Estadística Básica con R y R–Commander

(Versión Febrero 2008)

Autores:
A. J. Arriaza Gómez
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
S. Pérez Plaza
A. Sánchez Navas

UCA Universidad de Cádiz
Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN:

Depósito legal:

Estadística Básica con R y R-commander
(Versión Febrero 2008)
Autores: A. J. Arriaza Gómez, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, S. Pérez Plaza, A. Sánchez Navas

http://knuth.uca.es/ebrcmdr

Capítulo 6

Inferencia no paramétrica. Diagnosis del modelo

En este capítulo se aborda en primer lugar la realización de contrastes sobre la calidad de la muestra, a continuación se estudian test de bondad de ajuste, haciendo especial énfasis en los de normalidad y, por último, se dan alternativas no paramétricas para el caso de que las poblaciones no sean normales.

1. Pruebas de aleatoriedad

En esta sección se abordará el estudio de la calidad de la muestra extraída de la población, y aunque el procedimiento de obtención debería garantizar unos niveles mínimos de calidad, lo cierto es que en ocasiones los datos vienen impuestos sin que el investigador haya podido supervisar el procedimiento de extracción. No obstante y como en todo contraste, debe tenerse en cuenta que el test sólo desestimará la hipótesis si la evidencia muestral en su contra es muy fuerte.

En ocasiones, los elementos de la muestra se han obtenido en un marco territorial o temporal. Imagine por ejemplo mediciones de una cierta magnitud económica a lo largo de un periodo de tiempo o niveles de un determinado elemento químico en estudios de contaminación, bien en aire, agua o tierra. En estas situaciones es de esperar que las mediciones tomadas en un cierto entorno tengan ciertas analogías o pre-

senten tendencias. Para estudiar este tipo de situaciones se debe acudir a modelos específicos, como son las series temporales o los modelos geoespaciales, en ambos casos existe un elemento que sirve de variable de referencia o longitudinal: la fecha o el posicionamiento gps. Sin embargo, en otras situaciones donde no se contempla esa variable de referencia, las personas encargadas de realizar el muestreo, por comodidad o descuido, no adoptan las medidas para garantizar la independencia de las mediciones.

Ejemplo 6.1

Para analizar si existe autocorrelación entre los elementos de una muestra, se consideran los datos del PIB en billones de euros durante los últimos diez años: 13, 14, 18, 21, 22, 19, 20, 23, 27 y 30. Parece que debería existir influencia del PIB de años precedentes sobre los posteriores. Para comprobarlo se aplicará el test de autocorrelación de Ljung-Box, *contemplando autocorrelaciones de primer y segundo orden. Para la de primer orden, se fija la opción* `lag=1`.

```
> x<- c(13, 14, 18, 21, 22, 19, 20, 23, 27, 30)
> Box.test(x, lag = 1, type = c(``Ljung-Box''))
Box-Ljung test
data: x
X-squared = 4.2281, df = 1, p-value = 0.03976
```

Lo que indica, dado que $p = 0,03976$, *que para un* $\alpha = 0,05$ *se rechazaría la hipótesis de indepedencia lineal de primer orden, por lo que el valor del PIB del año* T *influye sobre la del año* $T + 1$. *Si se analiza la correlación de segundo orden,* `lag=2`, *se tiene:*

```
> Box.test(x, lag = 2, type = c(``Ljung-Box''))
Box-Ljung test
data: x
X-squared = 4.4046, df = 2, p-value = 0.1105
```

En esta ocasión y puesto que $p > 0,05$ *no se rechaza la hipótesis de independencia y se descarta la autocorrelación de segundo orden.*

Otra perspectiva desde la que analizar la aleatoriedad de la muestra, si ésta viene dada en forma de variable binaria, es comprobar si existen muy pocas o muchas rachas, entendiendo por racha al grupo de

valores consecutivos iguales interrumpido por uno de signo distinto. Si la variable no es de tipo binario, se la puede transformar para que lo sea asignando las clases de la dicotomía en función de que el elemento muestral esté por encima o por debajo de un determinado valor, típicamente la mediana.

Ejemplo 6.2

Para analizar la independencia de los mismos datos del PIB del ejemplo anterior se aplicará ahora el test de rachas. Previamente habrá que cargar el paquete `tseries` *de series temporales, bien desde el menú o con la instrucción* `library(''tseries'')`*. En este caso se realizará un contraste bilateral, rechazándose la hipótesis nula tanto si existen muchas rachas como si hay muy pocas, aunque las opciones de la función de* ***R*** *admitirían que se especificaran contrastes de carácter unilateral.*

```
> runs.test(as.factor(x>median(x)))
Runs Test
data: as.factor(x > median(x))
Standard Normal = -1.3416, p-value = 0.1797
alternative hypothesis: two.sided
```

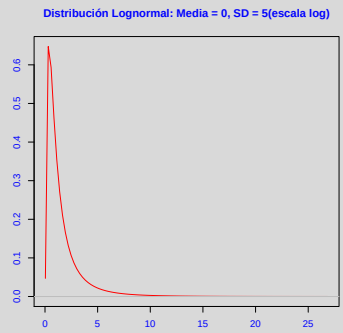
Con la orden `as.factor(x>median(x))` *se convierte a la variable* `x` *en dicotómica, dando códigos distintos en función de que el valor esté por debajo o por encima de la mediana (20,5). La salida del procedimiento indica, puesto que* $p > 0,05$*, que no hay evidencias para considerar los datos no aleatorios.*

2. Pruebas de bondad de ajuste

En este epígrafe se contrastará si la estructura de la población analizada se ajusta a una determinada distribución. En principio el procedimiento de obtención de la información deberá ofrecer pautas para decidir si la población tiene una u otra estructura probabilística. Así, en el caso que más nos interesa, si la variable se genera a partir de la medición objetiva de alguna característica, ésta será en general normal; la excepción se dará cuando se haya considerado un conjunto de individuos no homogéneos, mezclando grupos de edad, sexos, ... Si realmente

se han mezclado grupos de individuos, un análisis exploratorio arrojará una estructura probabilística multimodal, mientras que si, por el contrario, la población física es homogénea, la distribución presentará, si acaso, problemas de simetría; en algunas ocasiones estos problemas se pueden solucionar mediante transformaciones de los datos. También puede darse la circunstancia de que distribuciones que converjan a la normal en situaciones ideales y para muestras grandes, como es el caso de la binomial o la Poisson, necesiten alguna transformación para mejorar la simetría. Se analizará esta cuestión en el capítulo de Análisis de la Varianza. Por último, hay que indicar que en muchas ocasiones hay que realizar una operación de truncamiento para adaptar la distribución teórica al rango de valores de los datos en estudio.

Ejemplo 6.3



En problemas ecológicos es muy habitual que la abundancia de una especie tenga una distribución de tipo lognormal respecto a los parámetros ambientales, por tanto una transformación logarítmica convertiría a la abundancia en una variable normal. Como se puede ver, no se trata de una medición de una característica de los individuos, sino de una medida de su abundancia respecto a una variable ambiental.

A continuación se presentará un contraste específico de normalidad, como es el test de Shapiro-Wilk, y un par de test genéricos para evaluar la bondad del ajuste, uno para cuando los datos son continuos, el de Kolmogorov-Smirnov, y otro para variables categóricas, el test de la χ^2. En el caso de contrastes de normalidad, se recomienda el uso del test de Shapiro-Wilk para muestras pequeñas $n \leq 50$, mientras que si las muestras son grandes es preferible utilizar el test de Kolmogorov-Smirnov, salvo que los datos vengan dados en una distribución de frecuencias por intervalos donde se empleará la χ^2.

Ejemplo 6.4

El archivo de datos que se utilizará en este ejemplo es el `caracoles.dat` *que incluye las mediciones de dos variables, diámetro de las conchas (mm) y separación entre las espirales (μm), para un conjunto de 20 individuos adultos de una especie de caracoles. Dado el tamaño de la muestra, se contrastará la hipótesis de normalidad mediante el test de Shapiro-Wilk. Utilizando en este caso* **Rcmdr** *y marcando las opciones* `Estadísticos`→`Resúmenes`→ `Test de normalidad de Shapiro-Wilk...` *se obtiene el cuadro de diálogo, donde se selecciona la variable diámetro (*`Diam`*).*

En la ventana de resultados de **Rcmdr** *se tiene tanto la instrucción de* **R** *como la salida del procedimiento. En este caso el* `p-valor`$= 0,6869$ *viene a indicar que los datos se pueden considerar normales.*


```
>shapiro.test(Datos$Diam)
 Shapiro-Wilk normality test
data: Datos$Diam
W = 0.9668, p-value = 0.6869
```

Ejemplo 6.5

Se estudiará la normalidad de la variable peso del fichero `peso_altura.dat`*. Dado que el número de individuos es grande,* $n = 100$*, se utilizará el test de Kolmogorov-Smirnov. En primer lugar, con* **Rcmdr** *se calcula la media y la desviación típica del conjunto de datos, resultando* $\bar{x} = 73,37$ *y* $\sigma = 12,69$*. A continuación se computarán las diferencias entre la función de distribución empírica muestral y la distribución teórica* $N(73,37; 12,69)$*. Para ello se empleará el procedimiento* `ks.test`*.*

```
> ks.test(Datos$PESO,pnorm,73.37,12.69)
 One-sample Kolmogorov-Smirnov test
data: Datos$PESO
D = 0.136, p-value = 0.04939
alternative hypothesis: two-sided
```

En este caso y para un $\alpha = 0,05$ *se rechaza la hipótesis de que los pesos sigan una distribución normal.*

El test de Kolmogorov-Smirnov también se puede utilizar para comparar las distribuciones empíricas de dos conjuntos de datos, para ello en la instrucción se sustituiría la distribución a ajustar por la segunda variable.

Ejemplo 6.6

Se generan mediante instrucciones de **R** *dos muestras aleatorias de* 100 *y* 150 *elementos procedentes de distribuciones exponenciales de parámetros* 1 *y* 1,5*, respectivamente, mediante las instrucciones:*

```
x<-rexp(100,1); y<-rexp(150,1.5)
```

Aplicando de nuevo el test de Kolmogorov-Smirnov para comparar las funciones de distribución empírica de ambas muestras, se tendría:

```
>ks.test(x,y)
 Two-sample Kolmogorov-Smirnov test
data: x and y
D = 0.2833, p-value = 0.0001310
alternative hypothesis: two-sided
```

Se puede comprobar que el test rechaza la hipótesis de igualdad de funciones de distribución empíricas con un `p-valor`$= 0,00013$.

El análisis de la bondad de ajuste de una serie de datos a una distribución de probabilidad se estudia mediante el test de la chi-cuadrado de Pearson. Básicamente, el estadístico χ^2 evalúa las diferencias entre los valores observados y los valores ajustados por la ley de probabilidad. Se verán a continuación distintas situaciones y cómo se resuelven con **R**.

Ejemplo 6.7

Para contrastar si un dado no está trucado se lanza 60 *veces, obteniéndose los siguientes resultados:*

x_i	1	2	3	4	5	6
n_i	7	12	10	11	8	12

La hipótesis a contrastar es que $p_i = 1/6,\ \forall i$, *con lo que se tiene que* $E_i = 60(1/6) = 10,\ \forall i$.

Para resolver el contraste con **R** *basta introducir el vector de frecuencias,* $n = (7, 12, 10, 11, 8, 12)$, *y escribir las instrucciones de* **R**.

```
> n< −c(7,12,10,11,8,12)
>chisq.test(n)
 Chi-squared test for given probabilities
data: n
X-squared = 2.2, df = 5, p-value = 0.8208
```

A la vista del p-valor no se rechaza que el dado no está trucado.

El test Chi-cuadrado permite contrastar la hipótesis de independencia entre dos atributos organizados en tabla de contingencia.

Ejemplo 6.8

Se desea analizar la relación entre el nivel de estudios del padre y la orientación del alumno hacia las ciencias en un determinado instituto de bachillerato. Se cuenta para ello con la información obtenida en el centro.

	Estudios padre			
Orientación	*Ninguno*	*Básico*	*Medio*	*Superior*
Orientado	*23*	*12*	*34*	*32*
No orientado	*18*	*42*	*16*	*27*

Para contrastar esta relación se introduce la matriz de datos en **Rcmdr** *como se describe en el ejemplo 3.1, obteniéndose los siguientes resultados:*

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data: .Table
X-squared = 24.1629, df= 3, p-value = 2.31e-05
```

Lo que indica que se rechaza la hipótesis de independencia y existe una relación entre los estudios de los padres y la orientación hacia las ciencias de sus hijos.

Para el caso de tablas 2×2 se aplica el *test exacto de Fisher*, aunque existe la alternativa de aplicar el test Chi-cuadrado con la corrección de Yates. Para aplicar esta corrección bastaría especificar,`correct=TRUE`, en la instrucción de dicho test.

Ejemplo 6.9

En el conservatorio de música de una ciudad se pretende estudiar la relación existente entre el sexo del alumnado y su afición por los instrumentos de viento. Para ello, observados los 482 estudiantes se tiene:

	Hombre	*Mujer*
Aficionado	*150*	*97*
No aficionado	*123*	*112*

Se introduce la matriz de datos de la misma forma que en el ejemplo 3.1 seleccionando la opción de `Prueba exacta de Fisher`

```
>fisher.test(.Table)
Fisher's Exact Test for Count Data
data: .Table
p-value = 0.06655
alternative hypothesis: true odds ratio is not equal to 1
```

Por lo que para un nivel de significación $\alpha = 0,05$ no se rechaza, aunque con poca evidencia, la hipótesis de independencia entre el sexo y la afición a los instrumentos de viento.

Se analizará ahora la bondad de ajuste de unos datos a una distribución teórica no uniforme.

Ejemplo 6.10

Durante la Segunda Guerra Mundial los alemanes bombardearon en diversas ocasiones Londres. Al objeto de analizar si los bombardeos eran indiscriminados o se hacían con intención, se procedió a dividir la ciudad en cuadrículas y a contar el número de impactos en cada una de ellas. Los resultados se recogen en la siguiente tabla

Impactos	0	1	2	3	4	5
Número cuadrículas	229	211	93	35	7	1

Las hipótesis podrían ser expresadas, en términos probabilísticos, de la siguiente manera

$$H_0: \quad X \sim P(\lambda)$$
$$H_1: \quad X \not\sim P(\lambda)$$

puesto que si las bombas caen indiscriminadamente, lo hacen de forma independiente en un soporte continuo. Lo que, de ser cierto, indicaría que la variable que mide el número de impactos por cuadrículas debe ser Poisson.

En primer lugar, se estimará el parámetro de la Poisson a partir de la media muestral, resultando que $\hat{\lambda} = 0,929$. *A continuación se calcularán las probabilidades* $P(X = i)$, *con* $i = 0, 1, 2, 3, 4$ *y* $P(X \geq 5)$ *mediante* ***Rcmdr***.

Las probabilidades discretas se obtienen en: `Distribuciones`→`Distribuciones discretas`→`Distribución de Poisson`→`Probabilidades de Poisson...` *tomando* `media`= 0,929.

```
>.Table
    Pr
0 0.3949
1 0.3669
2 0.1704
3 0.0528
4 0.0123
5 0.0023
6 0.0004
7 0.0000
```

La probabilidad $P(X \geq 5)$ *se obtiene desde:* `Distribuciones`→ `Distribuciones discretas`→`Distribución de Poisson`→

`Probabilidades de Poisson acumuladas...`, *tomando* `valor(es) de la variable`$= 4$ *ya que* **Rcmdr** *realiza* $P(X > 4)$=$P(X \geq 5)$, *para la cola de la derecha y* `media`$= 0{,}929$, *resulta:*

```
> ppois(c(4), lambda=0.929, lower.tail=FALSE)

 [1] 0.002682857
```

Con objeto de comprobar si se verifica la restricción de que todos los valores esperados deben ser mayores a tres, se calcula $n \cdot P[X \geq 5] = 576 \cdot 0,0027 = 1{,}5552 < 3$, *por lo que debe procederse a una agrupación de clases y considerar ahora* $P(X \geq 4)$. *Se obtiene que* $n \cdot P[X \geq 4] = 576 \cdot 0{,}015 = 8{,}64 > 3$.

Se almacenan ahora estas probabilidades en un vector `p`, *las frecuencias de los valores que toma la variable en otro vector* `x` *y se aplica el test chi-cuadrado resultando:*

```
>p< −c(0.3949,0.3669,0.1704,0.0528,0.0150)
>x< −c(229,211,93,35,8)
>chisq.test(x,p=p,rescale.p=TRUE)
 Chi-squared test for given probabilities
data: x
X-squared = 1.0205, df = 4, p-value = 0.9067
```

Por lo que se puede afirmar de forma contundente, dado el valor de p, que los bombardeos alemanes fueron indiscriminados.

3. Contrastes de localización y escala

Si se desestima la hipótesis de normalidad de los datos, no son aplicables los test vistos en el capítulo anterior basados en dicha distribución, siendo necesario utilizar contrastes no paramétricos. Este tipo de test se basan en el análisis de la situación de los elementos de la muestra respecto a determinadas medidas de posición, muy en especial respecto a la mediana. De esta forma, se estudia si los datos muestrales están por encima o por debajo de la mediana, es decir, se analiza el signo de su diferencia con la mediana; o bien, se estudia la distancia ordenada a la que se encuentra de la mediana, es decir, se considera el rango o la posición que ocupa dicho elemento en la secuencia ordenada de las diferencias.

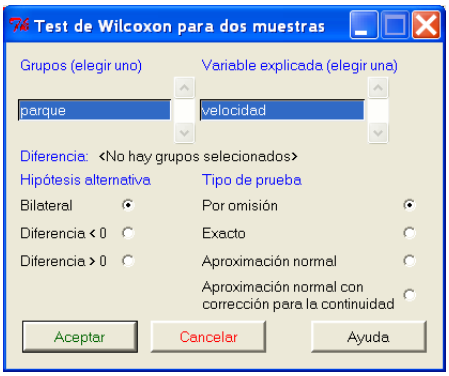


Figura 6.1: Test de Wilcoxon

En todo caso, las situaciones a analizar son las mismas del capítulo anterior: una muestra, dos muestras independientes y dos muestras apareadas, a las que se intentará dar respuesta con los ejemplos que siguen.

3.1. Dos muestras independientes

Ejemplo 6.11

Se estudiará mediante el test de Wilcoxon para muestras independientes si las dos ubicaciones del parque eólico, cuya información se encuentra en el archivo `eolico_apilado.dat`*, tienen la misma potencialidad eólica. Para ello, en el menú de* ***Rcmdr*** *se seleccionan las opciones de menú,* `Estadísticos`→`Test no paramétricos`→ `Test de Wilcoxon para dos muestras...`*, con lo que abre la ventana de diálogo 6.1.*

Seleccionados los únicos elementos de la base de datos, variable y factor, los resultados del análisis son:

```
> wilcox.test(velocidad~parque, alternative="two.sided",
data=Datos)
 Wilcoxon rank sum test with continuity correction
data: velocidad by parque
W = 276269.5, p-value = 0.2228
alternative hypothesis: true location shift is not equal to 0
```

Lo que implica el no rechazo de la hipótesis nula de igualdad de

medianas, siendo indistinta, desde esta óptica, la ubicación del parque eólico.

3.2. Una muestra

Ejemplo 6.12

Se desea contrastar la hipótesis nula, con $\alpha = 0,05$*, de que la separación mediana entre las espirales (variable* `Separ`*) de los caracoles del fichero* `caracoles.dat` *es menor o igual a* 110 μm*. Se supondrá que los datos son aleatorios pero no normales y se utilizará por tanto el test de Wilcoxon para una muestra. Trabajando directamente con* ***R*** *se tiene:*

```
> wilcox.test(Datos$Separ,alternative=c("greater"),mu=110)
 Wilcoxon signed rank test with continuity correction
data: Datos$Separ
V = 157, p-value = 0.006617
alternative hypothesis: true location is greater than 110
```

Por lo que se rechaza la hipótesis nula y se concluye que la separación mediana es superior a 110 μm*.*

3.3. Dos muestras pareadas

Ejemplo 6.13

Para documentar el caso de muestras pareadas se considera el mismo ejemplo que se usó en el capítulo anterior, la eficacia del tratamiento con fenofibrato, suponiendo ahora que la distribución de la diferencia de medias no es normal. En este caso se quiere probar la afirmación del fabricante de que el tratamiento durante un año con fenofibrato reduce el fibrinógeno en al menos 50 puntos. Se aplicará pues el `test de Wilcoxon para muestras pareadas`*. Para acceder al test, se ejecuta la secuencia de* ***Rcmdr****:*

Estadísticos→Test no paramétricos→Test de Wilcoxon para muestras pareadas...

Aunque las opciones de la ventana no admiten que se especifiquen diferencias, bastará con retocar mínimamente la instrucción añadiendo al final de la línea la opción `mu=50`.

```
> wilcox.test(Datos$FIB_A, Datos$FIB_D, alternative='greater',
paired=TRUE, mu=50)
 Wilcoxon signed rank test with continuity correction
data: Datos$FIB_A and Datos$FIB_D
V = 354, p-value = 0.01934
alternative hypothesis: true location shift is greater than 50
```

Así para $\alpha = 0,05$ *se rechaza la hipótesis de que* $\text{med}_A - \text{med}_D \leq 50$ *y se concluye que el medicamento produce una disminución de más de 50 puntos en el nivel de fenofibrato.*

4. Ejercicios

6.1 Contraste la normalidad de la variable separación entre las espirales (`Separ`) del fichero `caracoles.dat`.

6.2 Mediante el test de Kolmogorov-Smirnov, compruebe la hipótesis de igualdad de las funciones de distribución empírica de dos muestras de tamaño 200, procedentes de poblaciones N(0;1) y N(0;1,3) previamente generadas.

6.3 Compruebe la hipótesis de normalidad de la velocidad para cada una de las ubicaciones en el fichero `parque_eolico.dat`.

6.4 Contraste la hipótesis de que los datos siguientes, generados aleatoriamente mediante ordenador, procedan de una distribución Uniforme en el intervalo $[0,1]$ con un nivel de significación $\alpha = 0{,}05$.

$$0,582 \quad 0,501 \quad 0,497 \quad 0,026 \quad 0,132 \quad 0,561$$
$$0,642 \quad 0,994 \quad 0,948 \quad 0,081 \quad 0,179 \quad 0,619$$

6.5 En un grupo de 100 personas se estudian los atributos color del cabello (moreno, rubio y castaño) y color de los ojos (negro, marrón, azul y verde), obteniéndose la siguiente tabla de contingencia:

	Cabello		
Ojos	Moreno	Rubio	Castaño
Negros	20	8	4
Marrones	16	2	11
Azules	5	8	8
Verdes	10	5	3

¿Están relacionados dichos atributos?

6.6 Contraste si los datos de la siguiente muestra organizada como distribución de frecuencias proceden de una Normal.

$(L_{i-1}, L_i]$	n_i
$(0, 1]$	1
$(1, 2]$	3
$(2, 3]$	7
$(3, 4]$	12
$(4, 5]$	6
$(5, 6]$	2
$(6, 7]$	1

6.7 Estudie, utilizando el contraste χ^2 de bondad de ajuste, si la siguiente muestra de tamaño 30 procede de una Normal.

107	96	91	80	103	88	101	106	112	106
93	88	101	109	102	99	93	86	100	99
104	116	87	93	106	102	89	96	104	90

6.8 Con el fin de estudiar el tiempo de vida, en horas, de las baterías de 7 voltios, se extrae aleatoriamente un muestra de 10 de ellas, obteniéndose los siguientes resultados:

28.9	15.2	28.7	72.5	48.6
52.4	37.6	49.5	62.1	54.5

Proponga un modelo de distribución de probabilidad y estudie su ajuste.

6.9 Para medir la introversión se aplica a 12 individuos un test de personalidad en sus dos variantes, 1 y 2, que se supone la miden por igual. A partir de los datos de la siguiente tabla, compruebe mediante el test de rangos de Wilcoxon, con un nivel de significación del 5 %, si es cierto que las formas 1 y 2 miden por igual la introversión.

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
Forma 1	12	18	21	10	15	27	31	6	15	13	8	10
Forma 2	10	17	20	5	21	24	29	7	9	13	8	11

6.10 Para estudiar cuál de los dos tratamientos contra la artrosis es más eficaz se eligen aleatoriamente dos muestras de 10 y 22 pacientes

a los cuales se les somete a los tratamientos 1 y 2, respectivamente. Pasados tres meses se valoran ambos tratamientos de manera que el que tenga mayor puntuación será más eficaz. La tabla siguiente refleja los resultados obtenidos.

Tratamiento 1	12	15	21	17	38	42	10	23	35	28	
Tratamiento 2	21	18	42	25	14	52	65	40	43	35	18
	56	29	32	44	15	68	41	37	43	58	42

Utilice el test de Wilcoxon para evaluar si existen diferencias entre los dos tratamientos.