las componentes marginales, es decir, podemos encontrarnos ante multitud de situaciones en las que la muestra multivariante analizada carece de simetría, mientras que sus componentes marginales si lo son. Para ver este hecho de una forma más clara analizaremos el ejemplo

que se muestra a continuación.

$$M = \begin{pmatrix} & -5 & -2 & -1 & 1 & 2 & 5 \\ -5 & & & & & & * \\ -2 & * & & & & & \\ -1 & & & & & & * \\ 1 & & & * & & & \\ 2 & & * & & & & \\ 5 & & & & & * & \end{pmatrix}$$

La matriz anterior, representa una muestra multivariante cuyas marginales se corresponden con el eje de abscisas (columnas) y el eje de ordenadas (filas). Cabe destacar que los puntos  $(x, y)$  que deseamos representar, poseen un valor (\*) en la matriz. Podemos observar que la matriz  $M$  no es simétrica, pues  $M \neq M^t$ , donde  $M^t$  denota la matriz traspuesta. Por lo tanto, no existe simetría multivariante, mientras que si proyectamos los valores sobre el eje de abscisas observamos que a cada valor de  $x$  le corresponde un valor (\*). Lo mismo sucede con el eje de ordenadas, por lo que podemos deducir que existe simetría univariante para ambas componentes.

#### 1.4. Contrastes univariantes de dos muestras

En esta sección estudiaremos el problema de comparar muestras independientes de una variable numérica medida en dos poblaciones diferentes. Los datos consisten en  $n = n_1 + n_2$  observaciones, donde  $n_j$  denota el tamaño de la  $j$ -ésima muestra ( $j = 1, 2$ ). Consideramos el problema más típico de dos muestras: la comparación de tendencias centrales.

Sea  $X_1 = \{X_{1i}, i = 1, \dots, n_1\}$  la muestra de datos de la primera pobla-

ción y  $X_2 = \{X_{2i}, i = 1, \dots, n_2\}$  la muestra de datos de la segunda población. Los contrastes no paramétricos más importantes para el problema de localización de dos muestras son el contraste basado en rangos propuesto por Wilcoxon y el contraste basado en permutaciones. Estos procedimientos se describen a continuación.

#### 1.4.1. Contraste de Wilcoxon (Mann-Whitney)

El contraste de la suma de rangos de Wilcoxon se puede aplicar en presencia de muestras aleatorias independientes cuando la suposición de poblaciones normales no se cumple y la prueba  $t$ -paramétrica no se puede aplicar. Las suposiciones de esta prueba son las siguientes:

1. Los datos son realizaciones de variables aleatorias independientes y continuas y la independencia se asume tanto entre ambas muestras como dentro de las muestras.
2. Las variables aleatorias que generan datos de la misma muestra se encuentran idénticamente distribuidas.

Además, el contraste de Wilcoxon también se puede aplicar en presencia de datos categóricos ordenados. El propósito de este contraste es el de comparar las tendencias centrales de las dos muestras, para probar si los parámetros de localización de las poblaciones son iguales o no (contraste bilateral) o para probar si uno de ellos es mayor que el otro (contraste unilateral).

Una forma intuitiva de tratar este problema es combinar ambas muestras en un solo conjunto ordenado y asignar rangos crecientes a los valores muestrales, sin importar la población de la que proviene cada valor. A continuación se tomaría el estadístico definido a través de la suma de los rangos

asignados a la primera muestra. Los valores extremos del estadístico serían evidencias empíricas en favor de la hipótesis alternativa. La región de rechazo será determinada en función del tipo de alternativa, unilateral o bilateral, tal y como propuso Conover en 1999 (véase referencias en [1]).

Para construir formalmente el contraste, supongamos que las funciones de distribución de las poblaciones comparadas son  $F_1(x)$  y  $F_2(x)$  y sean  $X_{ji}$  generadas por la variable aleatoria  $Z_{ji}$ . Hay varias formas de especificar la prueba unilateral. Dos de ellas son el *modelo de efectos aleatorios* y el *modelo de efectos fijos*. En el primero de ellos se asume que  $Z_{1i} = \mu + \Delta_{1i} + \epsilon_{1i}$  y  $Z_{2i} = \mu + \epsilon_{2i}$ ,  $i = 1, \dots, n_j$ , donde  $\mu$  es una constante,  $\epsilon_{ji}$  ( $j = 1, 2$ ) son errores aleatorios intercambiables, con parámetro de localización igual a cero y el parámetro de escala igual a  $\sigma$ , y  $\Delta_{1i}$  son variables aleatorias no negativas. El modelo de efectos fijos es un caso especial del anterior en el que  $\Delta_{1i} = \theta$  con probabilidad 1, siendo  $\theta$  un parámetro constante y desconocido. De acuerdo con el modelo de efectos fijos las varianzas de las dos poblaciones comparadas son iguales (condición de homocedasticidad) y las dos distribuciones pueden diferir solo en la localización.

El contraste unilateral se puede formular como uno de dominancia estocástica de una de las variables respecto de la otra o como un contraste de localización sobre un parámetro de desplazamiento. En otras palabras, la hipótesis de que, por ejemplo, la primera población tiende a tomar valores mayores que la segunda, se puede representar como  $F_1(x) \leq F_2(x)$ ,  $\forall x \in \mathcal{R}$  y  $F_1(X) < F_2(x)$  para algún  $x$ , o dado el parámetro de localización  $\theta$ ,  $F_1(x) = F_2(x - \theta)$  con  $\theta > 0$ . La hipótesis nula del problema es  $H_0 : F_1(x) = F_2(x)$ ,  $\forall x \in \mathcal{R}$ , y la hipótesis alternativa es  $H_1 : F_1(x) \neq F_2(x)$  o, de acuerdo con el modelo de localización, podemos escribir  $H_0 : \theta = 0$  y  $H_1 : \theta > 0$ . Las dos representaciones son equivalentes si se asume el modelo de efectos fijos.

De ahora en adelante en esta subsección vamos a considerar este modelo.

El estadístico del contraste de la suma de rangos de Wilcoxon (equivalente al contraste de Mann-Whitney) se basa en la suma de los rangos de los elementos de la primera muestra

$$W = \sum_{i=1}^{n_1} \mathbb{R}(X_{1i})$$

donde  $\mathbb{R}(X_{1i})$  es el rango (creciente) de  $X_{1i}$  en la unión de las muestras  $\{X_{1i}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}\}$  por lo tanto, es igual a 1 si  $X_{1i}$  es el valor observado más pequeño, a 2 si  $X_{1i}$  es el segundo valor observado más pequeño, y así sucesivamente hasta  $n = n_1 + n_2$  para el valor más grande.

Si la hipótesis nula es verdadera, la suma de los rangos de la primera muestra será similar a la de la segunda muestra, por lo tanto, cuando  $W$  toma valores grandes,  $H_0$  se rechaza a favor de  $H_1$ . En  $H_0$ , la distribución del estadístico  $W$  no depende de parámetros desconocidos, pero depende de los tamaños de muestra  $n_1$  y  $n_2$ , por lo tanto de las propiedades de los rangos obtenemos

$$P\{\mathbb{R}(X_{11}), \dots, \mathbb{R}(X_{1n_1}) = (j_1, \dots, j_{n_1})\} = \frac{n_2!}{n!} \quad (1.3)$$

$\forall (j_1, \dots, j_{n_1})$  obtenido a partir de  $n_1$  elementos diferentes del conjunto  $(1, 2, \dots, n_1 + n_2)$ . El valor mínimo del estadístico  $W$  es  $w_{min} = 1 + \dots + n_1 = n_1(n_1 + 1)/2$  y el valor máximo es  $w_{max} = (n_2 + 1) + (n_2 + 2) + \dots + (n_2 + n_1) = n_1(2n_2 + n_1 + 1)/2$ . Por lo tanto,  $\forall k = w_{min}, \dots, w_{max}$

$$P\{W = k\} = N_k \frac{n_2!}{n!}$$

donde  $N_k$  es la cantidad de vectores  $(j_1, \dots, j_{n_1})$  que satisfacen la condición

$j_1 + \dots + j_{n_1} = k$ . Cuando la hipótesis alternativa es que la primera población toma valores más pequeños que la segunda, la hipótesis nula se rechaza para valores pequeños del estadístico. Finalmente, la hipótesis alternativa bilateral de la desigualdad en la distribución debería rechazarse si  $W \leq c_1$  o  $W \geq c_2$ , donde  $c_1$  y  $c_2$  son el número natural máximo y el número natural mínimo, respectivamente, que verifican las desigualdades,

$$\sum_{k=w_{min}}^{c_1} Pr\{W = k|H_0\} \leq \alpha/2$$

y

$$\sum_{k=c_2}^{w_{max}} P\{W = k|H_0\} \leq \alpha/2.$$

Bajo  $H_0$  las medias y varianzas de la suma de rangos  $W$  son

$$\mathbb{E}(W) = \frac{n_1(n+1)}{2}$$

y

$$\mathbb{V}(X) = Var(W) = \frac{n_1 n_2 (n+1)}{12}.$$

Un resultado importante propuesto por Bagdonavicius *et al.* (2011) [2], muestra que si las distribuciones de probabilidad de las poblaciones son absolutamente continuas, entonces cuando  $n \rightarrow \infty, n_1 \rightarrow p \in (0, 1)$  bajo la hipótesis nula

$$Z_{(n_1, n_2)} = \frac{W - \mathbb{E}(W)}{\sqrt{\mathbb{V}(W)}} \xrightarrow{\mathcal{L}} Z \sim N(0, 1).$$

Por lo tanto, para tamaños grandes de muestra, se puede usar la aproximación normal de la distribución de  $W$ . El contraste de Mann-Whitney se construye para el mismo problema de localización que se ha descrito. Para el contraste unilateral, donde se supone que la primera población toma valores

mayores que la segunda en la hipótesis alternativa, el estadístico es

$$U = \sum_{i=1}^{n_1} \sum_{s=1}^{n_2} D_{is},$$

donde  $D_{is} = \mathbb{I}(X_{1i} > X_{2s}) = 1$  si  $X_{1i} > X_{2s}$  y 0 en otro caso. De una forma similar, se puede derivar el contraste de Mann-Whitney para el contraste unilateral de cola inferior y el contraste bilateral.

El paquete básico de  $R$  contiene la función `wilcox.test`, la cual lleva a cabo el contraste estadístico de suma de rangos de Wilcoxon y calcula el  $p$ -valor para el caso de una o dos muestras. Por defecto, el  $p$ -valor se calcula si el tamaño muestral es inferior a 50 y no hay empates. En otro caso, se puede usar una aproximación normal.

### **Ejemplo de las calificaciones: continuación**

Para llevar a cabo el contraste, vamos a analizar los resultados obtenidos en la parte de teoría y problemas. En la Tabla 1.6 podemos observar los resultados del contraste de dos muestras que provienen de las calificaciones teóricas y las calificaciones prácticas de un grupo de estudiantes. Los tamaños muestrales son  $n_1 = n_2 = 38$ .

Deseamos saber si los alumnos poseen las mismas habilidades para responder cuestiones teóricas que problemas prácticos contra la hipótesis alternativa de que los estudiantes son más hábiles, es decir, obtienen mejores resultados en la parte teórica que en la parte práctica de una prueba. Sean  $X_T$  y  $X_P$  las variables aleatorias que representan el resultado obtenido por un estudiante en la parte de teoría y en la parte de problemas de una prueba, respectivamente. Las hipótesis para el contraste son  $H_0 : X_T \stackrel{d}{=} X_P$

y  $H_1 : X_T >^d X_P$ . El símbolo  $=^d$  denota la igualdad en distribución, que es la igualdad de las funciones de distribución. La notación matemática  $>^d$  denota la dominancia estocástica, que es que la función de distribución de la variable aleatoria  $X_T$  es menor o igual que la función de distribución de la variable  $X_P$  (y la desigualdad estricta es verdadera para algunos subconjuntos de  $\mathcal{R}$ ). En otras palabras, bajo la hipótesis alternativa los resultados de los estudiantes en la parte teórica tienden a ser mejores que los obtenidos en la parte práctica. El nivel de significación es  $\alpha = 0,01$ .

Al observar los Histogramas 1.5 de frecuencia de las dos muestras, parece que la distribución de los resultados de los estudiantes en teoría se desplaza hacia valores mayores que los resultados de los estudiantes en problemas. Para probar si esta conclusión basada en estadística descriptiva puede extenderse a las poblaciones correspondientes, podemos aplicar el contraste de Wilcoxon-Mann-Whitney.

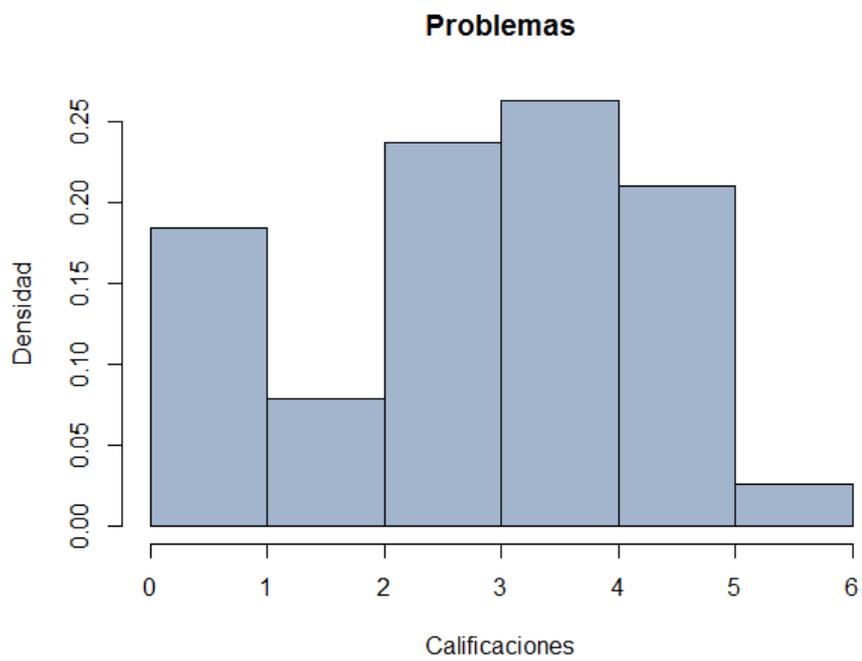
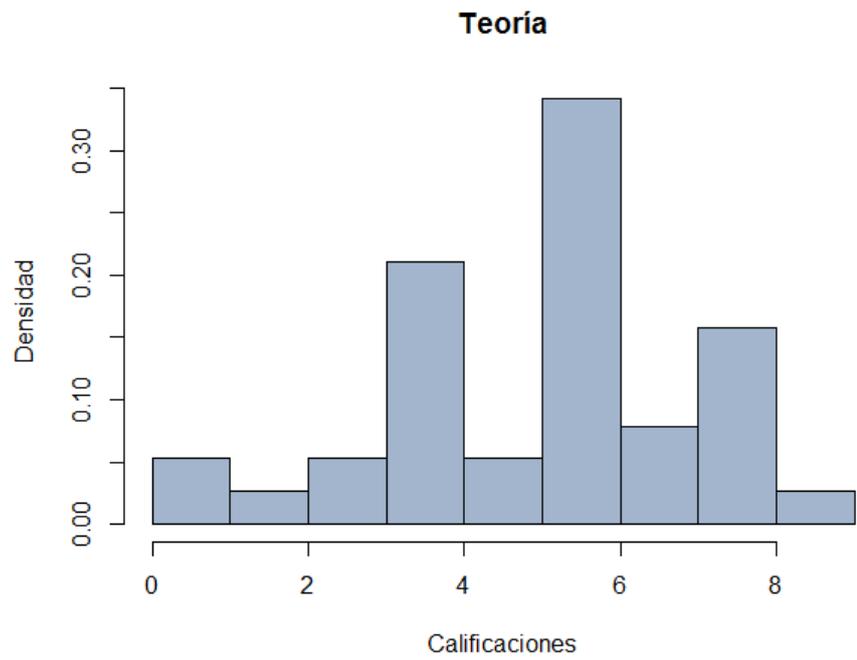


Figura 1.5: Histogramas de calificaciones parciales durante el curso académico 2015 – 16.

El código de *R* para el análisis es el siguiente:

```
teoria=c(3,5.5,3.5,0,5.5,6,3.5,6.5,5.5,7.5,4,5,  
         5.5,4.5,3.5,2,5.5,6,8,6,4,7,9,8,5.4,6,4,  
         5.5,6.5,8,3.5,1,3,3.5,7.5,6,7.5,6)
```

```
problemas=c(1.5,4.5,0,2.5,5,4,5,4,4.5,1,3,5.5,5,  
            3.5,3.5,1,3,4,3.5,4,2.5,5,5,3,1,4,1,  
            1.5,4,1,1.5,1,3,3,3,2.5,3.5,4.5)
```

```
wilcox.test(teoria,problemas,alternative ="greater")
```

Mediante el comando `wilcox.test(x,y,alternative="greater")` podemos llevar a cabo el contraste unilateral de Wilcoxon-Mann-Whitney para comparar  $x$  e  $y$  (vectores de datos de la primera y segunda muestra, respectivamente) para probar la hipótesis de que la primera población tiende a tomar valores mayores que la segunda. Para probar la hipótesis unilateral opuesta, el último argumento debería ser reemplazado por `alternative = "less"`. Para la alternativa bilateral de la desigualdad en la distribución la sintaxis es `alternative= "two.sided"`. La salida después de aplicar *R* al problema considerado es:

```
Wilcoxon rank sum test with continuity correction
```

```
data:  teoria and problemas
```

```
W = 1164.5, p-value = 2.026e-06
```

```
alternative hypothesis: true location shift is greater than 0
```

```
Warning message:
```

```
In wilcox.test.default(teoria, problemas, alternative ="greater"):  
cannot compute exact p-value with ties
```

En primer lugar, cabe destacar la presencia de una advertencia en la salida del programa  $R$ , debido a la presencia de empates, muy frecuente al tratarse de calificaciones. Por lo tanto, el  $p$ -valor calculado será una aproximación, aunque la evidencia empírica es que únicamente se pierde un poco de efectividad en la realización del contraste.

El valor observado del estadístico del contraste es  $W = 1164,5$  y el correspondiente  $p$ -valor (aproximado) es igual a  $2,026e - 06$  que lleva al rechazo de la hipótesis nula a favor de la hipótesis alternativa de que las calificaciones de los alumnos en la parte teórica es mejor que en la parte de problemas. El resultado obtenido demuestra que el examen no se encuentra equilibrado en cuanto a la dificultad del mismo, los estudiantes encuentran mayor dificultad a la hora de desarrollar sus habilidades en la parte práctica que en la parte teórica.

Un problema similar desde el punto de vista estadístico sería el de llevar a cabo un estudio basado en las calificaciones en el grupo de hombres y en el de mujeres de una misma asignatura. A diferencia del estudio anterior, vamos a considerar las calificaciones finales representadas en la Tabla 1.3. Para ello, tomaremos dos muestras. La primera está formada por las calificaciones correspondientes a mujeres, y la segunda está formada por la de los hombres.

El objetivo será estudiar si las calificaciones correspondientes a hombres son mayores que la de las mujeres. Sean las variables  $X^M$  y  $X^H$  las calificaciones obtenidas por mujeres y hombres, respectivamente. La hipótesis nula del problema es  $H_0 : X^M =^d X^H$ , contra la alternativa  $H_1 : X^M <^d X^H$ . Los tamaños muestrales son  $n_1 = 26$  y  $n_2 = 24$ , cuyos datos se encuentran recogidos en la Tabla 1.8.

Calificaciones de mujeres										
2.3	5	5.3	5	4.3	5	3.5	5	5.3	5.3	4
5	5	6	7	5.5	7.3	7.3	3.5	5.3	6.2	4.5
6.2	1	3.3	5.3							
Calificaciones de hombres										
7	1.8	1.3	6	5.3	7.7	4.3	9	3.5	1.5	4.3
5.8	3.3	6	7	2.8	5	2.5	5.2	2	6	3
6.1	5.5									

Tabla 1.8: Calificaciones finales por sexo de alumnos durante el curso académico 2015 – 16.

El código *R* para el contraste es:

```

mujeres=c(2.3,5,5.3,5,4.3,5,3.5,5,5.3,5.3,4,5,5,6,7,
          5.5,7.3,7.3,3.5,5.3,6.2,4.5,6.2,1,3.3,5.3)
hombres=c(7,1.8,1.3,6,5.3,7.7,4.3,9,3.5,1.5,4.3,5.8,
          3.3,6,7,2.8,5,2.5,5.2,2,6,3,6.1,5.5)
wilcox.test(mujeres,hombres,alternative ="less")

```

y la salida es:

```

Wilcoxon rank sum test with continuity correction

data:  mujeres and hombres
W = 332, p-value = 0.6551
alternative hypothesis: true location shift is less than 0

Warning message:
In wilcox.test.default(mujeres, hombres, alternative = "less") :
cannot compute exact p-value with ties

```

En primer lugar, cabe destacar un mensaje de advertencia debido a la existencia de empates en las calificaciones. La razón por la que existen em-

pates en las calificaciones es por el redondeo a la décima con el que son publicadas las calificaciones de acuerdo al formato que deberá ser introducido en las actas de la asignatura. Por este motivo, el contraste perderá un poco de efectividad.

La opción para la hipótesis alternativa es ahora "less". El valor del estadístico es  $W = 332$  y se corresponde con un  $p$ -valor igual a 0,6551, de manera que no hay evidencia para rechazar la hipótesis nula. Esto significa que no existen diferencias por sexo entre las calificaciones, es decir, las calificaciones obtenida son independientes de si el estudiante es hombre o mujer.

#### 1.4.2. Contrastes de permutación para tendencia central

El problema de comparar la tendencia central de dos muestras independientes en presencia de una variable numérica puede tratarse también a través del método basado en permutaciones. Supongamos homocedasticidad en la hipótesis nula y denotamos con  $F_1$  y  $F_2$  las distribuciones continuas comparadas, ambas de la misma familia  $F$ . Consideremos el problema de dominancia estocástica donde en la hipótesis alternativa se supone que la primera población toma mayores valores que la segunda. En otras palabras,  $H_1$  afirma la dominancia estocástica de la primera población sobre la segunda. Se debe tener en cuenta también que  $H_0$  implica la intercambiabilidad de los datos observados con respecto a los grupos, y los datos observados se pueden ver como si se asignaran aleatoriamente a dos grupos, pero provienen de la misma población.

El método de permutaciones no necesita la suposición de que las medias y las varianzas de las variables de respuesta sean finitas. Solo necesita que los parámetros de localización (medias, mediana u otros) sean finitos. A di-

ferencia del contraste de Wilcoxon, el contraste de permutación no requiere la continuidad de las variables de respuesta y puede aplicarse también en presencia de empates sin ninguna corrección o aproximación. Teniendo en cuenta que la transformación de la muestra en sus rangos no es biyectiva, en general no podemos esperar construir estadísticos suficientes. Una transformación de  $X$  es un estadístico suficiente si contiene toda la información necesaria para resolver el problema de inferencia estadística en  $F$ . Por tanto, el contraste de suma de rangos de Wilcoxon puede tener una potencia inferior al contraste basado en permutaciones debido a que el estadístico de este último sí se puede demostrar que es suficiente respecto a la muestra (Pesarin, 2001, véase [1]).

Consideremos  $X = X_1 \uplus X_2$  con  $\uplus$  que denota la concatenación vectorial, de modo que las dos muestras se agrupan en una y los primeros  $n_1$  elementos de  $X$  correspondan a  $X_1$  y los  $n_2$  elementos restantes a  $X_2$ .

Un estadístico de permutación adecuado es  $T^* = \bar{X}_1^* - \bar{X}_2^*$  donde  $\bar{X}_j^* = \sum_{i=1}^{n_j} X_{ji}^*/n_j, j = 1, 2$ , son las medias muestrales de los primeros  $n_1$  elementos y de los  $n_2$  elementos restantes de  $X^*$ , respectivamente, y  $X^*$  es un conjunto de datos permutados, que es el vector obtenido cambiando la posición de los elementos, o de manera equivalente asignando aleatoriamente  $n_1$  de los valores observados a la primera muestra y los restantes a la segunda. Como consecuencia de la intercambiabilidad, bajo la hipótesis nula, la distribución de  $T^*$  se puede estimar considerando  $B$  permutaciones independientes de los datos y calculando los valores del estadístico correspondiente a cada permutación (Pesarin y Salmaso, 2010,[7]). El  $p$ -valor de la prueba es  $\delta = \sum_{b=1}^B \mathbb{I}(T_{(b)}^* \geq T^0)/B$ , donde  $T_{(b)}^*$  es el valor del estadístico relacionado con la  $b$ -ésima permutación y  $T^0$  es el valor observado del estadístico correspondiente al conjunto de datos no permutados.

### Ejemplo de las calificaciones: continuación

Para ilustrar el procedimiento de permutaciones en la construcción de un contraste de igualdad de medias vamos a utilizar los datos del ejemplo de las calificaciones usado en secciones precedentes. En la Tabla 1.9 podemos observar los resultados del contraste de dos muestras que provienen de las calificaciones teóricas y las calificaciones prácticas de un grupo de estudiantes durante el curso académico 2015 – 16.

Calificaciones en teoría										
3	5.5	3.5	0	5.5	6	3.5	6.5	5.5	7.5	4
5	5.5	4.5	3.5	2	5.5	6	8	6	4	7
9	8	5.4	6	4	5.5	6.5	8	3.5	1	3
3.5	7.5	6	7.5	6						
Calificaciones en problemas										
1.5	4.5	0	2.5	5	4	5	4	4.5	1	3
5.5	5	3.5	3.5	1	3	4	3.5	4	2.5	5
5	3	1	4	1	1.5	4	1	1.5	1	3
3	3	2.5	3.5	4.5						

Tabla 1.9: Calificaciones parciales por grupo de alumnos durante el curso académico 2015 – 16.

En la Figura 1.5 podemos observar el histograma de frecuencias de ambas muestras. Los tamaños muestrales son  $n_1 = n_2 = 38$ . El código *R* para aplicación del contraste de permutación es el siguiente:

```
source("t2p.r")
source("perm_2samples.r")
teoria=c(3,5.5,3.5,0,5.5,6,3.5,6.5,5.5,7.5,4,5,
         5.5,4.5,3.5,2,5.5,6,8,6,4,7,9,8,5.4,6,4,
         5.5,6.5,8,3.5,1,3,3.5,7.5,6,7.5,6)
```

```

problemas=c(1.5,4.5,0,2.5,5,4,5,4,4.5,1,3,5.5,5,
            3.5,3.5,1,3,4,3.5,4,2.5,5,5,3,1,4,1,
            1.5,4,1,1.5,1,3,3,3,2.5,3.5,4.5)

```

```
data=c(teoria,problemas)
```

```
lab=rep(1:2,each=38)
```

```
data_mat=cbind(lab,data)
```

```
T=perm.2samples(data_mat,alt="greater",B=10000)
```

```
T$p.value
```

```
[1] 9.999e-05
```

La función `perm.2samples(data,alt,B)` definida en el archivo `perm.2samples.r` (la cual requiere el código escrito en el archivo `"t2p.r"` para el cálculo del  $p$ -valor), estima la distribución del estadístico usado en el contraste en base a las permutaciones generadas y requiere que los datos estén dispuestos en forma matricial, donde la primera columna contiene la etiqueta de los grupos y la segunda columna contiene los datos observados. Asimismo, podemos especificar el tipo de alternativa con las opciones `alt="greater"`, `alt="less"` o `alt="two.sided"`. El número de permutaciones por defecto es  $B = 1000$  pero para hacer uso de un número diferente podemos especificarlo en el tercer argumento (por ejemplo,  $B = 10000$ ). La matriz de datos es `data_mat`:

```

lab data
[1,]  1  3.0
[2,]  1  5.5
[3,]  1  3.5
[4,]  1  0.0
...  ...  ...

```

[38,]	1	6.0
[39,]	2	1.5
[40,]	2	4.5
[41,]	2	0.0
...	...	...
[76,]	2	4.5

El  $p$ -valor se calcula con la función `t2p`. El vector `t2p(T)` contiene los niveles de significación correspondientes a cada valor del estadístico en cada permutación y el primer elemento es el nivel de significación correspondiente a los valores observados del estadístico, que es el  $p$ -valor. El  $p$ -valor se puede obtener mediante la orden `T$p.value`. El valor  $9,999e - 05$  lleva al rechazo de la hipótesis nula a favor de la hipótesis alternativa de que los estudiantes obtienen mejores resultados en la parte de teoría que en la parte de problemas. El resultado, es entonces el mismo que el del contraste de Wilcoxon.

## 1.5. Contrastes multivariantes para dos muestras

En esta sección vamos a tratar la extensión multivariante del problema de localización de dos muestras. El conjunto de datos se compone de  $n = n_1 + n_2$  observaciones de dos poblaciones independientes de dimensión  $q$ , donde  $n_1$  y  $n_2$  son los tamaños de las dos muestras. La  $i$ -ésima observación multidimensional en el grupo  $j$ -ésimo se denota por  $X_{ij} = \{X_{ji1}, \dots, X_{jiq}\}$ . A continuación se construyen contrastes basados en rangos y en permutaciones.

### 1.5.1. Contrastes multivariantes basados en rangos

Consideremos el problema de probar la igualdad de dos distribuciones multivariantes  $F_1$  y  $F_2$ . Asumiremos que  $F_1$  y  $F_2$  tienen una forma similar no especificada pero posibles vectores de localización diferentes.

Sea  $F_j(x)$  la función de distribución acumulativa para la  $j$ -ésima población con  $j = 1, 2$ .

De acuerdo con el modelo de efectos fijos y si denotamos como  $Z_{ji}$  a la variable aleatoria multivariante a partir de la cual se genera  $X_{ji}$ , tenemos  $Z_{1i} = \mu + \theta + \epsilon_{1i}$  y  $Z_{2i} = \mu + \epsilon_{2i}$ ,  $i = 1, \dots, n_j, j = 1, 2$ , donde  $\mu \in \mathcal{R}^q$  es un vector constante,  $\epsilon_{ji}$  son errores aleatorios  $q$ -variantes intercambiables, con un conjunto de parámetros de localización igual al vector nulo y una matriz de varianzas y covarianzas igual a  $\Sigma$ , y  $\theta \in \mathcal{R}^q$  es el vector de parámetros que representan efectos fijos de tratamiento. La hipótesis es la siguiente:

$$H_0 : F_1(x) = F_2(x) = F(x), \forall x \in \mathcal{R}^q$$

contra la hipótesis alternativa bilateral

$$H_1 : F_1(x) \neq F_2(x).$$

Asimismo, la hipótesis nula puede ser escrita de la forma

$$H_0 : \theta = 0$$

mientras que la alternativa resulta

$$H_1 : \theta \neq 0.$$

La hipótesis alternativa establece que al menos un elemento del vector de efectos  $\theta$  no es igual a cero, es decir, al menos para una componente de la variable de respuesta tenemos un efecto no nulo.

Sea  $R_{jih}$  el rango de  $X_{jih}$  en el conjunto  $\{X_{11h}, \dots, X_{1n_1h}, X_{21h}, \dots, X_{2n_2h}\}$ ,  $h = 1, \dots, q$ . Tenemos que  $R_h$  denota al vector  $n$ -dimensional observado de los rangos relacionado con la  $h$ -ésima variable y consideremos la matriz de dimensión  $n \times q$ ,  $\mathbf{R} = \{\mathbf{R}_1, \dots, \mathbf{R}_q\}$ . Cada columna de la matriz se trata de una permutación de  $1, 2, \dots, n$ . Por lo tanto,  $\mathbf{R}$  puede considerarse como una realización de una matriz aleatoria de dimensión  $n \times q$  con  $(n!)^q$  realizaciones posibles. Como las  $q$  columnas de la matriz de rangos son en general estocásticamente dependientes, su distribución dependerá de la distribución desconocida  $F$  incluso cuando  $H_0 : F_1(x) = F_2(x) = F(x)$  se mantenga. Bajo  $H_0$  la distribución de la matriz de rangos condicionados a la matriz observada  $\mathbf{R}$  sobre el conjunto  $S(\mathbf{R}^*)$  de todas las permutaciones posibles de las filas  $\mathbf{R}^*$  es uniforme. La probabilidad de observar una de las  $n! = n \cdot (n-1) \cdot \dots \cdot 2 \cdot 1$  realizaciones posibles es  $1/n!$ .

Igual que en secciones precedentes, podemos utilizar un estadístico construido en base a diferentes calibradores de rango. En general se puede considerar la familia definida como

$$E^{(h)}(R) = g_h \left( \frac{R}{n+1} \right),$$

con  $h = 1, \dots, q$  y  $1 \leq R \leq n$ . Por lo tanto, la matriz  $\mathbf{R}$  se puede reemplazar por  $\mathbf{E} = [\mathbf{E}_1, \dots, \mathbf{E}_q]$  donde

$$\mathbf{E}_h = \left[ E^{(h)}(R_{11h}), \dots, E^{(h)}(R_{1n_1h}), E^{(h)}(R_{21h}), \dots, E^{(h)}(R_{2n_2h}) \right]'$$

con  $h = 1, \dots, q$ . Para cada muestra y para cada una de las  $q$  variables, el

calibrador promedio se puede calcular como

$$T_{j\bullet h} = \frac{\sum_{i=1}^{n_j} E^{(h)}(R_{jih})}{n_j}.$$

Bajo  $H_0$ , el promedio de los calibradores de rango deberían estar cerca de los valores medios totales  $\bar{T}_{\bullet\bullet h} = (n_1\bar{T}_{1\bullet h} + n_2\bar{T}_{2\bullet h})/n$  y los estadísticos  $(\bar{T}_{j\bullet h} - \bar{T}_{\bullet\bullet h})$  deben ser estocásticamente cercanos a cero para  $j = 1, 2$  y  $h = 1, \dots, q$ . En presencia de  $C \geq 2$  grupos, un estadístico adecuado para este problema podría ser

$$L = \sum_{j=1}^C n_j [\bar{T}_j - \bar{T}]' V^{-1} [\bar{T}_j - \bar{T}],$$

donde  $\bar{T}_j = [\bar{T}_{j\bullet 1}, \dots, \bar{T}_{j\bullet q}]'$ ,  $\bar{T} = [\bar{T}_{\bullet\bullet 1}, \dots, \bar{T}_{\bullet\bullet q}]'$  y  $V$  es la matriz de covarianza del vector  $\bar{T}_j - \bar{T}$  bajo  $H_0$ .  $L$  es el estadístico de tipo de Hotelling utilizado en un contraste paramétrico para una muestra procedente de una población con distribución normal multivariante. Por lo tanto, es un estadístico adecuado en condiciones normales. Cuando la distribución de los datos dista mucho de ser normal multivariante, este estadístico puede que no sea válido. Para dos muestras, una alternativa sería utilizar el estadístico

$$L = n_1 [\bar{T}_1 - \bar{T}]' V^{-1} [\bar{T}_1 - \bar{T}] + n_2 [\bar{T}_2 - \bar{T}]' V^{-1} [\bar{T}_2 - \bar{T}].$$

Para valores grandes de  $n$  y  $q$ , se tienen distribuciones asintóticas para  $L$ , esto fue propuesto por Puri y Sen en 1971 (véase referencias en [1]). La distribución asintótica de  $L$  tiende a ser chi-cuadrado con  $q(C-1)$  grados de libertad. Por lo tanto, para  $C = 2$ , la hipótesis nula debería rechazarse cuando  $L \geq \chi_{q;\alpha}^2$ , donde  $\alpha$  denota el nivel de significación. Especificando la forma concreta de los calibradores de rango en uso podemos obtener

diferentes contrastes para las hipótesis consideradas:

- Para construir un contraste de medianas en varias muestras multivariantes se puede usar el calibrador

$$E^{(h)}(R) = \begin{cases} 1 & \text{si } R \leq n/2 \\ 0 & \text{en otro caso} \end{cases} \quad (1.4)$$

- Para el contraste de suma de rangos en muestras multivariantes podríamos definir

$$E^{(h)}(R) = \frac{R}{n+1},$$

por lo tanto el estadístico  $T_{j \bullet h}$  es igual a  $(n+1)^{-1}$  veces el rango promedio en la  $j$ -ésima muestra para la  $h$ -ésima variable.

- Para un contraste con calibradores normales hacemos  $p = \frac{n-R+1}{n+1}$  y

$$E^{(h)}(R) = z_p,$$

donde  $z_p$  es el  $(1-p)$ -ésimo cuantil de la distribución normal estándar que es el valor tal que  $\Phi(z_p) = 1-p$ , con  $\Phi$  denotando la función de distribución de la distribución normal estándar.

El contraste de marcador normal se puede usar solo en problemas específicos, cuando se pueden reemplazar los rangos ordinarios por los calibradores normales relacionados. En cualquier otro caso, entre los procedimientos basados en rango, el contraste de la suma de rangos es la mejor opción.

### **Ejemplo de las calificaciones: continuación**

Para llevar a cabo un estudio bivalente, para cada muestra tomaremos dos componentes, estas se corresponden con las calificaciones en teoría y en

problemas. En la primera muestra (grupo A) las variables teoría y problemas tienen dimensión 20, mientras que en la segunda muestra (grupo B) estas tienen dimensión 18.

Queremos probar si la distribución bivalente de los resultados obtenidos en ambos grupos es la misma, contra la alternativa de que ambas distribuciones son diferentes. En otras palabras, queremos ver si las calificaciones obtenidas en la parte de teoría y en la parte de problemas son similares en los grupos A y B. El nivel de significación es  $\alpha = 0,10$ . Formalmente denotaremos con  $A$  y  $B$  las variables que representan ambas muestras. La hipótesis nula es

$$H_0 : (A_T, A_P) =^d (B_T, B_P),$$

la hipótesis nula es verdadera si la distribución bivalente de los resultados obtenidos en teoría y problemas es igual entre los grupos. La hipótesis alternativa es

$$H_1 : (A_T, A_P) \neq^d (B_T, B_P),$$

que muestra que los resultados en ambos grupos son diferentes. La tabla 1.10 muestra los datos para el problema.

Cabe señalar que el análisis multivariante nos aporta mayor información que si analizamos las variables marginales por separado. Esto es debido a que pueden existir relaciones entre las calificaciones que se pierden en el análisis marginal. Por ejemplo, puede existir un grupo (A o B) con correlación positiva entre las calificaciones en teoría y problemas, lo cual significa que aquellos individuos que estudian más teoría, poseen mayor conocimiento a la hora de resolver problemas. Mientras que puede ocurrir que el otro grupo posea correlación negativa entre teoría y problemas, es decir, cuanto más tiempo se le dedique a la parte teórica, menor tiempo de dedicación a

los problemas y, por tanto, peores resultados. Esta situación puede ser detectada mediante el análisis multivariante, mientras que mediante el análisis marginal no es posible detectar la dependencia de ambas variables.

	Grupo A									
<i>T</i>	3	5.5	3.5	0	5.5	6	3.5	6.5	5.5	7.5
<i>P</i>	1.5	4.5	0	2.5	5	4	5	4	4.5	1
<i>T</i>	4	5	5.5	4.5	3.5	2	5.5	6	8	6
<i>P</i>	3	5.5	5	3.5	3.5	1	3	4	3.5	4
	Grupo B									
<i>T</i>	4	7	9	8	4.5	6	4	5.5	6.5	8
<i>P</i>	2.5	5	5	3	1	4	1	1.5	4	1
<i>T</i>	2.5	1	3	3.5	7.5	6	7.5	6		
<i>P</i>	1.5	1	3	3	3	2.5	3.5	4.5		

Tabla 1.10: Calificaciones parciales por grupo de alumnos durante el curso académico 2015 – 16.

El paquete ICSNP de *R* contiene herramientas para análisis multivariantes no paramétricos. En particular, en este paquete se encuentra la función `rank.ctest(X,Y,scores)` que realiza el contraste de localización *C*-muestral ( $C \geq 2$ ) basado en los rangos marginales. Para el contraste de 2 muestras la función requiere de matrices  $n_j \times q$  de observaciones muestrales (*X*,*Y*) y un vector indicando las diferencias respecto a la media bajo la hipótesis nula ( $\mu$ ). `NULL` indica que no existen diferencias entre la media de ambos grupos. El argumento `scores` requiere el tipo de contraste que se realizará. Puede ser especificado mediante `"sign"`, para un contraste de signo, `"rank "` para un contraste de rango, o `"normal"` para un contraste de calibrador normal. El código *R* para el análisis es:

```
library("ICSNP")
X1=c(3,5.5,3.5,0,5.5,6,3.5,6.5,5.5,7.5,4,5,5.5,4.5,3.5,2,5.5,6,8)
X2=c(1.5,4.5,0,2.5,5,4,5.5,4,4.5,1,3,5.5,5,3.5,3.5,1,3,4,3.5)
X=cbind(X1,X2)
```

```
Y1=c(6,4,7,9,8,4.5,6,4,5.5,6.5,8,2.5,1,3,3.5,7.5,6,7.5,6)
Y2=c(4,2.5,5,5,3,1,4,1,1.5,4,1,1.5,1,3,3,3,2.5,3.5,4.5)
Y=cbind(Y1,Y2)
rank.ctest(X,Y,mu=NULL,scores="rank")
```

La salida es:

```
Marginal Two Sample Rank Sum Test
```

```
data: X and Y
```

```
T = 5.3435, df = 2, p-value = 0.06913
```

```
alternative hypothesis: true location difference is not equal to
c(0,0)
```

El programa nos muestra, entre otros, el valor del estadístico  $T$ , así como el  $p$ -valor asociado al contraste. Cabe destacar que únicamente se puede realizar el contraste cuando la hipótesis alternativa es bilateral. Para nuestro análisis, el  $p$ -valor encontrado es de 0,06913, lo cual nos lleva al rechazo de la hipótesis nula de igualdad de distribuciones, en favor de que ambas distribuciones son distintas. Al tratarse de un contraste complejo, es difícil saber de donde proviene las diferencias entre ambas distribuciones. Puede ocurrir que sea debido a una diferencia en la parte de problemas, en la parte de teoría, o bien en ambas simultáneamente.

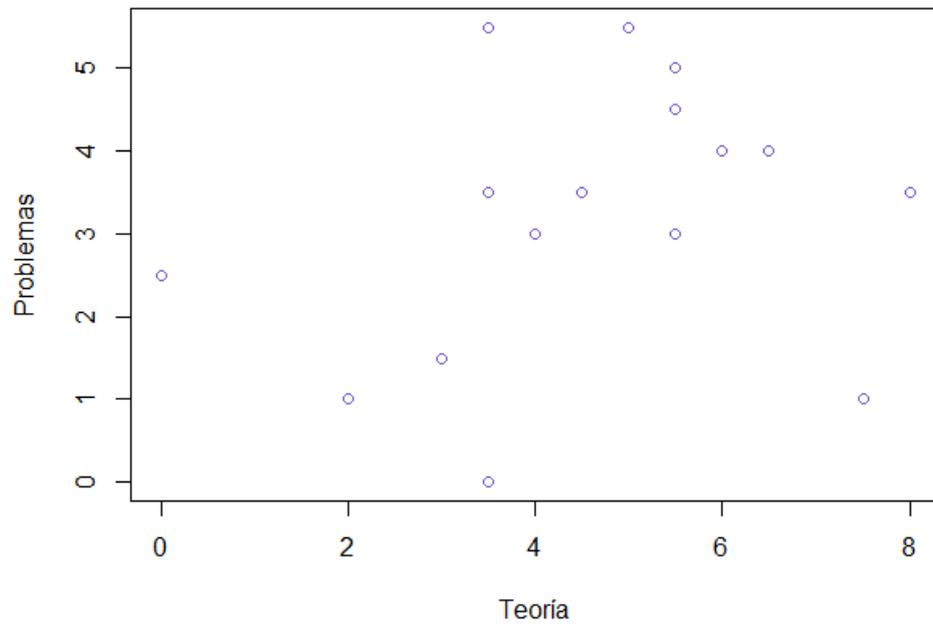


Figura 1.6: Nube de puntos del grupo A.

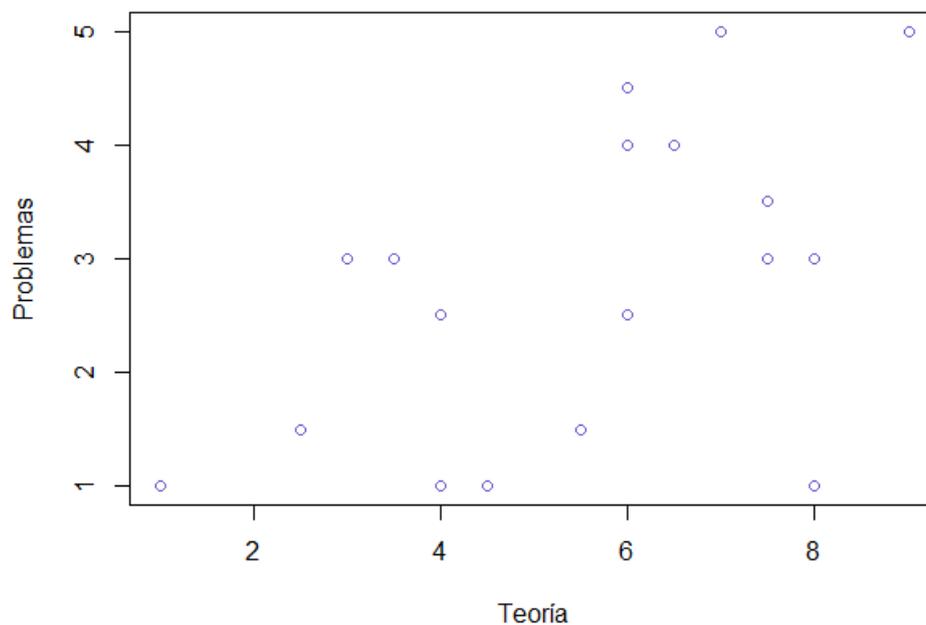


Figura 1.7: Nube de puntos del grupo B.

### 1.5.2. Contrastes de permutación para la tendencia central en muestras multivariantes

Para terminar este capítulo se presenta una extensión del contraste de permutación de dos muestras para problemas multivariantes. En este marco, los datos son  $q$ -dimensionales ( $q \geq 2$ ). A menudo, en los contrastes de hipótesis complejas, es decir, en presencia de muchas variables de respuesta o cuando se tratan varios aspectos de la distribución, el problema general se divide en un conjunto finito de  $k > 1$  diferentes pruebas parciales. Hay que tener en cuenta que el número  $q$  de respuestas no siempre coincide con  $k$ , aunque para la mayoría de los contrastes de localización de parámetros en

muestras multivariantes se verifica  $k = q$ .

Como en los problemas considerados en secciones precedentes, la hipótesis nula consiste en contrastar la igualdad en la distribución de dos vectores aleatorios, aunque se podría debilitar este contraste considerando la igualdad de las distribuciones de las variables marginales.

Usaremos la metodología de combinación no paramétrica descrita en la sección 1.1.2, para ello supondremos que  $H_0$  puede dividirse adecuadamente en un conjunto finito de sub-hipótesis  $H_{0i}$ , cada una apropiada para un aspecto parcial de interés o para una variable marginal (Pesarin y Salmaso, 2010, [7]). Por lo tanto,  $H_0$  y  $H_1$  vienen dadas como en la subsección 1.1.2. En este marco,  $H_1$  se llama hipótesis alternativa global.

Para cada contraste parcial univariante sobre la tendencia central, la diferencia de medias de la muestra puede ser un estadístico adecuado y se puede aplicar un contraste de permutación univariante sobre la tendencia central. Usando la metodología de combinación no paramétrica se combinan  $p$  valores parciales para obtener un estadístico univariante.

La principal ventaja de este procedimiento, además de la posibilidad de considerar variables de respuesta multivariante sin asumir ninguna distribución específica ni especificar la estructura de dependencia entre las variables (pero teniéndolo en cuenta implícitamente), es su gran flexibilidad que permite construir contrastes para hipótesis muy complejas. No se necesita una suposición de continuidad, por lo tanto, se puede aplicar a variables multivariantes continuas, discretas o mixtas. También se puede aplicar a alternativas unilaterales e incluso a alternativas complejas donde algunas de las sub-hipótesis parciales  $H_{1i}$  son bilaterales y otras unilaterales con diferentes direcciones posibles.

## Capítulo 2

# Contrastes de dispersión

### 2.1. Introducción

En muchos procesos de producción o en el análisis de los efectos producidos por un nuevo medicamento, por citar un par de ejemplos prácticos, aparecen frecuentemente problemas de contrastes sobre la variabilidad de los indicadores que se están analizando. Por ejemplo, podemos estar interesados en comparar estadísticamente las distorsiones producidas por dos técnicas de producción de un determinado objeto. Estas distorsiones pueden estar referidas a cualquiera de las características de dicho objeto, peso, longitud, calibrado, resistencia, etc. y pueden afectar decisivamente a la calidad de funcionamiento del mismo. En este contexto, la técnica de fabricación que produzca menos distorsiones frente a los valores de referencia será preferida por cuanto aporta un grado superior de fiabilidad en los resultados del proceso.

Para enfrentarnos a situaciones como la descrita anteriormente, vamos a suponer que comparamos dos variables y disponemos de una muestra de observaciones de cada una de ellas, cumpliéndose las propiedades habitua-

les de independencia e idéntica distribución, es decir, ambas son *muestras aleatorias simples* mutuamente independientes. En este capítulo se analizan diferentes técnicas no paramétricas para la construcción de contrastes sobre indicadores de escala, de forma individual o comparándolos de forma combinada con algún otro indicador de localización. El objetivo es determinar si las muestras proceden de poblaciones con idéntica varianza. Éste puede ser el fin último del análisis o un paso previo a la realización de otros estudios que requieren la propiedad de *homocedasticidad*. Las herramientas que usaremos en la construcción de los contrastes, al igual que en el tema anterior, emplearán estadísticos basados en rangos o en permutaciones.

Los contrastes que se presentan para el análisis de la varianza de dos muestras son:

- Contraste de rangos de Ansari-Bradley, que tiene como requisito previo la igualdad de medianas de ambas muestras.
- Contraste de permutaciones de Pan, que no requiere igualdad de medianas y está recomendado para distribuciones con colas pesadas o alto sesgo.
- Contraste de permutaciones de O'Brien, que tampoco requiere igualdad de medianas y está recomendado para distribuciones con colas livianas y simétricas.

En cuanto a los contrastes referidos a la detección simultánea de diferencias en localización y escala se presentarán los siguientes

- Contraste de Lepage, basado en la combinación de dos contrastes, uno de ellos para localización (Wilcoxon) y el otro para escala (Ansari-Bradley).

- Contraste de Cucconi, que presenta ciertas ventajas con respecto al anterior en cuanto a su potencia y sencillez de cálculo.

Finalmente, analizaremos dos contrastes más ambiciosos en los que se detectarán diferencias entre las características de las distribuciones en general, incluyendo diferencias de forma como la curtosis o el sesgo. Los test estudiados son:

- Contraste de Kolmogorov-Smirnov, que tiene como requisito previo la continuidad de las poblaciones y la independencia entre ambas.
- Contraste de Cramer-von Mises, que presenta una ventaja con respecto al anterior pues hace un uso más eficiente de los datos. Sin embargo, es más difícil de calcular.

## 2.2. Comparación de la dispersión

Los problemas relativos a la comparación de la dispersión son de gran interés en muchas áreas. Minimizar y controlar la dispersión es importante en el control de calidad. Esto fue estudiado por Nair y Pregibon en el año 1988, en los sistemas de producción agrícola, Fairfull et al. en el año 1985, en educación, Games *et al.* en el año 1972 y Olejnik y Algina en 1988 (véase referencias en [1]). Los contrastes para la igualdad de las varianzas son interesantes también como inicio al análisis de la varianza o el análisis discriminante.

Por otra parte, experimentalmente se ha observado que la prueba  $F$  no es robusta en el caso en que los datos estudiados carecen de normalidad. Por lo tanto, no parece sensato realizar un contraste muy sensible a la ausencia de normalidad para comprobar la homocedasticidad de las distribuciones antes de construir una prueba  $t$  de igualdad de medias para poblaciones normales.

Box, 1953 (véase referencias en [1]) ilustró con un símil muy gráfico esta idea “comenzar este análisis con el contraste de varianzas es como navegar en un bote de remos para ver si las condiciones atmosféricas están lo suficientemente en calma para que un transatlántico abandone el puerto!”.

Es importante resaltar que la comparación de varianzas u otras medidas de escala es mucho más difícil que la comparación de las medias u otras medidas de localización. Boos y Brownie, en el año 2004 (véase referencias en [1]) indicaron dos razones para este hecho. La primera razón es que el muestreo en poblaciones normales permite construir estadísticos para la detección de diferencias en los parámetros de localización cuyas distribuciones asintóticas son robustas a la ausencia de normalidad en los datos. La segunda razón es que para las comparaciones de parámetros de localización, la hipótesis de que las poblaciones pueden diferir solo en los parámetros de localización es apropiada, permitiendo así que el uso de métodos de permutación proporcione distribuciones exactas del estadístico empleado en el contraste. Por el contrario, para las comparaciones de escala, la hipótesis de que las poblaciones difieren solo en escala rara vez tiene sentido ya que usualmente están presentes las diferencias de localización. Teniendo en cuenta que en este caso es necesario realizar ajustes en las muestras de diferentes poblaciones mediante parámetros de escala que son desconocidos, el empleo de contrastes basados en permutaciones no suele conducir a distribuciones exactas de los estadísticos empleados.

En las secciones siguientes vamos a presentar diferentes contrastes. Entre ellos se encuentran los de rangos de Ansari-Bradley, de permutaciones de Pan y de O’Brien, para detectar otro tipo de cambios de escala; los contrastes de Lepage y de Cucconi, para la detección simultánea de localización y escala y por último los contrastes de Kolmogorov-Smirnov y de Cramér-von Mises

para detectar diferencias en las distribuciones.

### 2.2.1. Contraste de Ansari-Bradley

Dadas dos muestras de dos poblaciones con distribución desconocida y una vez estudiados los análisis relativos a los parámetros de localización y la simetría, sería interesante estudiar si existen diferencias en la dispersión de los datos considerados. Vamos a ilustrar esto mediante los datos correspondientes a la Tabla 1.3. Las muestras (Grupo A y Grupo B) poseen un valor de la mediana de 5 y 5,35, respectivamente.

Para comprobar la homocedasticidad de ambas muestras podemos hacer uso del contraste de Ansari-Bradley, que construye un estadístico basado en los rangos de las observaciones para comparar las escalas de dos muestras independientes bajo la suposición siguiente:

- Las medianas poblacionales de las muestras son iguales.

Sean  $F_1$  y  $F_2$  las funciones de distribución correspondientes a las poblaciones 1 y 2, respectivamente. La hipótesis nula es que ambas poblaciones tienen la misma distribución desconocida

$$H_0 : \{F_1(x) = F_2(x), \forall x \in \mathcal{R}\}.$$

Vamos a considerar el modelo de localización-escala que corresponde a tomar

$$F_1(x) = H\left(\frac{x - \mu_1}{\sigma_1}\right)$$

y

$$F_2(x) = H\left(\frac{x - \mu_2}{\sigma_2}\right),$$

$\forall x \in \mathcal{R}$ , donde  $H$  es la función de distribución de una variable continua

con mediana 0. Por lo tanto, los parámetros  $\mu_1$  y  $\mu_2$  denotan las medianas (parámetros de localización) de las poblaciones 1 y 2, respectivamente;  $\sigma_1$  y  $\sigma_2$  son los parámetros de escala de las poblaciones 1 y 2, respectivamente. El contraste bilateral se puede plantear de la forma siguiente:

$$H_0 : \{\sigma_1 = \sigma_2\}$$

contra

$$H_1 : \{\sigma_1 \neq \sigma_2\}.$$

Para el contraste anterior, el estadístico de Ansari-Bradley tendría la siguiente forma

$$AB = \sum_{i=1}^{n_1} AB_i,$$

donde  $AB_i$  es el calibrador de  $AB$  asignado a la variable  $X_{1i}$ . Para calcular los calibradores se usa la siguiente asignación de rangos, en primer lugar ordenamos los elementos de las dos muestras combinadas; en segundo lugar se le asigna el valor 1 tanto a los elementos mayores y menores, el valor 2 al segundo elemento menor y al segundo mayor, etc. Por lo tanto, los valores de  $AB$  son

$$1, 2, 3, \dots, \frac{n}{2}, \frac{n}{2}, \dots, 3, 2, 1, \text{ cuando } n \text{ es par, y}$$

$$1, 2, 3, \dots, \frac{n-1}{2}, \frac{n+1}{2}, \frac{n-1}{2}, \dots, 3, 2, 1, \text{ cuando } n \text{ es impar.}$$

El estadístico  $AB$  es la suma de los valores asignados a los elementos de  $\mathbf{X}_1$ .

Cabe destacar que el contraste de Ansari-Bradley se puede aplicar también en poblaciones con diferentes medianas, siempre que estas sean conocidas. En este caso, debemos restar  $\mu_1$  a los elementos de la primera muestra y

$\mu_2$  a los de la segunda, pues la mediana de las variables ajustadas satisfacen la hipótesis de igualdad de medianas.

La motivación del contraste es que si, por ejemplo,  $\sigma_1 > \sigma_2$ , los elementos de  $\mathbf{X}_1$  tienden a estar más dispersos que los elementos de  $\mathbf{X}_2$ . Por lo tanto, los valores de  $AB$  tenderían a ser más pequeños. En el caso en que  $\sigma_1 < \sigma_2$ , sucede lo contrario. Por lo tanto, las alternativas unilaterales

$H_1 : \{\sigma_1 > \sigma_2\}$  se rechazan para valores grandes del estadístico  $AB$ ; las alternativas unilaterales  $H_1 : \{\sigma_1 < \sigma_2\}$  por el contrario, se rechazarán para valores pequeños del estadístico  $AB$ ; la hipótesis alternativa bilateral  $H_1 : \{\sigma_1 \neq \sigma_2\}$  se rechaza ya sea para valores pequeños o grandes del estadístico  $AB$ .

En primer lugar, para aplicar este análisis a los datos de las calificaciones debemos restar las medianas respectivas de modo que los datos transformados verifiquen la hipótesis de partida. Para ello usaremos el programa  $R$ , cuyo código es el siguiente:

```
X1=c(2.3,7,5,1.8,1.3,5.3,6,5,4.3,5.3,5,7.7,4.3,5.3,5,7.7,4.3,
      3.5,5,5.3,9,5.3,4,3.5,1.5,4.3,5,5.8,5) #Grupo A
X2=c(6,3.3,6,7,5.5,7.3,7,2.8,7.3,5,2.5,5.2,3.5,5.3,6.2,4.5,6.2,
      2,1,6,3,3.3,6.1,5.3,9,5.4,4.3,5.5,5.5,5.3) #Grupo B
m1=median(X1)
m2=median(X2)
mediana1=c(rep(m1,length(X1)))
mediana2=c(rep(m2,length(X2)))

Y1= X1-mediana1 #Variable X1 tipificada
Y2= X2-mediana2 #Variable X2 tipificada
```

Mediante este procedimiento obtenemos los datos representados en la Tabla 2.1, donde se pueden observar los valores de las dos muestras resultantes  $\mathbf{Y}_1$  y  $\mathbf{Y}_2$ .

Grupo A									
-2.7	2.0	0.0	-3.2	-3.7	0.3	1.0	0.0	-0.7	0.3
0.0	2.7	-0.7	0.3	0.0	2.7	-0.7	-1.5	0.0	0.3
4.0	0.3	-1.0	-1.5	-3.5	-0.7	0.0	0.8	0.0	
Grupo B									
0.65	-2.05	0.65	1.65	0.15	1.95	1.65	-2.55	1.95	-0.35
-2.85	-0.15	-1.85	-0.05	0.85	-0.85	0.85	-3.35	-4.35	0.65
-2.35	-2.05	0.75	-0.05	3.65	0.05	-1.05	0.15	0.15	-0.05

Tabla 2.1: Datos transformados correspondientes al ejemplo de las calificaciones.

El paquete básico de  $R$  contiene la función `ansari.test` que calcula tanto el valor del estadístico  $AB$ , así como el  $p$ -valor. Si ambas muestras contienen menos de 50 valores y no hay empates, el  $p$ -valor calculado es exacto. En otro caso, se usa una aproximación normal asintótica. En presencia de empates se hace uso de la regla midrank. El código  $R$  para el análisis es el siguiente:

```
X1=c(2.3,7,5,1.8,1.3,5.3,6,5,4.3,5.3,5,7.7,4.3,5.3,5,7.7,4.3,
      3.5,5,5.3,9,5.3,4,3.5,1.5,4.3,5,5.8,5)
X2=c(6,3.3,6,7,5.5,7.3,7,2.8,7.3,5,2.5,5.2,3.5,5.3,6.2,4.5,6.2,
      2,1,6,3,3.3,6.1,5.3,9,5.4,4.3,5.5,5.5,5.3)
m1=median(X1)
m2=median(X2)
mediana1=c(rep(m1,length(X1)))
mediana2=c(rep(m2,length(X2)))
Y1= X1-mediana1
Y2= X2-mediana2
```

```
ansari.test(Y1,Y2,alternative="greater")
```

La salida es:

```
Ansari-Bradley test
```

```
data: Y1 and Y2
```

```
AB = 484, p-value = 0.8797
```

```
alternative hypothesis: true ratio of scales is greater than 1
```

```
Warning message:
```

```
In ansari.test.default(Y1, Y2, alternative = "greater") :
```

```
cannot compute exact p-value with ties
```

Como en análisis anteriores aparece un mensaje de advertencia debido a la presencia de empates a causa del redondeo en las calificaciones. Por este motivo nos devolverá el  $p$ -valor que resulta de hacer uso de una aproximación asintótica. El  $p$ -valor para el contraste de Ansari-Bradley es 0,8797. Debido a que se trata de un  $p$ -valor muy grande, no hay motivos para rechazar  $H_0$ . Este resultado nos muestra que la dispersión de las calificaciones en ambos grupos es la misma. En otras palabras, la variabilidad en las calificaciones es similar y pueden darse dos situaciones distintas. La primera es que la dispersión de los datos sea pequeña (varianzas pequeñas), lo cual implicaría que los estudiantes de cada grupo presentan un comportamiento homogéneo, es decir, poseen las mismas capacidades. Esto suele suceder cuando los alumnos estudian de manera conjunta. La segunda situación es que exista una variabilidad mayor, lo cual implica que, por ejemplo, existan varios subgrupos de estudio en cada uno de los grupos estudiados.

### 2.2.2. Contraste de permutaciones de Pan

Los datos reflejados en la Tabla 2.2 nos muestran las calificaciones en problemas, correspondientes a los grupos A y B, en el ejemplo de las calificaciones durante el curso académico 2015 – 16.

Grupo A									
1.5	4-5	0.0	2.5	5	4	5	4	4.5	1
3	5.5	5	3.5	3.5	1	3	4	3.5	4
Grupo B									
2.5	5	5	3	1	4	1	1.5	4	1
1.5	1	3	3	3	2.5	3.5	4.5		

Tabla 2.2: Calificaciones en problemas, separadas por grupo, en el ejemplo de las calificaciones durante el curso 2015 – 16.

El contraste basado en permutaciones de Pan podría ser adecuado a la hora de analizar la dispersión de las calificaciones en un grupo y en otro, pues como podemos observar en los histogramas de frecuencia de la Figura 2.1, ambas muestras son sesgadas. La primera es sesgada a la izquierda mientras que la segunda es sesgada a la derecha.

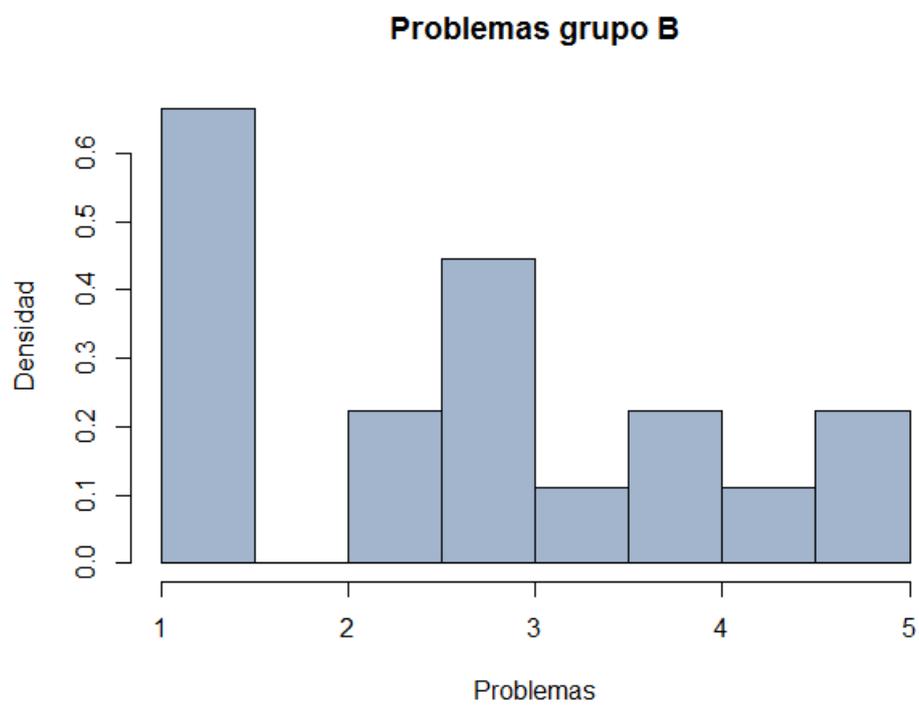
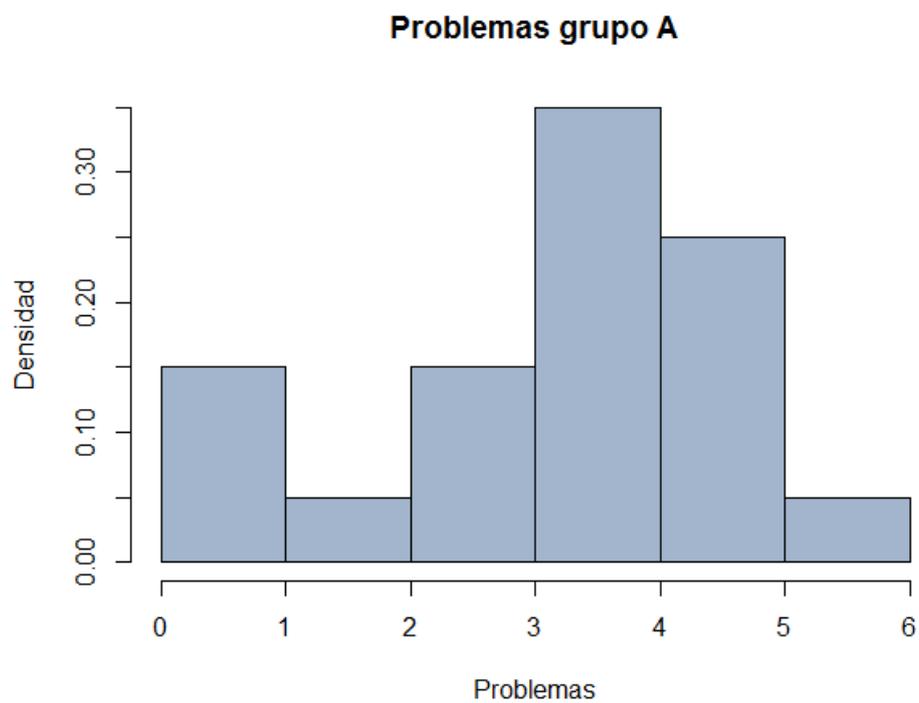


Figura 2.1: Histogramas de frecuencias de las calificaciones en problemas por grupo.

Un estadístico para el contraste de homocedasticidad que resulta apropiado para esta situación es la versión basada en permutaciones del contraste de Pan (1999) que fue propuesto por Marozzi (2011) [4]. Cabe destacar que el contraste Pan es una modificación del contraste de Levene que es la alternativa que usualmente se propone frente a la prueba F. Además la propuesta original del contraste de Levene posee varias versiones, por lo que a la hora de presentar el contraste Pan vamos a considerar la versión que consiste en una prueba  $t$ -Student de igualdad de medias para las diferencias en valor absoluto respecto a las medianas en cada muestra, es decir,  $Z_{ij} = |X_{ij} - \tilde{X}_j|$  donde  $\tilde{X}_j$  es la mediana de la  $j$ -ésima muestra.

A diferencia del contraste original de Levene, esta prueba es asintóticamente libre de distribuciones como fue observado por Miller en 1968, robusta frente a cambio en las distribuciones poblacionales como indicaron Brown y Forsythe en 1974 y garantiza una potencia del test considerable cuando las distribuciones subyacentes tienen cola pesada, como observó O'Brien en 1979 (véase referencias en [1]).

A pesar de las buenas propiedades de robustez y asintóticas exhibidas por este test, Pan en 1999 (véase referencias en [1]) encontró un inconveniente y es que su potencia nunca llega a alcanzar el valor 1 independientemente de lo diferente que sean los parámetros de escala que estamos comparando,  $\sigma_1$  y  $\sigma_2$ . Este hecho lleva a Pan a proponer una modificación para mitigar este problema, conservando las propiedades antes referidas.

El contraste de Pan está basado en el siguiente estadístico:

$$PAN = \frac{\ln \bar{Z}_1 - \ln \bar{Z}_2}{\sqrt{\frac{1}{n_1} \frac{V_1^2}{\bar{Z}_1^2} + \frac{1}{n_2} \frac{V_2^2}{\bar{Z}_2^2}}},$$

donde  $\bar{Z}_j$  y  $V_j^2$  denotan, respectivamente, la media y la varianza de  $Z_{ij}$ ,  $i =$

$1, \dots, n_j, j = 1, 2$ . La hipótesis nula  $H_0 : \{\sigma_1 = \sigma_2\}$  se rechaza a un nivel de significación  $\alpha$  a favor de la hipótesis alternativa  $H_1 : \{\sigma_1 \neq \sigma_2\}$  si  $|PAN| \geq t_{1-\alpha/2, n-2}$  donde  $t_{1-\alpha/2, n-2}$  es el  $100(1 - \alpha/2)$ -ésimo percentil de la distribución  $t$ -Student con  $n - 2$  grados de libertad. Cuando la hipótesis alternativa es unilateral de la forma  $H_1 : \{\sigma_1 > \sigma_2\}$ , la hipótesis nula se rechaza si  $PAN \geq t_{1-\alpha, n-2}$ . Asimismo, si  $H_1 : \{\sigma_1 < \sigma_2\}$ , la hipótesis nula se rechaza si  $PAN \leq t_{1-\alpha, n-2}$ . Marozzi (2011) [4], mostró que la nueva versión del contraste de Pan era mejor que la versión original y, por esta razón, consideramos el contraste de permutation Pan. Para llevar a cabo dicho contraste, vamos a considerar el marco de muestreo para los contrastes de escala descrito por Boos and Brownie en 1989 (véase referencias en [1]). Extraemos una muestra sin reemplazamiento de  $\mathbf{Y} = (X_{ij} - \bar{X}_j, i = 1, \dots, n_j, j = 1, 2)$ , donde  $\bar{X}_j$  es la media de la  $j$ -ésima muestra,  $B$  parejas de muestras de tamaños  $n_1$  y  $n_2$ , respectivamente y calculamos el estadístico para cada pareja obteniendo así  $PAN_b^*, b = 1, \dots, B$ . Sea  $PAN^0$  el valor observado del estadístico, es decir, estadístico  $PAN$  calculado en  $\mathbf{Y}$ .

- (i) *Contraste unilateral de cola superior.* Para comprobar  $H_0 : \{\sigma_1 = \sigma_2\}$  contra  $H_1 : \{\sigma_1 > \sigma_2\}$  a un nivel de significación  $\alpha$  se rechaza  $H_0$  si

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(PAN_b^* \geq PAN^0) < \alpha,$$

en otro caso no se rechaza.

- (ii) *Contraste unilateral de cola inferior.* Para comprobar  $H_0 : \{\sigma_1 = \sigma_2\}$  contra  $H_1 : \{\sigma_1 < \sigma_2\}$  a un nivel de significación  $\alpha$  se rechaza  $H_0$  si

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(PAN_b^* \leq PAN^0) < \alpha,$$

en otro caso no se rechaza.

(iii) *Contraste bilateral.* Para comprobar  $H_0 : \{\sigma_1 = \sigma_2\}$  contra  $H_1 : \{\sigma_1 \neq \sigma_2\}$  a un nivel de significación  $\alpha$  se rechaza  $H_0$  si

$$\frac{1}{B} \left[ \sum_{b=1}^B \mathbb{I}(PAN_b^* \geq -|PAN^0|) + \sum_{b=1}^B \mathbb{I}(PAN_b^* \leq |PAN^0|) \right] < \alpha,$$

en otro caso no se rechaza.

Se debe tener en cuenta que bajo  $H_0$  los elementos  $\mathbf{Y}$  no son intercambiables, por lo que el contraste de permutación no devuelve un  $p$ -valor exacto. Sin embargo, la *ley fuerte de los grandes números* asegura la convergencia de la estimación al  $p$ -valor cuando  $\min(n_1, n_2) \rightarrow \infty$  con  $n_1/n_2 \rightarrow C$ , donde  $C$  es una constante.

Para ilustrar este contraste de homocedasticidad usaremos el ejemplo de las calificaciones de las secciones anteriores. Vamos a considerar, en primer lugar, el caso (iii). En el caso en que no haya evidencia para rechazar la hipótesis nula, hemos terminado. En caso contrario, pasaríamos a analizar los casos (i) y (ii) para hacer un estudio más detallado y ver cuál sería el contraste adecuado.

Para llevar a cabo el contraste usando  $R$  descargamos el archivo "`pan_test.r`" el cual contiene la función `pan(x1,x2,alt,B)` donde `x1` y `x2` son las muestras, `alt` es el tipo de alternativa y `B` el número de permutaciones (por defecto `B = 10000`). La salida es el valor observado  $PAN^0$  y el  $p$ -valor del contraste. El código  $R$  para el desarrollo del contraste usando 10000 permutaciones para la estimación del  $p$ -valor es el siguiente:

```
source("pan_test.r")
problemasA = c(1.5,4.5,0,2.5,5,4,5,4,4.5,1,3,5.5,5,3.5,3.5,1,3,
              4,3.5,4)
```

```

problemasB=c(2.5,5,5,3,1,4,1,1.5,4,1,1.5,1,3,3,3,2.5,3.5,4.5)
pan.test=pan(problemasA,problemasB,alt="two.sided",B=10000)
pan.test

```

La salida de  $R$  es la siguiente:

```

$obs.value
[1] 0.1294308

$p.value
[1] 0.9074

```

El primer resultado corresponde al valor observado del estadístico y el segundo al  $p$ -valor del contraste. Como podemos observar, el  $p$ -valor es muy grande, 0,9074, por lo que no existe evidencia para rechazar  $H_0$ , de manera que la dispersión de calificaciones es igual en el grupo A que en el grupo B. Por lo tanto, no es necesario aplicar los casos (i) y (ii) a este contraste.

### 2.2.3. Contraste de permutaciones O'Brien

En la subsección anterior hemos descrito el test de Pan basado en permutaciones, que es una versión del test de Levene particularmente útil cuando las distribuciones subyacentes tienen colas pesadas o son altamente sesgadas. Otra versión del test de Levene sería útil en el caso en que las distribuciones poblacionales sean simétricas y con colas ligeras, el que se conoce como contraste de O'Brien basado en permutaciones. Fue utilizado por primera vez por Marozzi (2011) [4]. El contraste original de O'Brien es simplemente una prueba  $t$ -Student sobre la muestra

$$R_{ji} = \frac{(n_j - 1,5)n_j(X_{ji} - \bar{X}_j)^2 - 0,5S_j^2(n_j - 1)}{(n_j - 1)(n_j - 2)}$$

donde  $S_j^2$  denota la varianza muestral de  $X_j, j = 1, 2$ . Por lo tanto, el contraste O'Brien está basado en

$$OBR = \frac{\bar{R}_1 - \bar{R}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)W_1^2 + (n_2-1)W_2^2}{n-2}}},$$

donde  $\bar{R}_j$  y  $W_j^2$  denotan la media y la varianza muestrales de  $R_{ji}, i = 1, \dots, n_j, j = 1, 2$ , respectivamente. Tenemos que la hipótesis nula  $H_0 : \{\sigma_1 = \sigma_2\}$  se rechaza a un nivel de significación  $\alpha$  en favor de la hipótesis alternativa bilateral  $H_1 : \{\sigma_1 \neq \sigma_2\}$  si  $|OBR| \geq t_{1-\alpha/2, n-2}$ . Cuando la hipótesis alternativa es unilateral, es decir,  $H_0 : \{\sigma_1 > \sigma_2\}$ , la hipótesis nula se rechaza si  $OBR \leq -t_{1-\alpha, n-2}$ . Marozzi (2011) [4] mostró que esta versión de remuestreo del contraste de O'Brien es mejor que su versión original. Por esta razón, vamos a considerar el contraste de permutación de O'Brien. Para llevar a cabo el contraste, extraemos una muestra sin reemplazamiento de  $Y$  de  $B$  parejas de muestras de tamaños  $n_1$  y  $n_2$ , respectivamente, y calculamos el estadístico para cada pareja. Así obtenemos  $OBR_b^*, B = 1, \dots, B$ . Sea  $OBR^0$  el valor observado del estadístico, es decir, el estadístico  $OBR$  calculado en  $Y$ .

- (i) *Contraste unilateral de cola superior.* Para contrastar  $H_0 : \{\sigma_1 = \sigma_2\}$  contra  $H_1 : \{\sigma_1 > \sigma_2\}$  con un nivel de significación  $\alpha$ .  $H_0$  se rechaza si

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(OBR_b^* \geq OBR^0) < \alpha,$$

en otro caso no se rechaza.

- (ii) *Contraste unilateral de cola inferior.* Para contrastar  $H_0 : \{\sigma_1 = \sigma_2\}$

contra  $H_1 : \{\sigma_1 < \sigma_2\}$  con un nivel de significación  $\alpha$ .  $H_0$  se rechaza si

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(OBR_b^* \leq OBR^0) < \alpha,$$

en otro caso no se rechaza.

(iii) *Contraste bilateral*. Para contrastar  $H_0 : \{\sigma_1 = \sigma_2\}$  contra  $H_1 : \{\sigma_1 \neq \sigma_2\}$  con un nivel de significación  $\alpha$ .  $H_0$  se rechaza si

$$\frac{1}{B} \left[ \sum_{b=1}^B \mathbb{I}(OBR_b^* \leq |OBR^0|) + \sum_{b=1}^B \mathbb{I}(OBR_b^* \geq |OBR^0|) \right] < \alpha,$$

en otro caso no se rechaza.

### Ejemplo de las calificaciones: continuación

Los datos que vamos a considerar quedan reflejados en la Tabla 2.2. Como podemos observar en los histogramas 2.1, ambas muestras son sesgadas, por lo que ambas son asimétricas. Veamos, por lo tanto, si al aplicar a estos datos el contraste O'Brien, este nos devuelve el mismo resultado que al aplicar el contraste Pan, que sería el adecuado.

Para llevar a cabo el contraste O'Brien, vamos a considerar el caso (iii). Para el desarrollo del contraste usando  $R$  vamos a descargar el archivo ‘‘obrien\_test.r’’ que contiene la función `obrien(x1,x2,alt,B)`, donde `x1` y `x2` son las muestras, `alt` es el ipo de alternativa y `B` es el número de permutaciones. La salida nos muestra el valor observado  $OBR^0$  y el  $p$ -valor del contraste. El código  $R$  para el desarrollo del contraste usando 10000 permutaciones es el siguiente:

```
source("obrien_test.r")
problemasA = c(1.5,4.5,0,2.5,5,4,5,4,4.5,1,3,5.5,5,3.5,3.5,1,3,4,
```

3.5,4)

```
problemasB=c(2.5,5,5,3,1,4,1,1.5,4,1,1.5,1,3,3,3,2.5,3.5,4.5)
obrien.test=obrien(problemasA,problemasB,alt="two.sided",B=10000)
obrien.test
```

La salida de *R* es:

```
$obs.value
```

```
[1] 0.4560402
```

```
$p.value
```

```
[1] 0.657
```

El primer resultado nos muestra el valor observado del estadístico, mientras que el segundo corresponde al *p*-valor del contraste, cuyo valor es 0,657, por lo tanto, no hay evidencia para rechazar la hipótesis nula. Comparando ambos contrastes, el *p*-valor asociado al contraste Pan es mayor que el *p*-valor asociado al contraste O'Brien aunque en ambos casos no hay evidencias para rechazar la hipótesis nula. Sin embargo, si uno de ellos rechazara y el otro no la hipótesis nula al mismo nivel de significación deberíamos atender al resultado proporcionado por el test de Pan.

## 2.3. Comparación conjunta de tendencia central y dispersión

Comparar dos muestras es uno de los problemas más importantes en los contrastes estadísticos. Las distribuciones de las poblaciones pueden diferir tanto en la localización (ver Capítulo 1), como en el parámetro de escala (ver sección 2.2).

Para distribuciones normales y siempre que contemos con un contraste cuya hipótesis alternativa sea unilateral, las pruebas  $t$ -Student y  $F$  son las más adecuadas para los problemas de localización y escala, respectivamente. Además, la prueba  $t$ -Student es  $\alpha$  robusta para aquellas distribuciones que no son normales (excepto para las de cola muy pesada), mientras que la prueba  $F$  no lo es. En esta sección se tratarán aquellos contrastes con cambios en localización y escala simultáneamente. Éstos son particularmente útiles en campos tan dispares como el estudio de la dinámica del clima y el análisis de los precios de las acciones en bolsa.

En la práctica es muy difícil encontrar datos normales. Por este motivo, debemos buscar un contraste que sea útil para problemas con cambios de localización y escala donde los datos no son normales.

Vamos a considerar el esquema de contraste de dos etapas propuesto por Manly y Francis en 2002 (véase referencias en [1]). En la primera etapa, se lleva a cabo un *contraste de permutación exacto* para determinar si hay evidencia de que las distribuciones subyacentes correspondientes a las muestras que se comparan son diferentes. Si el contraste no es significativo, el procedimiento se detiene y se concluye que no hay evidencia de diferencias en la escala o localización. Si el contraste es significativo, entonces se lleva a cabo un *contraste de permutación aproximado y robusto* para detectar cambios en la localización junto con un procedimiento que indica si este contraste es fiable. Del mismo modo, también se llevará a cabo un contraste robusto para detectar cambios de escala.

Cabe destacar que este esquema de contraste controlaba la tasa de error de tipo I. Sin embargo, su uso no se generalizó pues esto sucedía debido al hecho de concluir con frecuencia que existía evidencia en cuanto a diferencias en las poblaciones sin indicar de dónde provenían dichas diferencias. Otro

inconveniente de este procedimiento es que resulta tener cierta complejidad computacional.

En esta subsección, abordaremos el problema de localización-escala dentro del marco permutación/rango. El contraste de rango más conocido se debe a Lepage en el año 1971, además consideramos también el contraste de Cucconi propuesto en 1968 (véase referencias en [1]), pero que no es tan conocido como el anterior.

### 2.3.1. El contraste de Lepage

En la literatura, los contrastes no paramétricos para detectar conjuntamente la localización y los cambios de escala, es decir, para probar

$$H_0 : \{\mu_1 = \mu_2 \cap \sigma_1 = \sigma_2\}$$

contra

$$H_0 : \{\mu_1 \neq \mu_2 \cup \sigma_1 \neq \sigma_2\}$$

se basan en la combinación de dos contrastes, uno de ellos para la localización y el otro para la escala. En general, la combinación se logra mediante la suma de los estadísticos estandarizados al cuadrado. Este es el caso del contraste Lepage, que se basa en el estadístico siguiente:

$$LEP = \frac{(W - \mathbb{E}_0(W))^2}{\mathbb{V}_0(W)} + \frac{(AB - \mathbb{E}_0(AB))^2}{\mathbb{V}_0(AB)},$$

donde  $W$  es el estadístico de Wilcoxon-Mann-Whitney, y  $AB$  el de Ansari-Bradley.  $\mathbb{E}_0(\cdot)$  y  $\mathbb{V}_0(\cdot)$  denotan el valor esperado y la varianza de  $W$  y  $AB$  bajo  $H_0$

$$\mathbb{E}_0(W) = n_1(n+1)/2, \quad \mathbb{V}_0 = n_1n_2(n+1)/2$$

$$\mathbb{E}_0(AB) = n_1(n+2)/4, \quad \mathbb{V}_0(AB) = n_1n_2(n+2)/48/(n-1)$$

cuando  $n$  es par,

$$\mathbb{E}_0(AB) = n_1(n+1)^2/4/n, \quad \mathbb{V}_0(AB) = n_1n_2(n+1)(n^2+3)/48/n^2$$

cuando  $n$  es impar.

En cuanto al contraste de Ansari-Bradley para la comparación de la dispersión, antes de la extensión de computadoras potentes y relativamente baratas, para poder aplicar el contraste de Lepage había que mirar la tabla correspondiente a los valores críticos o bien considerar una aproximación muestral. Además, en presencia de empates, la fórmula para calcularla es bastante pesada. Hoy en día, usando un software como *R*, el contraste se puede aplicar fácilmente como un contraste de permutación: primero se calcula el valor observado  $LEP^0$  del estadístico de Lepage, luego se simula la distribución de las permutaciones  $LEP^*$ ,  $b = 1, \dots, B$  del estadístico tomando una gran muestra aleatoria de permutaciones  $B$ , por ejemplo  $B = 10000$ , de la muestra agrupada  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$  y luego se calcula el  $p$ -valor correspondiente

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(LEP_b^* \geq LEP^0).$$

Para llevar a cabo el contraste mediante el programa *R*, cargamos el archivo "`lepage_text.r`" con la función `lepage(x1,x2,B)` donde  $x_1$  y  $x_2$  son las muestras y  $B$  representa el número de permutaciones (por defecto  $B=10000$ ). Cabe destacar que no hay distinción entre hipótesis alternativas, puesto que la hipótesis alternativa es siempre la negación lógica de la hipótesis nula. La salida de la función es el valor observado  $LEP^0$  y el  $p$ -valor del

contraste

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(LEP_b^* \geq LEP^0).$$

El contraste de Lepage tiene una gran aplicación en el campo de la medicina para analizar grupos de pacientes sometidos a diferentes tratamientos. Una aplicación interesante sería el análisis del subconjunto de datos obtenidos por Karpatkin et al. en 1981 (véase referencias en [1]) en su estudio sobre el efecto del tratamiento con esteroides maternos sobre el recuento de plaquetas en recién nacidos.

### 2.3.2. El contraste de Cucconi

El contraste de rango estándar para problemas de localización-escala de dos muestras es el contraste de Lepage que es una combinación del contraste de Wilcoxon para la localización y el contraste de Ansari-Bradley para la escala. Marozzi (2013) [5] revisó y comparó varios contrastes de localización-escala para dos muestras de tipo Lepage así como contrastes de Cucconi. El contraste de Cucconi no es tan conocido pero es de interés por varias razones. Primero, desde un punto de vista histórico, porque se propuso años antes que el contraste de Lepage. En segundo lugar, porque no es una combinación de un contraste de localización y un contraste de escala como los otros contrastes. En tercer lugar, como lo muestra Marozzi (2009 [3], 2013 [5]) se compara favorablemente con los contrastes de tipo Lepage en términos de potencia y probabilidad de error de tipo I y, lo que es muy importante, su cálculo es más sencillo.

Después de un período de más de 40 años, Marozzi (2009) [3] se centró en los contrastes de Cucconi calculando así por primera vez una tabla de valores críticos exactos (hay que tener en cuenta que Cucconi en el año 1968 proporcionó solo los valores críticos asintóticos) y estudiando su potencia y

probabilidad de error de tipo I. El contraste de Cucconi se basa en

$$CUC = \frac{U^2 + V^2 - 2\rho UV}{2(1 - \rho^2)},$$

donde

$$U = \frac{6 \sum_{i=1}^{n_1} R_{1i}^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}},$$

$$V = \frac{6 \sum_{i=1}^{n_1} (n+1 - R_{1i})^2 - n_1(n+1)(2n+1)}{\sqrt{n_1 n_2 (n+1)(2n+1)(8n+11)/5}},$$

$n_1 + n_2$ ,  $R_{ji}$  denotan los rangos de  $X_{ji}$  en la muestra  $X = (X_1, X_2)$  y

$$\rho = \frac{2(n^2 - 4)}{(2n+1)(8n+11)} - 1,$$

Como se puede observar  $U$  se basa en los cuadrados de los rangos  $R_{1i}$ , mientras que  $V$  se basa en los cuadrados de los rangos contrarios  $(n+1 - R_{1i})$  de la primera muestra. En ausencia de empates se verifica  $U' = -U$  y  $V' = -V$  donde  $U'$  y  $V'$  son  $U$  y  $V$  calculados en la segunda muestra y luego no importa si uno actúa en la primera o segunda muestra para calcular el estadístico  $CUC$ . En presencia de empates, el estadístico  $CUC$  calculado en la primera y segunda muestra da lugar a valores muy ligeramente diferentes. Este aspecto ya había sido observado por Cucconi en el año 1968 (véase referencias en [1]) y recientemente por Neuhäuser (2012)[5]. Para fines prácticos, lo mejor sería promediar los dos valores del estadístico  $CUC$ , como lo hace la función  $R$  que realiza el contraste de Cucconi que se considerará más adelante.

Bajo  $H_0 : \{\mu_1 = \mu_2 \cap \sigma_1 = \sigma_2\}$ ,  $\mathbb{E}(U) = \mathbb{E}(V) = 0$  y  $\mathbb{V}(U) = \mathbb{V}(V) = 1$  porque

$$\mathbb{E} \left( \sum_{i=1}^{n_1} R_{1i}^2 \right) = n_1(n+1)(2n+1)/6$$

y

$$\mathbb{V} \left( \sum_{i=1}^{n_1} R_{1i}^2 \right) = n_1 n_2 (n+1)(2n+1)(8n+11)/180.$$

Por supuesto

$$\mathbb{E} \left( \sum_{i=1}^{n_1} (n+1 - R_{1i})^2 \right) = \mathbb{E} \left( \sum_{i=1}^{n_1} R_{1i}^2 \right), \mathbb{V} \left( \sum_{i=1}^{n_1} (n+1 - R_{1i})^2 \right) = \mathbb{V} \left( \sum_{i=1}^{n_1} R_{1i}^2 \right),$$

y  $U$  y  $V$  son negativamente dependientes. Más precisamente,  $\text{Cor}(U, V)$  toma valores en el intervalo  $[-1, -7/8]$ . En efecto:

$$\text{Cor}(U, V) = \text{Cov}(U, V) = \frac{2(n^2 - 4)}{(2n+1)(8n+11)} - 1 = \rho,$$

el mínimo  $-1$  se alcanza en  $n = 2$ , mientras que el máximo se consigue haciendo tender  $n$  a infinito, es decir,

$$\lim_{n \rightarrow \infty} \rho = \frac{2(n^2 - 4)}{(2n+1)(8n+11)} - 1 = -\frac{7}{8} = \rho_0.$$

Bajo  $H_0$ ,  $(U, V)$  está centrado en  $(0, 0)$ , mientras que esto no se verifica bajo  $H_1$ .

Cuando  $\mu_1 \neq \mu_2$  y  $\sigma_1 = \sigma_2$ ,  $\sum_{i=1}^{n_1} R_{1i}^2$  tiende a ser más grande (menor) que  $n_1(n+1)(2n+1)/6$  cuando  $\mu_1 > \mu_2$  ( $\mu_1 < \mu_2$ ) por lo que  $U$  tiende a ser mayor (menor) que 0 mientras que  $V$  tiende a ser menor (mayor) que 0.

Cuando  $\mu_1 = \mu_2$  y  $\sigma_1 > \sigma_2$ , el rango de los elementos de la primera muestra tiende a los elementos extremos de la secuencia  $1, 2, \dots, n$ , mientras que cuando  $\sigma_1 < \sigma_2$ , estos tienden al valor medio.

En el primer (segundo) caso,  $\sum_{i=1}^{n_1} R_{1i}^2$  tiende a ser mayor (menor) que  $n_1(n+1)(2n+1)/6$ , luego tanto  $U$  como  $V$  tienden a ser mayor (menor) que 0. Cuando  $\mu_1 \neq \mu_2$  y  $\sigma_1 \neq \sigma_2$ ,  $(U, V)$ , de nuevo no está centrado en  $(0, 0)$ , por ejemplo, cuando  $\mu_1 > \mu_2$  y  $\sigma_1 < \sigma_2$ ,  $\mathbb{E}(U)$  puede estar cercano

a 0 mientras que  $\mathbb{E}(V)$  tiende a ser menor que 0. Además se verifica que si  $n_1, n_2 \rightarrow \infty$  y  $n_1/n \rightarrow \lambda \in (0, 1)$  entonces  $Pr(U \leq u) \rightarrow \Phi(u)$ , donde  $\Phi$  es la función de distribución normal estándar. Un resultado análogo se le aplica también a  $V$ . Además, si  $n_1, n_2 \rightarrow \infty$  y  $n_1/n \rightarrow \lambda \in (0, 1)$  entonces se verifica

$$Pr(U \leq u, V \leq v) \rightarrow \int_{-\infty}^u \int_{-\infty}^v \frac{1}{2\phi\sqrt{1-\rho_0^2}} \exp\left(-\frac{q^2 + r^2 - 2\rho_0qr}{2(1-\rho_0^2)}\right) dqdr.$$

Los puntos  $(u, v)$  dentro de la región de aceptación están cercanos a  $(0, 0)$ , esto quiere decir que se satisface

$$\frac{1}{2\phi\sqrt{1-\rho_0^2}} \exp\left(-\frac{u^2 + v^2 - 2\rho_0uv}{2(1-\rho_0^2)}\right) \geq k,$$

donde la constante  $k$  se elige para que la tasa de error de tipo uno sea  $\alpha$ . Sea  $k = \alpha(2\phi\sqrt{1-\rho_0^2})^{-1}$ , entonces se deduce que  $H_0$  debería ser aceptada si el punto  $(u, v)$  es tal que

$$\frac{u^2 + v^2 - 2\rho_0uv}{2(1-\rho_0^2)} < -\ln(\alpha).$$

Cabe señalar que la región de aceptación  $E$  del contraste es el conjunto de puntos  $(u, v)$  en el interior de la elipse

$$u^2 + v^2 - 2\rho_0uv = -2(1-\rho_0^2)\ln(\alpha).$$

Ya que es

$$\int \int_E \frac{1}{2\phi\sqrt{1-\rho_0^2}} \exp\left(-\frac{q^2 + r^2 - 2\rho_0qr}{2(1-\rho_0^2)}\right) dqdr = 1 - \alpha$$

el tamaño del contraste es  $\alpha$ . En la práctica, a no ser que estemos en presencia

de tamaños de muestras grandes,  $\rho_0$  puede ser reemplazada por  $\rho$ .

Cucconi en el año 1968 (véase referencias en [1]) observó que la tasa de convergencia a la normalidad era muy buena cuando  $n_1, n_2 > 6$  con  $n_1$  y  $n_2$  no muy diferentes.

Este contraste, al igual que el contraste de Lepage, es muy interesante en el campo de la medicina. En este caso, podemos considerar de nuevo el ejemplo citado en la subsección anterior. Para llevar a cabo el contraste de Cucconi, en primer lugar se calcula el valor observado  $CUC^0$  del estadístico de Cucconi. Posteriormente se simula la distribución de las permutaciones  $CUC_b^*, b = 1, \dots, B$  del estadístico tomando una muestra aleatoria grande, por ejemplo  $B = 10000$  permutaciones, de la muestra conjunta  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$  y luego calcular el correspondiente  $p$ -valor

$$\frac{1}{B} \sum_{b=1}^B \mathbb{I}(CUC_b^* \geq CUC^0).$$

Para llevar a cabo el contraste usando  $R$  hay que descargar el archivo `cucconi_test.r` que contiene la función `cucconi(x1,x2,B)` donde `x1` y `x2` son las muestras y `B` es el número de permutaciones. Cabe destacar que al igual que el contraste de Lepage, no hay distinción entre la hipótesis alternativa, esta es siempre la negación lógica de la hipótesis nula. La salida de la función es el valor observado  $CUC^0$  y el  $p$ -valor del contraste.

## 2.4. Comparación de distribuciones

El problema de comparar dos muestras se ha abordado en la sección 2.3 conjuntamente para contrastes con cambios de localización o escala. En esta sección, el problema se aborda desde el punto de vista más general, de hecho, estamos interesados en probar si existen diferencias entre las distribuciones

de las poblaciones subyacentes a las muestras. Los contrastes llevados a cabo en esta sección son sensibles a cualquier tipo de diferencias entre las funciones de distribución asociadas a las dos poblaciones, incluidas las diferencias en la forma, y por lo tanto en curtosis y/o asimetría. Por este motivo, cuando hay posibilidad de diferencias en la forma o en la localización y/o escala (entre las dos distribuciones) se sugiere utilizar los contrastes presentados en las siguientes subsecciones.

#### 2.4.1. Contraste de Kolmogorov-Smirnov

Un contraste familiar para el problema de dos muestras es el contraste de Kolmogorov-Smirnov, el cual requiere la suposición de las siguientes propiedades:

Dadas  $n = n_1 + n_2$  observaciones. Sean  $\mathbf{X}_1 = \{X_{1i}, i = 1, \dots, n_1\}$  y  $\mathbf{X}_2 = \{X_{2i}, i = 1, \dots, n_2\}$

- $\mathbf{X}_1$  y  $\mathbf{X}_2$  son muestras aleatorias que provienen de poblaciones continuas 1 y 2.
- $\mathbf{X}_1$  y  $\mathbf{X}_2$  son mutuamente independientes.

Además, este contraste está basado en las diferencias entre las funciones de distribución empíricas de ambas muestras. Si las variables aleatorias  $X_1$  y  $X_2$  subyacentes a las muestras son continuas, el contraste es exacto. En el caso discreto, o bien en presencia de empates, el test es conservativo, lo cual significa que su nivel de significación es menor o igual que  $\alpha$ . En otras palabras, la hipótesis nula del contraste no se rechaza con facilidad, para disminuir así el error de tipo I. Sean  $F_1$  y  $F_2$  las funciones de distribución correspondientes a las poblaciones 1 y 2, respectivamente. Por otra parte,  $\hat{F}_1$  y  $\hat{F}_2$  denotan las funciones de distribución empíricas de las muestras  $\mathbf{X}_1$

y  $\mathbf{X}_2$ . Por lo tanto, se tiene

$$\hat{F}_j(x) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{I}X_{1i} \leq x, j = 1, 2.$$

El estadístico difiere según el sistema de hipótesis (unilateral o bilateral) de interés. En el caso bilateral, la hipótesis nula es

$$H_0 : \{F_1(x) = F_2(x), \forall x \in \mathcal{R}\}$$

y la hipótesis alternativa es

$$H_1 : \{F_1(x) \neq F_2(x), \text{ para algún } x \in \mathcal{R}\}.$$

El estadístico del contraste es la norma del supremo de la diferencia entre  $\hat{F}_1$  y  $\hat{F}_2$ , es decir

$$KS = \sup_{x \in \mathcal{R}} |\hat{F}_1(x) - \hat{F}_2(x)|.$$

En el caso unilateral, cabe distinguir dos situaciones. La primera situación se presenta cuando la primera variable es estocásticamente mayor que la segunda en la hipótesis alternativa. En este caso, se tiene

$$H_0 : \{F_1(x) \leq F_2(x), \forall x \in \mathcal{R}\}$$

contra

$$H_1 : \{F_1(x) \geq F_2(x), \text{ para algún } x \in \mathcal{R}\}.$$

El estadístico del contraste es la máxima distancia vertical alcanzada por  $\hat{F}_1$  por encima de  $\hat{F}_2$ , es decir,

$$KS^+ = \sup_{x \in \mathcal{R}} (\hat{F}_1(x) - \hat{F}_2(x)).$$

La segunda situación sucede cuando la primera variable es estocásticamente menor que la segunda en la hipótesis alternativa. En este caso tenemos

$$H_0 : \{F_1(x) \geq F_2(x), \forall x \in \mathcal{R}\}$$

contra

$$H_1 : \{F_1(x) \leq F_2(x), \text{ para algún } x \in \mathcal{R}\}.$$

El estadístico del contraste es la máxima distancia vertical alcanzada por  $\hat{F}_2$  por encima de  $\hat{F}_1$ , es decir,

$$KS^- = \sup_{x \in \mathcal{R}} (\hat{F}_2(x) - \hat{F}_1(x)).$$

En la práctica, pueden ocurrir empates. Como la función de distribución empírica está bien definida en presencia de empates, los estadísticos de Kolmogorov-Smirnov no necesitan ningún ajuste. La única deficiencia es que el contraste se vuelve más conservador.

Por otra parte, dado que  $\hat{F}_1$  y  $\hat{F}_2$  son estimadores de  $F_1$  y  $F_2$ , respectivamente, los estadísticos de Kolmogorov-Smirnov pueden verse como estimadores de la norma del supremo de la diferencia entre  $F_1$  y  $F_2$ , que es 0 cuando la hipótesis nula es cierta. Los valores grandes del estadístico favorecen la hipótesis alternativa especificada.

Es importante tener en cuenta que el contraste de Kolmogorov-Smirnov, bajo los supuestos citados anteriormente es consistente contra cualquier diferencia entre  $F_1$  y  $F_2$ . La distribución nula exacta del estadístico de Kolmogorov-Smirnov se obtiene calculando el estadístico para todos los posibles ordenamientos de  $\mathbf{X}_1$  y  $\mathbf{X}_2$ , todos ellos igualmente probables bajo la hipótesis nula.

Cabe destacar que, en la práctica, para calcular un estadístico de Kolmogorov-Smirnov apropiado, podemos calcular las distancias en norma del supremo entre las funciones de distribución empíricas en las observaciones muestrales ordenadas en la muestra combinada  $X_{(1)} \leq \dots \leq X_{(n)}$ , en lugar de para todo  $x \in \mathcal{R}$ :

$$KS = \max_{i=1, \dots, n} |\hat{F}_1(X_{(i)}) - \hat{F}_2(X_{(i)})|$$

$$KS^+ = \max_{i=1, \dots, n} (\hat{F}_1(X_{(i)}) - \hat{F}_2(X_{(i)}))$$

$$KS^- = \max_{i=1, \dots, n} (\hat{F}_2(X_{(i)}) - \hat{F}_1(X_{(i)})).$$

### Ejemplo de las calificaciones: continuación

Disponemos de dos muestras correspondientes a las calificaciones obtenidas en teoría de los grupos A y B (ver Tabla 2.3).

Por otra parte, al observar los Histogramas 2.2 de densidad de las dos muestras, parece que las distribuciones de los resultados de los estudiantes perteneciente a cada uno de los grupos no presentan notables diferencias si únicamente nos limitamos al estudio cualitativo de ambas muestras. Para abordar un estudio cuyos resultados sean más consistentes y fiables podemos aplicar el contraste de Kolmogorov-Smirnov.

Grupo A									
3	5.5	3.5	0	5.5	6	3.5	6.5	5.5	7.5
4	5	5.5	4.5	3.5	2	5.5	6	8	6
Grupo B									
4	7	9	8	4.5	6	4	5.5	6.5	8
2.5	1	3	3.5	7.5	6	7.5	6		

Tabla 2.3: Calificaciones teóricas por grupo durante el curso académico 2015-16.

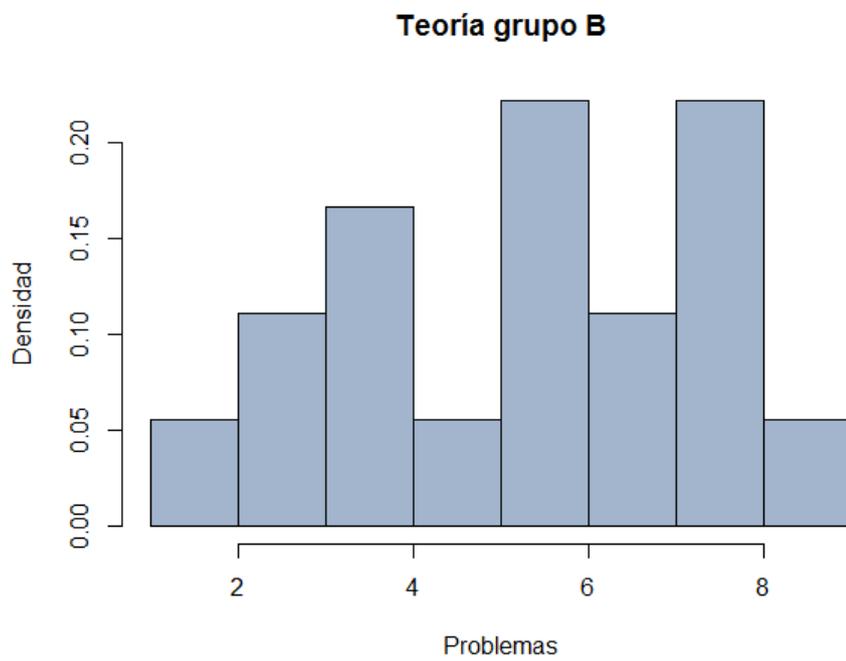
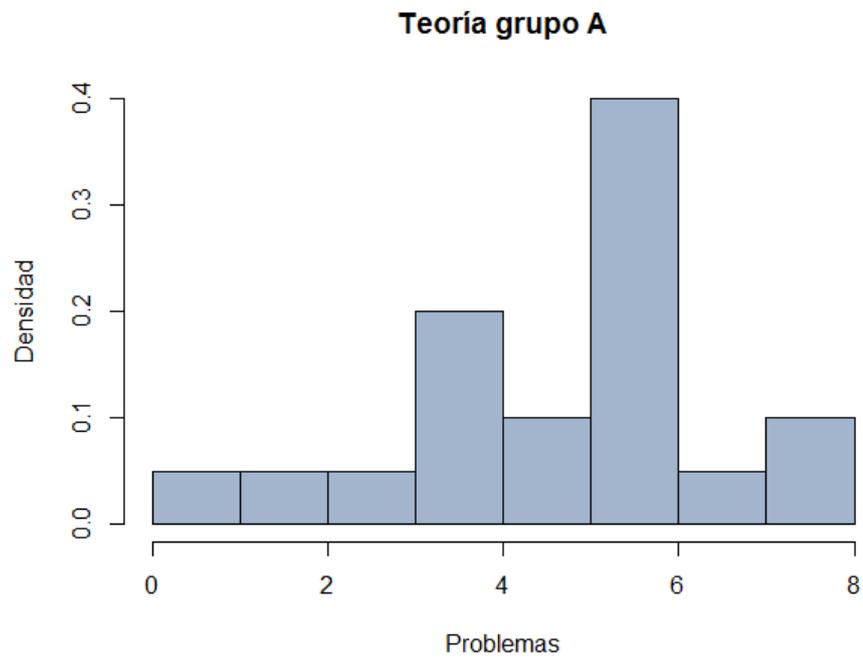


Figura 2.2: Histogramas de frecuencias de las calificaciones teóricas por grupos durante el curso académico 2015-16.

El paquete básico de *R* contiene la función `ks.test` que calcula tanto el estadístico de Kolmogorov-Smirnov, así como el *p*-valor. En el caso bilateral, si  $n_1 n_2 < 10000$  y en ausencia de empates, el *p*-valor calculado es exacto. En otros casos, se usan distribuciones asintóticas. Para el caso de muestras pequeñas, el contraste puede ser muy inexacto. Por otro lado, cabe señalar que en presencia de empates, *R* genera una advertencia. El código *R* utilizado es el siguiente:

```
teoriaA=c(3,5.5,3.5,0,5.5,6,3.5,6.5,5.5,7.5,4,5,5.5,4.5,3.5,2,
          5.5,6,8,6)
teoriaB=c(4,7,9,8,4.5,6,4,5.5,6.5,8,2.5,1,3,3.5,7.5,6,7.5,6)
ks.test(teoriaA,teoriaB,alternative="two.sided")
```

La salida de *R* es:

```
Two-sample Kolmogorov-Smirnov test
```

```
data:  teoriaA and teoriaB
D = 0.25556, p-value = 0.5661
alternative hypothesis: two-sided
```

```
Warning message:
```

```
In ks.test(teoriaA, teoriaB, alternative = "two.sided") :
cannot compute exact p-value with ties
```

En primer lugar, como en ocasiones anteriores, contamos con la presencia de una advertencia debido a la presencia de empates en las calificaciones. Esto se debe al redondeo a la décima. Por este motivo, el programa *R* devolverá un *p*-valor aproximado, por lo que el contraste perderá un poco de efectividad. Por otra parte, el programa nos proporciona el valor del estadístico, así

como el  $p$ -valor del contraste. Este último tiene un valor de 0,5661 lo cual supone que no existe evidencia para rechazar la hipótesis nula, de manera que la función de distribución en el grupo A es la misma que en el grupo B. En otras palabras, las calificaciones en ambos grupos se distribuyen de manera similar por lo que podemos pensar que el nivel de conocimiento en ambos grupos es el mismo.

Cabe destacar que el contraste de Kolmogorov-Smirnov es muy complejo pues resulta muy complicado analizar los resultados obtenidos en el caso en que las distribuciones en ambos grupos sean diferentes. En este tipo de análisis, cuando se obtiene este resultado, no podemos detectar la naturaleza de esta diferencia, sino que debemos de hacer un estudio más detallado para detectar la presencia en cambios de localización o de escala.

En el caso en que no hay evidencia para rechazar la hipótesis de igualdad de distribuciones (nuestro caso), el análisis es todo un éxito pues hemos conseguido bastante más información que en otros contrastes en los que únicamente se comparen parámetros de localización y/o escala.

#### **2.4.2. Contraste de Cramér-von Mises**

Esta última sección se dedicará al análisis de otro contraste familiar para problemas de dos muestras: el contraste de Cramér-von Mises. En muchas ocasiones éste es preferido al contraste de Kolmogorov-Smirnov debido a que lleva a cabo un uso más eficiente de los datos. Sin embargo posee una desventaja con respecto a este último, y es que el contraste de Cramér-von Mises es ligeramente más difícil de calcular.

Por otra parte, cabe señalar que se trata de un contraste bilateral, por

lo tanto, para el desarrollo del contraste se tomarán las siguientes hipótesis

$$H_0 : \{F_1(x) = F_2(x), \forall x \in \mathcal{R}\}$$

contra

$$H_1 : \{F_1(x) \neq F_2(x), \text{ para algún } x \in \mathcal{R}\}.$$

Además, el contraste de Cramér-von Mises requiere las mismas hipótesis que el contraste de Kolmogorov-Smirnov. Existen diferentes versiones para este contraste, aquí vamos a considerar la versión basada en la norma  $L_1$ , estudiada por Xiao et al. en 2006 (véase referencias en [1]). Se trata de una versión que tiene casi la misma potencia como que la que utiliza la norma  $L_2$ , pero es computacionalmente más eficiente. El estadístico para el contraste estudiado es el siguiente:

$$CVM = \frac{(n_1 n_2)^{1/2}}{n^3/3} \left( \sum_{i=1}^{n_1} |\hat{F}_1(X_{1i}) - \hat{F}_2(X_{1i})| + \sum_{i=1}^{n_2} |\hat{F}_1(X_{2i}) - \hat{F}_2(X_{2i})| \right).$$

Es interesante señalar que el estadístico  $CVM$  es una especie de estimación de la integral de las diferencias de las funciones  $\hat{F}_1$  y  $\hat{F}_2$  en valor absoluto. Valores grandes del estadístico nos da evidencia de la diferencia entre las funciones de distribución de ambas poblaciones.

El fundamento del contraste de Cramér-von Mises es similar al del contraste de Kolmogorov-Smirnov. Ambos contrastes son muy útiles cuando no hay evidencia de rechazo de  $H_0$  mientras que, en el caso contrario, al tratarse de contrastes muy complejos, poseen una muy difícil interpretación en el caso en que la hipótesis nula sea rechazada, dado que no podemos saber de dónde provienen estas diferencias. En este caso, se deberían hacer estudios más concretos para detectar la naturaleza de estas diferencias.

La distribución nula exacta del estadístico de Cramér-von Mises se obtiene calculando el estadístico para todos los posibles ordenamientos de  $\mathbf{X}_1$  y  $\mathbf{X}_2$ , todos ellos igualmente probables bajo la hipótesis nula.

# Bibliografía

- [1] S. Bonnini and L. Corain and M. Marozzi and L. Salmaso. Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R. John Wiley and Sons, 2014.
- [2] BAGDONAVICIUS,V.,KRUOPIS,J.,NIKULIN,M.S. *Non-parametric Tests for Complete Data*. John Wiley & Sons, Ltd, 2011.
- [3] MAROZZI,M. (2009). *Some notes on the location-scale cuconi test*. *Journal of Nonparametric Statistics*, 21, 5, 629-647.
- [4] MAROZZI,M. (2011). *Levene type for the ratio of two scales*. *Journal of Statistical Computation and Simulation*, 81, 815-826.
- [5] MAROZZI,M. (2013). *Nonparametric simultaneous test for location and scale testing: a comparison of several methods*. *Communication in Statistics-Simulation and Computation*, 42, 1298-1317.
- [6] NEUHÄUSER,M. *Nonparametric Statistical Tests: a Computational Approach*. CRC Press, 2012.
- [7] PESARIN,F.,SALMASO,L. *Permutation Tests for Complex Data:Theory , Applications and Software*. John Wiley & Sons, Ltd., 2010.