Distribución asintótica del estadístico T^+ de los rangos signados de Wilcoxon

deepseek.com

14 de abril de 2026

1. Introducción

Sea $X_1, \dots, X_n$ una muestra aleatoria simple de una distribución continua y simétrica respecto a 0 (bajo la hipótesis nula). Definimos el estadístico de Wilcoxon para una muestra como

$$T^+ = \sum_{i=1}^{n} R_i \, I_i,$$

donde R_i es el rango de $|X_i|$ entre $|X_1|, \dots, |X_n|$ (con $R_i \in \{1, \dots, n\}$) e $I_i = \mathbf{1}_{(0,\infty)}(X_i)$ es el indicador de que la observación es positiva.
El objetivo es demostrar que, bajo H_0,

$$\frac{T^+ - \mathbb{E}[T^+]}{\sqrt{\mathrm{Var}(T^+)}} \xrightarrow{d} \mathcal{N}(0, 1),$$

y proporcionar explícitamente la esperanza y la varianza.

2. Independencia de signos y rangos

Bajo H_0, la distribución de X_i es simétrica continua alrededor de 0. Entonces:

- Los signos $\mathrm{sgn}(X_i)$ son independientes entre sí y de los valores absolutos $|X_i|$.
- Los rangos R_i son una función únicamente de los valores absolutos, luego también son independientes de los signos.
- Además, los signos son equiprobables: $\mathbb{P}(I_i = 1) = 1/2$, e independientes.

Por tanto, el vector $(R_1, \dots, R_n)$ es una permutación aleatoria uniforme de $\{1, \dots, n\}$ independiente de $(I_1, \dots, I_n)$.

3. Esperanza y varianza de T^+

3.1. Esperanza

Usando la independencia,

$$\mathbb{E}[T^+] = \sum_{i=1}^{n} \mathbb{E}[R_i]\,\mathbb{E}[I_i] = \sum_{i=1}^{n} \frac{n+1}{2} \cdot \frac{1}{2} = \frac{n(n+1)}{4}.$$

3.2. Varianza

No podemos sumar simplemente las varianzas porque los R_i no son independientes. Sin embargo, aprovechamos que $\sum_{i=1}^{n} R_i = n(n+1)/2$ es constante.

$$\mathrm{Var}(T^+) = \sum_{i=1}^{n} \mathrm{Var}(R_i I_i) + \sum_{i \neq j} \mathrm{Cov}(R_i I_i, R_j I_j).$$

Dado que I_i y R_i son independientes:

$$\mathrm{Var}(R_i I_i) = \mathbb{E}[R_i^2]\mathbb{E}[I_i^2] - (\mathbb{E}[R_i]\mathbb{E}[I_i])^2 = \frac{1}{2}\mathbb{E}[R_i^2] - \frac{1}{4}(\mathbb{E}[R_i])^2.$$

Para $i \neq j$, usando la independencia entre (I_i, I_j) y (R_i, R_j),

$$\mathrm{Cov}(R_i I_i, R_j I_j) = \mathbb{E}[R_i R_j]\mathbb{E}[I_i I_j] - \mathbb{E}[R_i]\mathbb{E}[I_i]\mathbb{E}[R_j]\mathbb{E}[I_j] = \frac{1}{4}\,\mathrm{Cov}(R_i, R_j).$$

Entonces

$$\mathrm{Var}(T^+) = \sum_i \Big(\frac{1}{2}\mathbb{E}[R_i^2] - \frac{1}{4}(\mathbb{E}R_i)^2\Big) + \frac{1}{4}\sum_{i \neq j} \mathrm{Cov}(R_i, R_j).$$

Observamos que

$$\sum_{i \neq j} \mathrm{Cov}(R_i, R_j) = -\sum_i \mathrm{Var}(R_i),$$

pues $\mathrm{Var}(\sum_i R_i) = 0$ implica $\sum_i \mathrm{Var}(R_i) + \sum_{i \neq j} \mathrm{Cov}(R_i, R_j) = 0$. Sustituyendo:

$$\begin{aligned}
\mathrm{Var}(T^+) &= \sum_i \Big(\frac{1}{2}\mathbb{E}[R_i^2] - \frac{1}{4}(\mathbb{E}R_i)^2\Big) - \frac{1}{4}\sum_i \mathrm{Var}(R_i) \\
&= \sum_i \Big(\frac{1}{2}\mathbb{E}[R_i^2] - \frac{1}{4}(\mathbb{E}R_i)^2 - \frac{1}{4}\big(\mathbb{E}[R_i^2] - (\mathbb{E}R_i)^2\big)\Big) \\
&= \sum_i \frac{1}{4}\mathbb{E}[R_i^2].
\end{aligned}$$

Como los rangos son una permutación de $\{1, \ldots, n\}$, $\mathbb{E}[R_i^2] = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{(n+1)(2n+1)}{6}$. Por tanto,

$$\mathrm{Var}(T^+) = \frac{1}{4} \cdot n \cdot \frac{(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}.$$

4. Distribución exacta de T^+ bajo H_0

Un hecho clave: la distribución de T^+ no depende de la distribución subyacente de los $|X_i|$ y es idéntica a la de $\sum_{i=1}^n i\,B_i$, donde $B_1,\dots,B_n$ son independientes con $\mathbb{P}(B_i=1)=1/2$.

Demostración. El vector de rangos $(R_1,\dots,R_n)$ es una permutación aleatoria uniforme independiente de los signos. Sea σ la permutación tal que $R_{\sigma(i)}=i$. Entonces

$$T^+ = \sum_{i=1}^n R_i I_i = \sum_{i=1}^n i\, I_{\sigma(i)}.$$

Como σ es independiente de $(I_1,\dots,I_n)$ y los I_i son i.i.d. Bernoulli(1/2), la variable $(I_{\sigma(1)},\dots,I_{\sigma(n)})$ tiene la misma distribución que $(I_1,\dots,I_n)$. Luego $T^+ \overset{d}{=} \sum_{i=1}^n i\,B_i$ con B_i i.i.d. Bernoulli(1/2). $\square$

5. Normalidad asintótica mediante el TCL de Lindeberg

Consideremos la representación $T^+ = \sum_{i=1}^n iB_i$ con B_i i.i.d. Bernoulli(1/2). Definimos

$$Y_i = iB_i, \qquad \mu_i = \mathbb{E}[Y_i] = \frac{i}{2}, \qquad \sigma_i^2 = \operatorname{Var}(Y_i) = \frac{i^2}{4}.$$

Entonces

$$S_n = T^+ - \mathbb{E}[T^+] = \sum_{i=1}^n (Y_i - \mu_i), \qquad s_n^2 = \operatorname{Var}(S_n) = \sum_{i=1}^n \sigma_i^2 = \frac{n(n+1)(2n+1)}{24}.$$

Para aplicar el teorema central del límite de Lindeberg a la suma de variables independientes (pero no idénticamente distribuidas) $\frac{S_n}{s_n}$, debemos verificar que para todo $\varepsilon>0$,

$$\frac{1}{s_n^2}\sum_{i=1}^n \mathbb{E}\big[(Y_i-\mu_i)^2\mathbf{1}_{\{|Y_i-\mu_i|>\varepsilon s_n\}}\big] \longrightarrow 0 \quad (n\to\infty).$$

Observamos que $|Y_i-\mu_i| \le \frac{i}{2} \le \frac{n}{2}$. Por otra parte,

$$s_n = \sqrt{\frac{n(n+1)(2n+1)}{24}} \sim \frac{n^{3/2}}{\sqrt{12}} \quad (n\to\infty).$$

Luego, para cualquier $\varepsilon>0$ fijo, existe N tal que para todo $n\ge N$ se cumple $\frac{n}{2}<\varepsilon s_n$. En consecuencia, para n suficientemente grande, el conjunto $\{|Y_i-\mu_i|>\varepsilon s_n\}$ es vacío, y la suma del límite de Lindeberg es exactamente 0. Por tanto, la condición se satisface trivialmente.
Aplicando el TCL de Lindeberg, concluimos que

$$\frac{T^+ - \mathbb{E}[T^+]}{\sqrt{\operatorname{Var}(T^+)}} \xrightarrow{d} \mathcal{N}(0,1).$$

6. Conclusión

Hemos demostrado que bajo la hipótesis nula de simetría alrededor de cero:

- $\mathbb{E}[T^+] = \frac{n(n+1)}{4}$,
- $\mathrm{Var}(T^+) = \frac{n(n+1)(2n+1)}{24}$,
- $\frac{T^+ - \mathbb{E}[T^+]}{\sqrt{\mathrm{Var}(T^+)}}$ converge en distribución a una normal estándar.

La prueba se basa en la independencia entre signos y rangos, la representación exacta de T^+ como suma ponderada de Bernoulli independientes, y la aplicación directa del teorema de Lindeberg.