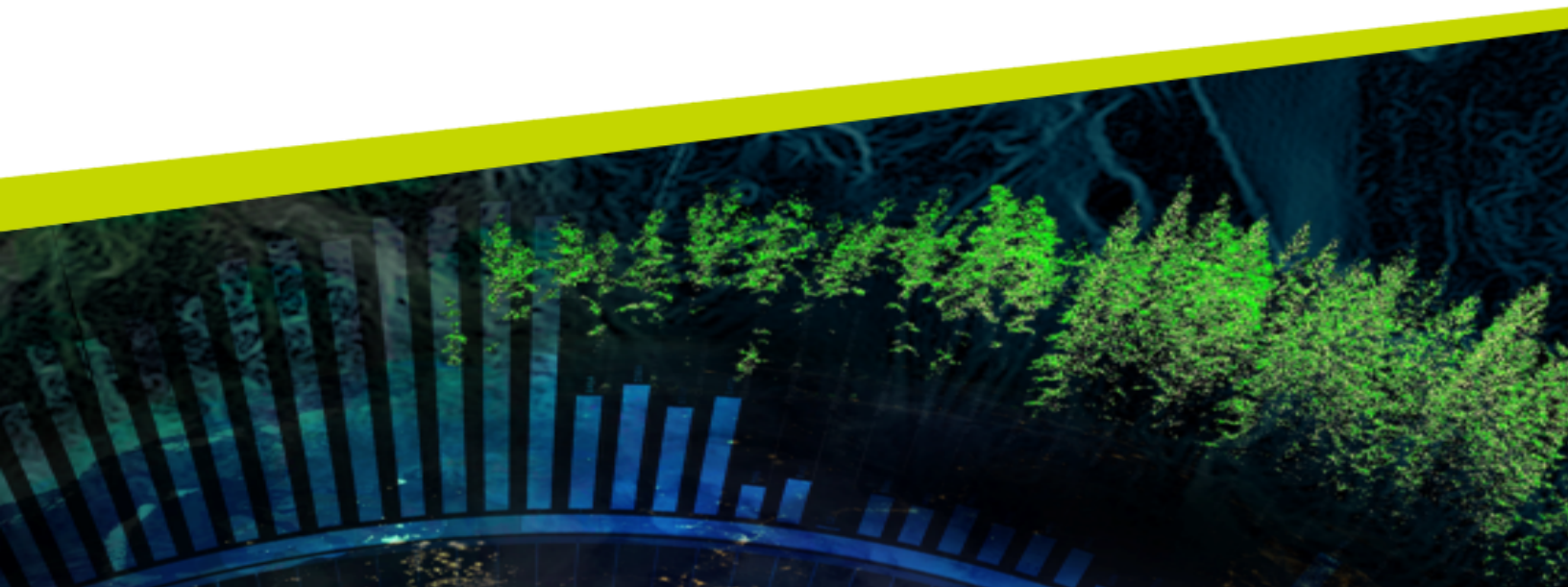
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



11.2 PRUEBAS NO PARAMÉTRICAS CON UNA Y DOS MUESTRAS NUMÉRICAS

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, la prueba paramétrica de Wilson. Mencionamos que este problema también puede ocurrir cuando se intenta 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).

11.2.1 Prueba de suma de rangos de Wilcoxon

En el capítulo 5 aprendimos que la prueba t de Student es adecuada para inferir sobre la media de una población a partir de una muestra de observaciones, siempre y cuando se verifiquen dos condiciones:

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

También vimos que esta prueba se podía extender a inferencias sobre la diferencia de las medias de dos poblaciones a partir de dos muestras independientes de ellas cuando se cumple que:

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

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

Como vimos en el capítulo 8, si usamos la prueba t en un escenario en que no se cumple alguna de estas condiciones, no hay garantías de que el resultado sea válido y, en consecuencia, las conclusiones que se obtengan a partir de él podrían ser equivocadas.

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 para dos muestras independientes[1]. Esta prueba 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").

En general, entonces, la prueba trabaja con dos muestras, que suelen llamarse "muestra A" y "muestra B" en los textos de estudio. Para darle mayor contexto, consideremos el siguiente ejemplo: 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, que llamaremos "índice de usabilidad", corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. La tabla 11.2 muestra los índices de usabilidad otorgados por cada participante.

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 los participantes consideran que 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 11.10) podemos observar que las distribuciones no se asemejan a una normal.

[1] Esta prueba también es alternativa para trabajar con una muestra

	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 11.2: muestras de índices de usabilidad para las interfaces de uso A y B.

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 (los valores se distribuyen de igual forma).
H_A: sí hay diferencia en la usabilidad de ambas interfaces (las distribuciones de los índices de usabilidad son distintas).

Notemos que, 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 los índices 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, posición en el *ranking* en chileno) 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 11.3 muestra el resultado de este proceso. Podemos notar que hay dos observaciones del 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[2].

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

$$S^A = 3{,}5 + 5{,}0 + 6{,}0 + 7{,}0 + 8{,}5 + 10{,}0 + 11{,}5 + 14{,}5 + 16{,}0 + 17{,}0 + 19{,}0 + 21{,}0 = 139$$

De manera análoga, para la muestra B se tiene:

$$S^B = 1{,}5 + 1{,}5 + 3{,}5 + 8{,}5 + 11{,}5 + 13{,}0 + 14{,}5 + 18{,}0 + 20{,}0 + 22{,}0 + 23{,}0 = 137$$

La suma de rangos para la muestra combinada está dada por la ecuación 11.6.

$$S^T = S^A + S^B = \frac{n^T\,(n^T + 1)}{2} \tag{11.6}$$

[2] En la literatura se menciona con frecuencia que otra condición para aplicar esta prueba es que la escala de medición sea intrínsecamente continua, con un número arbitrario de decimales. Esta condición permite suponer que **no habrá empates** al construir los rangos. Sin embargo, implementaciones actuales de la prueba son capaces de manejar la presencia de un **número acotado** de empates, realizando ciertos ajustes, por lo que también se puede aplicar a datos ordinales y discretos. Sin embargo, es importante tener en cuenta que cuando hay empates, las correcciones introducidas pueden afectar la precisión y potencia de la prueba.

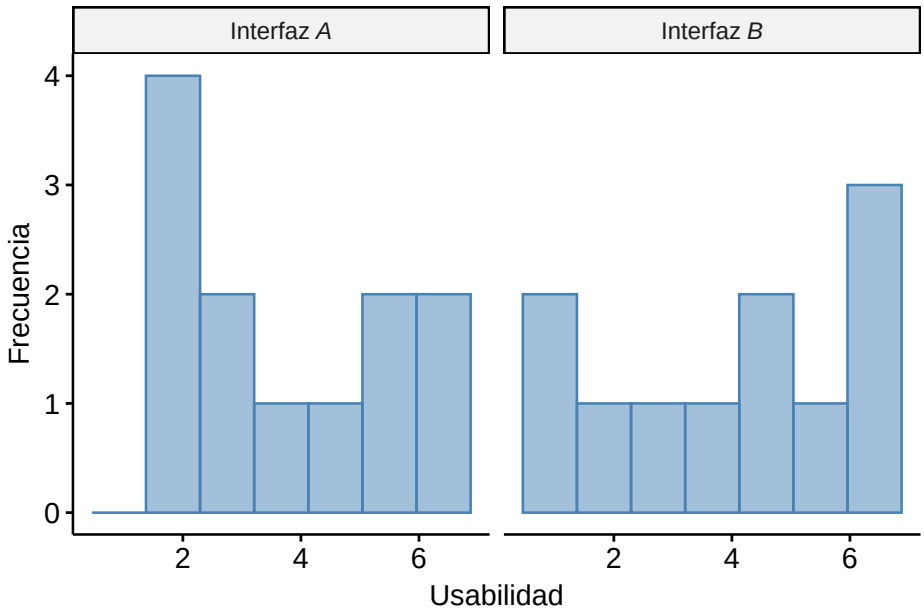


Figura 11.10: histogramas de las muestras descritas en la tabla 11.2.

Para el ejemplo:

$$S^T = 139 + 137 = \frac{23 \cdot (23+1)}{2} = 276$$

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 combinada se determina siempre mediante la ecuación 11.6 y el rango promedio es siempre como muestra la ecuación 11.7.

$$\overline{R} = \frac{S^T}{n^T} = \frac{n^T\ (n^T+1)}{2\ n^T} = \frac{n^T+1}{2} \tag{11.7}$$

Para el ejemplo:

$$\overline{R} = \frac{276}{23} = \frac{23+1}{2} = 12$$

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. Esto es equivalente a que la suma de los rangos de cada muestra contribuya de igual forma a la suma total, como presenta la ecuación 11.8.

$$\begin{aligned} S^A_{H_0} &= n^A\ \overline{R} = n^A\ \frac{n^T+1}{2} \\ S^B_{H_0} &= n^B\ \overline{R} = n^B\ \frac{n^T+1}{2} \end{aligned} \tag{11.8}$$

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 11.3: muestras combinadas con rango.

Para el ejemplo:

$$S^A_{H_0} = 12 \cdot \frac{(23+1)}{2} = 144$$
$$S^B_{H_0} = 11 \cdot \frac{(23+1)}{2} = 132$$

A partir de este punto, la prueba de Wilcoxon-Mann-Whitney tiene dos variantes, una para muestras grandes y otra para muestras pequeñas.

11.2.1.1 Prueba de suma de rangos de Wilcoxon para muestras grandes

Hasta ahora hemos determinado valores para las sumas de los rangos esperada en cada muestra cuando estas provienen de poblaciones con igual distribución (H_0). Así, los valores de las sumas de los rangos observados en las muestras pueden ser consideradas como estimaciones de estas sumas esperadas. En nuestro ejemplo:

$$S^A = 139 \triangleq S^A_{H_0} = 144$$
$$S^B = 137 \triangleq S^B_{H_0} = 132$$

Se ha demostrado que, para poblaciones con igual distribución, las sumas de los rangos tienen la misma desviación estándar, dada por la ecuación 11.9.

$$\sigma_S = \sqrt{\frac{n^A\ n^B\ (n^T+1)}{12}} \tag{11.9}$$

Con lo que para el ejemplo:

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

Cuando **ambas muestras tienen tamaño mayor o igual a 5**, siguiendo un procedimiento similar al descrito en la primera sección del capítulo 4, las distribuciones muestrales de S^A y S^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 S^A o S^B, dado por la ecuación 11.10, donde:

- S^M es cualquiera de los valores observados, S^A o S^B.
- $S^M_{H_0}$ es el valor nulo (la media de la distribución muestral de S^M si la hipótesis nula es cierta).
- σ_S es el error estándar de S^M (es decir, la desviación estándar de su distribución muestral).

$$z = \frac{(S^M - S^M_{H_0}) \pm 0{,}5}{\sigma_S} \tag{11.10}$$

Notemos que la ecuación 11.10 incluye un factor de corrección de continuidad, puesto que las distribuciones muestrales de las sumas de los rangos son intrínsecamente discretas (solo pueden asumir valores con decimales cuando existen rangos empatados). Este factor es negativo ($-0{,}5$) si $S^M > S^M_{H_0}$ y positivo ($0{,}5$) en caso contrario.

Volviendo al ejemplo, tenemos:

$$\begin{aligned} z^A &= \tfrac{(139-144)+0{,}5}{16{,}248} &&= -0{,}277 \\ z^B &= \tfrac{(137-132)-0{,}5}{16{,}248} &&= 0{,}277 \end{aligned}$$

Los valores z obtenidos a partir de S^A y S^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 la primera muestra, 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, que tiene una hipótesis alternativa bilateral, en R podemos calcular el valor p correspondiente 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 es posible descartar que las dos interfaces tienen niveles de usabilidad similares.

11.2.1.2 Prueba de suma de rangos de Wilcoxon para muestras pequeñas

Cuando las muestras son pequeñas (menos de 5 observaciones[3]), 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 similares a los obtenidos con la aproximación normal.

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 (ecuación 11.11), suponiendo alternadamente que cada una recibe los rangos más altos de la muestra combinada.

[3] Aunque algunos autores fijan en 10 e incluso ¡30 observaciones! como umbral para usar la aproximación normal.

$$
\begin{aligned}
S^A_{max} &= n^A\ n^B + \frac{n^T\ (n^T+1)}{2} \\
S^B_{max} &= n^B\ n^A + \frac{n^T\ (n^T+1)}{2}
\end{aligned}
\tag{11.11}
$$

Así, para el ejemplo:

$$
\begin{aligned}
S^A_{max} &= 12 \cdot 11 + \frac{12 \cdot (12+1)}{2} = 210 \\
S^B_{max} &= 11 \cdot 12 + \frac{11 \cdot (11+1)}{2} = 198
\end{aligned}
$$

Con esto podemos definir un nuevo estadístico de prueba U, como muestra la ecuación 11.12.

$$
\begin{aligned}
U^A &= S^A_{max} - S^A \\
U^B &= S^B_{max} - S^B \\
U &= \text{mín}(U^A, U^B)
\end{aligned}
\tag{11.12}
$$

Por lo que en el ejemplo:

$$
\begin{aligned}
U^A &= 210 - 139 = 71 \\
U^B &= 198 - 137 = 61 \\
U &= 61
\end{aligned}
$$

Si la hipótesis nula fuera verdadera, esperaríamos que:

$$
U^A = U^B = \frac{n^A\ n^B}{2}
$$

Para el ejemplo, bajo la hipótesis nula esperaríamos que $U^A = U^B = 66$.

En consecuencia, la pregunta asociada a la prueba de hipótesis es: si la hipótesis nula fuera verdadera, ¿qué tan probable es obtener un valor de U al menos tan pequeño como el observado?[4] Para el ejemplo esto sería: si no hay diferencias significativas en la usabilidad de ambas interfaces, ¿qué tan probable es obtener un valor $U \leq 61$?

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, para el ejemplo, ¡existen 676.039 combinaciones posibles! Afortunadamente existen tablas que permiten conocer el máximo valor U^{max} para el cual se rechaza la hipótesis nula para un nivel de significación dado, sin tener que revisar todas estas combinaciones.

Pero R no ofrece herramientas para calcular estos valores críticos, ni el valor p a partir del estadístico U observado, puesto que tiene implementadas funciones para la distribución del estadístico W, propuesto por Frank Wilcoxon en 1945, que lleva **a los mismos resultados**. Así, podemos aproximar el valor U^{max} (en rigor W^{max}), para una prueba bilateral, con la llamada `qwilcox(`$\alpha/2$`, `n^A`, `n^B`, lower.tail = TRUE)`. Sin embargo, debemos recordar que la distribución de U (como la de W) es discreta, por lo que el valor U^{max} devuelto por la llamada anterior está sobreestimado cuando su probabilidad no es exactamente $\alpha/2$. En este caso, debemos corregirlo restándole uno.

[4] Debemos notar que siempre se cumple que $U^A + U^B = n^A\ n^B$, por lo que en realidad podríamos haber escogido cualquiera de los valores U para realizar el procedimiento. Si se usara, en vez del menor, el mayor valor, es decir $U = \text{máx}(U^A, U^B)$, se debería responder qué tan probable es obtener un valor de U al menos tan grande como el observado.

Considerando el ejemplo, y un nivel de significación $\alpha = 0{,}05$, y como estamos realizando una prueba bilateral, obtenemos un primer valor para U^{max} con la llamada `qwilcox(0.05 / 2, 12, 11, lower.tail = TRUE)`, que devuelve 34. Para confirmar su exactitud, ejecutamos la llamada `pwilcox(34, 12, 11, lower.tail = TRUE)` para conocer más exactamente su probabilidad, lo que nos retorna el valor `0.02560947`. Como esta probabilidad no es exactamente $0,25$, debemos corregir y quedarnos con el valor $U^{max} = 33$ (que es el encontrado en las tablas de valores críticos para el estadístico U de Mann-Whitney).

Puesto que el valor observado es mayor que el valor crítico, $61 > 33$, fallamos al rechazar la hipótesis nula, por lo que concluimos con 95 % de confianza que no se puede descartar que la usabilidad de ambas interfaces sea la misma.

11.2.1.3 Prueba de suma de rangos de Wilcoxon en R

Como podría suponerse, 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 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, igual a cero por defecto.
- `conf.level`: nivel de confianza.

Notemos que al usar la función `wilcox.test()` con **una muestra** (no ejemplificado en este apunte), la hipótesis nula corresponde a que la población de origen se **distribuye simétricamente en torno al valor nulo** especificado (`mu` en la función). Cuando se usa con dos muestras independientes, como haremos para el ejemplo seguido, la hipótesis nula es que las poblaciones de origen **difieren por un desplazamiento igual al valor nulo** especificado.

Cuando la prueba es unilateral, la hipótesis alternativa es que el verdadero centro de la distribución poblacional, al trabajar con una muestra, o el verdadero desplazamiento entre las poblaciones, al trabajar con dos muestras independientes, se ubica a la izquierda o a la derecha del valor nulo considerado.

El script 11.5 muestra la aplicación de esta prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.11.

```
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

Warning message:
In wilcox.test.default(Interfaz_A, Interfaz_B, alternative = "two.sided",  :
  cannot compute exact p-value with ties
```

Figura 11.11: resultado de la prueba de Wilcoxon-Mann-Whitney para el ejemplo de las interfaces.

Script 11.5: prueba de Mann-Whitney para el ejemplo.

```
# Ingresar los datos.
Interfaz_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)
Interfaz_B <- c(5.0, 1.4, 5.6, 4.6, 6.7, 2.7, 1.3, 6.3, 3.7, 1.3, 6.8)

```

```r
# Establecer nivel de significación.
alfa <- 0.05

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

Podemos notar que la función `wilcox.test()` devuelve el mismo valor p (y el estadístico W con el mismo valor que el estadístico U) que calculamos anteriormente.

Vale la pena mencionar que cuando trabajemos con muestras es pequeñas, existen las funciones `wilcox_test()` del paquete `coin` y `wilcox.exact()` del paquete `exactRankTests`, entre otras opciones, para aplicar una prueba de Wilcoxon-Mann-Whitney exacta. Esta última usa, junto a algunos parámetros adicionales, los mismos argumentos que la función `wilcox.test()` mostrada aquí.

Notemos que la salida de la función `wilcox.test()` (figura 11.11) nos advierte que en el procedimiento se han producido empates, por lo que se hicieron ajustes y el valor p que reporta podría ser inexacto. La llamada `wilcox.exact(Interfaz_A, Interfaz_B)` permite conocer el valor p exacto: `0.7741`. Vemos que la diferencia es pequeña y no influye en la decisión tomada.

11.2.2 Prueba de rangos con signo de Wilcoxon

Recordemos que en el capítulo 5 también aprendimos una versión de la prueba t de Student para inferir sobre la media de las diferencias de dos muestras apareadas, siempre y cuando se verifiquen las siguientes dos condiciones:

1. Los pares de observaciones son independientes entre sí.
2. Las diferencias de observaciones apareadas siguen una distribución cercana a la normal.

Si no tenemos certeza de que se esté cumpliendo la segunda condición, recordando que esta no se cumple si la escala de medición no asegura intervalos iguales, podemos recurrir a la **prueba de rangos con signo de Wilcoxon**, que es conceptualmente similar a la prueba de suma de rangos de Wilcoxon presentada en la sección anterior, pero que en este caso corresponde a la alternativa no paramétrica a la prueba t de Student para **muestras apareadas**. Las condiciones que se deben cumplir para usar esta prueba son[5]:

1. Los pares de observaciones son independientes.
2. 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 de software desea evaluar la usabilidad de dos interfaces alternativas, A y B, para un nuevo producto, 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 cada participante da a la interfaz evaluada, que llamaremos "índice de usabilidad", corresponde al promedio simple de las puntuaciones de los 6 aspectos evaluados. Designados aleatoriamente, 5 participantes evaluaron primero la interfaz A, mientras que los otros 5 evaluaron primero la interfaz B. La tabla 11.4 muestra los índices de usabilidad otorgados por cada participante a cada una de las interfaces.

Formalmente, y notando que en este caso la hipótesis alternativa es unilateral, las hipótesis a contrastar serían:

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 (los

[5] La nota al pie sobre la condición de que la escala de medición de las observaciones sea intrínsecamente continua (pág. 182), también aplica para esta prueba.

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 11.4: muestra de los índices de usabilidad asignados por cada participante.

empates), 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 del par de observaciones. La tabla 11.5 ilustra el proceso descrito.

Participante	**Interfaz *A***	**Interfaz *B***	***A*−*B***	**$\lvert A-B\rvert$**	**Rango absoluto**	**Rango con signo**
4	4,7	4,7	0,0	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 11.5: asignación de rangos con signo para el ejemplo.

En teoría, una muestra de n pares distintos genera n **rangos no empatados** sin signo (columna "Rango absoluto" de la tabla 11.5). A su vez, cada uno de dichos rangos podría tomar valores positivos o negativos, por lo que se tienen ¡2^n posibles combinaciones de rangos con signos!

Por ejemplo, la tabla 11.6 muestra todas las posibles combinaciones de signos para $n = 3$ rangos, junto a las sumas de los rangos positivos (S^+), negativos (S^-) y en general (S^G). Podemos observar que la suma de los rangos positivos varía de 0 a 6, que la suma de los rangos negativos varía de -6 a 0, y que la suma general de los rangos con signo toma algunos valores de -6 a 6. Esto no es un accidente, puesto que para n pares, el rango máximo queda dado por la ecuación 11.13, que para $n = 3$ resulta $(3 \cdot 4)/2 = 6$.

$$R_{max} = \frac{n\,(n+1)}{2} \tag{11.13}$$

Si la hipótesis nula fuese cierta, los grupos presentarían valores similares para los rangos positivos y negativos y se distribuirían de manera homogénea, por lo que se esperaría que estas sumas tomaran los valores expresados en la ecuación 11.14, que corresponden a los valores nulos en el dominio de los rangos.

$$\begin{aligned} S^+_{H_0} = \quad S^-_{H_0} &= \quad R_{max}/2 \quad &= \frac{n\,(n+1)}{4} \\ S^G_{H_0} &= \quad S^+_{H_0} + S^-_{H_0} \quad &= 0 \end{aligned}$$

La figura 11.12 muestra las distribuciones muestrales de las sumas de los rangos positivos, negativos y en general, para distintos valores de n. En ella podemos apreciar que, a medida que el número de pares observados

Rango					
1	**2**	**3**	$\mathbf{S^+}$	$\mathbf{S^-}$	$\mathbf{S^G}$
+	+	+	6	0	6
+	+	-	3	-3	0
+	-	+	4	-2	2
+	-	-	1	-5	-4
-	+	+	5	-1	4
-	+	-	2	-4	-2
-	-	+	3	-3	0
-	-	-	0	-6	-6

Tabla 11.6: combinaciones de rangos positivos y negativos para una muestra de $n = 3$ pares.

aumenta, estas distribuciones **rápidamente** se aproximan cada vez más a distribuciones normales con medias en los valores nulos de la ecuación 11.14.

Aquí nos enfrentamos a un dilema, puesto que hay múltiples versiones de prueba de rangos con signo de Wilcoxon que usan estadísticos diferentes pero que llevan a resultados muy similares. Algunas usan como estadístico la suma de los rangos con signo (S^G), mientras que otras usan el menor valor absoluto de las sumas de los rangos positivos (S^+) y negativos (S^-). En este apunte usaremos el **estadístico V** que corresponde a la suma de los rangos con signo positivo, es decir $V = S^+$, sencillamente porque es el estadístico que reporta la función `wilcox.test()` que usaremos luego para aplicar esta prueba usando R. Se sabe que el error estándar para este estadístico está dado por la ecuación

$$\sigma_V = \sqrt{\frac{n\,(n+1)\,(2n+1)}{24}} \tag{11.14}$$

Volviendo al ejemplo (más grande) de la comparación de la usabilidad de dos interfaces (tabla 11.5), primero debemos notar que tenemos $n = 9$ pares de observaciones distintas (y un empate). Usando las definiciones anteriores, tendríamos:

$$\begin{aligned} V &= 3 + 5{,}5 + 5{,}5 + 7 + 8 + 9 && = 38 \\ S^+_{H_0} = V_0 &= \frac{9 \cdot (9+1)}{4} && = 22{,}5 \\ \sigma_V &= \sqrt{\frac{10 \cdot (10+1) \cdot (2 \cdot 10 + 1)}{24}} && \approx 8{,}441 \end{aligned}$$

La figura 11.12 presenta la distribución muestral para este caso (tercera fila, primera columna). Podemos notar que esta distribución comienza a parecerse bastante a una distribución normal con media 22,5 y desviación estándar alrededor de 8. Como ha sido tradicional, la prueba de hipótesis necesita responder si el valor observado $V = 38$ está lo suficientemente lejos del centro hipotético $V_0 = 22{,}5$ como para descartar una igualdad en la usabilidad de ambas interfaces.

Cuando la muestra de pares es grande, podemos trabaja bajo el supuesto de normalidad y calcular el estadístico de prueba z, de la forma que ha sido usual, dado por la ecuación 11.15.

$$z = \frac{V - V_0 \pm 0{,}5}{\sigma_V} \tag{11.15}$$

Al igual que para la prueba de suma de rangos de Wilcoxon, el estadístico de prueba incluye un factor de corrección de continuidad que es negativo si $V > V_0$ y positivo en caso contrario.

Así, para el ejemplo tenemos que:

$$z = \frac{38 - 22{,}5 - 0{,}5}{8{,}441} \approx 1{,}777$$

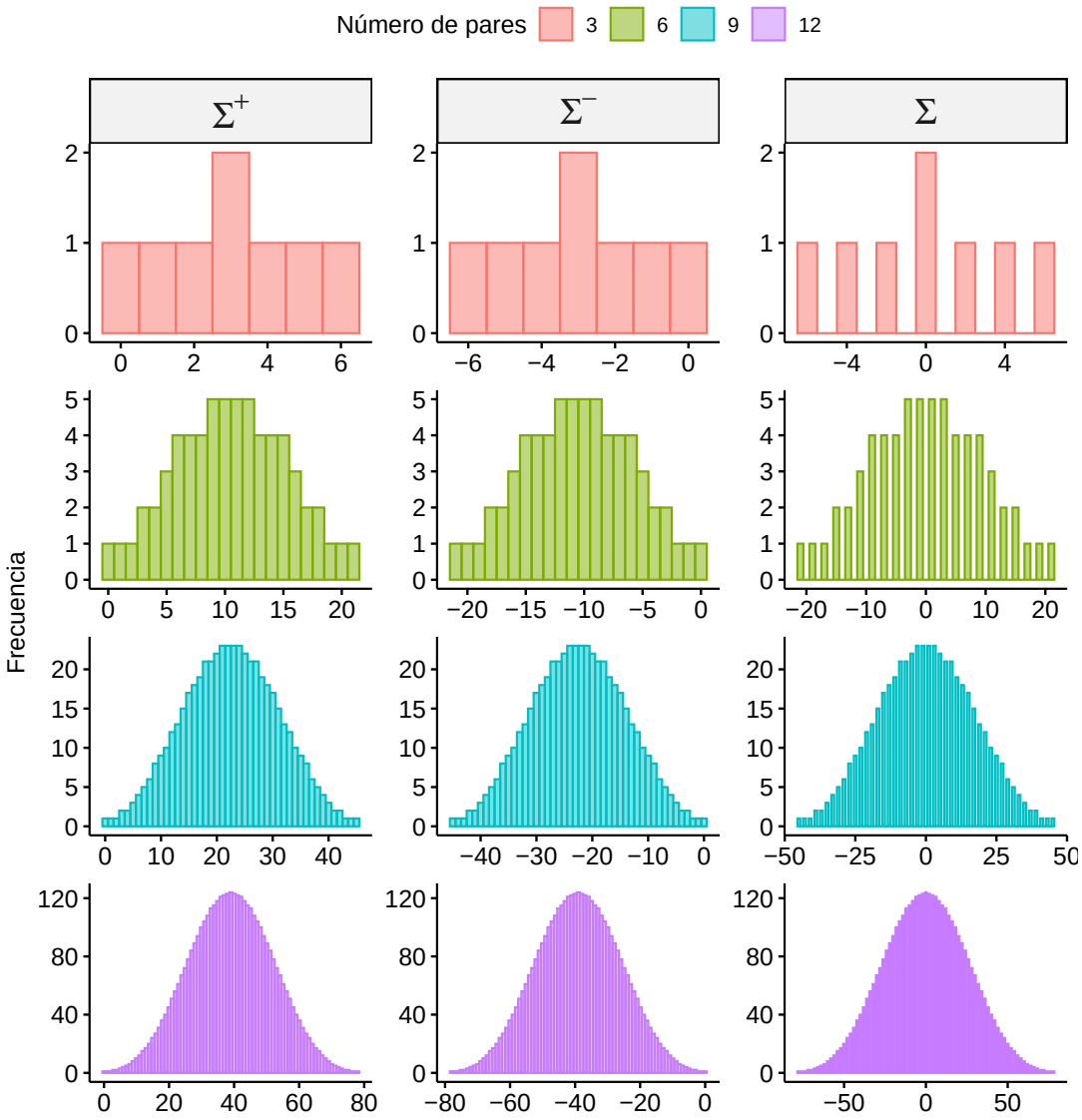


Figura 11.12: ejemplo de distribuciones muestrales de la sumas de los rangos con signo para muestras con 3, 6, 9 y 12 pares de observaciones.

Como hecho anteriormente, podemos obtener el valor p asociado a este estadístico de prueba mediante la llamada `pnorm(1.777, mean = 0, sd = 1, lower.tail = FALSE)` (no multiplicamos por 2, pues consideramos una prueba unilateral), obteniendo como resultado $p = 0{,}038$[6]. 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.

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 debemos asegurarnos de indicar que las muestras están apareadas a través de la llamada `wilcox.test(x, y, paired = TRUE, alternative, conf.level)`. Es decir, el valor por defecto para el parámetro `paired` es `FALSE` y este indica aplicar la prueba de suma de rangos de Wilcoxon a los datos, mientras que si explícitamente indicamos `paired = TRUE`, se aplica la prueba de rangos con signo de Wilcoxon. El script 11.6 muestra la aplicación de esta última prueba para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.13.

[6]Por supuesto, podríamos usar la llamada `pnorm(38 - 0.5, mean = 22.5, sd = 8.441, lower.tail = FALSE)` para obtener esta probabilidad, evitando la normalización.

```
 Wilcoxon signed rank test with continuity correction

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

Warning message:
In wilcox.test.default(Interfaz_A, Interfaz_B, paired = TRUE, alternative = "greater",  :
  cannot compute exact p-value with zeroes
```

Figura 11.13: resultado de la prueba de rangos con signo de Wilcoxon para el ejemplo.

Script 11.6: prueba de rangos con signo de Wilcoxon para el ejemplo.

```r
# Ingresar los datos.
Interfaz_A <- c(2.9, 6.1, 6.7, 4.7, 6.4, 5.7, 2.7, 6.9, 1.7, 6.4)
Interfaz_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(Interfaz_A, Interfaz_B, paired = TRUE,
                      alternative = "greater", conf.level = 1 - alfa)

print(prueba)
```

Vemos que el valor p entregado por la función `wilcox.test()` es el mismo que obtuvimos de forma manual. También observamos que la función nos advierte que tuvo que hacer correcciones por la presencia de empates (que aquí se les llama *zeroes*).

Además debemos tener en cuenta que esta función sigue el supuesto de normalidad, el que es válido para **n > 10 pares distintos**. Para nuestro ejemplo, o con muestras más pequeñas, tenemos que consultar una tabla de valores críticos para V o usar las funciones implementadas para su distribución (esto es, las funciones `psignrank(q, n, lower.tail)` y `qsignrank(p, n, lower.tail)`) o alguna implementación de una prueba exacta. En R existen, entre otras alternativas, las funciones `wilcoxsign_test()` del paquete `coin`, con argumentos `distribution = "exact"` y `zero.method = "Wilcoxon"` para reproducir el procedimiento visto aquí, o la función `wilcox.exact()` del paquete `exactRankTests`, indicando `paired = TRUE`. De hecho, usando la llamada `wilcox.exact(Interfaz_A, Interfaz_B, paired = TRUE, alternative = "greater")` permite conocer un p valor más confiable para nuestro ejemplo: $p = 0.037$ (que no es muy distinto al valor aproximado usando la suposición de normalidad).

11.2.3 Nota sobre las hipótesis e interpretación de las pruebas basadas en rangos

Es importante hacer una observación respecto a lo que se encuentra en Internet sobre las hipótesis e interpretación de las pruebas con rangos vistas en esta sección, ya que uno se topa con diferentes versiones que pueden llevar a confusión. Esto ocurre porque los estadísticos basados en rangos pueden tener significados ligeramente distintos dependiendo de los supuestos que se hagan sobre las distribuciones de las muestras (Fay & Proschan, 2010).

Una versión muy frecuente es que las pruebas con rangos comparan **medianas**. Cuando se trabaja con una muestra o la diferencia de dos muestras apareadas, la hipótesis nula es que la población de origen está centrada en torno al valor igual, menor o mayor que el valor nulo considerado (especificado con el argumento `mu` de la función `wilcox.test()`). Cuando, además, se supone que la población tiene una distribución **simétrica**, este valor nulo efectivamente corresponde al valor hipotético para la mediana de la población de origen.

Similarmente, al comparar muestras independientes, la hipótesis nula es que la diferencia entre los valores

en torno a los cuales están centradas las poblaciones de origen es igual, menor o mayor que el valor nulo considerado. Es necesario suponer que las poblaciones tienen distribuciones con la **misma forma simétrica** para poder hacer conclusiones sobre las diferencias de sus medianas.

Así, estas pruebas pueden resultar significativas por otras razones, como la presencia de asimetrías o diferencias en la dispersión, y solo descartando estas posibilidades, se les puede atribuir a diferencias en las medianas. Sin embargo, esto no es fácil de hacer con muestras pequeñas.

Por esta razón, en este apartado se ha usado la forma más robustas y libre de supuestos, considerando hipótesis que apuntan a detectar "diferencias" que no se refieren únicamente a las medianas, sino a si la población es dominantemente mayor o menor que el valor nulo hipotetizado o, en el caso de muestras independientes, que una de las distribuciones es dominantemente mayor o menor que la otra en general.

De aquí que aparece otra interpretación, relacionada con la probabilidad estocástica de que una observación aleatoria de la población de origen es mayor que el valor nulo, al trabajar con una muestra, o a otra observación aleatoria de la otra población. Es decir, se evalúa si la probabilidad $P(A > \mu_0^A)$ o $P(A > B)$, respectivamente, es mayor que 0,5.

Existen otras interpretaciones, pero que son encontradas con menor frecuencia.

11.2.4 Ejercicios propuestos (sección 11.2)

11.22 En la década del 1920 se hicieron los primeros estudios sobre la relación entre la velocidad de un automóvil con la distancia que necesita para detenerse. Los datos de estas pruebas se pueden encontrar el conjunto `cars` del paquete `datasets`. Con ellos, responde la siguiente pregunta: en promedio, la distribución de las distancias necesitadas para detener vehículos antiguos que viajaban a más de 10 millas por hora ¿se centra en un valor menor a 60 pies? No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.23 El conjunto `airquality` del paquete `datasets` contiene mediciones diarias de la calidad del aire en la ciudad de New York, EE.UU., registradas de mayo a septiembre de 1973. Verifica si la calidad del aire respecto del ozono es la misma los primeros 9 días de agosto que los primeros 9 días de septiembre. No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.24 El conjunto `ChickWeight` del paquete `datasets` contiene los resultados de un experimento del efecto de 4 tipos de dietas en el crecimiento temprano de pollitos. Verifica si las dietas 1 y 2 producen crecimientos similares. No olvides verificar si se cumplen las condiciones de la prueba que estás aplicando.

11.25 Da un ejemplo de una pregunta de investigación sobre las asignaturas comunes en ingeniería que requiera utilizar una prueba de Mann-Whitney para una muestra. Identifica bien las variables involucradas, justifica por qué no usar una prueba paramétrica equivalente, y enuncia las hipótesis a docimar.

11.26 Da un ejemplo de una pregunta de investigación sobre los conciertos realizados en Santiago que requiera utilizar una prueba Mann-Whitney para dos muestras debido a que la escala de la variable dependiente no permite usar una prueba t de Student. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.27 Da un ejemplo de una pregunta de investigación sobre el estado de la salud mental de estudiantes universitarios que requiera utilizar una prueba de sumas con signo de Wilcoxon por problemas con la escala de las mediciones. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.28 Da un ejemplo de una pregunta de investigación sobre el financiamiento de la educación superior en Chile que requiera utilizar una prueba de sumas con signo de Wilcoxon por problemas de normalidad. Identifica bien las variables involucradas y enuncia las hipótesis a docimar.

11.29 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?

11.30 Investiga qué alternativas existen para conocer el poder de una prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?

11.31 Investiga qué alternativas existen para estimar el tamaño del efecto para la prueba de sumas de rangos con signo de Wilcoxon. ¿Están implementadas en R?

11.32 Investiga qué alternativas existen para conocer el poder de una prueba de sumas de rangos con signo de Wilcoxon. ¿Están implementadas en R?