Test de Wilcoxon de rangos signados

Hemos visto que, con mínimas hipótesis sobre la distribución subyacente (única mediana y distribución continua), el test del signo es UMP para las hipótesis unilaterales. Veremos ahora que, si agregamos la hipótesis de simetría, es posible hallar un test más potente que el test del signo, un **test basado en rangos**.

La teoría de los tests basados en rangos es más complicada que la del test del signo. Bajo H_o, el estadístico de un test de rangos puede ser representado como una suma de v.a. independientes pero no idénticamente distribuidas y, bajo H_1, se pierde inclusive la independencia. Por ello, necesitaremos nuevas versiones del TCL.

¿Cómo justificar la hipótesis de simetría? En el problema de una muestra (posición) puede haber razones valederas para suponer simetría de la distribución subyacente. En el caso del diseño de datos apareados, sean (T,C) dos v.a. representando un tratamiento y un control, respectivamente y sea $F_{TC}(t,c)$ su función de distribución conjunta. Suponiendo que los sujetos son asignados a tratamiento o control en forma aleatoria e independiente, la hipótesis nula de no diferencia entre tratamiento y control, implica que $F_{TC}(t,c) = F_{TC}(c,t)$ y por lo tanto que X=T-C tiene distribución simétrica (es decir, $F(x) = 1 - F(-x)$). Si la alternativa especifica que el tratamiento agrega una constante al control, el problema se reduce a un problema de posición sobre X, es decir a testear

$$H_o: \theta = 0 \qquad \text{vs} \qquad H_1: \theta > 0$$

siendo θ el centro de simetría de X (será además la media si ésta existe).

Supondremos que $X_1,...,X_n$ es una muestra aleatoria de una distribución $F(x-\theta)$ con $F \in \Omega_s$, siendo

$$\Omega_s = \{F \;/\; F \text{ es absolutamente continua, simétrica y con única mediana en } 0\}$$

y que se desea testear

$$H_o: \theta = 0 \qquad \text{vs} \qquad H_1: \theta > 0$$

El test del signo se basa en información sobre el signo de las observaciones y no utiliza información sobre la distancia de las observaciones al cero. Sin embargo, si la distribución es simétrica alrededor de 0, el vector de valores absolutos $|X_1|, |X_2|,...,|X_n|$ es un estadístico suficiente y por lo tanto, parece razonable tratar de incorporar esta información.

Sea $|X|^{(1)} \leq |X|^{(2)} .... \leq |X|^{(n)}$, la muestra de valores absolutos ordenados y

R_j = rango ($|X_j|$) es decir $|X_j| = |X|^{(R_j)}$

D_j = j-ésimo antirango es decir $|X_{D_j}| = |X|^{(j)}$

<u>Estadístico del test</u>: el estadístico del test de Wilcoxon (1945), T^+, es la suma de los rangos de los valores absolutos de las observaciones mayores que 0 en la muestra original. Es decir, si definimos

$$W_j = \begin{cases} 1 & si \;\; |X|^{(j)} \; corresponde \; a \; una \; observación \; mayor \; que \; 0 \\ 0 & en \; caso \; contrario \end{cases}$$

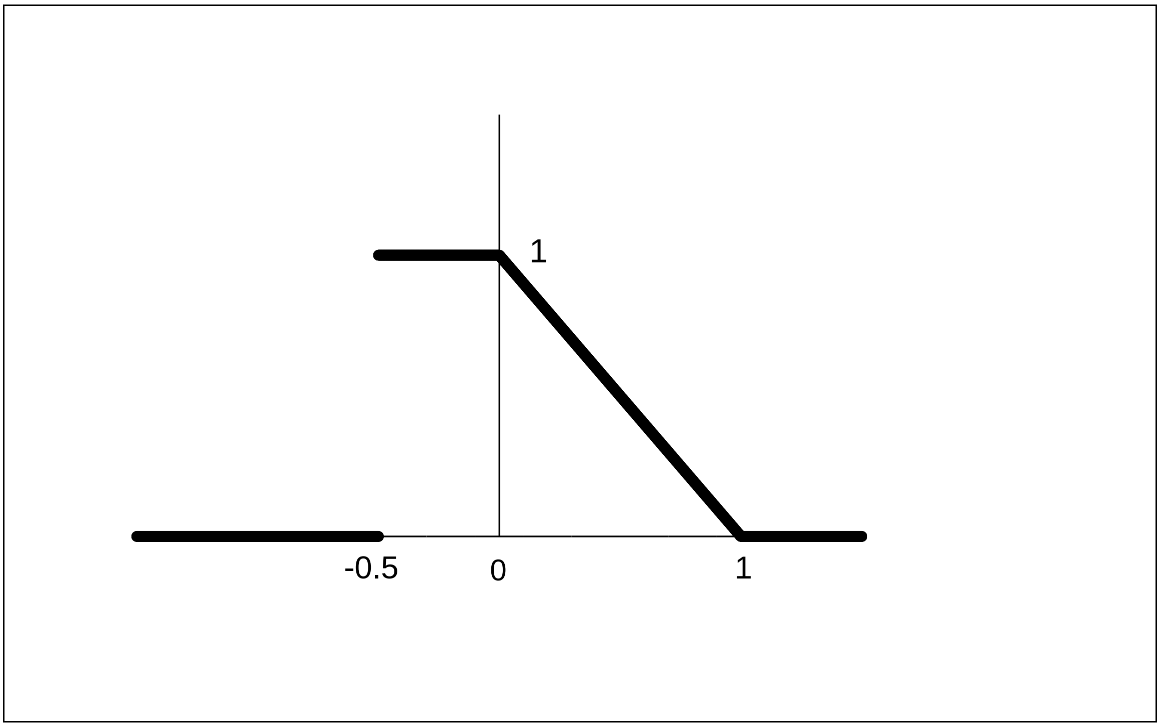
$$T^{+} = \sum_{j=1}^{n} j \, W_j = \sum_{j=1}^{n} R_j \, s(X_j)$$

siendo

$$s(x) = \begin{cases} 1 & si \; x > 0 \\ 0 & si \; x \le 0 \end{cases}$$

Observación: Si θ > 0 y la distribución simétrica se halla desplazada hacia la derecha, las observaciones positivas tienden a estar más alejadas del 0 que las negativas, entonces T^+ tiende a ser grande y se rechazaría H_o.

La mediana puede ser 0 aunque la distribución sea asimétrica. Observemos la siguiente función de densidad:



Es fácil ver que, en este caso, T^+ tenderá a ser grande aun cuando la mediana es 0. La hipótesis de simetría es necesaria para evitar interpretaciones erróneas de los valores grandes del estadístico. Si se conoce la mediana de la distribución, T^+ provee un test de simetría.

Hipótesis a testear y región de rechazo: Al testear

$$H_o: \theta = 0 \qquad \text{vs} \qquad H_1: \theta > 0$$

se rechazará H_o si $T^+ > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el percentil 1-α de la distribución exacta del estadístico, que ha sido tabulada.

Si las hipótesis a testear fuesen

$$H_o: \theta = 0 \qquad \text{vs} \qquad H_1: \theta < 0$$

se rechazaría H_o si $T^+ \leq w_\alpha$, donde w_α es el percentil α de la distribución exacta. También se podría definir T^- como la suma de los rangos de los valores absolutos de las observaciones menores que 0 en la muestra original. Es decir, definimos

$$\tilde{s}(x) = \begin{cases} 1 & \text{si } x < 0 \\ 0 & \text{si } x \geq 0 \end{cases}.$$

$$T^- = \sum_{j=1}^{n} R_j\, \tilde{s}(X_j) = \frac{n(n+1)}{2} - T^+$$

La última igualdad vale si no hay observaciones iguales a 0. Usando T^- se rechazaría H_o si $T^- > w_{1-\alpha}$.

En caso de producirse empates, se asigna a cada observación empatada el promedio de los rangos que tendrían si no fuesen empates. Por ejemplo, si las observaciones ordenadas son

1 1 2 3 4 4 4 5 6

los correspondientes rangos serían

1.5 1.5 3 4 6 6 6 8 9

¿Cómo se trabaja si hay observaciones iguales al valor a testear (que podemos suponer es 0)? . Hay dos propuestas:

- Eliminar los valores iguales a 0 y trabajar con el tamaño de muestra reducido, es decir con $(n - n_o)$.

- Pratt (1959) sugiere ordenar los datos incluyendo los ceros, calcular los rangos con todos los datos y luego, calcular T^+ como antes, es decir sin contar los ceros. Si se usa esta forma, se debe usar la aproximación Normal o las tablas exactas dadas por Rahe (1974, JASA, 368-373).

Si hay empates o si n es grande, conviene usar el siguiente estadístico:

$$T = \frac{T_o}{\left(\sum_{i=1}^{n} R_i^2\right)^{1/2}}$$

con $T_o = T^+ - T^-$. Si no hay empates,

$$T = \frac{T_o}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} = \frac{2\,T^+ - \frac{n(n+1)}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

El estadístico T tiene distribución asintótica Normal estándar.

Resumamos las hipótesis a testear y las zonas de rechazo.

A. H_o: $\theta = 0$ vs H_1: $\theta \neq 0$

Se rechaza H_o si $T^+ > w_{1-\alpha/2}$ o si $T^+ \leq w_{\alpha/2}$, o bien si hay empates o la muestra es grande, si $|T| > z_{\alpha/2}$.

B. H_o: $\theta = 0$ vs H_1: $\theta < 0$

Se rechaza H_o si $T^+ \leq w_\alpha$ (o equivalentemente si $T^- > w_{1-\alpha}$), o bien si hay empates o la muestra es grande, si $T \leq -z_\alpha$.

C. H_o: $\theta = 0$ vs H_1: $\theta > 0$

Se rechaza H_o si $T^+ > w_{1-\alpha}$, o bien si hay empates o la muestra es grande, si $T > z_\alpha$.

Distribución del estadístico de Wilcoxon bajo H_0: $\theta = 0$:

Teorema: Bajo H_o, y si $F \in \Omega_S$

a) $(s(X_1), ..., s(X_n))$ y $(R_1, ..., R_n)$ son independientes.

b) $W_1, ..., W_n$ son independientes e idénticamente distribuidos con $W_j \sim Bi(1,1/2)$

Entonces, $T^+ = \sum_{j=1}^{n} j\,W_j$ es combinación lineal de v.a. i.i.d. Bi(1,1/2) bajo H_o, y por lo tanto es distribución libre. Además

$$E(T^+) = \frac{n(n+1)}{4} \qquad V(T^+) = \frac{n(n+1)(2n+1)}{24}$$

Demostración: a) Como $(R_1,..., R_n)$ es función de $(|X_1|,...,|X_n|)$, y los pares $(s(X_i), |X_i|)$, i =1,...,n son independientes, es suficiente mostrar que $s(X_i)$ y $|X_i|$ son independientes.

$$P(s(X_i) = 1, |X_i| \leq x) = P(0 < X_i \leq x) = F(x) – F(0) =$$

$$= F(x) – ½ = ½ \; (2 F(x) – 1) = P(s(X_i) = 1) P(|X_i| \leq x)$$

Del mismo modo, se trabaja con $P(s(X_i) = 0, |X_i| \leq x)$.

Así como los signos son independientes de los rangos, lo son también de los antirangos, o sea $(s(X_1), ...,s(X_n))$ y $(D_1,...,D_n)$ son independientes.

b) Sea $D=(D_1,...,D_n)$ y $d=(d_1,...,d_n)$, entonces usando que $W_j = s(X_{D_j})$,

$$P(W_1 = w_1,...,W_n = w_n) = \sum_d P(s(X_{D_1}) = w_1,....,s(X_{D_n}) = w_n \;/\; D = d)\; P(D = d)$$

$$= \sum_d P(s(X_{d_1}) = w_1,....,s(X_{d_n}) = w_n)P(D = d) = \left(\frac{1}{2}\right)^n \sum_{\mathrm{d}} P(D = d) = \left(\frac{1}{2}\right)^n$$

Por lo tanto, $P(W_1 = w_1,...,W_n = w_n) = \prod_{i=1}^{n} P(W_i = w_i), \qquad P(W_i = w_i) = 1/2.$

Además, como $T^+ = \sum_{j=1}^{n} j\, W_j$, bajo H_o,

$$E(T^+) = \sum_{j=1}^{n} j\; \frac{1}{2} = \frac{n(n+1)}{4}$$

$$V(T^+) = \sum_{j=1}^{n} j^2 \frac{1}{4} = \frac{n(n+1)(2n+1)}{24}$$

Veamos cómo se obtiene la distribución exacta de T^+ con un ejemplo: Sea n=4. Los posibles rangos de los valores absolutos son 1, 2, 3 y 4. En la siguiente tabla se presentan las posibles asignaciones de signos a los rangos 1, 2, 3 y 4, con el valor asociado del estadístico T^+ . Recordemos que cada signo es + o – con probabilidad ½.

Rangos				T^+
1	2	3	4	
+	+	+	+	10
+	+	+	-	6
+	+	-	+	7
+	-	+	+	8
-	+	+	+	9
+	+	-	-	3
+	-	+	-	4
+	-	-	+	5
-	+	+	-	5
-	+	-	+	6
-	-	+	+	7
-	-	-	+	4
-	-	+	-	3
-	+	-	-	2
+	-	-	-	1
-	-	-	-	0

Como, bajo la hipótesis nula, cada configuración tiene probabilidad 1/16, podemos obtener $P(T^+ = k)$ para todo k. Por ejemplo,

$P(T^+ = 10) = 1/16$
$P(T^+ = 6) = 2/16$
$P(T^+ > 8) = 2/16$

Otra forma de obtener la distribución exacta es usando la función generadora de momentos.

Lema: Bajo H_o, la función generadora de momentos de T^+ es

$$M(t) = E\left(e^{t\,T^+}\right) = \frac{1}{2^n}\prod_{j=1}^{n}\left(1 + e^{t\,j}\right)$$

Por otro lado,

$$E\left(e^{t\,T^+}\right) = \sum_{k=0}^{\frac{n(n+1)}{2}} e^{t\,k}\,P(T^+ = k) = \sum_{k=0}^{\frac{n(n+1)}{2}} e^{t\,k}\,a_k$$

con $a_k = P(T^+ = k)$.

A partir de la función generadora de momentos podemos obtener los momentos de T^+ y su función de probabilidad puntual.

¿Cómo lo haríamos para nuestro ejemplo (n=4)?.

Si n=2,

$$M_2(t) = \frac{1}{2^2}(1+e^t)(1+e^{2t}) = \frac{1}{2^2}(1+e^t+e^{2t}+e^{3t})$$

entonces, P(T^+ = 0) = P(T^+ = 1) = P(T^+ = 2) = P(T^+ = 3) = ¼.

Si n=3,

$$M_3(t) = \frac{1}{2^3}(1+e^t)(1+e^{2t})(1+e^{3t}) = \frac{1}{2}M_2(t)(1+e^{3t})$$

Si n=4,

$$M_4(t) = \frac{1}{2^4}(1+e^t)(1+e^{2t})(1+e^{3t})(1+e^{4t}) = \frac{1}{2}M_3(t)(1+e^{4t})$$

y obtenemos, el siguiente desarrollo:

$$M_4(t) = \frac{1}{16}(e^{0t}+e^{1t}+e^{2t}+2e^{3t}+2e^{4t}+2e^{5t}+2e^{6t}+2e^{7t}+e^{8t}+e^{9t}+e^{10t})$$

y por lo tanto la siguiente función de probabilidad de T^+:

T^+	0	1	2	3	4	5	6	7	8	9	10
p	1/16	1/16	1/16	2/16	2/16	2/16	2/16	2/16	1/16	1/16	1/16

Es posible programar este algoritmo recursivo y obtener la distribución exacta de T^+ para cualquier valor de n.

Distribución asintótica del estadístico de Wilcoxon bajo H_0: θ = 0: Supongamos que n es grande. Dado que T^+ es una combinación lineal de los W_i, que bajo H_0 son independientes e idénticamente distribuidos, debemos utilizar la siguiente versión del Teorema Central del Límite (Lindeberg):

Teorema: Sean $V_1,...V_n$ v.a. i.i.d con $E(V_i)=0$ y $Var(V_i) = \sigma^2$, $0 < \sigma^2 < \infty$. Definamos

$$S = \frac{\sum_{i=1}^{n} a_i V_i}{\sqrt{n}}$$

Si $\frac{max|a_i|}{\sqrt{\sum_{i=1}^{n} a_i^2}} \rightarrow 0$, entonces

$$\frac{S}{[Var(S)]^{1/2}} \xrightarrow{d} Z \sim N(0,1) \qquad \text{con} \qquad Var(S) = \sigma^2 \sum_{i=1}^{n} a_i^2 / n$$

Demostración: Teorema A9, pag 301, Hettmansperger.

En nuestro caso,

$$T^+ = \sum_{j=1}^{n} j\, W_j$$

entonces, eligiendo $V_i = W_i - 1/2$ y valores adecuados de a_i y verificando las condiciones del teorema, se obtiene que

$$\frac{T^+ - n(n+1)/4}{\sqrt{\dfrac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} Z \sim N(0,1)$$

Observaciones: 1) Al usar la aproximación, puede aplicarse corrección por continuidad.

2) Hay algunas modificaciones que mejoran la aproximación, como por ejemplo una debida a Fellingham y Stocker (JASA, 1964), quienes mostraron que

$$P(T^+ \leq k) \cong \Phi(t) + \left[\frac{(3n^2 + 3n - 1)}{10n(n+1)(2n+1)}\right](t^3 - 3t)\varphi(t)$$

donde $t = (k + 0.5 - E(T^+)) / [Var(T^+)]^{1/2}$ y φ es la función de densidad normal standard.

Aplicación del test de Wilcoxon a datos apareados: Sean (X_1,Y_1),...., (X_n,Y_n) observaciones independientes. Podemos pensar que X_i es el resultado correspondiente al control e Y_i el correspondiente al tratamiento. Sea

$$D_i = Y_i - X_i \qquad \forall \quad i = 1,\ldots,n$$

El test de Wilcoxon para datos apareados consiste en la aplicación del test a las diferencias D_i, suponiendo que la distribución de las diferencias es simétrica.

Ejemplos: 1) Un fabricante de planchas, deseando probar la precisión del control del termostato en la posición de 500°F, da instrucciones a un ingeniero para que obtenga temperaturas reales a ese ajuste en 15 planchas, utilizando un termopar. Las mediciones obtenidas son:

Temperaturas		
529.77	541.08	467.91
462.75	486.04	542.15
464.20	539.94	495.13
489.17	489.07	531.73
531.09	500.65	527.88

Se desea testear

$$H_o: \theta = 500 \qquad \text{vs} \qquad H_1: \theta \neq 500$$

siendo θ la mediana de la distribución subyacente.

Restando 500 a cada observación y ordenando esas diferencias según sus valores absolutos, se obtiene

(+)0.65 (-)4.87 (-)10.83 (-)10.93 (-)13.96 (+)27.88 (+)29.77 (+)31.09

(+)31.73 (-)32.09 (-)35.80 (-)37.25 (+)39.94 (+)41.08 (+)42.15

Construimos el estadístico T^+, sumando los rangos correspondientes a las diferencias positivas

$$T^+ = 1 + 6 + 7 + 8 + 9 + 13 + 14 + 15 = 73$$

A nivel 0.05, se rechaza H_o si $T^+ > w_{0.975} = 120 - 26 = 94$ ó $T^+ \leq w_{0.025} = 26$. Estos valores se encuentran en la tabla A13 del libro de Conover. Por lo tanto no se rechaza H_o.

S-PLUS o R proveen el p-valor. Por ejemplo, la salida de R correspondiente a estos datos es la siguiente:

```
wilcox.test(ejemplo59,mu=500,alternative="two.sided")

        Wilcoxon signed rank test

data:  ejemplo59
V = 73, p-value = 0.4887
alternative hypothesis: true location is not equal to 500
```

2) Se realizó un estudio comparativo en el cual se evaluó la efectividad de dos métodos, uno tradicional y uno moderno de enseñanza del álgebra. En ese estudio 14 individuos fueron extraídos al azar de la población de interés y se formaron 7 pares en base a su IQ. Los miembros de cada par fueron asignados al azar a uno de los dos métodos de enseñanza, y posteriormente ambos grupos fueron instruidos durante 3 semanas. Todos los estudiantes rindieron el mismo examen al final del periodo de instrucción y los resultados obtenidos fueron los siguientes:

Par	Moderno	Tradicional	D_i	Rango (+)	Rango (-)
1	31	36	-5		3
2	42	38	4	2	
3	44	33	11	6	
4	48	36	12	7	
5	51	53	-2		1
6	57	49	8	4	
7	62	52	10	5	

Se deseaba testear

$$H_o: \theta = 0 \quad \text{vs} \quad H_1: \theta > 0$$

siendo θ la mediana de las diferencias D = Moderno – Tradicional. La zona de rechazo de nivel 0.05 para n = 7 es $T^+ > 24$, entonces a este nivel, como $T^+ = 24$ no se rechaza H_o. La correspondiente salida de R es

```
wilcox.test(ejemplo60[,1],ejemplo60[,2],alternative="greater",
            paired=TRUE)

    Wilcoxon signed rank test

data:  ejemplo60[, 1] and ejemplo60[, 2]
V = 24, p-value = 0.05469
alternative hypothesis: true location shift is greater than 0
```