Normalidad asintótica del estadístico de Wilcoxon

(por ChatGPT, DeepSeek y Claude)

1. Planteamiento

Sea $X_1, \ldots, X_n$ una muestra i.i.d. de una variable continua con mediana M bajo H_0.
Definimos:

$$Y_i = X_i - M, \quad \varepsilon_i = \mathbf{1}(Y_i > 0).$$

Bajo H_0 y continuidad:

$$\varepsilon_i \sim \text{Bernoulli}(1/2), \quad \text{i.i.d.}$$

Sea R_i el rango de $|Y_i|$. El estadístico de Wilcoxon es:

$$T^+ = \sum_{i=1}^{n} R_i \varepsilon_i.$$

Idea clave:

- Los signos ε_i son independientes.
- Los rangos R_i son dependientes (forman una permutación).
- **Los rangos y los signos son independientes entre sí.**

Justificación de la independencia rangos-signos.

Paso 1: para una sola variable, signo y valor absoluto son independientes.
Sea Y una variable aleatoria con densidad f continua y simétrica alrededor de cero, es decir, $f(y) = f(-y)$ para todo y. Queremos probar que $\varepsilon = \mathbf{1}(Y > 0)$ y $|Y|$ son independientes.
Basta comprobar que, para todo boreliano $A \subseteq (0, \infty)$:

$$\mathbb{P}(\varepsilon = 1,\ |Y| \in A) \ = \ \mathbb{P}(\varepsilon = 1) \cdot \mathbb{P}(|Y| \in A).$$

Calculamos cada factor:

$$\mathbb{P}(\varepsilon = 1,\ |Y| \in A) = \mathbb{P}(Y \in A) = \int_A f(y)\, dy,$$

$$\mathbb{P}(\varepsilon = 0,\ |Y| \in A) = \mathbb{P}(Y \in -A) = \int_{-A} f(y)\, dy = \int_A f(-y)\, dy = \int_A f(y)\, dy,$$

donde hemos usado el cambio de variable $y \mapsto -y$ y la simetría $f(-y) = f(y)$. Sumando:

$$\mathbb{P}(|Y| \in A) = 2 \int_A f(y)\, dy.$$

Por tanto:

$$\mathbb{P}(\varepsilon = 1) \cdot \mathbb{P}(|Y| \in A) = \frac{1}{2} \cdot 2 \int_A f(y)\, dy = \int_A f(y)\, dy = \mathbb{P}(\varepsilon = 1,\ |Y| \in A). \qquad \checkmark$$

El mismo argumento da $\mathbb{P}(\varepsilon = 0,\ |Y| \in A) = \mathbb{P}(\varepsilon = 0) \cdot \mathbb{P}(|Y| \in A)$, con lo que queda probada la independencia entre ε y $|Y|$.

Paso 2: del par individual al vector completo.
Aplicando el resultado anterior a cada Y_i, obtenemos que ε_i y $|Y_i|$ son independientes para cada i. Dado que los pares $(|Y_i|, \varepsilon_i)$ son i.i.d., el vector de rangos $(R_1, \ldots, R_n)$ —que es función únicamente de $(|Y_1|, \ldots, |Y_n|)$— es independiente del vector de signos $(\varepsilon_1, \ldots, \varepsilon_n)$. Este hecho es el cimiento de toda la demostración: permite condicionar en los rangos y tratar los signos como variables aleatorias libres.

2. Centramos el estadístico

Definimos variables centradas:

$$\xi_i := \varepsilon_i - \tfrac{1}{2}.$$

Entonces:

$$\mathbb{E}[\xi_i] = 0, \quad \mathrm{Var}(\xi_i) = \tfrac{1}{4}.$$

Reescribimos:

$$T^+ = \sum_{i=1}^{n} R_i \left(\xi_i + \tfrac{1}{2}\right) = \sum_{i=1}^{n} R_i \xi_i + \tfrac{1}{2} \sum_{i=1}^{n} R_i.$$

Como:

$$\sum_{i=1}^{n} R_i = \frac{n(n+1)}{2},$$

obtenemos:

$$T^+ - \mathbb{E}(T^+) = \sum_{i=1}^{n} R_i \xi_i.$$

Interpretación: hemos reducido el problema a estudiar una suma de variables centradas.

3. Estrategia: condicionar en los rangos

Fijamos los rangos $(R_1, \ldots, R_n)$.
Entonces:

- Los R_i son constantes.
- Las ξ_i siguen siendo i.i.d.

Por tanto:

$$S_n := \sum_{i=1}^{n} R_i \xi_i$$

es una suma de variables independientes (condicionalmente).

Idea clave:

Aunque globalmente hay dependencia, condicionalmente tenemos independencia.

4. Varianza

$$\mathrm{Var}(S_n \mid R) = \sum_{i=1}^{n} R_i^2 \mathrm{Var}(\xi_i) = \frac{1}{4} \sum_{i=1}^{n} R_i^2.$$

Como los rangos son una permutación de $1, \ldots, n$:

$$\sum_{i=1}^{n} R_i^2 = \sum_{r=1}^{n} r^2 = \frac{n(n+1)(2n+1)}{6}.$$

Luego:

$$\sigma_n^2 := \mathrm{Var}(S_n \mid R) = \frac{n(n+1)(2n+1)}{24}.$$

5. Aplicamos Lindeberg

Definimos:

$$X_{n,i} := R_i \xi_i.$$

Queremos aplicar el Teorema Central del Límite de Lindeberg al array triangular $\{X_{n,i}\}_{i=1}^{n}$, condicionado en R.

Paso 1: infinitesimalidad uniforme.

$$|X_{n,i}| = R_i |\xi_i| \leq \frac{R_i}{2} \leq \frac{n}{2}, \qquad \sigma_n \sim c\, n^{3/2},$$

por lo que:

$$\frac{\max_i |X_{n,i}|}{\sigma_n} = O(n^{-1/2}) \to 0.$$

Ningún término domina la suma: todos los sumandos son asintóticamente negligibles respecto a la desviación típica total.

Paso 2: verificación de la condición de Lindeberg.
Para cualquier $\varepsilon > 0$, acotamos la suma de Lindeberg usando el resultado del Paso 1. Dado que $|X_{n,i}| \leq n/2$ con probabilidad 1 (condicionado en R), el evento $\{|X_{n,i}| > \varepsilon\, \sigma_n\}$ solo puede ocurrir cuando $n/2 > \varepsilon\, \sigma_n$, es decir, para n suficientemente grande el indicador es idénticamente cero. Sin embargo, es más instructivo dar la cota explícita que vale para todo n:

$$\begin{aligned}
\frac{1}{\sigma_n^2} \sum_{i=1}^{n} \mathbb{E}\Big[X_{n,i}^2\, \mathbf{1}(|X_{n,i}| > \varepsilon \sigma_n) \;\Big|\; R \Big] \;&\leq\; \frac{1}{\sigma_n^2} \sum_{i=1}^{n} \mathbb{E}\Big[|X_{n,i}| \cdot |X_{n,i}|\, \mathbf{1}(|X_{n,i}| > \varepsilon \sigma_n) \;\Big|\; R \Big] \\
&\leq\; \frac{\max_i |X_{n,i}|}{\varepsilon\, \sigma_n} \cdot \frac{1}{\sigma_n^2} \sum_{i=1}^{n} \mathbb{E}\Big[X_{n,i}^2 \;\Big|\; R \Big] \\
&=\; \frac{\max_i |X_{n,i}|}{\varepsilon\, \sigma_n} \cdot 1 \;\longrightarrow\; 0,
\end{aligned}$$

donde hemos usado que $\mathbf{1}(|X_{n,i}| > \varepsilon\sigma_n) \leq |X_{n,i}|/(\varepsilon\sigma_n)$ en la segunda desigualdad, y que $\sigma_n^{-2} \sum_i \mathbb{E}[X_{n,i}^2 \mid R] = 1$ por definición de σ_n^2 en la última igualdad. La condición de Lindeberg queda así verificada.

6. Conclusión (CLT condicional)

Por el Teorema Central del Límite de Lindeberg (para arrays triangulares):

$$\frac{S_n}{\sigma_n} \;\Big|\; R \;\xrightarrow{d}\; \mathcal{N}(0,1).$$

6b. De la convergencia condicional a la incondicional

Hemos demostrado que, para casi toda realización fija de los rangos $(R_1, \ldots, R_n)$ —y en particular para cualquier permutación de $\{1, \ldots, n\}$, que es lo que los rangos siempre son—,

$$\frac{S_n}{\sigma_n} \,\Big|\, R \xrightarrow{d} \mathcal{N}(0,1).$$

Es decir, para todo $t \in \mathbb{R}$:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma_n} \le t \,\middle|\, R\right) = \Phi(t),$$

donde Φ es la función de distribución normal estándar. Queremos probar que la convergencia es también incondicional.

Justificación. Por la ley de probabilidad total, para cada n y cada t:

$$\mathbb{P}\left(\frac{S_n}{\sigma_n} \le t\right) = \mathbb{E}\left[\mathbb{P}\left(\frac{S_n}{\sigma_n} \le t \,\middle|\, R\right)\right].$$

La variable aleatoria $Y_n := \mathbb{P}\left(\frac{S_n}{\sigma_n} \le t \mid R\right)$ está acotada entre 0 y 1, y converge puntualmente (para casi toda realización de R) a $\Phi(t)$ cuando $n \to \infty$.
Aplicamos el **teorema de convergencia dominada**:

$$\lim_{n\to\infty} \mathbb{E}[Y_n] = \mathbb{E}\Big[\lim_{n\to\infty} Y_n\Big] = \mathbb{E}[\,\Phi(t)\,] = \Phi(t).$$

Por tanto:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma_n} \le t\right) = \Phi(t) \quad \forall\, t,$$

que es precisamente la convergencia en distribución a $\mathcal{N}(0,1)$.

Observación. *Este argumento funciona porque el límite condicional $\Phi(t)$ es* constante *(no depende de la realización de R) y la función de distribución límite es continua, lo que evita problemas en los puntos de discontinuidad. Así, la normalidad asintótica se traslada del mundo condicional al incondicional.*

7. Resultado final

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

8. Idea final (para recordar)

La normalidad aparece porque:

independencia signos-rangos (por simetría) + independencia condicional en los signos + pesos controlados + Lindeberg.