VERIFICACIÓN DE LAS CONDICIONES DE LA PRUEBA U DE MANN-WHITNEY

Dr. Ing. Jesús Rubén Azor Montoya
Facultad de Ingeniería – Universidad de Mendoza - Argentina
razor@telar.org
Cátedra Estadística Aplicada II

Palabras Clave: Suma de rangos, Prueba U de Mann-Whitney, Bondad de Ajuste, Distribución Normal.

Resumen – La *prueba U de Mann-Whitney* (también llamada *Prueba de Suma de Rangos de Wilcoxon*) es una alternativa no-paramétrica de la *Prueba t Bimuestral*. Mediante la misma, se puede probar si dos muestras independientes tienen la misma distribución. Como todas las pruebas no-paramétricas tiene la ventaja de no necesitar la distribución de la población en estudio, lo que le da robustez a las estimaciones que se hacen con dichas pruebas. En general, en los textos de Estadística para Ingenieros no se da una demostración para los estadísticos utilizados en la mencionada prueba. En el presente trabajo se realizarán una serie de simulaciones (con su correspondiente *prueba de Bondad de Ajuste*) que permitirán verificar la validez de los estadísticos que se dan en la literatura.

1 – La prueba U de Mann-Whitney

Las pruebas *no-paramétricas* se necesitan cuando no se tienen información sobre la composición de los datos poblacionales, esto es no se tiene conocimiento sobre su distribución de probabilidad. Se utilizan cuando no se cumplen las condiciones exigidas para la aplicación de las *paramétricas*. También cuando las muestras son pequeñas y falta información respecto de la densidad de probabilidad.

Como contrapartida, son menos eficientes que las paramétricas cuando se cumplen las condiciones para éstas.

Dentro de las pruebas no-paramétricas, una muy utilizada es la *Prueba de Suma de Rangos de Wilcoxon* o también llamada *prueba U de Mann-Whitney*, la cual es una alternativa a la paramétrica t bimuestral.

Con esta prueba se intenta probar la *Hipótesis Nula* que las muestras de tamaño n_1 y n_2 tomadas de dos poblaciones tienen la misma distribución, contra la *Hipótesis Alternativa* bilateral que establece en forma sencilla que las distribuciones no son las mismas.

La prueba se basa en las siguientes etapas:

- Se realiza una combinación de las n_1 y n_2 observaciones para formar un solo conjunto $n_1 + n_2$ observaciones arregladas en orden creciente de magnitud.
- Se asigna un rango numérico a cada observación comenzando desde 1. Si por alguna razón coincidiera más de un valor, para las distintas muestras, se asigna como rango el promedio de los mismos.
- Se suman los rangos correspondientes a cada muestra, obteniendo los valores R_1 y R_2.

Si las muestras aleatorias provienen de poblaciones que tienen la misma distribución, se espera que los rangos se encuentren lo suficientemente dispersos. De otro modo, cabe esperar que los rangos de las observaciones en cada muestra se encuentren muy agrupados en los extremos.

El estadístico utilizado se basa en la suma de los rangos asociados con las observaciones de cada una de las dos muestras, lo que lleva a los valores R_1 y R_2. .

2 – Simulaciones

Matlab ® tiene una función built-in que permite simular la aparición de los rangos sobre una de las muestras. Se trata de *randperm* que sólo necesita un argumento, para el caso en cuestión, $n_1 + n_2$. Una vez ejecutada, entrega un vector de longitud $n_1 + n_2$ que son las permutaciones, sin repetición, de los números del 1 al $n_1 + n_2$.

De modo que, a posteriori, si se trunca el vector resultante a los n_1 primeros elementos, simulará el juego de rangos aleatorios correspondientes a la primera muestra (si se tomaran los últimos n_2 elementos, se simularían los correspondientes a la segunda muestra).

Dado que cada número entero tiene igual posibilidad de aparición, se estaría en presencia de una variable aleatoria con distribución uniforme, con probabilidad $1/(n_1 + n_2)$ cuya media y varianza estarán dadas por

$$\mu = \sum_{i=1}^{n_1+n_2} \left(\frac{i}{n_1 + n_2} \right) \qquad \sigma^2 = \sum_{i=1}^{n_1+n_2} \left[\frac{(i-\mu)^2}{n_1 + n_2} \right]$$

Cuando se halla la suma de rangos de la primera muestra, se están sumando n_1 variables aleatorias con la misma función densidad de probabilidad (uniforme) y con la misma media y varianza.
Conforme al *Teorema Central del Límite,* la nueva variable aleatoria tiene una distribución normal con la misma media y cuya varianza es σ^2 / n_1.

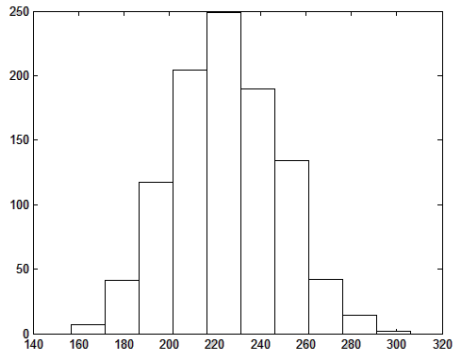
El siguiente segmento de programa de Matlab® permite generar una matriz *B* de 1000 filas (número de simulaciones) por 15 columnas (tamaño de la primera muestra, n_1) que representa 1000 casos distintos que se presentarían en una prueba de este tipo.

```
% Distribución simulada de rangos para la primera muestra
n1=15; % tamaño de la primera muestra
n2=14; % tamaño de la segunda muestra
N=1000; % Número de simulaciones
B=randperm(n1+n2);B=B(1:n1);
for z=1:N-1,
   A=randperm(n1+n2);A1=A(1:n1);
   B=[B; A1];
end
```

Si se suman todos los elementos de cada fila, se obtiene un vector *S* de 1000 elementos donde cada uno de ellos representa la *Suma de Rangos* de la primera muestra.

```
% Suma de rangos de la primera muestra
for i=1:N,S(i)=sum(B(i,:));end
hist(S,10)
```

En la última línea se indica realizar un histograma de 10 barras con el vector de Suma de Rangos obtenido. La salida gráfica tiene la siguiente forma



Donde se aprecia, con bastante nitidez, el comportamiento "normal" de la variable aleatoria Suma de Rangos.

Basado en lo anterior, Mann-Withney forma el estadístico U_1 que está dado por la siguiente expresión

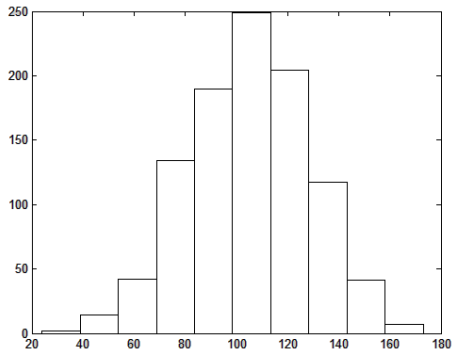
$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - R_1$$

El estadístico U_1, como se ve, es función de los tamaños muestrales n_1, n_2 y de R_1. Del mismo modo, se podría crear otro estadístico U_2.

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - R_2$$

Retomando el vector *S* obtenido del programa anterior, se hacen las transformaciones correspondientes para hallar un nuevo vector U1.

```
% Vector de valores simulados U1
for i=1:N,U1(i)=n1*n2+(n1*(n1+1)/2)-S(i);end
hist(U1,10)
```



Observando el nuevo histograma obtenido sobre *U1*, se ve que la distribución de la nueva variable aleatoria sigue siendo normal.

Según la teoría la media y varianza de *U1* son

$$\mu_{U1} = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - \mu_{R_1} = \frac{n_1 \cdot n_2}{2}$$

$$\sigma^2_{U1} = \sigma^2_{R_1} = n_1 \cdot n_2 \cdot \frac{(n_1 + n_2 + 1)}{12}$$

3 – Prueba de Bondad de Ajuste

Para poner a prueba la última aseveración, se someterá a los datos simulados a una *Prueba de Bondad de Ajuste* en cuanto a la "normalidad" de la distribución de U1.

En una prueba de este tipo, se trata de comparar una distribución de *frecuencia observada* con los valores correspondientes a una *distribución de esperada o teórica*.

En este caso simulado, la distribución esperada es una normal con media y varianza

$$\mu_{U1} = \frac{15 \cdot 14}{2} = 105 \qquad \sigma^2_{U1} = 15 \cdot 14 \cdot \frac{(15 + 14 + 1)}{12} = 525$$

En cuanto a la distribución observada, se puedeular la media y la varianza del vector U1 observado.

Mediante Matlab® estos valores se obtienen con los comandos

```
>> mean(U1)
ans =
  105.1900

>> var(U1)
ans =
  555.2612
```

Como se ve, los valores no están muy alejados de los teóricos.

Con el objeto de formalizar una prueba de significancia, se considerarán 10 intervalos comprendidos entre el mínimo y el máximo del vector U1.

min(U1) = 24 max(U1) = 173

Luego, se construye una tabla como la siguiente

Intervalos	fronteras de clase (en unidades normales estándar)	Probabilidad de estar en cada categoría (p)	Frecuencias esperadas =1000.p	Frecuencias observadas
24-38.9	-3.535 … -2.885	0.0018	1.8	2
38.9-53.8	-2.885 … -2.235	0.011	11	14
53.8-68.7	-2.235 … -1.584	0.044	44	42
68.7-83.6	-1.584 … -0.934	0.119	119	134
83.6-98.5	-0.934 … -0.284	0.213	213	190
98.5-113.4	-0.284 … 0.367	0.255	255	249
113.4-128.3	0.367 … 1.017	0.202	202	204
128.3-143.2	1.017 …. 1.667	0.107	107	117
143.2-158.1	1.667 … 2.317	0.037	37	41
158.1-173	2.317 … 2.968	0.009	9	7

La segunda columna se obtiene normalizando los extremos de los subintervalos conforme a la media y la varianza de la distribución teórica. A modo de ejemplo para el primer subintervalo

$$\frac{24-105}{\sqrt{525}} = -3.535$$

La tercera columna indica el área bajo la curva de la densidad de probabilidad normal estándar entre los límites del subintervalo

$$\int_{-3.535}^{-2.885} \frac{1}{\sqrt{2\cdot\pi}}\cdot e^{\frac{-z^2}{2}}\, dz = 1.753\times 10^{-3}$$

La cuarta columna, consiste simplemente en multiplicar por 1000 los valores de probabilidad teórica para obtener las *frecuencias esperadas* (e_i).

La quinta columna son las *frecuencias observadas* (o_i) de U1 en la simulación.

El estadístico que dirime esta prueba es el chi-cuadrado con $\nu=k-m-1$ grados de libertad.

$$\zeta^2 = \sum_{i=1}^{k}\left[\frac{(o_i-e_i)^2}{e_i}\right]$$

Donde k es el número de subintervalos y m es el número de parámetros de la distribución a verificar (en este caso $m=2$ ya que la normal tiene dos parámetros, la media y la varianza).

Un tema que no se puede soslayar es el del llamado "desflecamiento", esto es ninguna de las frecuencias esperadas debe ser inferior a 5. De ser así se deben fusionar los sub-intervalos en cuestión con el inmediato superior (o inferior).

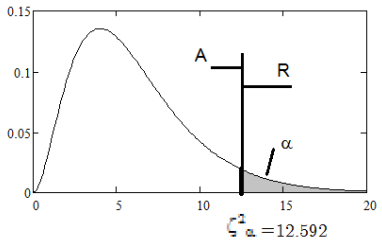
En este caso se fusionan el primer y segundo sub-intervalos, reduciéndose el valor de k a 9.

Con estos conceptos a la vista, se hace el planteo tradicional de una *Prueba de Hipótesis*.

1) Ho: La distribución de U es normal
   Ha: No lo es

2) Nivel de significancia. Se elige α=0.05
3) Criterio:

Se acepta la Hipótesis Nula si ζ^2 es menor que $\zeta^2_\alpha = 12.592$ con $\nu = 9-2-1=6$ grados de libertad, en caso contrario se acepta la Hipótesis Alternativa.



4) Cálculos:

$$\zeta^2 = \frac{(16-12.8)^2}{12.8} + \frac{(42-44)^2}{44} + \frac{(134-119)^2}{119} + \frac{(190-213)^2}{213} \ldots$$
$$+ \frac{(249-255)^2}{255} + \frac{(204-202)^2}{202} + \frac{(117-107)^2}{107} + \frac{(117-107)^2}{107} + \frac{(7-9)^2}{9} = 7.74$$

5) Decisión: ya que si $\zeta^2 = 7.74$ es menor que $\zeta^2_\alpha = 12.592$, se *Acepta* la Hipótesis nula y la distribución de U1 es normal con un nivel de significancia de 0.05.

4- Conclusiones

Conforme a lo presentado en el trabajo, se ha puesto de manifiesto el comportamiento de la variable aleatoria U1 que lleva a la elaboración de un estadístico para la prueba U de Mann-Whitney.

El comportamiento de la variable U2 es el mismo, con la salvedad de que los valores de media y varianza son distintos. También es necesario destacar que la repetición de valores en las muestras, mencionado más arriba y la forma en que se lauda, desmejora la exactitud de la prueba y debe ser considerado en caso de ser muy notable.

De este modo, con la probanza del comportamiento de la variable aleatoria, se da una verificación desde el punto de vista didáctico de esta Prueba más allá del tratamiento riguroso que se dá desde la Estadística Matemática.

Finalmente, un desarrollo en Java muy interesante, que permite realizar una Prueba U on-line en forma interactiva, se puede ver en http://faculty.vassar.edu/lowry/utest.html.

5 – Bibliografía

Kreyszig, E. "*Introducción a la Estadística Matemática*". Limusa, 1985

Miller, I. Freund, J y Jhonson, R. "*Probabilidad y Estadística para Ingenieros*" . Prentice-Hall Hispanoamericana. 1992

Canavos, G. "*Probabilidad y Estadística. Aplicaciones y Métodos*". Prentice-Hall. 1998