Revista Colombiana de Estadística
Junio 2010, volumen 33, no. 1, pp. 149 a 166

Un test de similitud entre dos secuencias dicotómicas ordenadas

A Similarity Test between Two Dichotomic Ordered Sequences

Ramón Giraldo Henao[a], Jimmy Corzo Salamanca[b]

Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Resumen

Se propone una prueba para la hipótesis de similitud de dos secuencias dicotómicas ordenadas. Un estudio de potencia basado en simulación indica que la prueba propuesta mantiene su tamaño bajo la hipótesis nula y que su potencia crece adecuadamente con el tamaño de muestra. Además la prueba tiene la misma potencia que la prueba del signo y supera en potencia a las pruebas basadas en las estadísticas de antirrachas y de Wilcoxon.

***Palabras clave*:** similitud, datos dicotómicos, potencia de una prueba.

Abstract

We propose a test for the hypothesis of similarity between dichotomic ordered sequences. A simulation study was carried out to estimate the power of the proposed test. It is shown that the test maintain its size under the null hypothesis and that its power increase with the considered alternative hypothesis. In addition the proposed test is as powerful as the sign test and overtakes the antiruns test and the Wilcoxon test.

***Key words*:** Similarity, Dichotomous data, Power of a test.

1. Introducción

En muchas áreas de la ciencia se llevan a cabo experimentos en los que se mide una variable respuesta de tipo dicotómico bajo dos tratamientos, en los que dicha respuesta tiene un orden específico asociado a bloques o covariables. Por ejemplo en farmacia cuando se realiza un bioensayo con el propósito de establecer diferencias entre machos y hembras respecto a su respuesta a un fármaco (vive/muere,

[a]Profesor Asociado. E-mail: rgiraldoh@unal.edu.co
[b]Profesor Asociado. E-mail: jacorzos@unal.edu.co

mejora/no mejora) y se considera como covariable la dosis del fármaco o la edad de los individuos. En psicología cuando se aplican dos test a los mismos individuos que aprueban o no aprueban el test y estos son de diferente edad o tienen diferente nivel de educación. En la industria cuando dos catadores califican como buenas o malas las muestras de algún producto y estas tienen diferente nivel de calidad. Es importante resaltar que en las tres situaciones descritas, aunque hay un orden en las respuestas, las observaciones pueden considerarse independientes puesto que estas mismas son evaluadas en individuos distintos. En este trabajo se presenta una estadística que permite realizar pruebas de hipótesis con información obtenida en experimentos análogos a los arriba mencionados, es decir en aquellos en los que la variable respuesta es binaria y tiene un orden implícito dado por una covariable, hay dos tratamientos y se asume independencia entre las observaciones.

El artículo se organiza como sigue: en la sección 2 se plantea el tipo de hipótesis de interés, se define la estadística de prueba y su distribución bajo la hipótesis nula. En la sección 3 se presentan teoremas respecto a los dos primeros momentos y se calcula la distancia entre la distribución de la estadística de prueba y la normal estándar para varios tamaños de sucesión. En la sección 4 se muestran los resultados de un estudio de potencia para la prueba propuesta y se compara esta con la potencia de otras pruebas útiles para la misma hipótesis . En la sección 5 se presentan aplicaciones de la metodología propuesta a dos conjuntos de datos reales correspondientes a mediciones de oxígeno disuelto en dos niveles de la columna de agua en la Ciénaga Grande de Santa Marta (CGSM)(IGAC 1973). Algunas conclusiones y propuestas de trabajo futuro se dan en la sección 6. El artículo finaliza con un apéndice en el que se presentan tablas de valores críticos y el código R (R Development Core Team 2005) usado.

2. Hipótesis y estadística de prueba

Sean $\eta_{11}, \ldots, \eta_{1n}$ y $\eta_{21}, \ldots, \eta_{2n}$ dos sucesiones dicotómicas observadas bajo los tratamientos 1, 2 y que están ordenadas según una covariable que tiene niveles $1, \ldots, n$. Si se define

$$\tau_k = \begin{cases} 0 & \text{si } \eta_{1k} = \eta_{2k} \\ 1 & \text{si } \eta_{1k} \neq \eta_{2k} \end{cases}, k = 1, \ldots, n \tag{1}$$

entonces $\tau_1, \ldots, \tau_n$ conforma una nueva sucesión dicotómica que en caso de contener muchos ceros indicará que las dos secuencias son símiles (concordantes) o por el contrario que las respuestas bajo los dos tratamientos son disímiles (discordantes) cuando esté compuesta por muchos unos. En términos de probabilidad se dirá que hay similitud entre las dos secuencias siempre que en la sucesión $\tau_1, \ldots, \tau_n$ la probabilidad del valor cero sea mayor o igual que la del valor uno. De otro lado, se estará bajo la hipótesis alterna (no similitud entre las dos secuencias) cada vez que la probabilidad del valor uno en la sucesión sea mayor que la del valor cero. De acuerdo con lo anterior, las hipótesis de interés pueden plantearsen de la siguiente

forma:

$$\begin{aligned} H_0 &: P(\tau_k = 0) \geq P(\tau_k = 1) \\ H_{\mathrm{a}} &: P(\tau_k = 0) < P(\tau_k = 1) \end{aligned} \tag{2}$$

A continuación se define la estadística de prueba propuesta para la hipótesis dada en (2). Esta se basa en el conteo del número de discordancias entre las sucesiones a comparar y en la posición de estas dentro de la sucesión de los τ_i. Adicionalmente incluye dos términos que facilitan su interpretación. La expresión de la estadística

$$GC(n) = \tau_\bullet + K + n(n-2) - h(n, \tau_\bullet) \tag{3}$$

donde n es el tamaño de la sucesión y los otros términos se definieron como

$$\tau_\bullet = \sum_{k=1}^{n} \tau_k, \tag{4}$$

$$K = \sum_{k=1}^{n} \delta_k, \ \text{ con } \delta_k = \begin{cases} k & \text{si } \tau_k = 1 \\ \tau_\bullet - n & \text{si } \tau_k = 0 \end{cases} \tag{5}$$

y

$$h(n, \tau_\bullet) = \begin{cases} -2n & \text{si } \tau_\bullet = 0 \\ 0 & \text{si } \tau_\bullet = 1 \\ a_m - b_m n & \text{si } \tau_\bullet = 2, \ldots, n, \ m = \tau_\bullet - 2 \end{cases} \tag{6}$$

En (6) $a_0 = 0$, $b_0 = -1$ y para $m \geq 1$ se utilizan las expresiones recursivas $a_m = a_{m-1} + (m^2 + m)$ y $b_m = b_{m-1} + (m-1)$.

La variable $\tau_\bullet$ definida en la ecuación (4) cuenta el número total de discordancias entre las dos sucesiones que están siendo comparadas, es decir, cuenta el número de unos de la sucesión $\tau_1, \ldots, \tau_n$. Esta variable tiene la misma expresión de la estadística usada en la prueba del signo (Conover 1999). En la ecuación (5) K indica la posición de los unos dentro de la secuencia $\tau_1, \ldots, \tau_n$ y su posición dentro de la misma. K es menor cuando las discordancias están al inicio de la secuencia que cuando se presentan hacia el final de la misma. Su valor mínimo se da cuando las dos secuencias originales son totalmente símiles y su valor máximo se obtiene cuando $\tau_k = 1$, para todo k, $k = 1, \ldots, n$, es decir cuando las dos secuencias son totalmente disímiles. Para facilitar la interpretación del indicador dado en (3) se incluyen los términos $n(n-2)$ y $h(n, \tau_\bullet)$. Con estos se logra, que independientemente del tamaño de la sucesión, la variable $GC(n)$ tome su mínimo en cero y aumente consecutivamente en la escala de los enteros positivos. En resumen, la estadística de prueba planteada detecta las diferencias entre dos secuencias binarias ordenadas y permite describir en qué posición del orden considerado es que estas mismas se presentan.

Tabla 1: Valores de $GC(4)$ en las posibles sucesiones dicotómicas de tamaño 4.

τ_1	τ_2	τ_3	τ_4	$\tau_\bullet$	δ_1	δ_2	δ_3	δ_4	K	$GC(4)$
0	0	0	0	0	−4	−4	−4	−4	−16	0
1	0	0	0	1	1	−3	−3	−3	−8	1
0	1	0	0	1	−3	2	−3	−3	−7	2
0	0	1	0	1	−3	−3	3	−3	−6	3
0	0	0	1	1	−3	−3	−3	4	−5	4
1	1	0	0	2	1	2	−2	−2	−1	5
1	0	1	0	2	1	−2	3	−2	0	6
0	1	1	0	2	−2	2	3	−2	1	7
1	0	0	1	2	1	−2	−2	4	1	7
0	1	0	1	2	−2	2	−2	4	2	8
0	0	1	1	2	−2	−2	3	4	3	9
1	1	1	0	3	1	2	3	−1	5	10
1	1	0	1	3	1	2	−1	4	6	11
1	0	1	1	3	1	−1	3	4	7	12
0	1	1	1	3	−1	2	3	4	8	13
1	1	1	1	4	1	2	3	4	10	14

A manera de ilustración del efecto de las expresiones (4), (5) y (6) en el valor del indicador propuesto en la ecuación (3), se presenta el cálculo de $GC(4)$ con todos los posibles arreglos de ceros y unos de una secuencia binaria de tamaño cuatro (tabla 1). Se observa que $GC(4)$ es una variable discreta, monótona creciente, con valores entre 0 y 14. Para este caso los valores del indicador definen puntualmente lo ocurrido respecto al número de unos y a la posición de los mismos dentro de las sucesiones, excepto cuando $GC(4)$ es igual a 7. Valores de $GC(4)$ entre 1 y 4 indican que hubo un solo uno en la sucesión y cada número revela la posición que este ocupa en la misma (1 si el uno está en la primera posición, 2 si está en la segunda, etc.). Valores entre 5 y 9 corresponden a sucesiones en las que hubo dos unos, con $GC(4)$ igual a 5 cuando los dos unos están en las primeras dos posiciones de la sucesión ordenada y a 9 cuando están en las dos últimas. Los valores 6, 7 y 8 reflejan la transición de los dos unos de las dos primeras a las dos últimas posiciones. Valores entre 10 y 13 indican que hubo tres unos y cada uno de estos valores corresponde a una única sucesión (no hay empates y por consiguiente definen explícitamente lo sucedido en la sucesión respecto a la posición de los unos). El valor de $GC(4)$ será 10 cuando los unos estén en las tres primeras posiciones e igual a 13 cuando estén en las tres últimas. Los valores 0 y 14 se obtendrán cuando en la sucesión no haya unos (las dos sucesiones originales son totalmente símiles) o todos los valores sean iguales a uno (las dos sucesiones originales son totalmente disímiles), respectivamente.

Los valores grandes de $GC(n)$ conducen a rechazar la hipótesis de similitud porque estos se presentan cuando la sucesión dicotómica $\tau_1, \ldots, \tau_n$ está compuesta por muchos unos, lo que indica, de acuerdo con el criterio de dicotomización dado en (1), que hay poca semejanza entre las dos sucesiones binarias originales. De lo anterior, la prueba basada en $GC(n)$ rechaza H_0 a favor de H_1 a un nivel de significancia α dado cuando $GC(n) \geq g_{1-\alpha}$, donde $g_{1-\alpha}$ es tal que $P_{H_0}(GC(n) \geq g_{1-\alpha}) = \alpha$. En la sección 6 se presentan los valores críticos para varios niveles

de significancia con tamaños de muestra entre 3 y 15. Usando el programa R (R Development Core Team 2005) dado en la sección 6, se puede hacer estimación de dichos valores críticos para cualquier tamaño de sucesión n. Los valores críticos también pueden obtenerse mediante una aproximación a la distribución normal, como se describe en la siguiente sección.

3. Propiedades de la distribución de $GC(n)$

En esta sección se enuncian dos teoremas referentes a los momentos de primer y segundo orden de la estadística de prueba y se estudia la aproximación de su distribución a una normal. La demostración de los teoremas puede consultarse en Giraldo (2003), disponible en http://www.docentes.unal.edu.co/rgiraldoh/docs/.

En el caso de hipótesis compuestas, del tipo

$$H_0 : \theta \in \Theta_0$$
$$H_a : \theta \in \Theta_1$$

con $\Theta_1 = \Theta - \Theta_0$ y Θ el espacio de parámetros, el nivel de significancia se define por $\alpha = \underset{\theta \in \Theta_0}{\text{máx}} P$ (rechazar H_0) (Dudewicz & Mishra 1988). La hipótesis nula de interés dada en (2) es compuesta y la distribución de la estadística de prueba $GC(n)$ bajo H_0 depende de la probabilidad que se asuma para $P(\tau_k = 0)$ en la sucesión $\tau_1, \ldots, \tau_n$. Bajo H_0 se tiene que $1/2 \leq P(\tau_k = 0) \leq 1$ y el α deseado se obtiene cuando $P(\tau_k = 0) = 1/2$ (Giraldo 2003). Por lo anterior, para el cálculo del valor esperado y la varianza de $GC(n)$ se supone que $P(\tau = 0) = P(\tau = 1) = 1/2$.

Teorema 1. *Sea $\tau_1, \ldots, \tau_n$ una sucesión dicotómica de tamaño n con $P(\tau_k = 1) = 0.5$ y $GC(n)$ definido como en (3); entonces:*

$$E(GC(n)) = \frac{n^3 + 5n}{12} \tag{7}$$

Teorema 2. *Sea $\tau_1, \ldots, \tau_n$ una sucesión dicotómica de tamaño n con $P(\tau_k = 1) = 0 : 5$ y $GC(n)$ definido como en (3); entonces:*

$$\begin{aligned} V(GC(n)) = & \frac{n(n-1)(n-2)(n-3)(4n^2+45n-4)}{576} \\ & + \frac{n(n-1)(90-303n+444n^2-27n^3)}{144} \\ & + \frac{(2n^5-34n^4+54n^3-26n^2+10n)}{12} - \left(\frac{n^3+5n}{12}\right)^2 \end{aligned} \tag{8}$$

En este trabajo no se realiza un estudio teórico de la distribución asintótica de la estadística $GC(n)$. Sin embargo a manera de exploración se estudia la convergencia a la normal calculando, para diferentes tamaños de sucesión n, la diferencia

máxima entre la frecuencia acumulada exacta y esta misma bajo la normal estándar. Para cada n se evalúa:

$$d_n = \max_x \left| F_n(x) - \Phi\left(\frac{x-\mu_n}{\sigma_n}\right)\right| \tag{9}$$

donde x corresponde a un valor de $GC(n)$, $F_n(x)$ es la distribución exacta de $GC(n)$, μ_n y σ_n^2 corresponden al valor esperado y a la varianza dados en (7) y (8), respectivamente, y $\Phi(\cdot)$ es la función de distribución de la normal estándar. Los cálculos de (9) en el total de la distribución y en las colas de la misma, con n entre 3 y 100, se dan en la tabla 2.

Tabla 2: Diferencias máximas entre la distribución de $GC(n)$ y la normal estándar en el total de la distribución y en las colas (al 5 %) de la misma, para tamaños de sucesión n entre 3 y 100.

n	d_n total	$F_n(x) \leq 0.05$	$F_n(x) \geq 0.95$
3	0.120		
4	0.081		
5	0.078	0.004	0.036
6	0.071	0.012	0.027
7	0.059	0.013	0.021
8	0.063	0.012	0.016
9	0.054	0.011	0.013
10	0.058	0.009	0.011
11	0.052	0.009	0.010
12	0.055	0.009	0.010
13	0.051	0.009	0.009
14	0.053	0.009	0.010
15	0.049	0.009	0.007
20	0.047	0.008	0.009
50	0.037	0.009	0.009
100	0.034	0.007	0.007

En la tabla 2 se puede observar que las diferencias en la distribución completa tienen un alto decrecimiento hasta $n = 7$ y que a partir de ahí las diferencias toman valores parecidos que varían entre 6.3 % y 4.9 %. En las colas se presentan comportamientos distintos. En la cola superior las diferencias toman su máximo para un tamaño de sucesión 5 (3.6 %) y de ahí en adelante disminuyen hasta alcanzar, para tamaños de sucesión mayores de 10, diferencias cercanas al 1 %. En la cola inferior cuando el tamaño de sucesión es pequeño ($n = 3$) se obtiene el valor mínimo (0.4 %), posteriormente las diferencias máximas se incrementan hasta 1.3 %, con $n = 7$, y cuando n aumenta se estabilizan en valores cercanos a 0.7 %. Los resultados muestran que en el total de la distribución el ajuste a la normal no es muy bueno, pero que en las colas las diferencias máximas definidas en (3) son relativamente pequeñas (menores del 1 % para tamaños de sucesión mayores de 13). En la figura 1 se presenta una comparación entre las funciones de distribución de la variable $GC(n)$ y la de la normal para varios tamaños de sucesión. Se observa que la distribución de $GC(n)$ es multimodal y que por ello esta se distancia de la normal y además que hacia las colas las diferencias son mucho menores. Para

propósitos de inferencia (prueba de hipótesis) solo se requiere que haya buen ajuste en las colas de la distribución, por ello el uso de una normal en este caso puede ser razonable. De acuerdo con lo anterior, el test dado en (2) podría realizarse de manera aproximada basado en la estadística:

$$Z_c = \frac{GC(n) - E(GC(n))}{\sqrt{V(GC(n))}}$$

donde el valor esperado y la varianza se obtienen de (7) y (8), respectivamente. La hipótesis nula se rechazaría al nivel α si $Z_c > z_{(1-\alpha)}$, con $z_{(1-\alpha)}$ el percentil $(1-\alpha)$ de la distribución normal estándar.

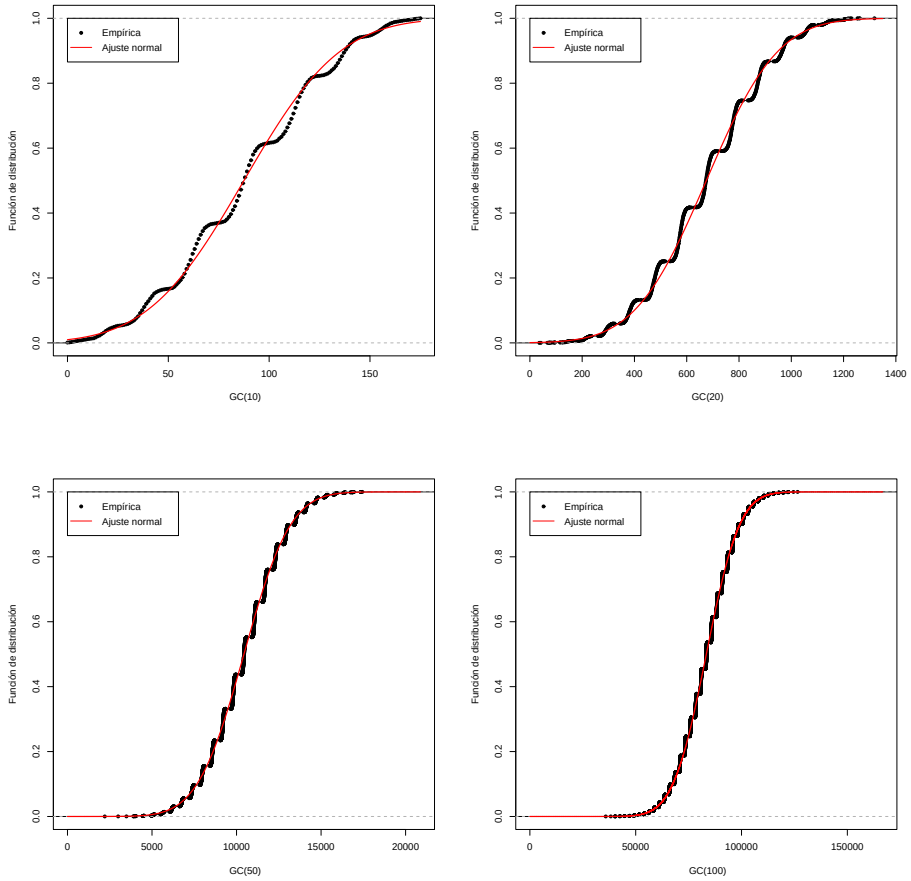

FIGURA 1: Comparación entre la función de distribución de $GC(n)$ y la función de distribución normal con parámetros $E(GC(n))$ y $V(GC(n))$. $n = 10$ (arriba izquierda); $n = 20$ (arriba derecha); $n = 50$ (abajo izquierda); $n = 100$ (abajo derecha). La función de distribución $F_n(x)$ de la variable $GC(n)$ se estima por simulación en el caso de n = 50, 100.

4. Potencia

El estudio de potencia tuvo dos enfoques: en el primero se consideró solo el indicador $GC(n)$ y se calculó la potencia de este para diferentes tamaños de la sucesión (n = 5, 10, 15, 20 y 30). Se usó en cada caso un α del 5 %. Para los valores de $n \leq 15$ se tomaron los valores críticos exactos, por exceso (ver sección 6). Con tamaños de sucesión 20 y 30 se estimaron los valores críticos a través de simulaciones de tamaño 10000 (con $P(\tau_k = 0) = 0.5$). Para estimar la potencia de la prueba se hicieron nuevamente simulaciones de tamaño 10000 bajo la hipótesis alterna, generando sucesiones dicotómicas de tamaño 5, 10, 15, 20 y 30, en las que las probabilidades fueron mayores de 0.5 ($P(\tau_k = 0)$ = 0.6, 0.7, 0.8, 0.9 y 0.99). En cada caso la potencia de la prueba se estimó calculando el número de veces que la estadística $GC(n)$ superó el correspondiente valor crítico antes calculado y dividiendo este sobre el número de simulaciones.

En el segundo enfoque se comparó la potencia de la estadística $GC(n)$ en la prueba de similitud con las obtenidas al usar adaptaciones, para este mismo fin, de las estadísticas $-C$ (Corzo 1990), del signo y de Wilcoxon (Conover 1999). Se emplearon tamaños de sucesión 5, 10 y 15 y en todos los casos se usó α del 5 %. Los valores críticos para $GC(n)$ y para la estadística $-C$ se presentan en la sección 6. Los de las estadísticas del signo y Wilcoxon se tomaron de Conover (1999) y Hollander & Wolfe (1999), respectivamente. Debido a que las estadísticas son discretas y no se consiguen valores críticos exactos para el nivel de significancia usado, se emplearon pruebas aleatorizadas (Dudewicz & Mishra 1988). Las correspondientes funciones críticas para los tres tamaños de sucesión considerados aparecen en Giraldo (2003). Para estimar las correspondientes potencias se simularon 10000 sucesiones dicotómicas de tamaño 5, 10 y 15, respectivamente, para valores de $P(\tau_k = 1)$ = 0.6, 0.7, 0.8, 0.9 y 0.99. Con cada sucesión se calcularon los valores de las cuatro estadísticas $GC(n)$, $-C$, del signo (S) y Wilcoxon (W) y se evaluaron las correspondientes funciones críticas. La potencia en cada caso resultó del cociente entre la suma de los valores obtenidos en las funciones críticas sobre 10000 (tamaño de la simulación).

4.1. Potencia de $GC(n)$

Una propiedad deseable de una prueba es que sea insesgada (Dudewicz & Mishra 1988), es decir que, para un tamaño de muestra fijo, la potencia de la prueba aumente en la medida en que haya alejamiento de la hipótesis nula y que bajo H_0 la probabilidad de rechazo sea pequeña. En la figura 2 se muestran las curvas de potencias estimadas para los 5 tamaños de sucesión considerados. Se observa que, para cada valor de n, la potencia tiene una tendencia creciente en la medida en que se incrementa (es decir cuando hay alejamiento de la hipótesis nula) y que en ningún caso la potencia estimada bajo H_0 excede el valor 0.05, lo cual permite concluir, desde el punto de vista de la simulación, que la prueba basada en $GC(n)$ es insesgada.

Por otra parte, en la figura 3 se muestra el comportamiento de la potencia como función del tamaño de la sucesión. Se puede comprobar allí, gráficamente,

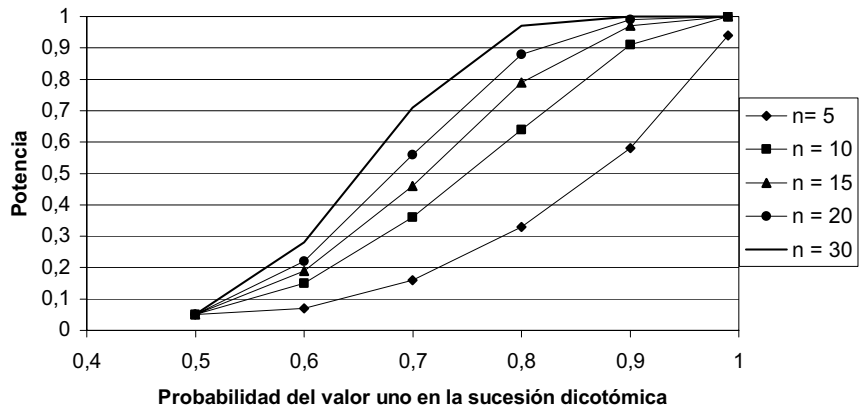

FIGURA 2: Potencia en la prueba de similitud de dos sucesiones dicotómicas con base en $GC(n)$ para varias alternativas (n corresponde al tamaño de la sucesión).

que la potencia es una función creciente en términos del tamaño de la sucesión (a mayor tamaño de sucesión, mayor potencia), lo que indica, con base en resultados de simulación, que esta es consistente (Hollander & Wolfe 1999).

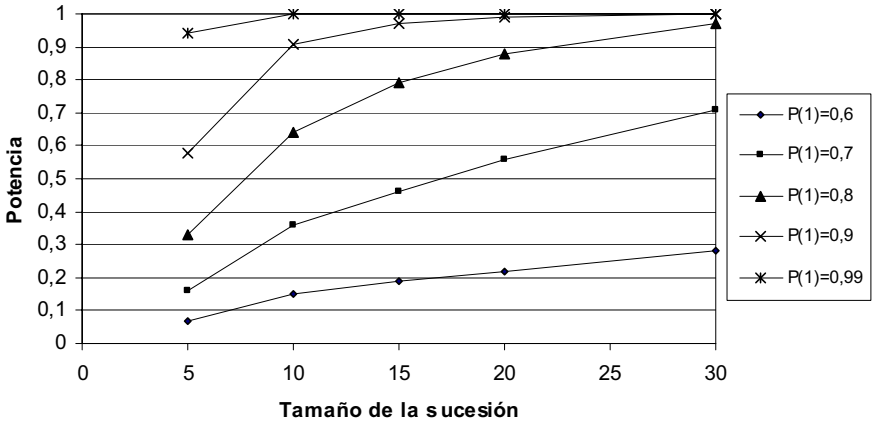

FIGURA 3: Potencia en la prueba de similitud de sucesiones dicotómicas con la estadística $GC(n)$, según el tamaño de la muestra para varias alternativas. P(1) corresponde a $P(\tau_k = 1)$.

4.2. Estudio comparativo de potencia

Con base en el procedimiento de simulación antes mencionado, se obtuvieron los valores de potencia dados en las tablas 3 a 5. Los resultados muestran que para el

tamaño de sucesión más pequeño (n= 5) las cuatro estadísticas consideradas tienen igual potencia (tabla 3); que para los otros dos tamaños de sucesión estudiados (n = 10 y 15) las pruebas basadas en la estadística $GC(n)$ y en la del signo tienen mayor potencia que las dos restantes; y que la prueba basada en la estadística de Wilcoxon resulta ser la menos potente entre las cuatro consideradas.

Tabla 3: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 5 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.11	0.11	0.11	0.11
0.7	0.21	0.21	0.21	0.21
0.8	0.38	0.38	0.38	0.38
0.9	0.63	0.63	0.63	0.63
0.99	0.95	0.95	0.95	0.95

Los valores de potencia estimados se redondearon a dos cifras significativas.

Tabla 4: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 10 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.16	0.16	0.16	0.14
0.7	0.36	0.33	0.35	0.29
0.8	0.65	0.61	0.65	0.58
0.9	0.91	0.88	0.91	0.85
0.99	0.99	0.99	0.99	0.99

Los valores de potencia estimados se redondearon a dos cifras significativas.

Al comparar los resultados de en las tablas 2 a 4 se puede establecer que las cuatro estadísticas producen pruebas insesgadas (en la subsección anterior se estableció esto solamente para la estadística $GCij(n)$). Los resultados descritos, aunque no permiten la deducción de conclusiones desde un punto de vista formal, puesto que se basaron en simulación y se obtuvieron solo para tres tamaños de sucesión, sí hacen posible intuir que la prueba basada en la estadística $GCij(n)$ puede tener la misma potencia que la realizada con base en la estadística del signo (tablas 3 a 5). Esto resulta muy relevante teniendo en cuenta que esta última es la más potente para hipótesis de similitud como la planteada en la ecuación (2) (Randles & Wolfe 1979), cuando la información original es dicotómica (mínima escala de medida).

TABLA 5: Potencia de las pruebas de similitud de dos sucesiones dicotómicas basada en $GC(n)$, y las estadísticas $-C$, del signo (S) y de Wilcoxon (W), para tamaño de sucesión 15 y nivel de significancia del 5 %.

$P(\tau_i = 1)$	$GC(n)$	$-C$	S	W
0.5	0.05	0.05	0.05	0.05
0.6	0.19	0.18	0.19	0.17
0.7	0.47	0.42	0.46	0.38
0.8	0.80	0.75	0.79	0.70
0.9	0.98	0.96	0.98	0.94
0.99	0.99	0.99	0.99	0.99

Los valores de potencia estimados se redondearon a dos cifras significativas.

5. Aplicación: riesgo de muerte de aerobios en la CGSM de acuerdo con el nivel de oxígeno

Una función para medir la influencia del oxígeno disuelto en el riesgo de muerte de aerobios en sistemas tropicales costeros se presentan en la figura 4 (Mancera et al. 1996).

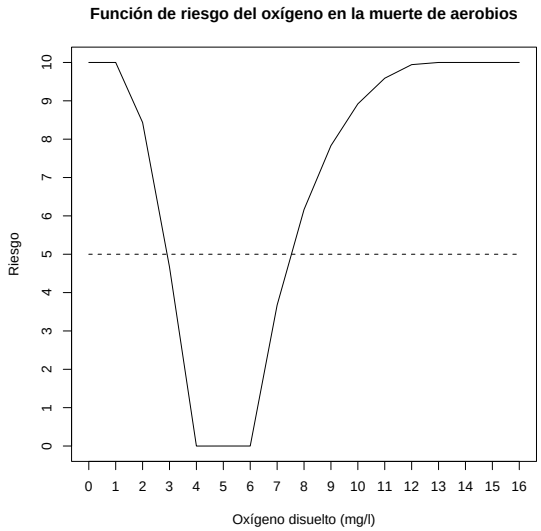

FIGURA 4: Función de riesgo del oxígeno disuelto en el cálculo del indicador de riesgo de mortandad de aerobios (IRMA). La línea punteada corresponde a riesgo igual a 5.

Magnitudes de oxígeno por debajo de 3 (mg/l) o por encima de 7.5 (mg/l) corresponden a valores de la función de riesgo superiores a 5 (figura 4) e implican un riesgo moderado o alto de muerte de peces y otros organismos. En este trabajo se usa dicha función para convertir en datos binarios los registros de oxígeno disuelto obtenidos en 114 sitios de la Ciénaga Grande de Santa Marta en un muestreo llevado a cabo el 8 de marzo de 1997 (Giraldo et al. 2000) y para probar con base

en estos si el oxígeno del sistema se encontraba en un nivel de riesgo normal el día del muestreo. Las observaciones se tomaron en dos niveles de la columna de agua (superficie y fondo).

Un test de igualdad de medias basado en los datos originales permitió establecer que el día de la muestra había diferencias significativas ($P < 0.05$) entre los dos niveles de la columna de agua. En la figura 5 se muestran los correspondientes intervalos de confianza del 95 % para las medias. De acuerdo con esta figura se concluye, como era de esperarse, que la media de oxígeno en superficie es mayor que la media de oxígeno en el fondo de la columna. Este resultado, aunque muy relevante desde el punto de vista biológico, no permite establecer en términos globales si la variable en cuestión está en un nivel normal o de riesgo para la vida de los organismos dentro del sistema. La comparación de las medias de los dos niveles respecto a los puntos críticos 3.5 mg/l y 7 mg/l, aunque útil desde un punto de vista descriptivo, tampoco permite hacer una prueba formal de esta hipótesis. Por ello el uso de la prueba propuesta en este trabajo resulta de interés con la información considerada.

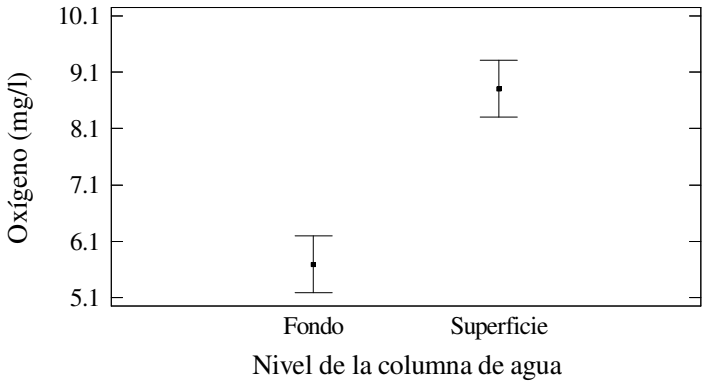

FIGURA 5: Intervalos de confianza del 95 % para la media de oxígeno en dos niveles de la columna de agua en la Ciénaga Grande de Santa Marta. Datos medidos en marzo de 1997 en 114 sitios del sistema.

En todos los casos los valores en superficie fueron mayores que los del fondo. Valores de oxígeno en fondo menores de 3 mg/l y de oxígeno superficial mayores de 7.5 mg/l se codificaron con el valor 1. Aunque hubo diferencias entre superficie y fondo mayores de 4.5 mg/l, en ningún sitio se dio el caso de que simultáneamente el oxígeno de fondo fuera menor de 3 mg/l y el de superficie mayor de 7.5 mg/l, es decir las concordancias en 1 no se presentaron. Con esta dicotomización las concordancias en ceros implican ausencia de riesgo (oxígeno de fondo mayor de 3 mg/l y oxígeno de superficie menor de 7.5 mg/l). De acuerdo con esto, las hipótesis de interés son

$$H_0 : P(\tau_k = 0) = P(\tau_k = 1) = 1/2$$
$$H_\mathrm{a} : P(\tau_k = 0) < P(\tau_k = 1)$$

En la tabla 6 se muestran los tres primeros y los tres últimos datos de las secuencias binarias obtenidas. Las sucesiones dicotómicas están ordenadas de acuerdo con la batimetría (m), teniendo en cuenta que a mayor profundidad menor nivel de oxígeno.

Tabla 6: Esquema de las sucesiones dicotómicas.

Profundidad	OSB (η_{in})	OFB (η_{jn})	Diferencia (τ_{ijn})
0.25	0	0	0
0.40	1	0	1
0.50	0	0	0
⋮	⋮	⋮	⋮
2.00	0	0	0
2.10	1	0	1
2.50	1	0	1

Obtenidas con los siguientes criterios:
Oxígeno superficial (1: mayor de 7.5 mg/l; 0: menor de 7.5 mg/l)
Oxígeno en el fondo (1: menor de 3 mg/l; 0: mayor de 3 mg/l)
Las secuencias están ordenadas según la batimetría.
Los datos originales de oxígeno se tomaron en 114 sitios de
la Ciénaga Grande de Santa Marta en marzo de 1997.
OSB: oxígeno en superficie convertido en binario
OFB: oxígeno en fondo convertido en binario

En 70 de los 114 sitios se obtuvo un 1 para la sucesión τ_n, es decir que en un 61 % de los sitios de la muestra hay un nivel de riesgo por encima de 5. De los 70 sitios con valores uno en la secuencia dicotómica τ_n, solo dos corresponden a sitios con oxígeno de fondo menor de 3 mg/l. Los restantes unos de dicha sucesión se deben a niveles de oxígeno mayores de 7.5 mg/l en superficie. Esto indica de manera descriptiva que el nivel de oxígeno en el sistema estaba el día de la toma de la muestra en condiciones no favorables para la vida de organismos aerobios, específicamente porque los niveles de oxígeno eran mayores de 7.5 mg/l (en 37 de los 114 sitios hubo valores mayores de 7.5 mg/l tanto en superficie como en fondo).

Usando el programa R del apéndice se obtuvo el valor de la estadística de prueba $GC(114)$ y una estimación del correspondiente valor crítico para dos colas con un α del 10 %. Estos fueron respectivamente 164836 y 152479. El valor de la estadística de prueba estandarizado es 2.4027, el cual es mayor que el valor crítico estimado para la prueba de dos colas 1.6843 (ligeramente mayor que el de la normal para el mismo nivel de significancia). Con el mismo programa se obtuvo un error del 0.00717 en la aproximación por la normal. Por lo tanto se rechaza con un nivel de significancia del 10 % la hipótesis de que el nivel de oxígeno del sistema se encontraba en un nivel de riesgo normal. Desde un punto de vista práctico puede concluirse que el 8 de marzo de 1997 los organismos aerobios de la CGSM estaban expuestos a niveles de oxígeno de riesgo moderado o alto (de acuerdo con la función de riesgo dada en la figura 4) especialmente en los sitios de mayor profundidad, puesto que el valor muestral de la estadística está a la derecha de la media de la distribución (123509). El conjunto total de datos y el código R para el análisis de los datos puede obtenerse en la página http://www.docentes.unal.edu.co/rgiraldoh/docs/.

6. Conclusión

La estadística propuesta para probar la hipótesis de similitud de dos secuencias binarias ordenadas es insesgada y consistente y tiene según los resultados de simulación la misma potencia de la prueba basada en la estadística del signo. Su ventaja radica en las posibilidades de interpretación en los casos en los que se rechaza la hipótesis de interés. Con la estadística propuesta es posible establecer si las diferencias tienden a estar al comienzo o al final de la sucesión. El análisis de datos realizado muestra la utilidad práctica de la prueba planteada.

[Recibido: abril de 2009 — Aceptado: mayo de 2010]

Referencias

Conover, W. (1999), *Practical Nonparametric Statistics*, John Wiley & Sons, New York.

Corzo, J. (1990), 'Teoría de rachas', *Revista Colombiana de Estadística* **19-20**, 80–93.

Dudewicz, E. & Mishra, N. (1988), *Modern Mathematical Statistics*, John Wiley & Sons, New York.

Giraldo, R. (2003), Construcción de un indicador para el estudio conjunto de la distribución espacial de múltiples variables binarias, Tesis de maestría, Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia.

Giraldo, R., Méndez, N. & Troncoso, W. (2000), 'Geoestadística: una herramienta para la modelación en estuarios', *Revista Académica Colombiana Ciencia* **24**(90), 57–92.

Hollander, T. & Wolfe, D. (1999), *Nonparametric Statistical Methods*, John Wiley & Sons, New York.

Mancera, E., Giraldo, R. & Salazar, J. (1996), IRMA: indicador de riesgo de mortandad de aerobios. Instituto de Investigaciones Marinas (INVEMAR).

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Randles, R. & Wolfe, D. (1979), *Introduction to the Theory of Non-parametric Statistics*, John Wiley & Sons, New York.

Apéndice

Valores críticos de la estadística $GC(n)$

Valores críticos $GC(n)$, de una cola superior, para muestras dicotómicas de tamaños 3 a 15. Se reportan los valores de $GC(n)$ cuyo nivel de significancia exacto, α^*, es más cercano al teórico por defecto (línea superior) o por exceso (línea inferior). El asterisco significa que el estadístico $GC(n)$ no tiene valores con probabilidad inferior al nivel de significancia dado.

	α			
Tamaño de la sucesión	**0.05**	**0.025**	**0.01**	**0.005**
3	*	*	*	*
	7 (0.125)	7 (0.125)	7 (0.125)	7 (0.125)
4	*	*	*	*
	14 (0.0625)	14 (0.0625)	14 (0.0625)	14 (0.0625)
5	25 (0.0313)	*	*	*
	24 (0.0625)	25 (0.0313)	14 (0.0625)	14 (0.0625)
6	39 (0.0469)	41 (0.0156)	*	*
	38 (0.0625)	40 (0.0313)	41 (0.0156)	41 (0.0156)
7	58 (0.0469)	61 (0.0234)	63 (0.0078)	*
	57 (0.0547)	60 (0.0313)	62 (0.0156)	63 (0.0078)
8	82 (0.0430)	87 (0.0234)	91 (0.0078)	92 (0.0039)
	81 (0.0508)	86 (0.0273)	90 (0.0117)	91 (0.0078)
9	114 (0.0430)	118 (0.0234)	125 (0.0098)	128 (0.0039)
	113 (0.0508)	117 (0.0273)	124 (0.0117)	127 (0.0059)
10	152 (0.0488)	159 (0.0025)	166 (0.0098)	171 (0.0049)
	151 (0.0508)	158 (0.0264)	135 (0.0107)	170 (0.0059)
11	193 (0.0479)	208 (0.0249)	216 (0.0088)	222 (0.0049)
	192 (0.0527)	207 (0.0269)	215 (0.0103)	221 (0.0054)
12	250 (0.0498)	258 (0.0249)	277 (0.0093)	282 (0.0046)
	249 (0.0535)	257 (0.0269)	276 (0.0105)	281 (0.0054)
13	303 (0.0494)	328 (0.0233)	347 (0.0098)	355 (0.0048)
	302 (0.0508)	327 (0.0254)	346 (0.0101)	354 (0.0054)
14	378 (0.0481)	407 (0.0246)	420 (0.0097)	439 (0.0049)
	377 (0.0511)	406 (0.0255)	419 (0.0106)	438 (0.0052)
15	468 (0.0473)	481 (0.0247)	516 (0.0095)	524 (0.0049)
	467 (0.0507)	480 (0.0262)	515 (0.0103)	523 (0.0053)

Los valores entre paréntesis son los valores de significancia exactos, es decir $\alpha* = P(GC(n) \geq x)$, con x un valor de $GC(n)$.

Valores críticos de la estadística de antirrachas

Valores críticos de la estadística de antirrachas, de una cola superior, para muestras dicotómicas de tamaños 3 a 15. Se reportan los valores de $-C$ cuyo nivel de significancia exacto, α^*, es más cercano al teórico por defecto (línea superior) o por exceso (línea inferior). El asterisco significa que el estadístico $-C$ no tiene valores con probabilidad inferior al nivel de significancia dado.

	α			
Tamaño de la sucesión	**0.005**	**0.01**	**0.025**	**0.05**
3	* 2.00 (0.125)	* 2.00 (0.125)	* 2.00 (0.125)	* 2.00 (0.125)
4	* 2.50 (0.0625)	* 2.5 (0.0625)	* 2.5 (0.0625)	* 2.5 (0.0625)
5	* 3.00 (0.0313)	* 3.00 (0.0313)	* 3.00 (0.0313)	3.00 (0.0313) 2.25 (0.0625)
6	* 3.50 (0.0156)	* 3.50 (0.0156)	3.50 (0.0156) 2.80 (0.0313)	2.80 (0.0313) 2.50 (0.0938)
7	* 4.00 (0.0078)	4.00 (0.0078) 3.33 (0.0156)	3.33 (0.0156) 3.00 (0.0547)	3.33 (0.0156) 3.00 (0.0547)
8	4.50 (0.0039) 3.86 (0.0078)	3.86 (0.0078) 3.50 (0.0313)	3.86 (0.0078) 3.50 (0.0313	3.00 (0.0430) 2.80 (0.0625)
9	4.37 (0.0039) 4.00 (0.0195)	4.37 (0.0039) 4.00 (0.0195)	3.43 (0.0234) 3.37 (0.0254)	3.12 (0.0391) 3.00 (0.0703)
10	4.56 (0.0029) 4.50 (0.0107)	4.56 (0.0029) 4.50 (0.0107)	3.60 (0.0225) 3.50 (0.0439)	3.50 (0.0439) 3.43 (0.0508)
11	5.10 (0.0015) 5.00 (0.0059)	4.50 (0.0073) 4.31 (0.0117)	4.11 (0.0127) 4.00 (0.0308)	3.37 (0.0449) 3.36 (0.0571)
12	5.00 (0.0042) 4.89 (0.0063)	4.55 (0.0095) 4.50 (0.0183)	4.10 (0.0220) 4.00 (0.0269)	3.56 (0.0398) 3.50 (0.0562)
13	5.17 (0.0039) 5.10 (0.0051)	5.10 (0.0051) 5.00 (0.0107)	4.09 (0.0248) 4.00 (0.0408)	3.57 (0.0480) 3.56 (0.0514)
14	5.58 (0.0028) 5.50 (0.0062)	5.00 (0.0095) 4.89 (0.0105)	4.55 (0.0184) 4.50 (0.0259)	3.82 (0.0494) 3.80 (0.0502)
15	5.54	5.17 (0.0088) 5.10 (0.0105)	4.46	4.09 (0.0378) 4.00 (0.0504)

Los valores entre paréntesis son los valores de significancia exactos, es decir $\alpha* = P(-C \geq x)$, con x un valor de $-C$.

Código R para cálculo de valores críticos de la estadística $GC(n)$

En esta sección se muestra el código R usado en la aplicación y en la estimación de potencias de la sección 2. El código puede obtenerse en http://www.docentes.unal.edu.co/rgiraldoh/docs/.

```
################################################################
# 1. Programa general. Este usa las funciones GC, GCCritico y
#    decisión definidas de 2 a 4. Para compilar este programa
#    con sus datos cambie oxigeno.txt por su archivo de datos
#    el cual debe contener una columna con la secuencia dico-
#    tómica de diferencias nombrada como suc.
################################################################

rm(list=ls())
source("GC.R")
source("GCCritico.R")
source("Decision.R")
```

```
suc<-read.table("oxigeno.txt", head=T)
attach(suc) suc<-suc$suc
Estadistica<-GC(suc)
Valor.Critico<-GCCritico(length(suc))
Decision(Estadistica,Valor.Critico)

#########################################################
2. Función para el cálculo del indicador
#########################################################
GC<-function(suc)
{
suc<-as.vector(suc)
tau.punto<-sum(suc)
n<-length(suc)

############################
# 2.1. Cálculo de K
############################
delta<-rep(0,n)
for (i in 1:n)
     {
     if(suc[i]==1) delta[i]<-i else (delta[i]<-tau.punto-n)
     }
K<-sum(delta)
##############################
# 2.2. Cálculo de h(n, tau)
##############################
a<-0
b<--1
am<-0
bm<--1
for (i in 1:(n-2))
     {
     a<-a+((i^2)+i)
     am<-rbind(am,a)
     b<-b+(i-1)
     bm<-rbind(bm,b)
     }
ab<-as.matrix(cbind(am,bm))
m<-tau.punto-2
if(tau.punto==0) h<--2*n
if(tau.punto==1) h<--0
if(tau.punto>1 ) h<-ab[m+1,1]-(ab[m+1,2]*n)
##############################
# 2.3 Valor de la estadística
##############################
GC<- tau.punto+K+(n*(n-2))-h
```

```
return(list(tau.punto=tau.punto, K=K,h=h, GC=GC))
}

################################################
# 3. Función para el cálculo del valor crítico
################################################
GCCritico<-function(n)
{
 critico<-NULL
 for(i in 1:10000)
    {
    suc<-rbinom(n,1,0.5) # Cambiar 0.5 por 0.6,...,.99
                         # para calcular potencia
    GCal<-GC(suc)
    critico<-rbind(critico,GCal$GC)
    }
 valor.critico<-quantile(critico,0.95)
 return(valor.critico)
}

#########################
# 4. Función de decisión
#########################

Decision<-function(Estadistica, Valor.Critico)
{
 rechazo<-paste("SE", "RECHAZA", "H0")
 no.rechazo<-paste("NO","HAY", "EVIDENCIA",
                   "PARA", "RECHAZAR", "H0")
 decision<-ifelse (Estadistica$GC>Valor.Critico,
                   rechazo, no.rechazo)
 return(list(Estadistica=Estadistica$GC,ValorCritico=Valor.Critico,
            Decision=decision))
}
```