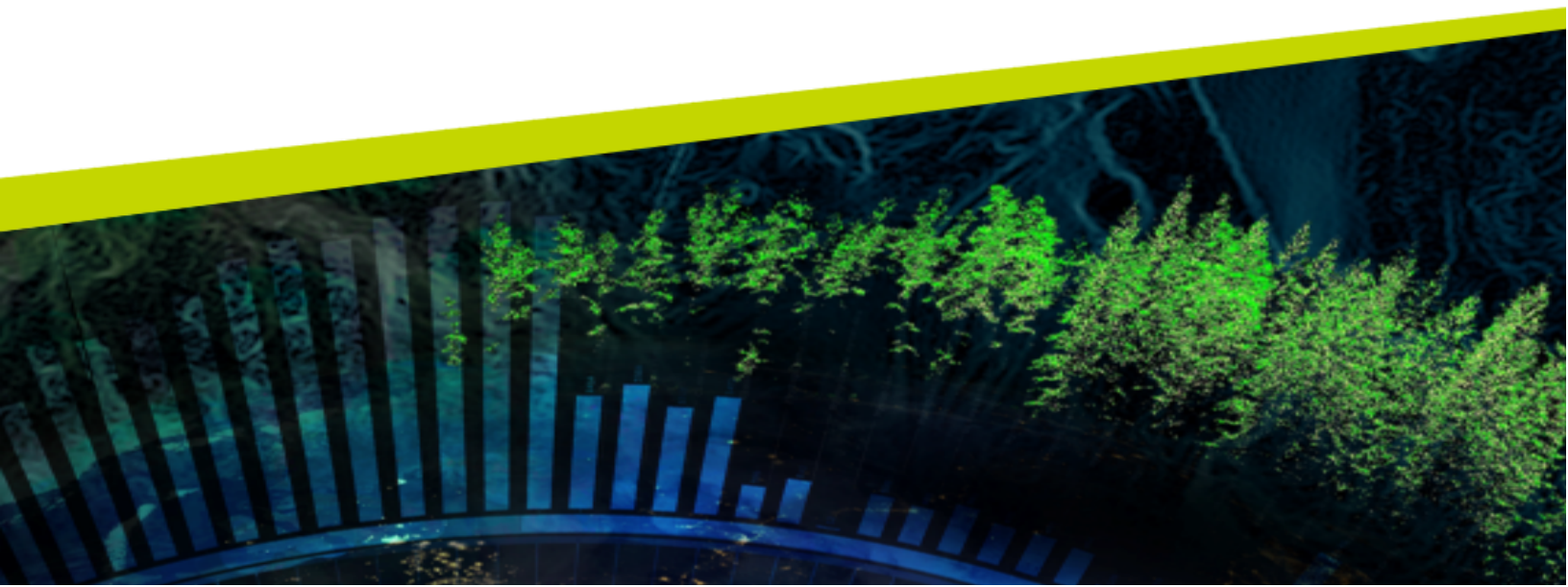
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



11.2 PRUEBAS NO PARAMÉTRICAS CON DOS MUESTRAS DE UNA VARIABLE NUMÉRICA

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).

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.

11.2.1 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. 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").

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 11.2 muestra las evaluaciones realizadas 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 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 11.11) podemos observar que las distribuciones no se asemejan a una normal.

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

	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: evaluación de las interfaces de usuario A y B.

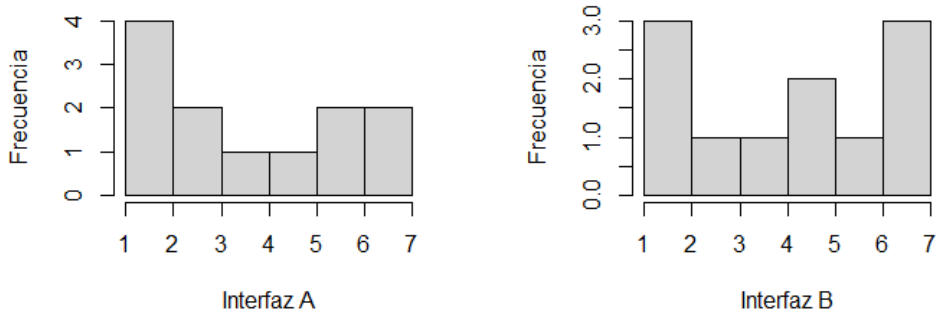


Figura 11.11: histogramas de las muestras.

H_0: no hay diferencia en la usabilidad (en términos de tendencia central) de ambas interfaces.
H_A: sí hay diferencia en la usabilidad (respecto de la tendencia central) de ambas interfaces.

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 11.3 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 $5,0$.

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:

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

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.

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

$$T_T = \frac{n_T \cdot (n_T + 1)}{2} \tag{11.6}$$

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 11.6 y la media de rangos de la muestra combinada es siempre como muestra la ecuación 11.7.

$$\mu = \frac{n_T \cdot (n_T + 1)}{2} \cdot \frac{1}{n_T} = \frac{n_T + 1}{2} \tag{11.7}$$

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.

11.2.1.1 Prueba de suma de rangos de Wilcoxon para muestras grandes

Hasta ahora, hemos determinado que:

- El valor observado $T_A = 139$ proviene de una distribución muestral con media $\mu_A = 144$.
- El valor observado $T_B = 137$ proviene de 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 11.8.

$$\sigma_T = \sqrt{\frac{n_A \cdot n_B \cdot (n_T + 1)}{12}} \tag{11.8}$$

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 11.9, 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).

$$z = \frac{(T_{obs} - \mu_{obs}) \pm 0,5}{\sigma_T} \tag{11.9}$$

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, 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[1]), 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 11.10. 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.

$$T_{x^{max}} = n_x \cdot n_y + \frac{n_x \cdot (n_x + 1)}{2} \tag{11.10}$$

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 11.11.

$$U_x = T_{x^{max}} - T_x \tag{11.11}$$

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 11.12, por lo que podemos escoger cualquiera de los valores U obtenidos para realizar el resto del procedimiento.

$$U_A + U_B = n_A \cdot n_B \tag{11.12}$$

Si la hipótesis nula fuese cierta, esperaríamos que:

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

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

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

Notemos que cuando se hace una prueba de suma de rangos de Wilcoxon con **una muestra** (no ejemplificado en este apunte), la hipótesis nula corresponde a que la población de origen tiene una localización centrada en el valor nulo especificado (`mu` en la función). Cuando se comparan dos muestras, como en el ejemplo que seguimos en esta sección, el valor nulo corresponde a la diferencia entre las localizaciones de tendencia central de las poblaciones de origen. El valor cero, entonces, se usa para hipotetizar la igualdad de las tendencias centrales de ambas poblaciones.

Similarmente, el argumento `alternative` permite indicar si la hipótesis alternativa apunta a que la verdadera localización o la verdadera diferencia entre localizaciones, al trabajar con una o dos muestras, respectivamente, se ubica a la izquierda o 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 11.12.

```
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 11.12: resultado de la prueba de Wilcoxon-Mann-Whitney para el ejemplo en uso.

Script 11.5: 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)
```

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 el que calculamos anteriormente.

Valdría la pena mencionar que cuando la muestra es pequeña, 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.

11.2.2 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 corresponde a la alternativa no paramétrica a la prueba t de Student con **muestras apareadas**. 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. Designados aleatoriamente, 5 participantes evaluaron primero la interfaz A, mientras que los otros 5 evaluaron primero la interfaz B. La tabla 11.4 muestra las valoraciones 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 11.4: evaluación de las interfaces de usuario/a A y B.

Formalmente, y notando que en este caso la hipótesis alternativa es unilateral, las hipótesis son:

H_0: las mismas personas no perciben diferencia en la usabilidad de ambas interfaces (igual localización de tendencia central).

H_A: las mismas personas consideran que la interfaz A tiene mejor usabilidad que la interfaz B (la localización de tendencia central de la interfaz A está a la derecha de la de 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 11.5 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	-	-
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.

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 para $n = 3$, junto a las sumas de los rangos positivos (Σ^+), negativos (Σ^-) y en general (Σ).

Rango					
1	**2**	**3**	**Σ^+**	**Σ^-**	**Σ**
+	+	+	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.

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, y los grupos presentaran valores similares para los rangos positivos y negativos 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.

$$\Sigma^+ \approx -\Sigma^- \approx \frac{R^{max}}{2} \approx \frac{n(n+1)}{4} \text{ y, en consecuencia, } \Sigma \approx 0 \tag{11.14}$$

La figura 11.13 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 aumenta, estas distribuciones **rápidamente** se aproximan cada vez más a distribuciones normales con medias en los valores nulos de la ecuación 11.14.

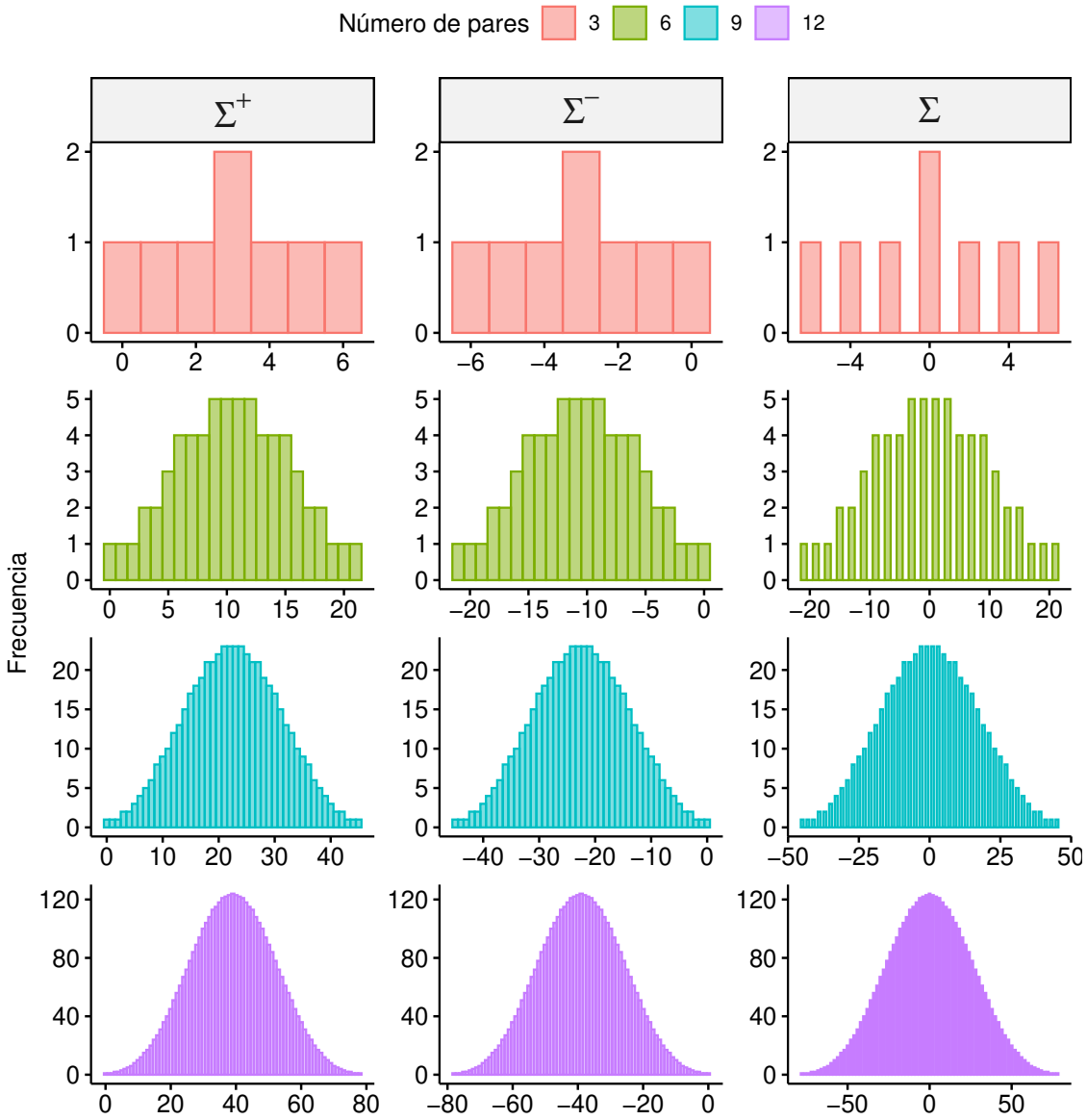


Figura 11.13: ejemplo de distribuciones muestrales de la sumas de los rangos con signo para muestras con 3, 6, 8 y 12 pares de observaciones.

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 (Σ), mientras que otras usan el menor valor absoluto de las sumas de los rangos positivos (Σ^+) y negativos (Σ^-). En este apunte usaremos el estadístico V que corresponde a la suma de los rangos con signo positivo, es decir $V = \Sigma^+$, sencillamente porque es el estadístico que reporta la función `wilcox.test()` que usaremos luego para aplicar esta prueba usando R.

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. La figura 11.13 presenta la distribución muestral para este caso (tercera fila, primera columna), cuya media es el valor nulo $V_0 = 9(9+1)/4 = 22{,}5$ y error estándar dado por la ecuación 11.15.

$$\sigma_V = \sqrt{\frac{n(n+1)\cdot(2n+1)}{24}} \tag{11.15}$$

Que para el caso de $n = 9$ resulta:

$$\sigma_V = \sqrt{\frac{9\cdot(9+1)\cdot(2\cdot 9+1)}{24}} = 8{,}441$$

En nuestro ejemplo, el estadístico V vale

$$V = 3 + 5{,}5 + 5{,}5 + 7 + 8 + 9 = 38.$$

¿Está 38 lo suficientemente lejos del valor 22,5 como para rechazar la posible igualdad de 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 11.16.

$$z = \frac{V - V_0}{\sigma_V} \tag{11.16}$$

Así, para el ejemplo tenemos que:

$$z = \frac{38 - 22{,}5}{8{,}441} = 1{,}836$$

Como en la prueba vista anteriormente, podemos obtener el valor p asociado a este estadístico de prueba mediante la llamada `pnorm(1.836, mean = 0, sd = 1, lower.tail = FALSE)` (no multiplicamos por 2, pues consideramos una prueba unilateral), obteniendo como resultado $p = 0{,}038$. 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 la prueba de rangos con signo de Wilcoxon para el ejemplo, obteniéndose los resultados que se presentan en la figura 11.14.

```
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 11.14: 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.

```
# 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, 5.5, 4.9)
b <- c(6.0, 2.8, 1.3, 4.7, 3.1, 1.8, 2.9, 4.0, 2.3, 1.6, 3.3, 2.8)
```

```r

# 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)
```

Notemos que el valor p resultado entregado por la función `wilcox.test()` es un poco distinto al conseguido de forma manual. Esto se debe a que esta función, como se observa en la figura 11.14, está aplicando una corrección de continuidad al calcular el estadístico de prueba z, descuenta 0,5 del numerador de la ecuación 11.16.

También 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 la distribución de V o usar 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`.

11.2.3 Nota sobre las hipótesis

Es importante hacer una observación respecto a lo que se encuentra en Internet sobre las hipótesis de las pruebas con rangos vistas en esta sección, ya que con frecuencia se menciona que estas comparan "medianas" o "distribuciones".

Debemos tener en claro que estas pruebas comparan **sumas de rangos**. Para que podamos interpretar la prueba de Mann-Whitney como una comparación de medianas, debemos hacer una suposición adicional: que las distribuciones de las dos poblaciones tienen la **misma forma** (aunque estén desplazadas en localizaciones distintas). Con esta suposición, o mejor aún **comprobación** (aunque esto no es fácil con muestras pequeñas), si la prueba de Mann-Whitney sugiere rechazar la hipótesis nula, se podría concluir que diferencias significativas de las medianas son la causa de este rechazo.

El argumento `mu` de la función `wilcox.test()` define el valor nulo de la prueba. Cuando se trabaja con una muestra (no ejemplificado en este capítulo) o la diferencia de dos muestras apareadas (subsección 11.2.2), se prueba la hipótesis nula que la distribución de origen es simétrica en torno al valor `mu`. Esto equivale a decir que `mu` es el valor nulo para **la mediana** de la distribución de origen, en el primer caso, o de la distribución de las diferencias de las variables de origen, en el segundo.

Cuando se comparan dos grupos independientes, se prueba la hipótesis que los parámetros de localización de las distribuciones de `x` e `y` difieren en `mu`. Como se dijo anteriormente, cuando estas distribuciones de origen tienen la misma forma (igual simetría y varianza), esto sería equivalente a verificar que poblaciones tienen **las mismas medianas**. Pero cuando la prueba es unilateral, solo se revisa si el parámetro de localización de la distribución de `x` está a la izquierda (`alternative = "less"`) o a la derecha (`alternative = "greater"`) del de la distribución de `y`.

11.2.4 Ejercicios propuestos

1. 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.

2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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?
8. Investiga qué alternativas conocer el poder de una prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?
9. 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?
10. Investiga qué alternativas conocer el poder de una prueba de sumas con rangos de Wilcoxon-Mann-Whitney. ¿Están implementadas en R?