La prueba de Mann-Whitney no compara medianas (salvo que lo haga)

José Ruiz-Tagle Maturana[*]

Fundación Instituto Profesional DUOC UC, Santiago, Chile
Millennium Nucleus for the Evaluation and Analysis of Drug Policies (nDP), Santiago, Chile

Resumen

En la investigación biomédica es habitual utilizar la prueba de Mann-Whitney como alternativa no paramétrica a la prueba t de Student, especialmente cuando no se cumple el supuesto de normalidad. Sin embargo, es frecuente interpretar esta prueba como una comparación de medianas, lo que puede inducir a error. Mann-Whitney no contrasta directamente la igualdad de medianas, sino la probabilidad de que una observación de un grupo exceda a una del otro, es decir, evalúa superioridad estocástica. En este artículo, mediante ejemplos simulados, mostramos situaciones en que la prueba resulta significativa aún cuando las medianas son iguales, y otras en que no lo es pese a que las medianas difieren. Se discuten las condiciones bajo las cuales puede interpretarse válidamente como prueba de diferencia de medianas, así como las limitaciones de su uso en contextos aplicados. Este trabajo aportará claridad conceptual para mejorar la interpretación estadística en estudios de salud.
Palabras clave: Prueba de Mann-Whitney; Mediana; Métodos no paramétricos.

¿Qué es lo que contrasta realmente Mann-Whitney?

Es común que en la investigación biomédica se utilice la prueba U de Mann-Whitney para comparar −supuestamente− medianas entre dos grupos independientes (Ibarrondo et al., 2022; López et al., 2024; Peruga et al., 2021). La prueba de Mann-Whitney se basa en los rangos de los datos, no en los valores originales, y su hipótesis de nulidad suele expresarse como que ambas muestras provienen de poblaciones con la misma distribución. De forma equivalente, si X e Y representan valores aleatorios de los grupos A y B respectivamente, la hipótesis de nulidad implica que la probabilidad de que un valor de X sea mayor que uno de Y es igual a la probabilidad opuesta, es decir, $\mathrm{P}(X > Y) = \mathrm{P}(Y > X)$. Bajo esta hipótesis, cada grupo tiene aproximadamente un 50 % de probabilidades de producir un valor más grande que el otro en un sorteo aleatorio. En consecuencia, la prueba realmente evalúa si un grupo tiende a arrojar valores más altos que el otro (Fagerland & Sandvik, 2009).

Otra manera de interpretarlo es preguntarse: "¿Cuál es la probabilidad de que, al elegir un individuo al azar de cada grupo, el del grupo A tenga un valor mayor que el del grupo B?" La prueba Mann-Whitney esencialmente verifica si esa probabilidad difiere significativamente de 50 %. Si una de las distribuciones tiende a generar números mayores, los rangos acumulados de ese grupo serán sistemáticamente más altos que los del otro, resultando en un estadístico U extremo (y un p-valor pequeño). Entonces, ¿por qué se habla tanto de medianas? Ocurre que si las dos poblaciones difieren solamente en su posición (es decir, una está "desplazada" respecto a la otra pero tienen la misma forma y dispersión), la mediana de la distribución más alta también será mayor, y Mann-Whitney efectivamente estaría contrastando esa diferencia de localización. En ese escenario ideal, un p-valor pequeño sí indica que una mediana es significativamente mayor que la otra. Sin embargo, en la práctica es poco frecuente que dos grupos difieran únicamente en la mediana manteniendo formas idénticas. Cuando las distribuciones difieren en su forma −por ejemplo, en su dispersión o asimetría− Mann Whitney deja de ser "una prueba de medianas" y pasa a contrastar en general si $\mathrm{P}(X > Y) = 0,5$, o si, por el contrario, un grupo domina estocásticamente al otro (McElduff et al., 2010). Es justamente esta naturaleza flexible la que obliga a ser cauteloso con su interpretación (Dexter, 2013).

Ejemplo 1: distribuciones con la misma mediana, pero *U* significativa

Supongamos que tenemos dos grupos de datos con distribuciones marcadamente diferentes en forma, aunque curiosamente comparten la misma mediana. El Grupo A tiene una distribución sesgada a la derecha (algunos valores mucho más altos que el resto), mientras que el Grupo B está sesgado a la izquierda (algunos valores mucho más bajos). Para simplificar, usemos valores discretos: en el Grupo A, seis observaciones toman el valor 5 y el resto se reparten entre 7, 8, 9 y 10;

[*] J.ruiztagle@profesor.duoc.cl.

en el Grupo B, la mayoría de las observaciones también rondan 5 pero algunas son tan bajas como 1, 2, 3 y 4. Ambas muestras, por construcción, tienen mediana 5.

En la Figura 1 se observa que las distribuciones simuladas de dos grupos (A en azul, B en anaranjado) tienen una mediana idéntica de 5. Cada punto marca un valor observado. Obsérvese que, aunque las medianas coinciden, el Grupo A presenta varios valores altos (hasta 10) mientras el Grupo B incluye varios valores pequeños (hasta 1). Esta diferencia en la forma y colas de las distribuciones es detectada por la prueba U de Mann-Whitney, que arroja un resultado significativo ($p \approx 0,007$), indicando que efectivamente un grupo tiende a producir valores más altos que el otro.

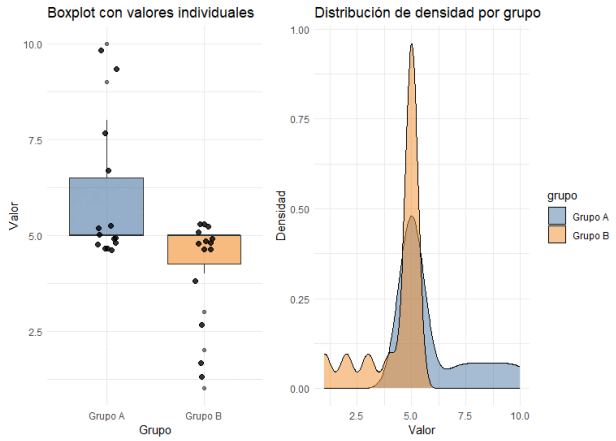

Figura 1: Visualización de la igualdad de medianas y diferencia de forma entre dos grupos.

En este ejemplo, encontramos evidencia estadística de diferencia entre grupos a pesar de que las medianas muestrales eran iguales en ambos casos, esto ocurre porque la U de Mann-Whitney no estaba realmente contrastando "medianas iguales *vs* medianas distintas", sino "ningún grupo domina al otro *vs* un grupo tiende a ser mayor". Dado que el Grupo A tenía varios valores notablemente más grandes que cualquier valor del Grupo B, hay una probabilidad elevada de que, al comparar elementos al azar, A resulte mayor que B. La prueba captura esa superioridad estocástica del Grupo A sobre el B. Este escenario nos enseña que igualdad de medianas no garantiza un resultado no significativo. Si las distribuciones difieren en otros aspectos (asimetría, colas, varianza, etc.), la prueba puede "saltar" ante esas diferencias.

Ejemplo 2: distribuciones con medianas distintas, pero *U* no significativa

De forma contraria, es posible que la prueba no detecte una diferencia significativa, a pesar de que las medianas sean completamente diferentes. Esto puede ocurrir cuando las distribuciones se sobreponen o tienen variabilidades extremas. Imaginemos que el Grupo A tiene la mayoría de sus valores muy bajos y unos pocos valores extremadamente altos. El Grupo B, en cambio, tiene la mayoría de sus valores más altos, pero también incluye algunos valores inusualmente bajos. En concreto, digamos que en una muestra del Grupo A el 70 % de las observaciones son alrededor de 1 (cercanas a cero) y el otro 30 % son valores grandes (alrededor de 20). Por su parte, en el Grupo B el 70 % de los datos rondan 10 (bastante mayores que los de A) pero hay un 30 % de valores pequeños cercanos a 0. En este caso, la mediana de A podría ser cercana a 1 y la de B cercana a 10, claramente diferentes.

En la Figura 2 se puede observar que, a pesar de la gran brecha entre las medianas, cada grupo tiene una minoría de valores extremos que se "cuelan" en el rango del otro: el Grupo A posee algunos datos muy altos que superan a la mayoría del Grupo B, mientras que el Grupo B tiene algunos datos inferiores a casi todos los de A. Debido a este solapamiento, la prueba no encuentra una diferencia en los rangos ($p \approx 0,97$). Esto ocurre porque, independiente de que una mediana sea mayor que la otra, Mann-Whitney puede no detectar diferencia si no hay una dominancia clara de un grupo sobre el otro. Para la prueba, las "victorias" de B en la zona central son contrarrestadas por las "victorias" de A en los valores extremos, resultando en un empate técnico. Este caso ilustra que un p-valor alto de Mann-Whitney no indica que las medianas sean iguales, sólo indica que no hay evidencia de que un grupo tienda a dar valores mayores que el otro de forma consistente.

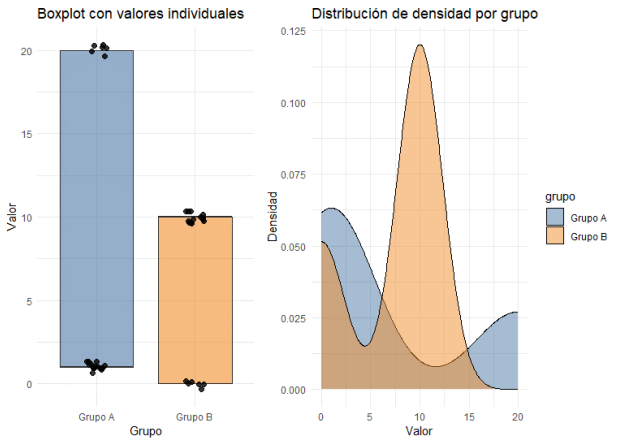

Figura 2: Visualización de la diferencia de medianas y diferencia de forma entre dos grupos.

Conclusión

Paradójicamente, todos los ejemplos aquí presentados transgreden uno de los supuestos fundamentales de la prueba de Mann-Whitney: la igualdad de forma entre las distribuciones comparadas (Sheskin, 2011). Si bien la prueba se suele justificar como una alternativa robusta a la prueba t de Student cuando no se cum-

ple normalidad, pocas veces se discute que su validez —incluso como contraste de superioridad estocástica— requiere que las distribuciones sean comparables en forma. De algún modo, esto lleva a una pregunta ineludible: si no vamos a examinar los supuestos distribucionales, ¿para qué utilizar esta prueba? La respuesta no es desecharla, sino recordar que su utilidad depende de entender lo que realmente está contrastando, y no de proyectar sobre ella lo que quisiéramos que hiciera. Si el objetivo del análisis es exclusivamente comparar las medianas, Mann-Whitney puede no ser la herramienta adecuada. Existen métodos más robustos como la regresión cuantil que permitiría estimar diferencias en torno al percentil 50 (Waldmann, 2018). Si se cumple el supuesto de igualdad de forma entre distribuciones, la prueba de Mann-Whitney puede interpretarse válidamente como una comparación de medianas. Pero si ese supuesto no se verifica —o ni siquiera se evalúa—, el test pierde su fundamento interpretativo. En ese escenario, no sólo deja de comparar medianas: deja de tener sentido usarlo.

Referencias

Dexter, F. (2013). Wilcoxon-Mann-Whitney Test Used for Data That Are Not Normally Distributed. *Anesthesia & Analgesia*, *117*(3), 537. https://doi.org/10.1213/ANE.0b013e31829ed28f

Fagerland, M., & Sandvik, L. (2009). The Wilcoxon–Mann–Whitney test under scrutiny. *Statistics in Medicine*, *28*(10), 1487-1497. https://doi.org/10.1002/sim.3561

Ibarrondo, O., Lizeaga, G., Martínez-Llorente, J., Larrañaga, I., Soto-Gordoa, M., & Álvarez-López, I. (2022). Health care costs of breast, prostate, colorectal and lung cancer care by clinical stage and cost component. *Gaceta Sanitaria*, *36*(3), 246-252. https://doi.org/10.1016/j.gaceta.2020.12.035

López, M., Fu, M., Fernández, E., Henderson, E., & Continente, X. (2024). ¿Se cumple la ley de control del tabaquismo en las terrazas de hostelería? *Gaceta Sanitaria*, *38*, 102422. https://doi.org/10.1016/j.gaceta.2024.102422

McElduff, F., Cortina-Borja, M., Chan, S., & Wade, A. (2010). When t-tests or Wilcoxon-Mann-Whitney tests won't do. *Advances in Physiology Education*, *34*(3), 128-133. https://doi.org/10.1152/advan.00017.2010

Peruga, A., Fu, M., Molina, X., & Fernández, E. (2021). Night entertainment venues comply poorly with the smoke-free law in Chile. *Gaceta Sanitaria*, *35*(4), 402-404. https://doi.org/10.1016/j.gaceta.2020.04.016

Sheskin, D. (2011). *Handbook of parametric and nonparametric statistical procedures* (5th ed.). CRC Press.

Waldmann, E. (2018). Quantile regression: A short story on how and why. *Statistical Modelling*, *18*(3-4), 203-218. https://doi.org/10.1177/1471082X18759142