Revista de Matemática: Teoría y Aplicaciones 2(1): 39–44 (1995)

OBTENCIÓN DE LA FÓRMULA DE LA VARIANCIA DEL ESTADÍSTICO DE LA DÓCIMA DE RANGOS DE WILCOXON-MANN-WHITNEY

José Francisco Pastrana[1]

Resumen

En este artículo se deriva la fórmula de $Var(S)$, en donde S es el estadístico de la dócima de Wilcoxon-Mann-Whitney, para probar la hipótesis de igualdad de las distribuciones de dos poblaciones. El propósito es brindar detalles del tipo de pruebas propias de la Estadística No Paramétricas que a menudo son omitidos en los libros de texto especializados. El fundamento principal del procedimiento que se ofrece para derivar la fórmula, es el hecho de que bajo la hipótesis de igualdad de las distribuciones, los rangos $R(x_1), R(x_2), \ldots, R(x_m)$, la muestra de una de las dos poblaciones constituye una muestra simple al azar sin reemplazo del conjunto de rangos $\{1, 2, \ldots, m+n\}$ en donde m y n son los tamaños de las dos muestras independientes obtenidas.

1 Introducción

La Dócima de Rangos de Wilcoxon-Mann-Whitney, es un procedimiento para probar, empleando rangos, la hipótesis nula $H_0 : f(x) = g(x)$ versus la hipótesis alternativa $H_1 : f(x) = g(x+\theta)$, en donde f y g son funciones de densidad ($f.d.$) no especificadas, de las variables aleatorias continuas X y Y respectivamente. El parámetro θ tampoco es especificado.

El procedimiento o prueba parte de la consideración de dos muestras aleatorias independientes: una $X_1, X_2, \ldots, X_m$ de la población con $f.\,d.\,f$ y otra $Y_1, Y_2, \ldots Y_n$ de la población con $f.\,d.\,g$. Las observaciones x_i y y_j, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, son ordenadas de menor a mayor en una única secuencia y los rangos correspondientes, $R(x_i)$ y $R(y_j)$, establecidos. El estadístico que permite rechazar o no rechazar H_0 es S= Suma de los rangos de las variables X_i, o sea,

$$S = \sum_{muestras} R(X_i)$$

con valor particular $s = \sum R(x_i)$, y, asumiendo que H_0 es cierta:

$$\text{Media de } S = E[S] = m(m+n+1)/2 \text{ y}$$

[1]Escuela de Estadística, Universidad de Costa Rica

$$\text{Variancia de } S = \text{ Var}(S) = mn(m+n+1)/12$$

Existen otros procedimientos para probar H_0 versus H_1: La Dócima de Homogeneidad, que requeriría de la agrupación de las observaciones de cada muestra en un número dado de intervalos; la Dócima de Smirnov-Kolmogorov para dos muestras aleatorias y, bajo el supuesto de normalidad e igualdad de variancias, la Dócima de la T de Student con $m+n-2$ grados de libertad.

El propósito de este artículo es presentar la obtención de la fórmula $\text{Var}(S) = mn(m+n+1)/12$, para lo cual hay que obtener a su vez, la fórmula $E(S) = m(m+n+1)/2$. El conocimiento de ambas fórmulas es vital al probar H_0 versus H_1 empleando la Dócima de Rangos de Wilcoxon-Mann-Whitney, independientemente de si aplica la distribución exacta de S (caso en que m y n son relativamente pequeños) o la distribución aproximada o asintótica de S (que consiste en la distribución normal para el caso en que m y n son suficientemente grandes).

El procedimiento que se presenta para obtener las fórmulas de la Media y la Variancia de S, es original y tiene como fundamento principal, el hecho de que, bajo la hipótesis nula H_0, ambas muestras son obtenidas de la misma distribución (población), de modo que $R(X_1), R(X_2), \ldots, R(X_m)$ constituye una muestra simple al azar sin reemplazo, de la población de rangos dada por $\{1, 2, \ldots, m+n\}$.[2]

La justificación de este artículo está en la omisión común de los detalles o la prueba completa de la fórmula de la Variancia de S, en los textos especializados (por ejemplo: [1, 2, 3]) y en la falta de ilustraciones de este tipo de pruebas en general en los textos de Estadística No Paramétrica.

2 Fórmula de la Variancia de S

2.1 Fórmula de definición de $Var(S)$

Por definición $Var(S) = E[S - E[S]]^2$. Ahora, $E[S - E[S]]^2 = E[S^2 - \{E[S]\}^2$. La obtención de $\{E[S]\}^2$ es fácil tan pronto se tiene $E(S)$.

2.2 Obtención de la media de S

La media de S está dada por:

$$E[S] = \sum_{muestras} sP(s)$$

Para un entero arbitrario pero fijo m, el número de muestras sin reemplazo que pueden ser obtenidas de la población de rangos $\{1, 2, \ldots, m+n\}$ es: $\binom{n+m}{m}$

[2]El caso de dos o más valores x_i (o y_j) iguales, o sea, el caso de empates dentro de las muestras, es irrelevante, pues basta asignarles el rango promedio; de esta manera, el valor s de S no es afectado por los empates dentro de las muestras.

Sin embargo, cuando uno o más valores x_i iguala a uno o más valores y_j o sea, en el caso de empates intramuestras, es preciso, ya sea modificar la fórmula dada de la Variancia de S, si se asigna el correspondiente rango promedio a los valores x_i y y_j iguales o aplicar la prueba dos veces, la primera vez asignando los rangos menores a las x_i y la segunda vez asignándoles los rangos mayores. En este último caso, la prueba puede resultar inconclusa.

Por lo tanto, la probabilidad $P(s)$ de obtener una muestra sin reemplazo, de tamaño m fijo, de esa población de rangos, está dada por:

$$\frac{1}{\binom{m+n}{m}} = \frac{m!\,n!}{(m+n)!}$$

De esta manera:

$$P(\text{muestra sin reemplazo de tamaño } m \text{ arbitrario pero fijo}) = P(S)$$

$$= E(S) = \frac{m!\,n!}{(m+n)!} \sum_{muestras} s = \frac{m!\,n!}{(m+n)} \sum_{muestras} \sum_{i=1}^{m} R(x_i)$$

El número de muestras de tamaño fijo m en que aparece un rango j de $\{1, 2, \ldots, m+n\}$, arbitrario pero fijo es $\binom{m+n-1}{m-1}$

Así:

$$E[S] = \frac{m!\,n!}{(m+n)!} \binom{m+n-1}{m-1} \sum_{j=1}^{m+n} j$$

de donde:

$$\begin{aligned} E[S] &= \frac{m!\,n!}{(m+n)!}\ \frac{(m+n-1)!}{(m-1)!\,n!}\ \frac{1}{2}\,(m+n)\,(m+n+1) \\ &= \frac{m(m+n+1)}{2} \end{aligned}$$

Ahora:

$$\{E[S]\}^2 = \left\{\frac{m(m+n+1)}{2}\right\}^2 = \frac{m^2}{4}\,(m+n+1)^2 = \frac{m^2}{4}\,\{(m+n)^2 + 2(m+n) + 1\} \quad (1)$$

2.3 Obtención de $E[S^2]$

Por definición,

$$E[S^2] = \sum_{muestras} s^2\,P(s). \quad \text{En este caso} \quad P(s) = \frac{m!\,n!}{(m+n)!}.$$

Entonces:

$$\begin{aligned} E[S^2] &= \frac{m!\,n!}{(m+n)!} \sum_{muestras} s^2 = \frac{m!\,n!}{(m+n)!} \sum_{muestras} \left\{\sum_{i=1}^{m} R(x_i)\right\}^2 \\ &= \frac{m!\,n!}{(m+n)!} \sum_{muestras} \left\{\sum_{i=1}^{m} (R(x_i))^2 + 2\sum_{i<j}^{m} R(x_i)\,R(x_j)\right\} \end{aligned} \quad (2)$$

El número de muestras sin reemplazo de tamaño fijo m en que aparecen rangos i, j de $\{1, 2, \ldots, m+n\}$, arbitrarios pero fijos, es: $\binom{m+n-2}{m-2}$.

De ahí:

$$\sum_{muestras} \{2 \sum_{1<j=1}^{m} R(x_i)\, R(x_j)\} = \binom{m+n-2}{m-2} \sum_{i<j=1}^{m} ij.$$

Así, a partir de (2), se obtiene:

$$E[S^2] = \frac{m!\, n!}{(m+n)!} \left[\binom{m+n-1}{m-1} \sum_{j=1}^{m+n} j^2 + 2 \binom{m+n-2}{m-2} \sum_{i<j=1}^{m+n} ij \right] \tag{3}$$

pero:

$$\sum_{j=1}^{m+n} j^2 = \frac{1}{6}\,(m+n)\,(m+n+1)\,(2(m+n)+1), \tag{4}$$

$$\begin{aligned} \sum_{i<j=1}^{m+n} ij = \sum_{i=1}^{m+n-1} i \sum_{j=i+1}^{m+n} j &= \sum_{i=1}^{m+n-1} i \frac{1}{2}\,(m+n-i)\,(m+n+i+1) \\ &= \frac{1}{2} \sum_{i=1}^{m+n-1} i\,[(m+n)^2 - i^2 + m + n - i]. \end{aligned} \tag{5}$$

Sustituyendo (4) y (5) en (3):

$$\begin{aligned} E[S^2] &= \frac{m!\, n!}{(m+n)!} \left\{ \binom{m+n-1}{m-1} \frac{1}{6}\,(m+n)\,(m+n+1)\,(2(m+n+1)) \right. \\ &\quad \left. +2 \binom{m+n-2}{m-2} \frac{1}{2} \sum_{i=1}^{m+n-1} i[(m+n)^2 - i^2 + m + n - i] \right\} \\ &= \frac{m!\, n!}{(m+n)!} \frac{(m+n-1)!}{n!\,(m-1)!} \frac{1}{6} (m+n)\,(m+n+1)\,(2m+2n+1) \\ &\quad + \frac{m!\, n!}{(m+n)!} \frac{(m+n-2)!}{n!\,(m-2)!} \left\{ (m+n)^2 \sum_{i=1}^{m+n-1} i - \sum_{i=1}^{m+n-1} i^3 \right. \\ &\quad \left. +(m+n) \sum_{i=1}^{m+n-1} i - \sum_{i=1}^{m+n-1} i^2 \right\} \\ &\quad = \frac{m(m+n+1)\,(2m+2n+1)}{6} + \frac{(m-1)\,m}{(m+n-1)\,(m+n)} \times \\ &\quad \times \left\{ [(m+n)^2 + (m+n)] \sum_{i=1}^{m+n-1} i - \sum_{i=1}^{m+n-1} i^2 - \sum_{i=1}^{m+n-1} i^3 \right\}. \end{aligned} \tag{6}$$

Como $\sum_{i=1}^{a} i^3 = \frac{1}{4} a^2\,(a+1)^2$, sustituyendo en (6) se obtiene:

$$E[S^2] = \frac{m\,(m+n+1)\,(2m+2n+1)}{6}$$

$$
\begin{aligned}
&+\frac{(m-1)\,m}{(m+n-1)\,(m+n)}\left\{(m+n)\,(m+n+1)\,\frac{1}{2}\,(m+n-1)\,(m+n)\right.\\
&\left.-\frac{1}{6}\,(m+n-1)\,(m+n)\,(2m+2n-1)\;-\frac{1}{4}\,(m+n-1)^2\,(m+n)^2\right\}\\
=\;&\frac{m\,(m+n+1)\,(2m+2n+1)}{6}\\
&+\frac{(m-1)\,m}{(m+n-1)\,(m+n)}\{\frac{1}{12}\,(m+n)\,(m+n-1)\,[6\,(m+n)\,(m+n+1)\\
&-4\,(m+n)+2-3\,(m+n)\,(m+n-1)]\}\\
=\;&\frac{m\,(m+n+1)\,(2m+2n+1)}{6}+\frac{(m-1)\,m}{12}\left\{3\,(m+n)^2+5\,(m+n)+2\right\}\\
=\;&\frac{m}{12}\{2\,(m+n+1)\,(2m+2n+1)+(m-1)\,[3(m+n)^2+5(m+n)+2]\}
\end{aligned}
$$

De esta manera se llega a:

$$E[S^2]=\frac{m}{12}\{2(m+n+1)\,(2m+2n+1)+(m-1)\,[3(m+n)^2+5\,(m+n)+2]\} \qquad (7)$$

3 Obtención de $Var(S^2)$

Sustituyendo (7) y (1) en la fórmula de definición de la variancia, se obtiene:

$$
\begin{aligned}
Var(S^2) &= E[S^2]-\{E(S)\}^2\\
&= \frac{m}{12}\,\{2\,(m+n+1)\,(2m+2n+1)+(m-1)[3\,(m+n)^2+5\,(m+n)+2]\}\\
&\quad -\frac{m^2}{4}\{(m+n)^2+2\,(m+n)+1\}\\
&= \frac{m}{12}\,\{2\,(m+n+1)\,(2m+2n+1)+(m-1)[3\,(m+n)^2+5\,(m+n)+2]\\
&\quad -3m\,(m+n)^2-6m\,(m+n)-3m\}\\
&= \frac{m}{12}\{4\,(m+n)^2+6\,(m+n)+2+(m-1)\,[3\,(m+n)^2+5(m+n)+2]\\
&\quad -3m\,(m+n)^2-6m\,(m+n)-3m\}\\
&= \frac{m}{12}\{[4+3\,(m-1)-3m]\,(m+n)^2+[6+5\,(m-1)-6m]\,(m+n)\\
&\quad +2\,(m-1)-3m+2\}\\
&= \frac{m}{12}\{(m+n)^2+(1-m)\,(m+n)-m\}\\
&= \frac{m}{12}\,(m^2+2mn+n^2+m+n-m^2-mn-m)\\
&= \frac{m}{12}\,(m^2+mn+n)=\frac{mn\,(m+n+1)}{12}
\end{aligned}
$$

Por lo tanto

$$Var(S^2)=\frac{mn\,(m+n+1}{12}$$

que es la fórmula que se quería obtener.

4 Conclusión

La derivación de la fórmula de la variancia de S, estadístico de la dócima de Wilcoxon-Mann-Whitney, ha hecho posible la ilustración de los métodos de prueba propios de la Estadística No Paramétrica que a menudo son omitidos en los libros de texto correspondientes.

Referencias

[1] Conover, W.J. (1971) *Practical Non Parametric Statistics*. John Wiley & Sons Inc., New York.

[2] De Groot, M.H. (1975) *Probability and Statistics*. Addison-Wesley Publishing Company.

[3] Lehmann, E.L. (1975) *Nonparametric Statistical Methods Based on Ranks*. Holden-Pay, Inc.