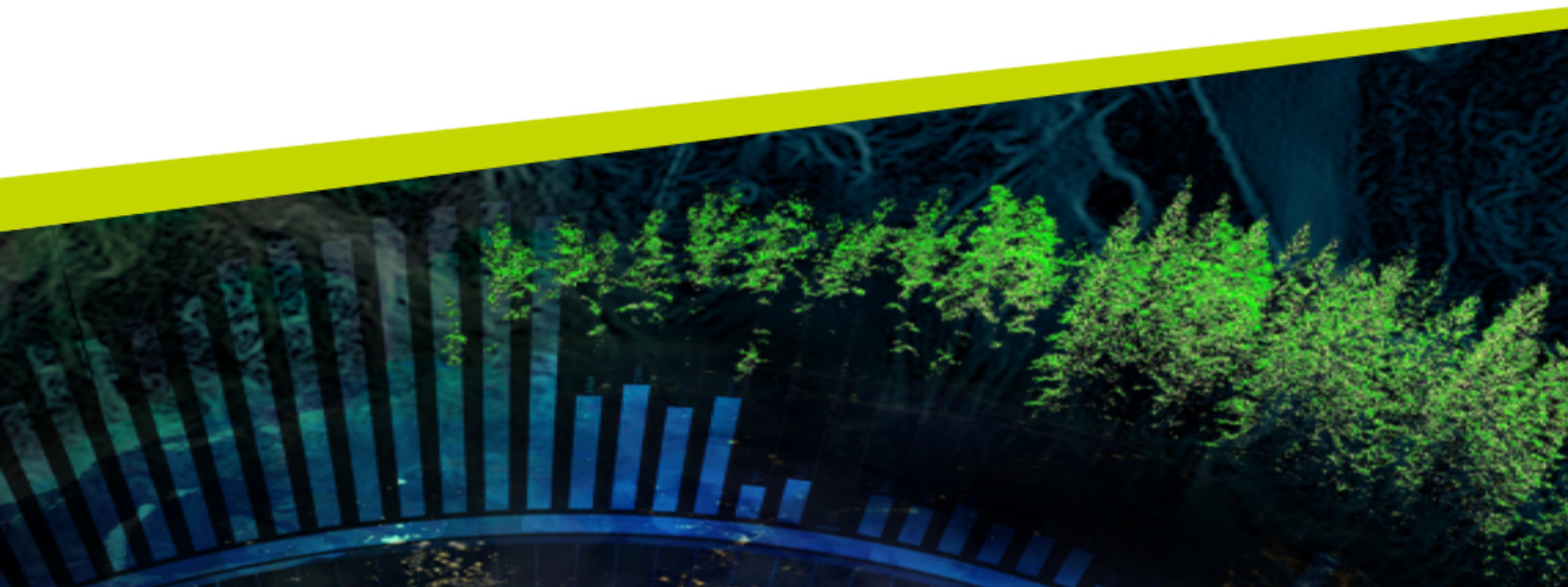
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 11. INFERENCIA NO PARAMÉTRICA CON MEDIANAS

En el capítulo 8 conocimos algunos métodos no paramétricos que podemos usar para inferir sobre frecuencias cuando nuestro conjunto de datos no cumple con las condiciones para poder usar, por ejemplo, las pruebas de Wald o de Wilson. Mencionamos también que este problema también puede ocurrir para el caso de inferir con medias, por lo que en este capítulo conoceremos alternativas no paramétricas para las pruebas t de Student (para una y dos medias) y ANOVA (para más de dos medias). Para ello nos basaremos principalmente en Lowry (1999, caps. 11a, 12a, 14a, 15a), Glen (2021) y Lærd Statistics (2020).

11.1 PRUEBAS PARA UNA O DOS MUESTRAS

En el capítulo 5 aprendimos que la prueba t de Student es adecuada para inferir acerca de una o dos medias muestrales, siempre y cuando se verifiquen algunas condiciones. En el caso de la prueba t de una muestra (o de la diferencia de dos muestras pareadas):

1. Las observaciones son independientes entre sí.
2. Las observaciones provienen de una distribución cercana a la normal.

En el caso de dos muestras independientes:

1. Cada muestra cumple las condiciones para usar la distribución t.
2. Las muestras son independientes entre sí.

Es importante mencionar también que la distribución normal es continua, de donde se desprende que la escala de medición empleada para la medición de las muestras debe ser de intervalos iguales.

Como ya vimos en el capítulo 8, si usamos la prueba t en un escenario en que no se cumple alguna de estas condiciones, el resultado no sería válido pues carecería de sentido y, en consecuencia, también lo harían las conclusiones que se obtengan a partir de él.

11.1.1 Prueba de suma de rangos de Wilcoxon

La **prueba de suma de rangos de Wilcoxon**, también llamada **prueba U de Mann-Whitney** o **prueba de Wilcoxon-Mann-Whitney**, es una alternativa no paramétrica a la prueba t de Student con muestras independientes. Pese a ser no paramétrica, requiere verificar el cumplimiento de las siguientes condiciones:

1. Las observaciones de ambas muestras son independientes.
2. La escala de medición empleada debe ser a lo menos ordinal, de modo que tenga sentido hablar de relaciones de orden (“igual que”, “menor que”, “mayor o igual que”).

Consideremos el siguiente contexto para estudiar la aplicación de esta prueba: una empresa de desarrollo de software desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto de software. Con este fin, la empresa ha seleccionado al azar a 23 voluntarias y voluntarios, quienes son asignados de manera

aleatoria a dos grupos, cada uno de los cuales debe probar una de las interfaces ($n_A = 12$, $n_B = 11$). Cada participante debe evaluar 6 aspectos de usabilidad de la interfaz, cada uno de los cuales se mide con una escala Likert de 7 puntos, donde 1 significa "muy malo" y 7, "muy bueno". La valoración que cada participante da a la interfaz evaluada corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 11.1 muestra las evaluaciones realizadas por cada participante.

	Interfaz A	**Interfaz B**
	2,7	5,0
	6,6	1,4
	1,6	5,6
	5,1	4,6
	3,7	6,7
	6,1	2,7
	5,0	1,3
	1,4	6,3
	1,8	3,7
	1,5	1,3
	3,0	6,8
	5,3	
Media	3,65	4,13

Tabla 11.1: evaluación de las interfaces de usuario A y B.

En este caso, si bien se cumple la condición de independencia de la prueba t de Student, no podemos usar esta prueba por dos razones: primero, no todas las escalas Likert pueden asegurar que son de igual intervalo. En el ejemplo, si dos participantes califican un aspecto de la interfaz A con notas 3 y 5, mientras que dos participantes califican esos aspectos con notas 4 y 6 para la interfaz B, ¿se podría asegurar que en ambos casos existe la misma diferencia de usabilidad (2 puntos)? Pocas escalas Likert tienen estudios de reproducibilidad que aseguren esta consistencia, por lo que no podríamos asumir que la escala es de intervalos iguales en este ejemplo. En segundo lugar, al revisar los histogramas para las muestras (figura 11.1) podemos observar que las distribuciones no se asemejan a una normal.

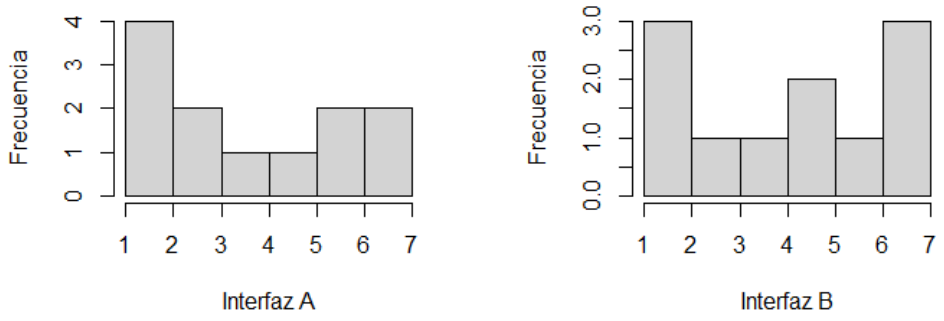


Figura 11.1: histogramas de las muestras.

Como alternativa, podemos usar la prueba no paramétrica de Wilcoxon-Mann-Whitney, cuyas hipótesis para el ejemplo son:

H_0: no hay diferencia en la usabilidad de ambas interfaces (se distribuyen de igual forma).
H_A: sí hay diferencia en la usabilidad de ambas interfaces (distribuciones distintas).

Al igual que en el caso de la prueba χ^2 de Pearson, estas hipótesis no hacen referencia a algún parámetro de

una supuesta distribución para las poblaciones de puntuaciones de usabilidad, es decir, nos entregan menos información que la prueba paramétrica equivalente.

El primer paso de la prueba consiste en combinar todas las observaciones en un único conjunto de tamaño $n_T = n_A + n_B$ y ordenarlo de menor a mayor. A cada elemento se le asigna un **valor de rango** (*rank* en inglés) de 1 a n_T, de acuerdo a la posición que ocupa en el conjunto ordenado. En caso de que un valor aparezca más de una vez, cada repetición toma como valor el rango promedio de todas las ocurrencias del valor. La tabla 11.2 muestra el resultado de este proceso. Podemos notar que hay dos observaciones con valor $1,3$ a las que le corresponderían los rangos 1 y 2, por lo que, en consecuencia, ambas reciben el mismo valor de rango, igual al promedio $1,5$. Esto también ocurre para las puntuaciones $1,4$; $2,7$; $3,7$ y $5,0$.

Observación	**Muestra**	**Rango**
1,3	B	1,5
1,3	B	1,5
1,4	A	3,5
1,4	B	3,5
1,5	A	5,0
1,6	A	6,0
1,8	A	7,0
2,7	A	8,5
2,7	B	8,5
3,0	A	10,0
3,7	A	11,5
3,7	B	11,5
4,6	B	13,0
5,0	A	14,5
5,0	B	14,5
5,1	A	16,0
5,3	A	17,0
5,6	B	18,0
6,1	A	19,0
6,3	B	20,0
6,6	A	21,0
6,7	B	22,0
6,8	B	23,0

Tabla 11.2: muestras combinadas con rango.

A continuación, se suman los rangos asociados a las observaciones de cada muestra, y para la muestra combinada. Así, para la muestra A obtenemos:

$$T_A = 3,5 + 5,0 + 6,0 + 7,0 + 8,5 + 10,0 + 11,5 + 14,5 + 16,0 + 17,0 + 19,0 + 21,0 = 139$$

De manera análoga, para la muestra B se tiene:

$$T_B = 1,5 + 1,5 + 3,5 + 8,5 + 11,5 + 13,0 + 14,5 + 18,0 + 20,0 + 22,0 + 23,0 = 137$$

La suma de rangos para la muestra combinada siempre está dada por la ecuación 11.1.

$$T_T = \frac{n_T \cdot (n_T + 1)}{2} \tag{11.1}$$

Para el ejemplo:

$$T_T = \frac{23 \cdot (23+1)}{2} = 276$$

Trabajar con los rangos en lugar de las observaciones nos ofrece dos ventajas: la primera es que el foco solo está en las relaciones de orden entre las observaciones, sin necesidad de que estas provengan de una escala de intervalos iguales. La segunda es que esta transformación facilita conocer de manera sencilla algunas propiedades del conjunto de datos. Por ejemplo, la suma de rangos de la muestra se determina siempre mediante la ecuación 11.1 y la media de rangos de la muestra combinada es siempre como muestra la ecuación 11.2.

$$\mu = \frac{n_T \cdot (n_T+1)}{2} \cdot \frac{1}{n_T} = \frac{n_T+1}{2} \tag{11.2}$$

Para el ejemplo:

$$\mu = \frac{23+1}{2} = 12$$

En consecuencia, la hipótesis nula en el dominio de los rangos es que las medias de los rangos de las dos muestras son iguales. Si la hipótesis nula fuera cierta, las observaciones en ambas muestras serían similares, por lo que, al ordenar la muestra combinada, ambas muestras se mezclarían de manera homogénea. En consecuencia, deberíamos esperar que los promedios de rangos para cada muestra se aproximen al rango promedio de la muestra combinada, es decir, que T_A y T_B se aproximen a los siguientes valores:

$$\mu_A = n_A \cdot \frac{(n_T)+1}{2} = 12 \cdot \frac{(23+1)}{2} = 144$$
$$\mu_B = n_B \cdot \frac{(n_T)+1}{2} = 11 \cdot \frac{(23+1)}{2} = 132$$

La prueba de Wilcoxon-Mann-Whitney tiene dos variantes, una para muestras grandes y otra para muestras pequeñas, que se diferencian a partir de este punto.

11.1.1.1 Prueba de suma de rangos de Wilcoxon para muestras grandes

Hasta ahora, hemos determinado que:

- El valor observado $T_A = 139$ pertenece a una distribución muestral con media $\mu_A = 144$.
- El valor observado $T_B = 137$ pertenece a una distribución muestral con media $\mu_B = 132$.

Bajo el supuesto de que la hipótesis nula sea verdadera, podríamos demostrar que las distribuciones muestrales de T_A y T_B tienen la misma desviación estándar, dada por la ecuación 11.3.

$$\sigma_T = \sqrt{\frac{n_A \cdot n_B \cdot (n_T+1)}{12}} \tag{11.3}$$

Con lo que:

$$\sigma_T = \sqrt{\frac{12 \cdot 11 \cdot (23+1)}{12}} = 16,248$$

Cuando **ambas muestras tienen tamaño mayor o igual a 5**, siguiendo un procedimiento similar al descrito en la primera sección del capítulo 4, podemos demostrar que las distribuciones muestrales de T_A y T_B tienden a aproximarse a la distribución normal. En consecuencia, una vez conocidas la media y la desviación estándar de una distribución normal para la muestra, podemos calcular el estadístico z para T_A o T_B, dado por la ecuación 11.4, donde:

- T_{obs} es cualquiera de los valores observados, T_A o T_B.
- μ_{obs} es la media de la distribución muestral de T_{obs}.
- σ_T es la desviación estándar de la distribución muestral de T_{obs} (es decir, el error estándar).

$$z = \frac{(T_{obs} - \mu_{obs}) \pm 0,5}{\sigma_T} \tag{11.4}$$

Puesto que las distribuciones muestrales de T son intrínsecamente discretas (solo pueden asumir valores con decimales cuando existen rangos empatados), debemos emplear un factor de corrección de continuidad:

- $-0,5$ si $T_{obs} > \mu_{obs}$.
- $0,5$ si $T_{obs} < \mu_{obs}$.

Volviendo al ejemplo, tenemos:

$$z_A = \frac{(139 - 144) + 0,5}{16,248} = -0,277$$
$$z_B = \frac{(137 - 132) - 0,5}{\sigma_T} = 0,277$$

Los valores z obtenidos a partir de T_A y T_B siempre tienen igual valor absoluto y signos opuestos, por lo que no importa cuál de ellos usemos para la prueba de significación estadística. No obstante, debemos tener muy claro el significado del signo de z: si para el ejemplo tuviésemos como hipótesis alternativa que la interfaz A es mejor que la interfaz B, entonces esperaríamos que las observaciones de mayor rango estuvieran en el grupo A, por lo que z_A tendría que ser positivo.

El valor z obtenido permite calcular el valor p para una hipótesis alternativa unilateral (pues solo delimita la región de rechazo en una de las colas de la distribución normal estándar subyacente). Así, para el ejemplo, cuya hipótesis alternativa es bilateral, podemos calcular el valor p en R mediante la llamada `2 * pnorm(-0.277, mean = 0, sd = 1, lower.tail = TRUE)`, obteniéndose como resultado $p = 0,782$.

Evidentemente, el valor p obtenido es muy alto, por lo que fallamos al rechazar la hipótesis nula. En consecuencia, podemos concluir que no hay diferencia significativa en la usabilidad de las dos interfaces.

11.1.1.2 Prueba de suma de rangos de Wilcoxon para muestras pequeñas

Cuando las muestras son pequeñas (menos de 5 observaciones), no podemos usar el supuesto de normalidad del apartado anterior, por lo que necesitamos una vía alternativa. Este método sirve también para muestras más grandes, con resultados equivalentes a los ya obtenidos.

Aprovechando una vez más las ventajas de considerar los rangos en lugar de las observaciones originales, podemos calcular el máximo valor posible para la suma de rangos de cada muestra como indica la ecuación 11.5. Fijémonos en que el valor máximo para la suma de rangos de una muestra se produce cuando esta contiene los n_x rangos mayores de la muestra combinada.

$$T_{x[max]} = n_x \cdot n_y + \frac{n_x \cdot (n_x + 1)}{2} \tag{11.5}$$

Así, para el ejemplo:

$$T_{A[max]} = 12 \cdot 11 + \frac{12 \cdot (12 + 1)}{2} = 210$$
$$T_{B[max]} = 11 \cdot 12 + \frac{11 \cdot (11 + 1)}{2} = 198$$

Con esto podemos definir un nuevo estadístico de prueba U, como muestra la ecuación 11.6.

$$U_x = T_{x[max]} - T_x \tag{11.6}$$

Por lo que:

$$U_A = 210 - 139 = 71$$
$$U_B = 198 - 137 = 61$$

El valor del estadístico de prueba es el mínimo entre U_A y U_B, por lo que $U = 61$.

Debemos notar que siempre se cumple la identidad presentada en la ecuación 11.7, por lo que podemos escoger cualquiera de los valores U obtenidos para realizar el resto del procedimiento.

$$U_A + U_B = n_A \cdot n_B \tag{11.7}$$

Si la hipótesis nula fuese cierta, esperaríamos que:

$$U_A = T_{A[max]} - \mu_A = 210 - 144 = 66$$
$$U_B = T_{B[max]} - \mu_B = 198 - 132 = 66$$

Formalmente, entonces, si la hipótesis nula fuera verdadera, esperaríamos que:

$$U_A = U_B = \frac{n_A \cdot n_B}{2}$$

En consecuencia, la pregunta asociada a la prueba de hipótesis es: si la hipótesis nula es verdadera (no hay diferencias significativas en la usabilidad de ambas interfaces), ¿qué tan probable es obtener un valor de U al menos tan pequeño como el observado ($U = 61$)? Para responder a esta pregunta, seguimos un procedimiento similar al que ya conocimos para la prueba exacta de Fisher (capítulo 8): se calculan todas las formas en que n_T rangos podrían combinarse en dos grupos de tamaños n_A y n_B, y luego se determina la proporción de las combinaciones que produzcan un valor de U al menos tan pequeño como el encontrado. Pero ¡existen 676.039 combinaciones posibles!

Aunque R no ofrece herramientas para calcular el valor p a partir del estadístico U (pues utiliza el estadístico W, propuesto por Frank Wilcoxon en 1945, que lleva a los mismos resultados), afortunadamente existen tablas que permiten conocer el máximo valor de U para el cual se rechaza la hipótesis nula para un nivel de significación dado sin tener que revisar todas las combinaciones. Considerando $\alpha = 0,05$ para una prueba bilateral, el valor crítico es $U = 33$ (Real Statistics Using Excel, s.f.). Puesto que $61 > 33$, fallamos al rechazar la hipótesis nula, por lo que concluimos con 95 % de confianza que no existe una diferencia estadísticamente significativa en la usabilidad de ambas interfaces.

11.1.1.3 Prueba de suma de rangos de Wilcoxon en R

Como ya dijimos, la implementación de esta prueba en R usa el estadístico W (introducido por Wilcoxon) en lugar del estadístico U empleado por Mann y Whitney. Es por ello que que esta prueba se realiza mediante la función `wilcox.test(x, y, paired = FALSE, alternative, mu, conf.level)`, donde:

- `x, y`: vectores numéricos con las observaciones. Para aplicar la prueba con una única muestra, `y` debe ser nulo (por defecto, lo es).
- `paired`: booleano con valor falso para indicar que las muestras son independientes (se asume por defecto).
- `alternative`: señala el tipo de hipótesis alternativa: bilateral ("two.sided") o unilateral ("less" o "greater").
- `mu`: valor nulo para la media (solo para la prueba con una muestra).
- `conf.level`: nivel de confianza.

El script 11.1 muestra la aplicación de esta prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.2.

```
Wilcoxon rank sum test with continuity correction

data:  a and b
W = 61, p-value = 0.7816
alternative hypothesis: true location shift is not equal to 0
```

Figura 11.2: resultado de la prueba de Mann-Whitney (en rigor, de la prueba para el ejemplo.

Script 11.1: prueba de Mann-Whitney para el ejemplo.

```
# Ingresar los datos.
a <- c(2.7, 6.6, 1.6, 5.1, 3.7, 6.1, 5.0, 1.4, 1.8, 1.5, 3.0, 5.3)
b <- c(5.0, 1.4, 5.6, 4.6, 6.7, 2.7, 1.3, 6.3, 3.7, 1.3, 6.8)

# Establecer nivel de significación.
alfa <- 0.05

# Hacer la prueba de Mann-Whitney.
prueba <- wilcox.test(a, b, alternative = "two.sided", conf.level = 1 - alfa)
print(prueba)
```

11.1.2 Prueba de rangos con signo de Wilcoxon

La **prueba de rangos con signo de Wilcoxon** es, conceptualmente, parecida a la prueba de suma de rangos de Wilcoxon presentada en la sección anterior. Sin embargo, en este caso es la alternativa no paramétrica a la prueba t de Student con muestras pareadas. Las condiciones que se deben cumplir para usar esta prueba son:

1. Los pares de observaciones son independientes.
2. La escala de medición empleada para las observaciones es intrínsecamente continua.
3. La escala de medición empleada para ambas muestras debe ser a lo menos ordinal.

Consideremos ahora un nuevo contexto para la aplicación de esta prueba. Una empresa de desarrollo desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto de software, a fin de

determinar si, como asegura el departamento de diseño, es mejor la interfaz A. Para ello, la empresa ha seleccionado a 10 participantes al azar, quienes deben evaluar 6 aspectos de usabilidad de cada interfaz, cada uno de los cuales se mide con una escala Likert de 7 puntos, donde 1 significa "muy malo" y 7, "muy bueno". La valoración que un participante da a la interfaz evaluada corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 11.3 muestra las evaluaciones realizadas por cada participante a cada una de las interfaces.

Participante	**Interfaz A**	**Interfaz B**
1	2,9	6,0
2	6,1	2,8
3	6,7	1,3
4	4,7	4,7
5	6,4	3,1
6	5,7	1,8
7	2,7	2,9
8	6,9	4,0
9	1,7	2,3
10	6,4	1,6

Tabla 11.3: evaluación de las interfaces de usuario A y B.

Formalmente, las hipótesis son:

H_0: las mismas personas no perciben diferencia en la usabilidad de ambas interfaces.
H_A: las mismas personas consideran que la interfaz A tiene mejor usabilidad que la interfaz B.

La mecánica inicial para esta prueba consiste en calcular las diferencias entre cada par de observaciones y obtener luego su valor absoluto. Generalmente se descartan aquellas instancias con diferencia igual a cero, pues no aportan información relevante al procedimiento. A continuación se ordenan las diferencias absolutas en orden creciente y se les asignan rangos de manera correlativa del mismo modo que en la prueba de Wilcoxon-Mann-Whitney. Una vez asignados los rangos, se les incorpora el signo asociado a la diferencia. La tabla 11.4 ilustra el proceso descrito.

Participante	**Interfaz A**	**Interfaz B**	**A−B**	**|A−B|**	**Rango absoluto**	**Rango con signo**
4	4,7	4,7	0,0	0	-	-
7	2,7	2,9	-0,2	0,2	1	-1
9	1,7	2,3	-0,6	0,6	2	-2
8	6,9	4,0	2,9	2,9	3	+3
1	2,9	6,0	-3,1	3,1	4	-4
2	6,1	2,8	3,3	3,3	5,5	+5,5
5	6,4	3,1	3,3	3,3	5,5	+5,5
6	5,7	1,8	3,9	3,9	7	+7
10	6,4	1,6	4,8	4,8	8	+8
3	6,7	1,3	5,4	5,4	9	+9

Tabla 11.4: asignación de rangos con signo.

Una vez realizado este proceso, se calcula el estadístico de prueba W, correspondiente a la suma de los rangos con signo. Debemos notar que, tras eliminar aquellas observaciones con diferencia igual a 0, el tamaño de las muestras para el ejemplo es $n = 9$. Así:

$$W = -1 + -2 + 3 + -4 + 5,5 + 5,5 + 7 + 8 + 9 = 31$$

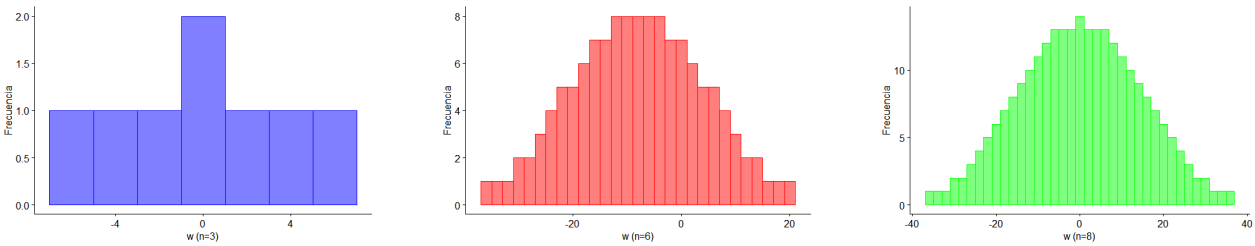
Desde luego, el máximo valor posible para W, W_{max} corresponde a la suma de los n rangos sin signo (todos positivos) (ecuación 11.2) y, además, $W_{min} = -|W_{max}|$.

Para entender mejor la distribución de W, una muestra de tamaño n genera n rangos no empatados sin signo (columna "Rango absoluto" de la tabla 11.4). A su vez, cada uno de dichos rangos podría tomar valores positivos o negativos, por lo que para W se tienen 2^n combinaciones para los signos de los rangos. La tabla 11.5 muestra todas las posibles combinaciones para $n = 3$. Si la hipótesis nula fuese cierta, los rangos positivos y negativos se distribuirían de manera homogénea, por lo que esperaríamos que el valor de W fuese cercano a 0 (hipótesis nula en el dominio de los rangos).

La figura 11.3 muestra la distribución de W para distintos valores de n. En ella podemos apreciar que, a medida que n aumenta, la distribución de W se aproxima cada vez más a una distribución normal centrada en $\mu_W = 0$.

Rango			
1	**2**	**3**	W
+	+	+	6
+	+	-	0
+	-	+	2
+	-	-	-4
-	+	+	4
-	+	-	-2
-	-	+	0
-	-	-	-6

Tabla 11.5: valores que puede adoptar el estadístico W para $n = 3$.



(a) $n = 3$ (b) $n = 6$ (c) $n = 8$

Figura 11.3: distribución de W.

La desviación estándar de la distribución muestral de W está dada por la ecuación 11.8.

$$\sigma_W = \sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{6}} \tag{11.8}$$

Para el ejemplo:

$$\sigma_W = \sqrt{\frac{9 \cdot (9+1) \cdot (2 \cdot 9+1)}{6}} = 16,882$$

Puesto que estamos trabajando bajo el supuesto de normalidad, calculamos el estadístico de prueba z, dado por la ecuación 11.9.

$$z = \frac{W - 0,5}{\sigma_W} \tag{11.9}$$

Así, para el ejemplo tenemos que:

$$z = \frac{31 - 0,5}{16,882} = 1,807$$

Una vez conocido el estadístico de prueba, podemos obtener el valor p mediante la llamada `pnorm(1.807, mean = 0, sd = 1, lower.tail = FALSE)` (no multiplicamos por 2, pues es una prueba unilateral), obteniendo como resultado $p = 0,035$. Considerando un nivel de significación $\alpha = 0,05$, rechazamos la hipótesis nula en favor de la hipótesis alternativa. En consecuencia, concluimos con 95 % de confianza que la usabilidad de la interfaz A es mejor que la de la interfaz B.

Siempre debemos tener en cuenta que el supuesto de normalidad es válido únicamente para $n > 10$, por lo que en caso de que las muestras sean más pequeñas, tenemos que consultar la tabla de valores críticos para la distribución W.

En R, la prueba de rangos con signo de Wilcoxon está implementada en la misma función que en el caso de muestras independientes, pero ahora la llamada es `wilcox.test(x, y, paired = TRUE, alternative, conf.level)`, donde:

- `x, y`: vectores numéricos con las observaciones.
- `paired`: booleano con valor verdadero para indicar que las muestras son pareadas.
- `alternative`: señala el tipo de hipótesis alternativa: bilateral ("two.sided") o unilateral ("less" o "greater").
- `paired`: indica si las muestras están o no pareadas.
- `conf.level`: nivel de confianza.

Así, el valor por defecto para el parámetro `paired` es `FALSE`, en cuyo caso se efectúa la prueba de suma de rangos de Wilcoxon; mientras que si explícitamente indicamos `paired = TRUE`, se aplica la prueba de rangos con signo de Wilcoxon.

El script 11.2 muestra la aplicación de la prueba de rangos con signo de Wilcoxon para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.4. Es importante tener en cuenta en cuenta que R usa una variante ligeramente diferente. En lugar del estadístico de prueba W, calcula el estadístico V, correspondiente a la suma de los rangos con signo positivo.

```
Wilcoxon signed rank test with continuity correction

data:  a and b
V = 38, p-value = 0.03778
alternative hypothesis: true location shift is greater than 0
```

Figura 11.4: resultado de la prueba de rangos con signo de Wilcoxon para el ejemplo.

Script 11.2: prueba de rangos con signo de Wilcoxon para el ejemplo.

```
# Ingresar los datos.
a <- c(2.9, 6.1, 6.7, 4.7, 6.4, 5.7, 2.7, 6.9, 1.7, 6.4)
b <- c(6.0, 2.8, 1.3, 4.7, 3.1, 1.8, 2.9, 4.0, 2.3, 1.6)

# Establecer nivel de significación.
alfa <- 0.05

# Hacer la prueba de rangos con signo de Wilcoxon.
prueba <- wilcox.test(a, b, alternative = "greater", paired = TRUE,
                      conf.level = 1 - alfa)

print(prueba)
```

11.2 PRUEBAS PARA MÁS DE DOS MUESTRAS

Al igual que existen alternativas no paramétricas para inferir con una o dos medias muestrales, también las hay para cuando se tienen más de dos muestras. Conoceremos ahora alternativas no paramétricas para el procedimiento ANOVA de una vía, tanto para muestras independientes como para muestras correlacionadas.

11.2.1 Prueba de Kruskal-Wallis

En el capítulo 9 estudiamos el procedimiento ANOVA de una vía para $k > 2$ muestras independientes, el cual requiere el cumplimiento de los siguientes supuestos:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las k muestras son obtenidas de manera aleatoria e independiente desde la(s) población(es) de origen.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. Las k muestras tienen varianzas aproximadamente iguales.

Si bien ANOVA es usualmente robusto ante desviaciones leves de las condiciones (excepto la segunda) cuando las muestras son de igual tamaño, no ocurre lo mismo cuando los tamaños de las muestras difieren. En este caso, una alternativa es emplear la **prueba de Kruskal-Wallis**, cuyas condiciones son:

1. La variable independiente debe tener a lo menos dos niveles (aunque, para dos niveles, se suele usar la prueba de Wilcoxon-Mann-Whitney).
2. La escala de la variable dependiente debe ser, a lo menos, ordinal.
3. Las observaciones son independientes entre sí.

Para ilustrar esta prueba, tomemos el ejemplo de un ingeniero que cuenta con cuatro algoritmos (A, B, C y D) para resolver un determinado problema (en iguales condiciones y para instancias de tamaño fijo) y desea comparar su eficiencia. Para cada algoritmo, selecciona una muestra aleatoria independiente de instancias y registra el tiempo de ejecución (en milisegundos) del algoritmo en cuestión para cada una de las instancias de la muestra correspondiente, obteniendo las siguientes mediciones:

- Algoritmo A: 21, 22, 22, 23, 23, 23, 23, 24, 24, 24, 25, 26
- Algoritmo B: 15, 17, 18, 18, 19, 19, 20, 20, 21
- Algoritmo C: 9, 10, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 14, 15
- Algoritmo D: 15, 15, 16, 16, 16, 18, 18, 18

Las hipótesis a contrastar son, entonces:

H_0: todos los algoritmos son igual de eficientes (o, de manera similar, ningún algoritmo es menos ni más eficiente que los demás).
H_A: al menos uno de los algoritmos presenta una eficiencia diferente a al menos algún otro algoritmo.

El procedimiento de la prueba de Kruskal-Wallis tiene elementos similares a los descritos en las pruebas no paramétricas para una y dos medias. El primer paso consiste en combinar las muestras y luego asignar el rango a cada elemento, obteniéndose para el ejemplo el resultado de la tabla 11.6.

A continuación se calcula la suma (T_x) y la media (M_x) de los rangos en cada grupo y en la muestra combinada. La tabla 11.7 presenta los valores obtenidos para el ejemplo, incluyendo además el tamaño muestral (n_x).

De manera similar a ANOVA, se requiere determinar la diferencia entre las medias grupales. Para ello se calculan las desviaciones cuadradas de las medias grupales de los rangos con respecto a la media total de los rangos. Así, la variabilidad entre grupos está dada por la ecuación 11.10.

Observaciones				Ranking de obs.			
A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
21	15	9	15	31,5	15,5	1,0	15,5
22	17	10	15	33,5	21,0	3,5	15,5
22	18	10	16	33,5	24,0	3,5	19,0
23	18	10	16	36,5	24,0	3,5	19,0
23	19	10	16	36,5	27,5	3,5	19,0
23	19	11	18	36,5	27,5	8,5	24,0
23	20	11	18	36,5	29,5	8,5	24,0
24	20	12	18	40,0	29,5	8,5	24,0
24	21	12		40,0	31,5	8,5	
24		12		40,0		8,5	
25		12		42,0		8,5	
26		13		43,0		12,0	
		14				13,0	
		15				15,5	

Tabla 11.6: asignación de rangos a la muestra combinada.

	A	**B**	**C**	**D**	**Combinada**
n	12	9	14	8	43
T	449,50	230,00	106,50	160,00	946,00
M	37,46	25,56	7,61	20,00	22,00

Tabla 11.7: resumen de los rangos.

$$SS_{bg(R)} = \sum_{i=1}^{k} n_i \cdot (M_i - M_T)^2 \tag{11.10}$$

Para el ejemplo, entonces:

$$SS_{bg(R)} = n_A \cdot (M_A - M_T)^2 + n_B \cdot (M_B - M_T)^2 + n_C \cdot (M_C - M_T)^2 + n_D \cdot (M_D - M_T)^2 = \\ 12 \cdot (37,46 - 22)^2 + 9 \cdot (25,56 - 22)^2 + 14 \cdot (7,61 - 22)^2 + 8 \cdot (20 - 22)^2 = 5.913,21$$

La hipótesis nula, llevada al dominio de los rangos, es que los rangos medios de los distintos grupos no serán muy diferentes entre sí. Podría esperarse que el valor nulo para $SS_{bg(R)}$ fuera 0, no obstante, no es así. Supongamos por un momento que tenemos 3 muestras con dos observaciones cada una, con lo que tendríamos un total de 6 rangos. Dichos rangos pueden combinarse de 90 maneras distintas para formar tres grupos con dos elementos. La distribución muestral de $SS_{bg(R)}$ estaría dada, entonces, por los valores de $SS_{bg(R)}$ obtenidos para cada una de las 90 combinaciones, de los cuales únicamente 6 son iguales a 0 y todos los restantes, mayores que 0 (recuerde que es matemáticamente imposible obtener desviaciones cuadradas con valor negativo). La media de la distribución muestral para $SS_{bg(R)}$ está dada por la ecuación 11.11.

$$\mu_{SS} = (k-1)\frac{n_T \cdot (n_T + 1)}{12} \tag{11.11}$$

Para el ejemplo, entonces, tenemos que el valor nulo es:

$$\mu_{SS} = (4-1)\frac{43 \cdot (43+1)}{12} = 473$$

Llegado este punto, se define el estadístico de prueba H, el cual se construye en torno al valor obtenido para $SS_{bg(R)}$ y parte de la fórmula empleada para calcular el valor nulo, como muestra la ecuación 11.12.

$$H = \frac{SS_{bg(R)}}{\frac{n_T \cdot (n_T+1)}{12}} = \frac{12 \cdot SS_{bg(R)}}{n_T \cdot (n_T+1)} \tag{11.12}$$

En consecuencia, el valor del estadístico de prueba para el ejemplo es:

$$H = \frac{12 \cdot 5.913,21}{43 \cdot (43+1)} = 37.5$$

Cuando cada uno de los k grupos tiene a lo menos 5 observaciones, el estadístico de prueba H sigue una distribución χ^2 con $\nu = k - 1$ grados de libertad. Así, podemos calcular el valor p para el ejemplo (en R) mediante la llamada `pchisq(37.5, 3, lower.tail = FALSE)`, obteniéndose como resultado $p = 3.606 \cdot 10^{-8}$. Este valor indica que la evidencia es suficientemente fuerte como para rechazar la hipótesis nula en favor de la hipótesis alternativa, incluso para un nivel de significación $\alpha = 0,01$. En consecuencia, podemos concluir con 99 % de confianza que existen diferencias significativas entre los tiempos promedio de ejecución de los algoritmos A, B, C y D.

Fijémonos en que, al igual que ANOVA, la prueba de Kruskal-Wallis es de tipo ómnibus, por lo que no entrega información en relación a cuáles grupos presentan diferencias. En consecuencia, una vez más es necesario efectuar un análisis post-hoc cuando se detectan diferencias significativas. De manera similar a la estudiada en el capítulo 9, podemos hacer comparaciones entre pares de grupos con la prueba de Wilcoxon-Mann-Whitney (equivalentes a las realizadas con la prueba t de Student para ANOVA de una vía para muestras independientes), usando alguno de los factores de corrección que ya conocimos en el capítulo 8 (por ejemplo, Holm y Bonferroni) (Amat Rodrigo, 2016b).

En R, podemos ejecutar la prueba de Kruskal-Wallis mediante la función `kruskal.test(formula, data)`, donde:

- `formula`: tiene la forma <variable dependiente>∼ <variable independiente (factor)>.
- `data`: matriz de datos en formato largo.

Para los procedimientos post-hoc, las pruebas de Bonferroni y Holm pueden realizarse mediante la función `pairwise.wilcox.test(x, g, p.adjust.method, paired = FALSE)`, donde:

- `x`: vector con la variable dependiente.
- `g`: factor o agrupamiento.
- `p.adjust.method`: puede ser "holm" o "bonferroni", entre otras alternativas.
- `paired`: valor booleano que indica si la prueba es pareada (verdadero) o no. Para la prueba de Kruskal-Wallis debe ser `FALSE`.

El script 11.3 muestra la realización de la prueba de Kruskal-Wallis para el ejemplo e incorpora el procedimiento post-hoc de Holm. Los resultados se presentan en la figura 11.5. Podemos ver que el valor p difiere ligeramente al obtenido anteriormente, debido a errores de redondeo. A partir de los resultados del procedimiento post-hoc, considerando un nivel de significación $\alpha = 0,01$, podemos concluir con 99 % de confianza que existen diferencias significativas entre los tiempos promedio de ejecución de todos los pares de algoritmos con excepción de los algoritmos B y D.

Script 11.3: prueba de Kruskal-Wallis y el procedimiento post-hoc de Holm para el ejemplo.

```
# Construir la mariz de datos.
A <- c(24, 23, 26, 21, 24, 24, 25, 22, 23, 22, 23, 23)
B <- c(17, 15, 18, 20, 19, 21, 20, 18, 19)
C <- c(10, 11, 14, 11, 15, 12, 12, 10, 9, 13, 12, 12, 10, 10)
D <- c(18, 16, 18, 15, 16, 15, 18, 16)
Tiempo <- c(A, B, C, D)

```

```
        Kruskal-Wallis rank sum test

        data:  Tiempo by Algoritmo
        Kruskal-Wallis chi-squared = 37.714, df = 3,
        p-value = 3.249e-08


        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

        data:  datos$Tiempo and datos$Algoritmo

          A       B       C
        B 0.00060 -       -
        C 9.3e-05 0.00042 -
        D 0.00060 0.02738 0.00060

        P value adjustment method: holm
```

Figura 11.5: resultado de la prueba de Kruskal-Wallis y el procedimiento post-hoc de Holm para el ejemplo.

```r
Algoritmo <- c(rep("A", length(A)),
               rep("B", length(B)),
               rep("C", length(C)),
               rep("D", length(D)))

Algoritmo <- factor(Algoritmo)

datos <- data.frame(Tiempo, Algoritmo)

# Establecer nivel de significación
alfa <- 0.01

# Hacer la prueba de Kruskal-Wallis.
prueba <- kruskal.test(Tiempo ~ Algoritmo, data = datos)
print(prueba)

# Efectuar procedimiento post-hoc de Holm si se encuentran diferencias
# significativas.
if(prueba$p.value < alfa) {
  post_hoc <- pairwise.wilcox.test(datos$Tiempo,
                                   datos$Algoritmo,
                                   p.adjust.method = "holm",
                                   paired = FALSE)

  print(post_hoc)
}
```

Notemos que `pairwise.wilcox.test()` solo reporta los p valores ajustados. Si queremos conocer el tamaño del efecto de las diferencias detectadas, debemos realizar las correspondientes pruebas de Wilcoxon-Mann-Whitney para todos los pares de grupos que presenten diferencias significativas.

11.2.2 Prueba de Friedman

Como es natural suponer, podemos considerar la **prueba de Friedman** como una alternativa no paramétrica al procedimiento ANOVA de una vía con muestras correlacionadas descrito en el capítulo 10. Sin embargo, debemos saber que no es exactamente una extensión de esta prueba, puesto que no considera las diferencias relativas entre sujetos (como lo hace ANOVA y la prueba de rangos con signo de Wilcoxon), y en consecuencia, como señala Baguley (2012), el poder estadístico es bastante menor.

Recordemos las condiciones que se deben verificar para poder aplicar la prueba ANOVA para muestras correlacionadas:

1. La escala con que se mide la variable dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las mediciones son independientes al interior de cada grupo.
3. Se puede suponer razonablemente que la(s) población(es) de origen sigue(n) una distribución normal.
4. La matriz de varianzas-covarianzas es esférica. Como explica Horn (2008, p. 1), esta condición establece que las varianzas entre los diferentes niveles de las medidas repetidas deben ser iguales.

Existen situaciones en las que no podemos comprobar que la escala de medición de la variable dependiente sea de intervalos iguales:

- Cuando las observaciones se miden en una escala logarítmica (por ejemplo, la escala de pH para medir la acidez o la escala de Richter para medir la intensidad de los sismos).
- Cuando las mediciones provienen de una escala ordinal, por ejemplo, un orden de preferencia.
- Cuando las mediciones de base provienen de una escala ordinal. Por ejemplo, cuando se suman o promedian puntajes de diversos elementos evaluados con una escala Likert.

Las condiciones requeridas por la prueba de Friedman son las siguientes:

1. La variable independiente debe ser categórica y tener a lo menos tres niveles.
2. La escala de la variable dependiente debe ser, a lo menos, ordinal.
3. Los sujetos son una muestra aleatoria e independiente de la población.

Como ejemplo para esta prueba, supongamos ahora que un equipo de desarrolladores desea establecer qué interfaz gráfica (A, B o C) resulta más atractiva para los usuarios de un nuevo sistema, por lo que han seleccionado una muestra aleatoria representativa de los distintos tipos de usuarios y les han solicitado evaluar 6 aspectos de cada interfaz con una escala Likert de 5 puntos, donde el valor 1 corresponde a una valoración muy negativa y 5, a una muy positiva. La tabla 11.8 muestra las puntuaciones totales asignadas por cada participante a las diferentes interfaces. En consecuencia, las hipótesis a contrastar son:

H_0: las interfaces tienen preferencias similares.
H_A: al menos una interfaz obtiene una preferencia distinta a las demás.

Usuario	A	B	C
1	21	6	13
2	10	21	25
3	7	18	18
4	21	7	20
5	24	24	24
6	27	13	8
7	17	13	29

Tabla 11.8: evaluación realizada por los usuarios a cada una de las distintas interfaces.

El primer paso del proceso consiste en asignar rangos a las observaciones de cada sujeto. La interfaz con puntuación más baja recibe un rango de 1 y la más alta, un rango de 3 (generalizando, si se tienen k

observaciones pareadas, se asignan rangos con valores 1 a k). En caso de empate, se asigna el promedio de los rangos correspondientes. La tabla 11.9 muestra el resultado de este proceso.

	Originales			**Rangos**		
Usuario	**A**	**B**	**C**	**A**	**B**	**C**
1	21	6	13	3	1	2
2	10	21	25	1	2	3
3	7	18	18	1	2,5	2,5
4	21	7	20	3	1	2
5	24	24	24	2	2	2
6	27	13	8	3	2	1
7	17	13	29	2	1	3

Tabla 11.9: ranking de las interfaces por usuario.

La hipótesis nula para la prueba de Friedman es que, los rangos promedio de cada interfaz serán muy similares. Si denotamos el rango promedio de un grupo (interfaz) por M_x, para cada grupo esperamos, entonces, que se cumpla la igualdad de la ecuación 11.13, donde k es la cantidad de grupos.

$$M_x = \frac{k+1}{2} \tag{11.13}$$

A continuación se calculan algunas estadísticas de resumen, donde n corresponde al tamaño de cada muestra y M, a la media de los rangos (tabla 11.10).

	A	**B**	**C**	**Combinada**
n	7	7	7	21
M	2,14	1,64	2,21	2

Tabla 11.10: resumen de los rangos.

Con estos valores, podemos definir una medida para la variabilidad de los grupos agregados, dada por la ecuación 11.14.

$$SS_{bg(R)} = \sum_{i=1}^{k} n_i \cdot (M_i - M_T)^2 \tag{11.14}$$

Haciendo el cálculo para el ejemplo, tenemos:

$$SS_{bg(R)} = 7 \cdot [(2,14-2)^2 + (1,64-2)^2 + (2,21-2)^2] = 1,357$$

Con el resultado anterior, podemos ahora calcular el estadístico de prueba (11.15), que sigue una distribución χ^2 con $k-1$ grados de libertad.

$$\chi^2 = \frac{SS_{bg(R)}}{\frac{k\cdot(k+1)}{12}} = \frac{12 \cdot SS_{bg(R)}}{k \cdot (k+1)} \tag{11.15}$$

Para el ejemplo:

$$\chi^2 = \frac{12 \cdot 1,357}{3 \cdot (3+1)} = 1,357$$

Una vez más, calculamos el valor p mediante la llamada `pchisq(1.357, 2, lower.tail = FALSE)`, obteniéndose $p = 0,507$. Considerando un nivel de significación $\alpha = 0,05$, se falla al rechazar la hipótesis nula. En consecuencia, concluimos con 95 % de confianza que no hay diferencias significativas de preferencia entre las distintas interfaces.

En este caso no es necesario realizar un procedimiento post-hoc, pues la prueba ómnibus no encontró diferencias estadísticamente significativas. No obstante, si fuese necesario, podemos efectuar una prueba de rangos con signo de Wilcoxon por cada par de grupos y aplicar algún factor de corrección.

Para hacer la prueba de Friedman en R, podemos usar la función `friedman.test(formula, data)`, donde:

- `formula`: tiene la forma <variable dependiente>∼ <variable independiente>| <identificador de sujeto o bloque>.
- `data`: matriz de datos en formato largo.

Para los procedimientos post-hoc, podemos aplicar los ajustes de Bonferroni y Holm mediante la función `pairwise.wilcox.test()`, del mismo modo descrito para la prueba de Kruskal-Wallis, cuidando en este caso que el argumento `paired` debe tomar forzosamente el valor `TRUE`. Si además queremos conocer el tamaño del efecto detectado para aquellos pares identificados como relevantes, debemos realizar las correspondientes pruebas de rangos con signo de Wilcoxon para todos los pares de grupos que presenten diferencias significativas (Amat Rodrigo, 2016a).

El script 11.4 muestra la realización de la prueba de Friedman para el ejemplo, cuyo resultado se presenta en la figura 11.6, e incorpora el procedimiento post-hoc de Holm por fines académicos, ya que solo debería realizase si la prueba ómnibus encuentra diferencias significativas.

Script 11.4: prueba de Friedman y el procedimiento post-hoc de Holm para el ejemplo.

```
# Construir la mariz de datos.
A <- c(21, 10, 7, 21, 24, 27, 17)
B <- c(6, 21, 18, 7, 24, 13, 13)
C <- c(13, 25, 18, 20, 24, 8, 29)

Puntuacion <- c(A, B, C)

Interfaz <- c(rep("A", length(A)),
              rep("B", length(B)),
              rep("C", length(C)))

Sujeto <- rep(1:7, 3)

Interfaz <- factor(Interfaz)

datos <- data.frame(Sujeto, Puntuacion, Interfaz)

# Establecer nivel de significación
alfa <- 0.05

# Hacer la prueba de Friedman.
prueba <- friedman.test(Puntuacion ~ Interfaz | Sujeto, data = datos)
print(prueba)

# Efectuar procedimiento post-hoc de Holm si se encuentran diferencias
# significativas.
if(prueba$p.value < alfa) {
  post_hoc <- pairwise.wilcox.test(datos$Puntuacion,
                                   datos$Interfaz,
                                   p.adjust.method = "holm",
                                   paired = TRUE)

```

```
  print(post_hoc)
}
```

```
        Friedman rank sum test

data:  Puntuacion and Interfaz and Sujeto
Friedman chi-squared = 1.6522, df = 2, p-value = 0.4378
```

Figura 11.6: valores p obtenidos en las pruebas t para cada par de grupos mediante los métodos de Bonferroni y Holm.

11.3 EJERCICIOS PROPUESTOS

1. ¿Qué riesgos se corren si se aplica la prueba t de Student con dos muestras que no cumplen con las suposiciones que hace esta prueba?
2. La prueba de Wilcoxon-Mann-Whitney es una alternativa no paramétrica ¿para qué versión de la prueba t de Student?
3. ¿Qué suposiciones hace la prueba de Wilcoxon-Mann-Whitney?
4. Explica cómo la prueba de Wilcoxon-Mann-Whitney construye el ranking de los datos.
5. ¿Qué estadístico usa la prueba de Wilcoxon-Mann-Whitney y cómo se calcula?
6. ¿Por qué a la prueba de Wilcoxon-Mann-Whitney también se le conoce como U-test?
7. La prueba de los rangos con signo de Wilcoxon es una alternativa no paramétrica ¿para qué versión de la prueba t de Student?
8. ¿Qué suposiciones hace la prueba de los rangos con signo de Wilcoxon?
9. Explica cómo la prueba de los rangos con signo de Wilcoxon construye el ranking de los datos.
10. ¿Qué estadístico usa la prueba de los rangos con signo de Wilcoxon y cómo se calcula?
11. ¿Cuándo es más relevante preocuparse de las violaciones de las condiciones del procedimiento ANOVA para muestras independientes?
12. Explica cómo la prueba de Kruskal-Wallis construye el ranking de los datos.
13. ¿Qué estadístico usa la prueba de Kruskal-Wallis y cómo se calcula? ¿Qué distribución sigue dicho estadístico?
14. ¿Cuál es la hipótesis nula de la prueba de Kruskal-Wallis?
15. ¿Qué suposiciones hace la prueba de Kruskal-Wallis?
16. Explique cómo la prueba de Friedman construye el ranking de los datos.
17. ¿Qué estadístico usa la prueba de Friedman y cómo se calcula? ¿Qué distribución sigue dicho estadístico?
18. ¿Cuál es la hipótesis nula de la prueba de Friedman?
19. ¿Qué suposiciones hace la prueba de Friedman?

REFERENCIAS

Amat Rodrigo, J. (2016a). *Test de Friedman*. Consultado el 29 de mayo de 2021, desde https://www.cienciadedatos.net/documentos/21_friedman_test

Amat Rodrigo, J. (2016b). *Test Kruskal-Wallis*. Consultado el 29 de mayo de 2021, desde https://www.cienciadedatos.net/documentos/20_kruskal-wallis_test

Baguley, T. (2012). *Beware the Friedman test!* Consultado el 13 de diciembre de 2021, desde https://seriousstats.wordpress.com/2012/02/14/friedman/

Glen, S. (2021). *Kruskal Wallis H Test: Definition, Examples & Assumptions*. Consultado el 5 de junio de 2021, desde https://www.statisticshowto.com/kruskal-wallis/

Horn, R. A. (2008). *Sphericity in repeated measures analysis*. Consultado el 11 de mayo de 2021, desde http://oak.ucc.nau.edu/rh232/courses/EPS625/Handouts/RM-ANOVA/Sphericity.pdf

Lærd Statistics. (2020). *Friedman Test in SPSS Statistics* [Lund Research Ltd.]. Consultado el 5 de junio de 2021, desde https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php

Lowry, R. (1999). *Concepts & Applications of Inferential Statistics*. Consultado el 3 de mayo de 2021, desde http://vassarstats.net/textbook/

Real Statistics Using Excel. (s.f.). *Mann-Whitney Table*. Consultado el 28 de mayo de 2021, desde https://www.real-statistics.com/statistics-tables/mann-whitney-table/