Revista Colombiana de Estadística
Nº 8 - 1983.

PRUEBAS DE HIPOTESIS SOBRE PARAMETROS DE LOCALIZACION Y DISPERSION BASADAS EN SECUENCIAS

Jorge Ortiz P.

Profesor Asistente
Universidad Nacional

1.1. Introducción. Las pruebas basadas en secuencias (llamadas comunmente rachas) son ampliamente conocidas en estadística no paramétrica. Su mayor utilidad se encuentra en estudios de aleatoriedad de series de datos, aunque pueden encontrarse aplicaciones mas específicas en problemas de comparación de medidas de localización y de dispersión.

Las dos estadísticas mas conocidas a través de la literatura disponible son el número de secuencias R y la longitud de la secuencia más

grande L_{max} . Ellas permiten efectuar pruebas de hipótesis muy generales para las cuales son muy escasas las familias de pruebas disponibles. Sin embargo al restringir las hipótesis aparecen pruebas mas específicas que van desplazando por sus mejores características a las arriba mencionadas. En este sentido se puede decir que la noción de secuencia no ha sido suficientemente aprovechada. El propósito de este artículo es el de proponer una metodología que permita utilizar la noción de secuencia en campos mas específicos.

Se tratará el caso particular de dos muestras independientes; sin embargo los conceptos son de carácter mas general y pueden aplicarse a casos de más de dos muestras.

2.1. La noción de agrupamiento. Esta noción es intuitiva y corresponde a la concentración de datos de un mismo tipo en algún lugar observable. Las hipótesis estadísticas en su expresión más general están íntimamente relacionadas con esta noción y pueden expresarse de la siguiente manera:

H^0: No existen agrupamientos sistemáticos en los valores de las variables a observar. (Hipótesis Nula).

H^1: Existen agrupamientos sistemáticos en los valores de las variables a observar. (Hipótesis Alternativa).

Los patrones de agrupamiento que se quieran estudiar determinan el tipo específico de hipótesis asociado a una aplicación particular:

A) ***LOCALIZACION.*** Es obvio que si los datos de una muestra están sistemáticamente agrupados en uno de los extremos y si las muestras reflejan situaciones poblacionales entonces la medida de localización de la población corrrespondiente estará muy probablemente en ese extremo. Por ejemplo, agrupamientos sistemáticos de las observaciones de la muestra X en el extremo izquierdo permiten concluir que muy probablemente la población X tiene medida de localización inferior a la de la población Y.

En general los agrupamientos sistemáticos en *uno* de los extremos de una serie de datos ordenados permiten sacar conclusiones con respecto a las medidas de localización.

B) *DISPERSION.* Para los estudios de comparación de dispersiones de dos muestras los agrupamientos que interesan son los que pueden presentarse simultáneamente en *los dos* extremos de la serie de datos ordenados. La muestra que presente agrupamientos mayores en *los dos* extremos revelará que la población correspondiente tiene mayor dispersión.

C) *ALEATORIEDAD.* Esta es una de las nociones más generales que utiliza la estadística. Aleatorio significa no sistemático. Esto implica que algún tipo de agrupamiento sistemático observado en cualquier lugar de una serie de datos es síntoma de no aleatoriedad. Se identifican así los estudios sobre localización y dispersión como casos particulares de los estudios sobre aleatoriedad. No son estos los únicos casos particulares: Si se considera que por sistemático se entiende algo que tiene un patrón de comportamiento reconocible entonces una vez que se especifica un modelo las diferencias (o residuos) de las observaciones con respecto a dicho modelo deberán tener un comportamiento no identificable (no sistemático) para que el modelo pueda ser considerado como no mejorable bajo las condiciones dadas.

3.1. Pruebas basadas en secuencias. El concepto de secuencia (serie de datos del mismo tipo precedidos y seguidos por datos de tipo diferente) es el que está más directamente relacionado con la noción de agrupamiento y resulta un tanto extraño que este hecho no haya sido trabajado con mayor asiduidad.

Se dan enseguida algunas pautas generales de utilización de las relaciones presentadas arriba.

Luego de la discusión del parágrafo 2 es evidente que las estadísticas R y L_{max} no son las más indicadas para estudios de localización o dispersión. Las dos consideran agrupamientos en cualquier lugar de la serie de datos observada; en consecuencia su mejor desempeño se encuentra en estudios de aleatoriedad para los cuales fueron precisamente diseñadas; pero ninguna es sensible específicamente a agrupamientos en uno de los extremos o en ambos, hecho que las hace poco recomendables para estudios de localización o dispersión.

La principal deficiencia de las pruebas construídas a partir de R y L_{max} para hipótesis sobre localización o dispersión consiste en el

no aprovechamiento de la información concerniente a la posición de los agrupamientos observados: el número de secuencias puede ser lo suficientemente pequeño para indicar no aleatoriedad pero no señala esto cual es el patrón específico detectado; este número pequeño puede deberse a agrupamientos de los datos de un tipo en los dos extremos (mostrando así diferencia de dispersiones) o en un solo extremo (mostrando diferencia en localización) o en cualquier otro lugar de la serie observada (mostrando un tipo de no aleatoriedad diferente). Igual comentario puede hacerse para L_{max}. Se corre por lo tanto el riesgo de hacer análisis completamente diferentes a los deseados cuando se utilizan estas estadísticas para pruebas en problemas de localización o dispersión.

Una posible solución consiste en:

- Ordenar las secuencias según algún criterio determinado. Este puede ser el orden natural creciente si las variables son numéricas u ordinales o el orden de aparición en un proceso si los datos son de naturaleza cualitativa.

- Asignar a cada longitud de secuencia una ponderación que corresponda a su posición dentro de la serie ordenada.

El tipo de ponderación depende directamente del parámetro en consideración como se verá en lo que sigue.

A) ***LOCALIZACION.*** Según lo discutido en el parágrafo 2.1.A, los agrupamientos que interesan son los que pueden aparecer en *uno solo* de los extremos; así cualquier tipo de ponderación que permita identificar cada uno de los extremos resulta adecuado. En particular las ponderaciones que asignen puntajes pequeños al extremo izquierdo y puntajes grandes al derecho convienen en el estudio de comparación de medidas de localización. Una expresión sencilla que refleja estas ideas es la siguiente:

$$T_L = \sum_{i=1}^{R} (-1)^{1-d_X^i} \, i \, L_i / (R-1) \ ,$$

donde la notación utilizada es

R = Número de secuencias de tipo X o de tipo Y observadas.

d_X^i = Función indicadora cuyo valor es 1 cuando la secuencia i está conformada por datos de tipo X y 0 cuando la secuencia i está conformada por datos de tipo Y.

i = Posición de la secuencia en la serie ordenada según un criterio determinado como se dijo antes.

L_i = Longitud de la secuencia i (número de elementos que la conforman).

Obsérvese que la longitud L_i está ponderada por $i/(R-1)$ esta ponderación asigna puntajes mayores a medida que las secuencias se encuentran localizadas mas a la derecha. El otro coeficiente $(-1)^{1-d_i}$ simplemente da signo positivo a las secuencias de tipo X y negativo a las de tipo Y de manera que cuando haya un dominio de agrupamientos grandes de los datos de tipo X en la parte derecha de la serie observada entonces T_L tomará valores positivos y en caso contrario T_L tomará valores negativos.

B) *DISPERSION.* Puesto que las diferencias en dispersión se manifiestan cuando los agrupamientos más grandes se presentan en *los dos* extremos de la serie de datos, la ponderación debe ser tal que los puntajes mayores sean asignados precisamente a las secuencias más alejadas del centro. Una estadística apropiada sería entonces:

$$T_D = \sum_{i=1}^{R} (-1)^{1-d_X^i} L_i (i - \frac{R+1}{2})^2 / (R-1)^2$$

la notación es la misma que se utilizó para localización.

4.1. Familias de Pruebas. Las dos estadísticas presentadas son casos particulares de familias más generales cuyas formas son:

$$T_L = \sum_{i=1}^{R} (-1)^{1-d_X^i} Q(i, L_i) \quad \text{para localización}$$

donde $Q(i, L_i)$ es una función de parámetros i y L_i creciente en cada uno de ellos.

$$T_D = \sum_{i=1}^{R} (-1)^{1-d_X^i} W(i, L_i) \quad \text{para dispersión}$$

donde $W(i, L_i)$ es una función de parámetros i y L_i creciente en L_i y en el valor absoluto de $(i - \frac{R+1}{2})$, señalando esto último que a medida que la secuencia se aleja más del centro su ponderación será mayor.

5.1. Resultados. Se expondrán aquí algunos resultados obtenidos para la estadística T_L del parágrafo 3.1.A. La Tabla 1 muestra los valores de los cuantiles más próximos a los usuales.

La Tabla 2 muestra la potencia de la prueba para algunos valores de $n_X = n_Y$. Aunque, debido a las diferencias en niveles de significación con respecto a los de la estadística de Wilcoxon, no es posible establecer una comparación adecuada se puede observar que en términos generales el comportamiento de las dos estadísticas es muy similar en el caso de la distribución normal. En otros casos no ha sido aún estudiada. Sin embargo la partición inducida por T_L en el espacio de los rangos tiene un mayor número de clases que la inducida por la estadística de Wilcoxon lo cual hace esperar que el comportamiento difiera en alguna distribución. Para el cálculo de la potencia se utilizaron las tablas de Milton (1970) donde se puede encontrar también algunos cálculos de potencia de la prueba de Wilcoxon y de la prueba de la mediana. Es de notar que cuando los niveles de significación coinciden las potencias de las tres pruebas son iguales; las ventajas se encuentran entonces en la mayor disponibilidad de niveles de significa

ción; esta se tiene precisamente con la estadística T_L, como lo muestra la Tabla 3.

*

T A B L A 1

Valores críticos de T_L .

$n_X=n_Y$ \ w	$w.01$	$w.025$	$w.05$	$w.10$
3			-3.00(.0500)	-1.67(.1000)
4	-4.00(.0143)	-2.67(.0286)	-2.00(.0571)	-1.60(.0714) -1.33(.1143)
5	-3.67(.0079) -3.00(.0159)	-2.60(.0198) -2.33(.0317)	-2.20(.0397) -1.80(.0595)	-1.50(.0952) -1.40(.1190)
6	-3.33(.0087) -3.20(.0108)	-2.50(.0238) -2.40(.0325)	-2.29(.0346) -2.00(.0671)	-1.56(.0963) -1.50(.1050)
7	-3.25(.0093) -3.00(.0163)	-2.71(.0210) -2.60(.0262)	-2.17(.0975) -2.14(.0545)	-1.63(.0945) -1.57(.1066)
8	-3.33(.0096) -3.20(.0117)	-2.75(.0231) -2.67(.0264)	-2.25(.0483) -2.22(.0517)	-1.71(.0979) -1.67(.1033)

T A B L A 2

Potencia de la prueba basada en T_L

(Prueba unilateral contra alternativas de diferencia d en localización)

$n_X=n_Y$ \ d	.0	.2	.4	.6	.8	1.0	1.5	2.0	3.0
3	.050 .100	.075 .144	.108 .198	.150 .263	.201 .337	.260 .418	.437 .624	.624 .797	.890 .967
4	.014 .029 .057	.025 .048 .092	.041 .077 .140	.065 .117 .202	.098 .170 .278	.141 .235 .365	.293 .441 .602	.489 .660 .803	.830 .932 .977
5	.0079 .0159 .0198 .0317 .0397 .0595	.016 .030 .051 .071 .084 .115	.029 .054 .065 .096 .116 .151	.050 .089 .106 .152 .179 .236	.083 .140 .162 .225 .260 .328	.128 .206 .235 .314 .356 .432	.301 .431 .470 .574 .620 .689	.530 .676 .709 .799 .831 .871	.889 .952 .960 .982 .986 .990
6	.0206	.041	.076	.130	.204	.299	.585	.828	.991
7	.009 .019 .048 .083	.020 .041 .094 .149	.042 .079 .162 .244	.081 .141 .260 .365	.140 .228 .385 .500	.224 .339 .520 .634	.514 .656 .814 .886	.792 .887 .957 .982	.990 .997 .999 9999

T A B L A 3

Número de valores diferentes de la Estadística hasta el cuantil W_α

A) T_L

$n_X = n_Y$ \ α	100%	10%	5%	1%
3	14	2 (.10)	1 (.05)	
4	27	5 (.114)	3 (.057)	1 (.014)
5	60	12 (.119)	7 (.060)	2 (.008)
6	81	18 (.105)	13 (.067)	6 (.011)
7	162	36 (.107)	24 (.048)	12 (.009)
8	197	47 (.103)	38 (.052)	19 (.01)

B) *Wilcoxon*

α	100%	10%	5%	1%
3	10	2 (.10)	1 (.05)	
4	16	4 (.10)	3 (.057)	1 (.014)
5	26	7 (.111)	5 (.048)	2 (.008)
6	36	10 (.09)	8 (.048)	4 (.008)
7	51	15 (.104)	12 (.049)	7 (.009)
8	65	21 (.113)	17 (.052)	10 (.009)

BIBLIOGRAFIA

Milton Roy, C., *Rank Order Probabilities*, John Wiley, New York, 1970.

* *