PESQUIMAT, Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. I - N°2 Pgs. 83-104 Lima – Perú Dic. 1998

MÉTODOS NO PARAMÉTRICOS PARA EL ANÁLISIS DE DATOS CENSURADOS (CASO DE DOS MUESTRAS)

Mg. Antonio Bravo Quiroz
Universidad Nacional Mayor de San Marcos

INTRODUCCIÓN

En los estudios de los tiempos de sobrevivencia, además del problema de censuramiento de dichos tiempos, un problema que puede surgir es comparar dos poblaciones **X** y **Y** con distribuciones F_x, F_Y respectivamente. Esto puede ser el caso de comparar el tiempo de sobrevivencia de pacientes con dos tipos de tratamiento, por ejemplo, uno nuevo y el vigente (tratamiento vs. Control), grupos raciales, etc.

En el presente trabajo, estudiaremos las pruebas no paramétricas para el caso de dos muestras, como son la de Wicoxon, de Gehan y de Mantel-Haenszel.

Sean la muestras tomadas de la población **X** y **Y**, respectivamente, donde

$$X_1, X_2, ..., X_m \text{ es una muestra iid de } F_x$$

$$Y_1, Y_2, ..., Y_m \text{ es una muestra iid de } F_Y \qquad (0.01)$$

de modo que la muestra combinada de los X´s e Y´s es

$$X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_m \qquad (0.02)$$

que las podemos denotar por

$$Z_1, Z_2, ..., Z_m, Z_{m+1}, Z_{m+2}, ..., Z_{n+m} \qquad (0.03)$$

cuya estadística de orden es

$$Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n+m)} \qquad (0.04)$$

luego, el rango de $Z_{(K)}$ es K.

1. LA ESTADÍSTICA DE WILCOXON

La estadística de Wilcoxon, es la estadística más conocida en el análisis no paramétrico de datos estadísticos, en situaciones, por ejemplo, donde deseamos comparar dos tipos de tratamientos: uno nuevo vs. otro común. En este caso, la hipótesis en consideración es de la forma:

$H_0 : F_X = F_Y \; vs. \;\; H_0 : F_X$ es estocástic amente mayor que F_Y

$$\left(F_X \; \alpha \; F_Y\right) \qquad (1.01)$$

Bajo la H_0, la estadística de test de Wilcoxon es definido por

$$w = \sum_{K=m+1}^{m+n} \Re_K$$

donde :

$$\Re_K = \text{El rango de } Y_K \text{ en la muestra combinada.}$$

$$= j \qquad si \;\; Y_K = Z_{(j)} \; y \;\; 1 \leq j \leq m+n$$

Una estadística equivalente a la Wilcoxon (1.02) es la de Mann-Whitney, que es definida por

$$U = \#\left\{\left(X_{i1}, Y_j\right) : X_i < Y_j; i = 1, \ldots\ldots m, \;\; j = 1, \ldots\ldots n\right\}$$

$$= \sum_{j=1}^{n} U_j \qquad (1.03)$$

donde

$$U_j = \#\left\{X_i : X_i < Y_j; i = 1, \ldots\ldots m\right.$$

$$j = 1, \ldots, n$$

Para hallar la equivalente entre (1.03) y (1.02), observemos el siguiente argumento: Sea $S_1 < S_2 < ... < S_n$ números enteros tal que S_j es el Rango de $Y_{(j)}$ en la muestra combinada. Luego,

$$U_j = Rango\ \{Y_{(j)}\} - \#\{X_i : X_i < Y_j ; i = 1, ..., m$$
$$= S_j - j$$

de modo que (1.03) puede ser escrita en la forma

$$U = \sum_{j=1}^{n} U_j = \sum_{j=1}^{n} (S_j - j) = \sum_{j=1}^{n} S_j - \frac{n(n+1)}{2}$$

pero de (1.02) tenemos que

$$U = W - \frac{n(n+1)}{2} \qquad (1.04)$$

Luego, la regla de decisión, bajo la H_0, es comparar U vs. U_t, donde U_t es el valor tabular para m, $n \leq 8$ y $m < n$. *Si* $n, m > 8$, lo podemos usar la aproximación normal.

1.1. LA MEDIDA Y VARIANZA DE U:

De las definiciones (1.03) y (1.04), podemos observar que la estadística U de Mann – Whitney es una variable aleatoria, que depende exclusivamente del órden de los elementos de la muestra combinada. La esperanza y varianza, puede ser circulada a partir de los siguientes argumentos:

Sea La función indicadora h_{ij}, donde

$$h_{ij} = \begin{cases} 1 & Si \quad X_i < Y_j \\ 0 & Otro \quad caso \end{cases} \qquad (1.05)$$

que es una variable aleatoria (ensayo) de Bernoulli, con

$$E_\theta(h_{ij}) = P(X_i < Y_j) \overset{def}{=} p$$
$$Var_\theta(h_{ij}) = E_\theta(h_{ij}^2) - \{E_\theta(h_{ij})\}^2$$
$$= p - p^2 = p(1-p)^2 \qquad (1.06)$$

donde $\theta = (F_x, F_Y)$, indica que en el cálculo de la esperanza, la probabilidad $P(X_i < Y_j)$ es calculado utilizando las distribuciones de X e Y.

La estadística U la podemos expresar en términos de la función h_{ij} dado que

$$U_j = \#\{X_i : X_i < Y_j ; i = 1, ..., m\} = \sum_{i=1}^{m} h_{hij}$$

luego

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij} \qquad (1.07)$$

por tanto,

$$E_\theta(U) = E_\theta\left(\sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij}\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} E_\theta(h_{ij}) = mnp \qquad (1.08)$$

$$E_\theta(U^2) = E_\theta\left(\sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij}\right)^2$$
$$= E_\theta\Big(\sum_{i=1}^{m}\sum_{j=1}^{n} h_{ij} + \sum_{i=1}^{m}\sum_{j\neq l}^{n} h_{ij}h_{il}$$
$$+ \sum_{j=1}^{n}\sum_{i\neq k}^{m} k_{ij}h_{kj} + \sum_{i\neq k}^{m}\sum_{j\neq l}^{n} h_{ij}h_{kl}\Big)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n}(E_{ij}^2) + \sum_{i=1}^{m}\sum_{j\neq l}^{n} E_\theta(h_{ij}h_{il})$$
$$+ \sum_{j=1}^{n}\sum_{i\neq k}^{m} E_\theta(h_{ij}h_{kj}) + \sum_{i\neq k}^{m}\sum_{j\neq l}^{n} E_\theta(h_{ij}h_{kl})$$

pero

$$E_\theta(h_{ij}^2) = E_\theta(h_{ij}) = p(X_i < Y_j) \overset{def}{=} p$$

$$E_\theta(h_{ij}h_{il}) = p(X_i < Y_j, X_i < Y_l) = p(X_i < Min(Y_j, Y_l)) \overset{def}{} q$$
$$E_\theta(h_{ij}h_{kj}) = p(X_i < Y_j, X_k < Y_j) = p(Max(X_i, X_j) < Y_j) \overset{def}{} r$$
$$E_\theta(h_{ij}h_{kl}) = E_\theta(h_{ij})E_\theta(h_{kl}) = p^2$$

de modo que

$$E_\theta(U^2) = \sum_{i=1}^{m}\sum_{j=1}^{n} p + \sum_{i=1}^{m}\sum_{j\neq;}^{n} q + \sum_{j=1}^{n}\sum_{i\neq k}^{m} r + \sum_{i\neq k}^{m}\sum_{j\neq l}^{n} p^2$$
$$= mnp + m(n-1)q + mn(m-1)r + mn(mn-m-n+1)p^2 \qquad (1.09)$$

de modo que

$$Var_\theta(U) = E_\theta(U^2) - \{E_\theta(U)\}^2$$
$$= mnp + mn(n-1)q + nm(m-1)r + mn(mn-m-n+1)p^2 - (mp)^2$$
$$= mnp(1-p) + mn(n-1)(q-p^2) + nm(m-1)(r-p^2) \qquad (1.10)$$

el problema que tenemos es que no conocemos los valores de las probabilidades p, q y r. Pero, bajo la hipótesis nula, podemos evaluar dichas probabilidades. Esto es:

$$\text{si } H_o : F_x = F_y = F$$

y asumiendo que F es continua, tenemos:

$$p = P(X_i < Y_j) = P((X_i, Y_j) \in B_1)$$

donde

$$E_1 = \{(x, y) : x < y\}$$

Luego

$$p = P(X_i < Y_j) = \int_{B_1} f_{XY}(x,y)dydx = \int_{-\infty}^{\infty}\int_{X}^{\infty} f_{XY}(x,y)dydx$$
$$= \int_{-\infty}^{\infty} f_X(x)\left(\int_{X}^{\infty} f_Y(y)dy\right)dx = \int_{-\infty}^{\infty}(1-F(x))f_X(x)dx = \frac{1}{2}$$
$$q = P(X_i < min(Y_j, Y_l)) = p(X_i < Y_{(1)} \leq Y_{(2)})$$
$$= p((X_i, Y) \in B_2)$$

donde

$$B_2 = \{(x,y,z): x<y<z \vee x<z<y\}$$

Luego,

$$q = P(X_i < Min(Y_j, Y_l)) = \int_{B_2} f_{XYz}(x,y,z)dydxdz$$

$$= \int_{-\infty}^{\infty}\int_{x}^{\infty}\int_{y}^{\infty} f_{Xy}(dydxdz) + \int_{-\infty}^{\infty}\int_{x}^{\infty}\int_{z}^{\infty} f_{Xy}(x,y,z)dydzdy$$

$$= 2\int_{-\infty}^{\infty}\int_{x}^{\infty}\int_{y}^{\infty} f_{xy}(x,y)dydxdz = \frac{1}{3}$$

$$r = P(Max(X_i, X_k) < Y_j) = P(X_{(1)} \le X_{(2)} \le Y_j)$$

$$= P(X_i, Y_j) \in B_3$$

donde

$$B_3 = \{(x,y,z): x<y<z \vee x<z<y$$

como B_3 es equivalente a B_2, entonces $q = r$.

Esto es

$$q = P\left(Max\left(X_i, X_k\right) < Y_j\right) = \frac{1}{3}$$

Luego, bajo la hipótesis nula, tenemos que:

$$E_\theta(U) = nmp = \frac{mn}{2}$$

$$\mathrm{Var}_\theta(U) = nm\left(\frac{1}{4}\right) + n\left(\frac{1}{3}-\frac{1}{4}\right) + mn(n-1)\left(\frac{1}{3}-\frac{1}{4}\right) \qquad (1.11)$$

$$= \frac{mn(m+n+1)}{12}$$

1.2. DISTRIBUCIÓN ASINTÓTICA DE U

De la ecuación (1.07), podemos observar que la estadística U es una variable aleatoria definida como la suma de m variables indicadoras h_{ij} constituyen ensayos independientes de Bernoulli. Luego, bajo la hipótesis nula $H_o : F_X = F_Y$ y asumiendo que F_X y F_Y son continuas con $0 < P(X < Y) < 1$, por el teorema Central del límite, tenemos

$$\frac{U - (mn/2)}{\sqrt{\frac{mn(m+n+1)}{12}}} \xrightarrow{\text{d}} N(0,1) \quad \text{si Mín}(m,n) \to \infty \qquad (1.12)$$

1.3. PARAMETROS DE LA ESTADISTICA W DE WICOXON

La ecuación (1.04), define una forma equivalente a (1.02). Esto es,

$$W = \sum_{k=m+1}^{m+n} R_k = U + \frac{n(n+1)}{2}$$

por (1.08), (1.10) y (1.11), tenemos

$$E_\theta(W) = E_\theta(U) + \frac{n(n+1)}{2}$$

$$= \frac{mn}{2} + \frac{n(n+1)}{2} = \frac{n(m+n+1)}{2} \qquad (1.13)$$

$$\text{Var}_\theta(W) = \text{Var}_\theta(U) = \frac{mn(m+n+1)}{12} \qquad (1.14)$$

y la distribución asintótica de W es

$$\frac{W - n(m+n+1)/2}{\sqrt{\frac{mn(m+n+1)}{12}}} \xrightarrow{\text{d}} N(0,1) \quad \text{si Mín}(m,n) \to \infty \qquad (1.15)$$

EL TEST DE GEHAN

Este test es una extensión del test de Wilcoxon, donde se consideran, en las muestras X e Y, el problema del censuramiento de los datos.

2.1. NOTACIÓN

Bajo el cesuramiento aleatorio (censuramiento por la derecha), la notación que adoptaremos s la siguiente:

MUESTRA – 01:

Sean $T_1, T_2, ..., T_m$ los tiempos de sobrevivencia de los pacientes en la muestra – 01, que son m variables aleatorias iid con función de distribución común F_1.

Sean $C_1, C_2, ..., C_m$ los tiempos de censuramiento de los pacientes de la muestra - 01, que son m variables aleatorias iid con alguna función común G_1. Además, asumiremos que las variables que las variables aleatorias T y C son independientes.

Entonces, los datos censurados en la muestra son

$$(X_1, \delta_1), (X_2, \delta_2), ..., (X_m, \delta_m) \qquad (2.01)$$

donde

$$X_k = Mín\ (T_k, C_k) \qquad \text{y} \qquad \delta_k = I_{[T_k \leq C_k]}$$

MUESTRA – 02:

Sean $U_1, U_2, ..., U_N$ los tiempos de sobrevivencia de los pacientes en la muestra – 02, que son n variables aleatorias iid con función de distribución común F_2.

Sean $D_1, D_2, ..., D_N$ los tiempos de sobrevivencia de los pacientes en la muestra – 02, que son n variables aleatorias iid con función de distribución común G_2. Además, asumiremos que las variables aleatorias U y D independientes.

Entonces, los datos censurados en la muestra son

$$(Y_k, \varepsilon_1), (Y_2, \varepsilon_2), ..., (Y_n, \varepsilon_n) \qquad (2.02)$$

donde

$$Y_k = Mín(U_k, D_k) \qquad \text{y} \qquad \varepsilon_k = I_{[U_k \le D_k]}$$

2.2. LA ESTADÍSTICA DE GEHAN

Dadas las muestras observadas de la forma de (2.01) y (2.02), que provienen de las poblaciones F_1 y F_2 respectivamente. Entonces, la función de sobrevivencia en cada población, son $S_1(t) =$ y $S_2(t)$, donde

$$\begin{aligned} S_1(t) &= P(T > t) = 1 - F_1(t) \\ S_2(t) &= P(V > t) = 1 - F_2(t) \end{aligned} \tag{2.03}$$

Entonces, al comparar las dos muestras, la hipótesis natural que podemos considerar es:

$$H_o : S_1(t) = S_2(t) \tag{2.04}$$

que, por (2.03), podemos observar que la hipótesis nula (2.04) es equivalente a:

$$H_o : F_1(t) = F_2(t) \tag{2.05}$$

de modo que la hipótesis alternativas son:

$$\begin{aligned} H_1 &: F_1(t) > F_2(t) \\ H_2 &: F_1(t) > F_2(t) \\ H_3 &: F_1(t) \neq F_2(t) \end{aligned} \tag{2.06}$$

Gehan (1965), como una generalización del test de Wicoxon, propone una estadística de test para el caso existan observaciones censuradas en las muestras, que viene a ser una adaptación del test de Wilcoxon al caso de datos censurados.

Sea la muestra combinada

$$(X_1, \delta_1),\ (X_2, \delta_2), \ldots, (X_m, \delta_m)\quad (Y_1, \varepsilon_1),\ (Y_2, \varepsilon_2), \ldots, (Y_n, \varepsilon_n)$$

que los podemos denotar por

$$(Z_1, \xi_1), \quad (Z_2, \xi_2), \ldots, (Z_{m+n}, \xi_{m+n})$$

cuya estadística de orden de $Z_{(1)}$, $Z_{(2)}$, ..., Z_{m+n} es

$$Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(m+n)}$$

donde

$$\begin{aligned} Z_{(1)} &= Min(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Z_{(m+n)} &= Máx(X_1, \ldots, X_m, Y_1, \ldots, Y_n) \end{aligned}$$

Sea

$$R_{1k} = Rango\ (X_k) \text{ en la muestra combinada}$$
$$k = 1, 2, \ldots, m$$

luego, por (1.02, la estadística de Wilcoxon para datos censurados es definido por

$$R_1 = \sum_{k=1}^{m} R_{1k} \qquad (2.07)$$

que es la estadística de test de Gehan. La regla de decisión del test, bajo la $H_0 : \quad F_1 = F_2$, que es R_2 es pequeño o grande se debe rechazar la hipótesis nula.

2.2.1 CASO DE DATOS NO CENSURADOS

En el caso donde las observaciones de las dos muestras no están afectadas por el censuramiento, la estadística R_1 defina en (2.07) es la misma que (1.02), la definida en la primera parte. Una variación que propone Gehan (1965) es que la estadística U de Mann-Whitney puede ser mejorado si definimos los escores U_{ij} de la siguiente forma del siguiente lema.

LEMA: La forma de la estadística de Mann-Whitney del test de Wilcoxon definido por

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} \overset{def}{\longrightarrow} \sum_{i=1}^{m} \sum_{j=1}^{n} U(X_i, Y_j) \qquad (2.08)$$

donde

$$U_{ij} \overset{def}{\longrightarrow} U(X_i, Y_j) = \begin{cases} +1 & \text{Si} \quad T_i > U_j \\ 0 & \text{Si} \quad T_i = U_j \\ -1 & \text{Si} \quad T_i < U_j \end{cases} \qquad (2.09)$$

Podemos observar que la expresión (2.08) es la misma que (1.03) y (2.10) es la equivalente a (1.02) y (1.04).

En la demostración del lema, debemos observar los siguientes argumentos:

Sea $X_1, X_2, ..., X_m$ la muestra observada en la Muestra – 01, donde

$$X_k = T_k \quad \text{y} \quad \delta_k = 1 \quad ; \quad k = 1, 2, ..., m$$

Sea $Y_1, Y_2, ..., Y_n$ la muestra observada en la Muestra – 02, donde

$$Y_k = U_k \quad \text{y} \quad \varepsilon_k = 1 \quad ; \quad k = 1, 2, ..., n$$

la muestra combinada de las dos muestras es

$$X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$$

o lo que es lo mismo

$$X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n$$
$$Z_1, Z_2, ..., Z_m, Z_{m+1}, Z_{m+2}, ..., Z_{m+n}$$

luego, la estadística de orden de la muestra combinada es

$$Z_{(1)} \leq Z_{(2)} \leq ... \leq Z_{(m+n)}$$

En la muestra combinada podemos definir los rangos de X_k o de $X_{(k)}$ donde

R_{1k}=Rango de X_k en la muestra combinada

y

R^*_{1k} =Rango de $X_{(k)}$ en la muestra combinada

de modo que la expresión (2.07), que en este caso es la estadística de Wilcoxon, puede ser escrita como

$$R_1 = \sum_{k=1}^{m} R_{1k} \quad = \quad \sum_{k=1}^{m} R^*_{1k} \tag{2.11}$$

y por (2.09), la estadística de Mann-Whitney como

$$U \quad = \sum_{i=1}^{m} \quad \sum_{j=1}^{n} U_{ij} \quad = \quad W_{XY} - W_{YX} \tag{2.12}$$

donde

$$W_{XY} = \#\left\{\text{Pares}\left(X_i, Y_j\right): X_i > Y_j\right.$$
$$W_{YX} = \#\left\{\text{Pares}\left(X_i, Y_j\right): X_i < Y_j\right.$$

tal que

$$W_{XY} + W_{YX} = mn \tag{2.13}$$

Entonces, observando que:

$X_{(1)} : R^*_{11}$ = Rango de $X_{(1)}$

En la muestra $\{Y, ..., Y, X_{(1)}$..... existen $R^*_{11} - 1$ pares $\left(X_{(1)}, Y\right)$ tal que $X_{(1)} > Y$.

$X_{(2)}: R_{12}^{*}$ = Rango de $X_{(2)}$

En la muestra $\{Y, ..., Y, X_{(1)}, y, ...Y, X_{(2)}, ...$ existen $R_{12}^{*}-1$ observaciones menores que $X_{(2)}$, de los cuales hay $R_{12}^{*}-2$ pares $(X_{(2)}, Y)$, tal que $X_{(2)} > Y$.

$X_{(m)}: R_{1m}^{*}$ = Rango de $X_{(m)}$

En la muestra $\{Y,...,Y, X_{(1)}, y, ...Y, X_{(m)}, ..\}$ existen $R_{1m}^{*}-1$ observaciones menores que $X_{(m)}$, de los cuales hay $R_{1m}^{*}-m$ pares $(X_{(m)}, Y)$, tal que $X_{(m)} > Y$.

Tenemos

$$\begin{aligned} W_{XY} &= \sum_{k=1}^{m} \left[R_{1k}^{*} - k\right] = \sum_{k=1}^{m} R_{k=1}^{*} - \sum_{k=1}^{m} k \\ &= \sum_{k=1}^{m} R_{1k}^{*} - \frac{m(m+1)}{2} \end{aligned} \tag{2.14}$$

de modo que

$$\begin{aligned} U &= W_{XY} - W_{YX} = W_{XY} - (mn - W_{XY}) \\ &= \sum_{K=1}^{m} R_{1K}^{*} - \frac{m(m+1)}{2} - \left(mn - \sum_{k=1}^{m} R_{1k}^{*} + \frac{m(m+1)}{2} \right) \\ &= 2\sum_{k=1}^{m} R_{1k}^{*} - m(m+1) - mn \\ &= 2R_{1} - m(m+1) - mn \end{aligned}$$

por tanto, resolviendo la ecuación para R_1 obtenemos el resultado (2.10) del lema. Esto es,

$$R_1 = \frac{U}{2} + \frac{m(m+n+1)}{2}$$

Este resultado muestra que existe una forma alternativa para expresar la estadística de Wilcoxon. Asimismo, a partir de este resultado, bajo la hipótesis nula $H_0 : F_1(t) = F_2(t)$, podemos obtener la media y varianza de U y luego de R_1.

$$E_0(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} E_0(U_{ij})$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} (P(X > Y) - P(X > Y)) = 0$$

$$\text{Var}_0(U) = E_0(U^2) = E_0\left(\sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} \right)^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} E_0(U_{ij}^2) + \sum_{i=1}^{m} \sum_{j\neq 1}^{n} E_0(U_{ij}U_{il})$$

$$+ \sum_{j=1}^{n} \sum_{i\neq k}^{m} E_0(U_{ij}U_{kj}) + \sum_{i\neq k}^{m} \sum_{j\neq 1}^{n} E_0(U_{ij}U_{kl})$$

$$= mn + \frac{mn(n-1)}{3} + \frac{mn(m-1)}{3} + 0$$

$$= \frac{mn(m+n+1)}{3} \qquad (2.15)$$

de modo que las expresiones (1.13) y (1.14) están por

$$E_0(R_1) = \frac{1}{2} E_0(U) + \frac{m(m+n+1)}{2} = \frac{m(m+n+1)}{2}$$

$$\text{Var}(R_1) = \frac{1}{4} \text{Var}_0(U) = \frac{mn(m+n+1)}{12}$$

2.2.2. CASO DE DATOS CENSURADOS

En el caso donde el censuramiento está presente en algunas de las observaciones de las dos muestras, Gehan (1965)propone una generalización del test de Wilcoxon que está condicionado por el censuramiento observado en cada una de las muestras. Esto implica

una modificación de la estadística R_1 definida en (2.07=, donde la estadística U de Mann-Whitney, de la forma.

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} \quad \underline{\text{def}} \quad \sum_{i=1}^{m} \sum_{j=1}^{n} U(X_1, Y_j) \tag{2.16}$$

es modificado, definiendo los escores U_{ij} de manera similar a lo que se definió en (2.09), esto es

$$U_{ij} \quad \underline{def} \quad U(X_i, Y_j) = \begin{cases} +1 & \text{Si} \quad T_i > U_j \\ 0 & \text{Si} \quad T_i = U_j \\ -1 & \text{Si} \quad T_i < U_j \end{cases} \tag{2.17}$$

Donde, en este caso, debemos fijar un criterio para determinar la condición de $>$, $=$ y $<$ en cada par (T_i, T_j). Esto es,

$$[T_1 > U_j] \Leftrightarrow [X_i > Y_j, \varepsilon_j = 1] \vee [X_1 = Y_j, \delta_i = 0, \varepsilon_j = 1]$$

$$[T_1 > U_j] \Leftrightarrow [X_i < Y_j, \varepsilon_j = 1] \vee [X_1 = Y_j, \delta_i = 1, \varepsilon_j = 0]$$

$$[T_1 = U_j] \Leftrightarrow \text{en otros casos} \tag{2.18}$$

De modo que la estadística de Mann-Whitney (2.16), para el caso de datos censurados, puede ser expresado en la forma de (2.12), donde se contabilizan los pares que satisfacen cada una de las condiciones (2.18), de acuerdo a (2.17). Esto es

$$\begin{aligned} U &= \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij} \\ &= \#\{\text{Pares}(X_i, Y_j): \quad X_i > Y_j, \varepsilon_j = 1\} \\ &\quad + \#\{\text{Pares}(X_i, Y_j): \quad X_i = Y_j, \delta_i = 0, \varepsilon_j = 1 \\ &\quad - \#\{\text{Pares}(X_i, Y_j): \quad X_i < Y_j, \delta_i = 1\} \\ &\quad - \#\{\text{Pares}(X_i, Y_j): X_i = Y_j, \delta_i = 1, \varepsilon_j = 0\} \end{aligned} \tag{2.19}$$

El test de la hipótesis nula $H_0 : F_1(t) = F_2(t)$, esta se rechazará si U 0 $|U|$ es grande. La distribución asintótica de la estadística U es la normal., para calcular los parámetros de esta distribución requerimos conocer la media y varianza de U, esto es, los momentos de primer y segundo orden de U .

Para el cálculo de la media y varianza de U, Gehan (1965) hace uso de la teoría de las permutaciones bajo una hipótesis nula más restrictiva. Esto es

$$H_0^* : \quad F_1(t) = F_2(t) \quad \text{y} \quad G_1(t) = G_2(t) \tag{2.20}$$

Una forma alternativa para expresar la estadística de Gehan, es la resultante a partir de la muestra combinada, denotada por

$$(X_1, \delta_1), \ \ldots, \ (X_m, \delta_m), \ (Y_1, \varepsilon_1), \ \ldots, \ (Y_n, \varepsilon_n)$$

o lo que es lo mismo

$$(Z_1, \xi_1), \ldots, (Z_m, \xi_m), (Z_{m+1}, \xi_{m+1}), \ldots, (Z_{m+n}, \xi_{m+n})$$

Luego, el problema equivalente, consiste en obtener una muestra de tamaño m , sin reemplazo, de una urna que contiene $m+n$ bolas distinguibles denotadas por $(Z_1, \xi_1), \ldots, (Z_{m+n}, \xi_{m+n}$. La muestra obtenida los denotaremos por $(X_1, \delta_1), \ldots, (X_m, \delta_m)$ y las observaciones no muestreadas las denotaremos por $(Y_1, \varepsilon_1), \ldots, (Y_n, \varepsilon_n)$. Entonces, Mantel (1967) propuso una forma alternativa de la estadística U , denotada por U_{ij}^* es definido como

$$U_{ij}^* \overset{def}{=} U^*\left[(Z_i, \xi_i), (Z_j, \xi_j)\right]$$

$$= \begin{cases} +1 & \text{Si } Z_i > Z_j, \xi_j = 1 \,;\, Z_i + Z_j, \xi_i = 0, \xi_j = 1 \\ 0 & \text{Otro Caso} \\ -1 & \text{Si } Z_i < Z_j, \xi_i = 1 \,;\, Z_i = Z_j, \xi_i = 1, \xi_j = 0 \end{cases} \tag{2.21}$$

de modo que

$$U_i^* = \sum_{\substack{j=1 \\ j \neq i}} U_{ij}^*$$

y la estadística de Gehan es de la forma

$$U_{\text{Gehan}} = \sum_{j=1} U_i^* I_{[i \in I_1]} \tag{2.22}$$

donde I_1 es un conjunto de enteros que indica la muestra 1. Esto es,

$$I_1 = \text{Subconjunto tamaño } \{1, 2, ..., m+n\}$$

Para observar este resultado, primeramente podemos estudiar el siguiente caso particular. Sea la muestra combinada

$$(Z_1, \xi_1), ..., (Z_m, \xi_m), (Z_{m+1}, \xi_{m+1}), ..., Z_{m+n}, \xi_{m+n}$$

$$(X_1, \xi_1), ..., (X_m, \xi_m), (Y_1, \varepsilon_1), ..., (Y_n, \varepsilon_n)$$

donde $I_1 = \{1, 2, ..., m$. Luego, tenemos que

$$U = \sum_{i=1}^{m} U_i^* \tag{2.23}$$

pero observemos el siguiente argumento particular

$$
\begin{array}{l}
U^*_{11} + U^*_{12} \;\; + U^*_{13} + \ldots + U^*_{1m-1} + U^*_{1m} \ldots \;\; + \ldots + U^*_{1m+n} \\
U^*_{21} + U^*_{22} \;\; + U^*_{23} + \ldots + U^*_{2m-1} + U^*_{2m} \ldots \;\; + \ldots + U^*_{2m+n} \\
U^*_{31} + U^*_{32} \;\; + U^*_{33} + \ldots + U^*_{3m-1} + U^*_{3m} \ldots \;\; + \ldots + U^*_{3m+n} \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
U^*_{m1} + U^*_{m2} + U^*_{m3} + \ldots + U^*_{mm-1} + U^*_{mm} \ldots + \ldots + U^*_{mm+n}
\end{array}
$$

en el lado izquierdo del arreglo anterior, podemos observar que:

$$U^*_{ij} = U^*_{ji} \text{ de modo que } U^*_{ij} + U^*_{ji} = 0$$

de modo que,

$$U = \sum_{i=1}^{m} U^*_i = \sum_{i=1}^{m} \sum_{j=1}^{n} U^*_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} U^*_{im+j}$$

donde

$$U^*_{im+j} = \begin{cases} +1 & \text{Si } Z_i > Z_j, \xi_j = 1 ; Z_i = Z_j, \xi_i = 0, \xi_j = 1 \\ 0 & \text{Otro caso} \\ -1 & \text{Si } Z_i < Z_j, \xi_i = 1 ; Z_i = Z_j, \xi_i = 1, \xi_j = 0 \end{cases}$$

Este resultado particular puede ser generalizado mediante el siguiente argumento:

- Sea $I_1 =$ Subconjunto de tamaño m de $\{1, 2, ..., m+n\}$
- Si $k_1, k_2 \in I_1$, entonces

$$U^*_{k_1k_2} = -U^*_{k_1k_2} \tag{2.24}$$

De modo que

$$U^*_k = \sum_{\substack{t=1 \\ t\neq k}}^{m+n} U^*_{kt} = \sum_{\substack{t:t\in I_1 \\ t\neq k}}^{n} U^*_{kt} + \sum_{\substack{t:t\in I_1^c \\ t\neq k}}^{n} U^*_{kt} \tag{2.25}$$

Luego, la forma alternativa de la expresión (2.22) es,

$$U_{\text{Gehan}} = \sum_{k=1}^{m+n} U_i^* I_{[i\in I_1]}$$

$$= \sum_{k=1}^{m+n} \left[\sum_{\substack{t:t\in I_1 \\ t\neq k}}^{n} U_{kt}^* + \sum_{\substack{t:t\in I_1^c \\ t\neq k}}^{n} U_{kt}^* \right] I_{[i\in I_1]}$$

$$= \sum_{k=1}^{m+n} \left[\sum_{\substack{t:t\in I_1^c \\ t\neq 1}}^{n} U_{kt}^* \right] I_{[i\in I_1]} \qquad (2.26)$$

En esta última expresión podemos observar que, bajo la hipótesis nula de la forma de (2.20), tenemos que las observaciones $(Z_k, \xi_k), k = 1,2,...,m+n$ son idénticamente distribuidos, mas no son independientes, dado que por la definición de los Z's se pierde la independencia. Esta condición implica que los escores U_{kt}^* y U_k^* son idénticamente distribuidos.

Luego, pensando en el problema de dos muestras, dado los $m+n$ escores $U_1^*, U_2^*, ..., U_{m+n}^*$ ocurren con probabilidad $1/m+n$, bajo la hipótesis nula

$$H_0^* : \mathrm{F}_1(t) = \mathrm{F}_2(t) \quad \text{y} \quad \mathrm{G}_1(t) = \mathrm{G}_2(t)$$

Podemos calcular la esperanza y varianza de U^*. De (2.23),

$$E_{H^*}(U^*) = E_{H^*}\left[\sum_{i=1}^{m} U_i^* \right] = m\ E_{H^*}\left[U_i^*\right]$$

$$= m\ \frac{1}{m+n}\left[U_1^* + ... + U_{m+n}^*\right] = 0$$

dado que $\sum_{i=1}^{m+n} U_1^* = 0$

Dado que los son idénticamente distribuidos, pero no independientes, tenemos que

$$\mathrm{Var}_{H^*}\left(U^*\right)=\mathrm{Var}_{H^*}\left(\sum_{i=1}^{m} U_i^*\right)$$

$$=m\,\mathrm{Var}_{H^*}\left(U_i^*\right)+m(m-1)\,\mathrm{Cov}_{H^*}\left(U_i^*,U_j^*\right)$$

donde

$$\mathrm{Var}_{H^*}\left(U_k^*\right)=\frac{1}{m+n}\left(U_1^{*2}+\ldots+U_{m+n}^{*2}\right)=\frac{1}{m+n}\left(\sum_{i=1}^{m+n} U_1^{*2}\right)$$

Para el cálculo de la covarianza, podemos hacer uso de los resultados conocidos del muestreo en poblaciones finitas. Para esto usaremos el siguiente argumento de las urnas: Dado que existen valores, deseamos efectuar retiros sin reemplazo, a fin de conocer la distribución de la suma de dichos valores. Esto es,

$$U=\sum_{k=1}^{m+n} U_i^*\,\mathrm{I}_{[i\in \mathrm{I}_1]}$$

Sea $\mathrm{N}=m+n$. Si n = 0, N = m, entonces

$$\mathrm{Var}(U)=0\Leftrightarrow U=\sum_{k=1}^{m+n} U_i^*$$

De modo que

$$0=\mathrm{Var}(U)=(m+n)\,\mathrm{Var}_{H^*}\left(U_k^*\right)+(m+n)(m+n-1)\,\mathrm{Cov}_{H^*}\left(U_k^*,U_t^*\right)$$

$$=(m+n)\frac{1}{m+n}\sum_{i=1}^{m+n} U_i^{*2}+(m+n)(m+n-1)\,\mathrm{Cov}_{H^*}\left(U_k^*,U_t^*\right)$$

y

$$\mathrm{Cov}_{H^*}\left(U_k^*,U_t^*\right)=\frac{1}{(m+n)(m+n-1)}\sum_{i=1}^{m+n} U_i^{*2}$$

por tanto

$$\mathrm{Var}_{H^*}(U^*) = \mathrm{m}\,\mathrm{Var}_{H^*}(U_i^*) + \mathrm{m}(\mathrm{m}-1)\,\mathrm{Cov}_{H^*}(U_i^*, U_j^*)$$

$$= \frac{\mathrm{m}}{\mathrm{m}+\mathrm{n}} \sum_{i=1}^{\mathrm{m}+\mathrm{n}} U_i^{*2} - \frac{\mathrm{m}(\mathrm{m}-1)}{(\mathrm{m}+\mathrm{n})(\mathrm{m}+\mathrm{n}-1)} \sum_{i=1}^{\mathrm{m}+\mathrm{n}} U_i^{*2}$$

$$= \frac{\mathrm{mn}}{(\mathrm{m}+\mathrm{n})(\mathrm{m}+\mathrm{n}-1)} \sum_{i=1}^{\mathrm{m}+\mathrm{n}} U_i^{*2}$$

por la equivalencia entre la U_{Gehan} y la U^*_{Mantel}, bajo la hipótesis nula (2.20) de los resultados anteriores, tenemos que

$$E_{H^*}(U) = 0 \tag{2.27}$$

$$\mathrm{Var}_{H^*}(U) = \frac{\mathrm{mn}}{(\mathrm{m}+\mathrm{n})(\mathrm{m}+\mathrm{n}-1)} \sum_{i=1}^{\mathrm{m}+\mathrm{n}} U_i^{*2} \tag{2.28}$$

De acuerdo al teorema central del límite, para el caso de poblaciones finitas, la distribución asintótica de U es aproximadamente normal. Esto es,

$$\frac{U - E_{H^*}(U)}{\sqrt{\mathrm{Var}_{H^*}(U)}} \approx \mathrm{N}(0, 1) \tag{2.29}$$

BIBLIOGRAFÍA

1. ANDERSON, S., AUQUIER, A., HAUCH, W., OAKES, D., VENDAELE, W., WEISVBERG, H. (1980): Statistical Methods for Comparative Studies. Ed. J. Willey – N.Y.
2. BICKEL, P., DOCKSUM, K. (1977): Mathematical Statistics. Ed. Holden- Day Inc. – San Farancisco.
3. COX, D. R. (1972): Regresion Models and Life Tables. JRSS Series B, Vol 34
4. GEHAN, J. (1965): A generalizee wilcoxon test for comparing arbitrarily singly – censored samples Biometrika, Vol. 52.
5. GEHAN, E. & THOMAS D. (1969): The performance of some two sample test in small samples with and without censoring. Biometrika, vol. 56
6. MILLER, R. (1981): Survival Analysis. De. J. Willey N.Y.
7. MILLER, EFRON, BROWN & MOSES (1980): Biostatistics Casebook. John Willey & Sons. N.Y.
8. RAO, C. R. (1973): Linear Statistical Inference and Its Aplications. 2da. Edic. Ed. J. Willey N.Y.
9. SOARES, J., BARTMAN, F., (1983): Metodos Estadisticos em Medicina e Biologia. IMPA R. J. – Brazil.