Revista Colombiana de Estadística
Nº 8 - 1983.

UNA PRUEBA DE DISPERSION BASADA EN SECUENCIAS (*)

Jimmy Corzo S.
Instructor Asociado
Universidad Nacional

Jorge Ortiz P.
Profesor Asistente
Universidad Nacional

Este artículo contiene los resultados obtenidos en el estudio de las aplicaciones del concepto de secuencia(**) en problemas de comparación de dispersiones de dos muestras independientes. Como logro, se propone una estadística que en los casos estudiados presenta características muy interesantes con respecto a las prue-

---

(*) Este artículo es un *resumen* del trabajo de tesis de M.S. presentato por Jimmy Corzo bajo la direción de Jorge Ortiz.

(**) Una secuencia o racha es una sucesión de elementos de un mismo tipo seguidos o precedidos por elementos de otro tipo o por ningún elemento.

bas no paramétricas más conocidas para el mismo efecto como son la de Mood y la de Klotz. (Gibbons, 1971).

Introducción.

Uno de los problemas corrientes en el trabajo estadístico aplicado es el de comparar la dispersión de dos muestras independientes, con el objeto de establecer cual de ellas es más dispersa. Cuando se puede justificar adecuadamente el supuesto de normalidad de las poblaciones de donde provienen las muestras e incluso en algunos otros casos la prueba más adecuada es la conocida Y del cociente de las varianzas. Sin embargo, puede ocurrir que no se pueda justificar tal supuesto o que la información disponible no se encuentra en forma cuantitativa sino ordinal o nominal y en cualquiera de estos casos no debe usarse esta prueba, pues pierde sensibilidad en el primero y no puede calcularse el valor de la estadística en los otros dos.

Resulta más conveniente entonces, recurrir a los métodos no paramétricos pues éstos no requieren tan fuertes supuestos, como tampoco exigen la información en forma cuantitativa. Estos

métodos en su mayoría han sido construídos con base en el concepto de rango, como son por ejemplo las estadísticas de Mood, Klotz, etc. Surge así la inquietud de utilizar otro criterio para detectar dispersión, que en este caso es el de secuencia, para lo cual como se verá es necesario interpretar en términos de dispersión el significado de no aleatoriedad. Estas ideas junto con la presentación de la estadística propuesta y algunos de los resultados obtenidos en su estudio serán desarrolladas en lo que sigue.

El concepto de dispersión y la hipótesis de aleatoriedad.

La manera más común de interpretar el concepto de dispersión es la de fluctuaciones de las observaciones alrededor de su centro de masa o punto de localización del equilibrio; sin embargo, cuando el tipo de información disponible es por ejemplo de tipo nominal, esta manera de explicar tal término pierde sentido, pues en este caso no es posible medir las fluctuaciones alrededor del centro. or tanto parece razonable hablar de dispersión en términos relativos en el siguiente sentido: si los datos estan en

forma nominal ellos pueden encontrarse clasificados en solo dos categorías y el estudio de dispersión referirse a la observación de cual de las dos categorías tiene sus elementos más separados entre si con respecto a los de la otra. En este momento tiene lugar la hipótesis de aleatoriedad pues en forma general este término se refiere a la no presentación agrupada de elementos de cualquiera de las categorías en alguna parte de la serie ordenada de datos.

Para el caso de interés aquí, las dos categorías se refieren a las observaciones de dos muestras independientes que han sido combinadas en orden de magnitud (muestra combinada), encontrándose en la combinación, los elementos de cada muestra mezclados con los de la otra. Así por ejemplo para dos muestras de tamaño 5 una presentación cualquiera de los elementos de estas puede ser

A A B A B B A B A B

donde las letras A y B se refieren a cada una de las muestras y la correspondiente hipótesis de aleatoriedad establece que no se encuentre en la muestra combinada ningún agrupamiento sistemático de elementos de alguna de las muestras.

La hipótesis alterna de no aleatoriedad se refiere simplemente a la presentación sistemática de los elementos de alguna de las muestras en cualquier lugar de la muestra combinada, o de otra manera a la tendencia de los símbolos que representan las muestras a mostrar algún patrón definido de aparición. Sin embargo, si se tiene en cuenta el lugar de la muestra combinada donde se presenta el agrupamiento sistemático, surgen dos interpretaciones diferentes de la no aleatoriedad, una de las cuales concierne al concepto de dispersión(*). Así cuando los agrupamientos sistemáticos de elementos de la misma muestra se encuentran en los dos extremos de la muestra combinada esto constituye un indicativo de que los elementos de la primera muestra están más separados o más dispersos entre sí que los de la segunda como se observa en la siguiente configuración de dos muestras independientes de tamaño 5.

A A B B B A B B A A

en la que hay más elementos de la muestra B separando a los de la A que elementos de la mues-

---

(*) La otra interpretación se refiere a localización y se encuentra explicada en Ortiz, J. (Artículo anterior).

tra A separando a los de la B.

Actualmente entre los métodos no paramétricos, se encuentran algunas pruebas basadas en secuencias para detectar no aleatoriedad pero en cualquier lugar de la serie y no este tipo específico de no aleatoriedad que indica diferencias en dispersión. Las pruebas más conocidas para detectar dispersión la de Mood, la de Klotz, estan basadas en los rangos de las observaciones y en la información que reporta una de las muestras. Entonces se propone una estadística basada en el concepto de secuencia, que involucre información de las dos muestras en cuestión en espera de que ésta resulte más potente que las dos estadísticas basadas en rangos.

La Estadística propuesta.

Dada la información de dos muestras independientes del mismo tamaño N, la posición de las diferentes secuencias de los elementos de cada una de ellas en la muestra combinada de tamaño $2N$, será como ya se discutió arriba un indicador de diferencias en dispersión. También sus longitudes L_i serán otro indicador, pues si son *todas* las secuencias pequeñas, los elementos de la co-

rrespondiente muestra estarán repartidos en varias secuencias cortas, distribuídas posiblemente de manera uniforme en la muestra combinada, lo cual implica que los elementos de la otra muestra deben encontrarse distribuídos necesariamente de la misma manera, indicando esto similitud en las dispersiones de ambas muestras.

Finalmente el número de secuencias *NR* es otro indicador, pues cuando por ejemplo *NR* = 2 necesariamente todos los elementos de una de las muestras estarán primero, formando una secuencia y luego estarán los elementos de la otra muestra formando la otra secuencia, lo cual indicará que posiblemente las dos son igualmente dispersas. En cambio cuando *NR* = 3, la confifuración de la muestra combinada tiene que ser: una parte de los elementos de una de las muestras formando la primera secuencia, después todos los elementos de la otra muestra haciendo la segunda secuencia y finalmente el resto de los elementos de la primera muestra constituyendo la tercera secuencia; pero esta configuración indica que la muestra cuyos elementos estan en la primera y última secuencias, debe ser más dispersa que la otra pues sus elementos están separados por *todos* los de la otra. En la medida que *NR* comienza a aumentar las diferencias en dispersión

comenzarán a depender también de las longitudes de las rachas; por lo que resulta conveniente mezclar dos indicadores. Así por ejemplo, cuando $NR = 4$ y todas las secuencias son de la misma longitud (esto implica tamaños de muestra pares), tendremos mayores posibilidades de que las dos muestras sean igualmente dispersas mientras que si por ejemplo las dos primeras secuencias sean mucho más largas que las dos últimas (el caso extremo es la 1ª secuencia de longitud $N-1$, segunda secuencia de longitud $N-1$, tercera y cuarta de longitud 1) esto indicará que la muestra cuyos elementos forman la 1ª secuencia posiblemente estará más dispersa que la otra, pues sus elementos estarán separados por una secuencia muy larga de elementos de la otra muestra, no ocurriendo lo mismo con esta última cuyos elementos estan separados por una secuencia de pocos elementos de la primera muestra.

Todas estas ideas reunidas sugieren que los cuadrados de las distancias entre las posiciones de las secuencias y la posición mediana $(NR+1)/2$, ponderadas por sus correspondientes longitudes son un buen indicativo de las diferencias en dispersión y por esto se propuso la siguiente estadística de prueba

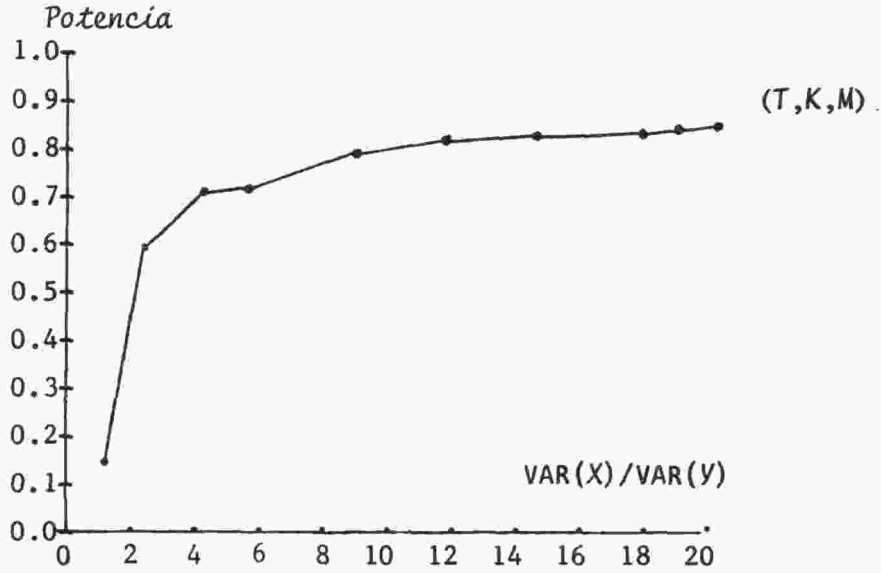
$$T = \frac{1}{(NR-1)^2} \sum_{i=1}^{NR} \left(i - \frac{NR+1}{2}\right)^2 L_i (-1)^{\delta_i^A}$$

donde la función δ_i^A toma el valor de 1 si la i-ésima secuencia es de elementos de la muestra A y es cero si es de elementos de la muestra B; ésta es una manera de dar signos a las diferentes distancias que garantiza que el valor de las estadísticas aumente o disminuya según que la correspondiente secuencia sea de la muestra B ó de la muestra A respectivamente. Por último el denominador $(NR-1)^2$ fué ajustado después de ensayar otros, al observar que con él, el valor de la estadística detectaba más cambios pequeños en la varianza de una de las muestras.

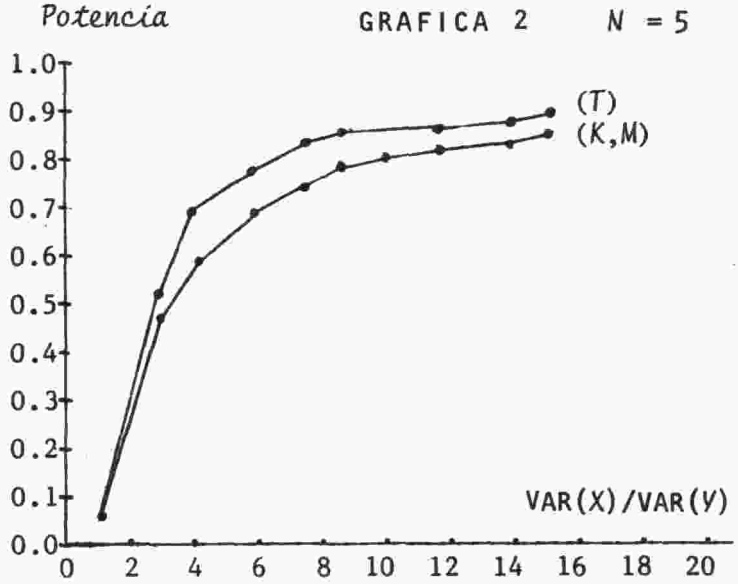
Estudio de potencia para tamaños de muestra pequeños de las dos muestras.

El estudio de potencia se hizo para tamaños de muestra entre 4 y 8. En los casos $N = 4$ y $N = 5$ la estadística tuvo el mismo comportamiento que el presentado por las pruebas propuestas (T) por Mood (M) y Klotz (K) en los otros tres casos rechazó en una proporción mayor la hipótesis nula falsa.

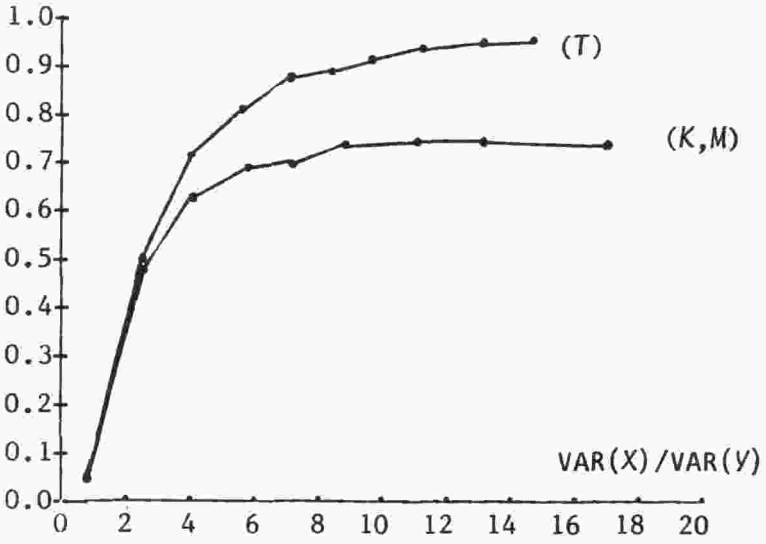
GRAFICA 1 $N = 4$



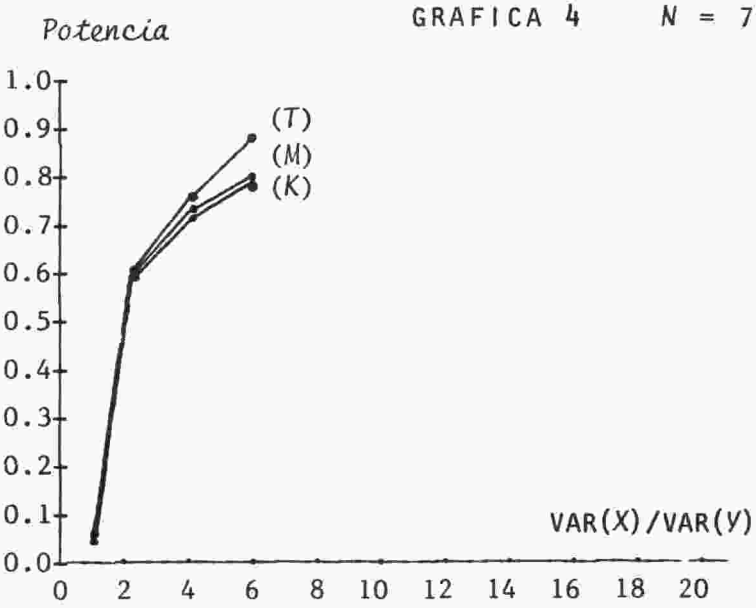
GRAFICA 2 $N = 5$



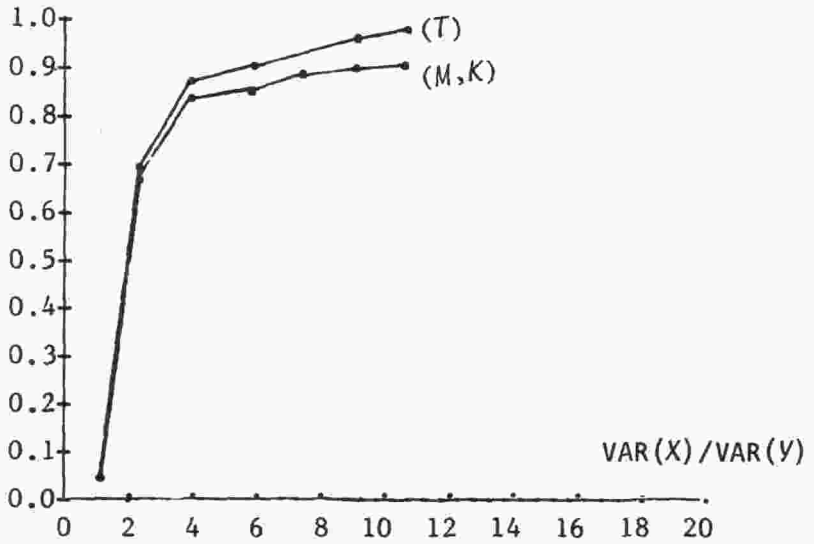
GRAFICA 3 N = 6



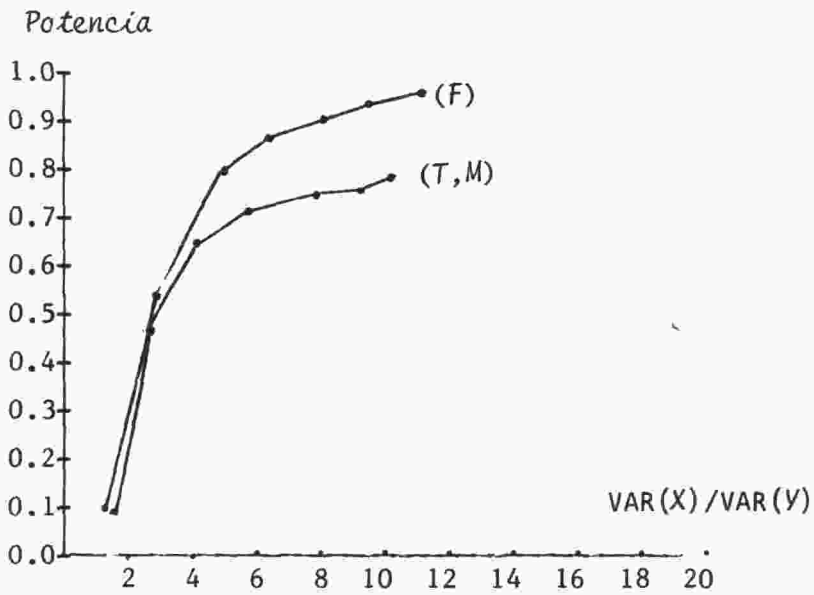
GRAFICA 4 N = 7



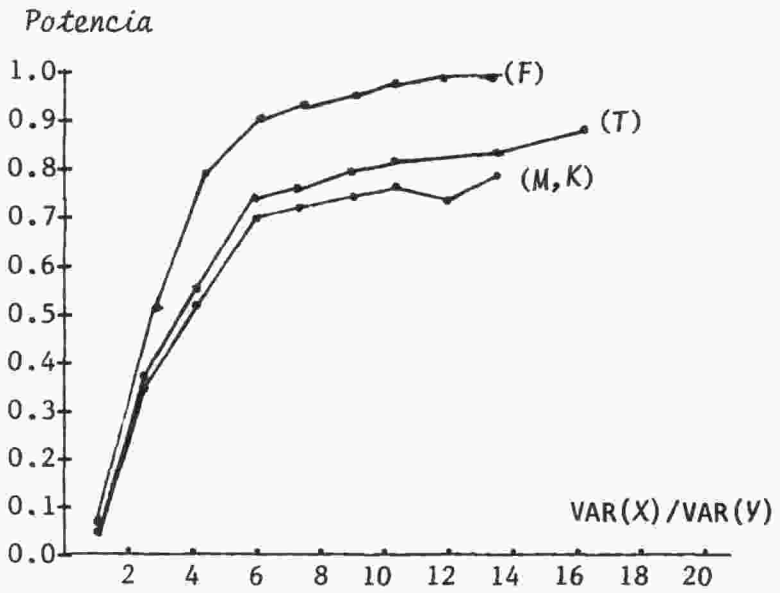
GRAFICA 5 $N = 8$



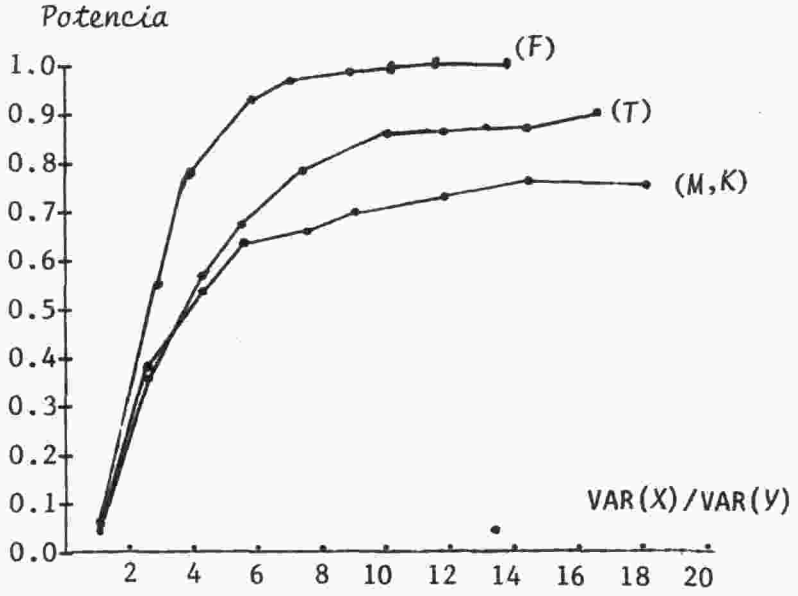
GRAFICA 6 $N = 4$



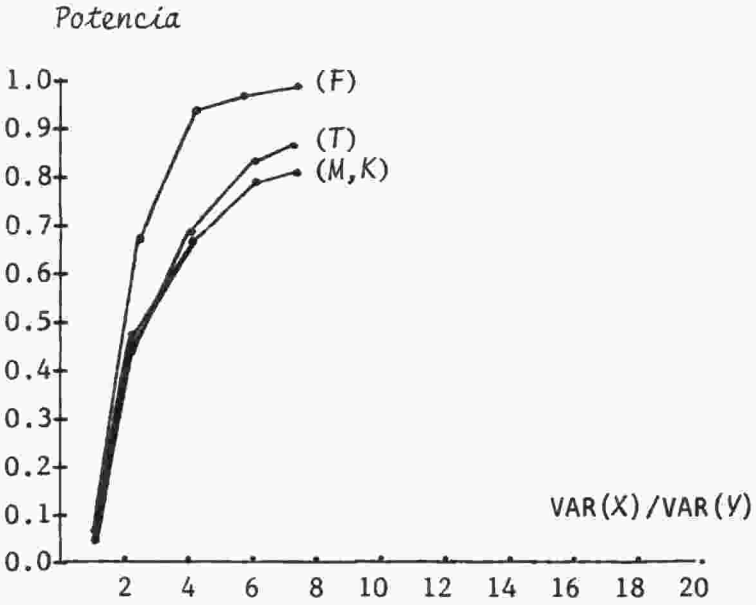
GRAFICA 7 $N = 5$



GRAFICA 8 $N = 6$



GRAFICA 9 $N = 9$



Como se observa en la Gráfica 1, las tres estadísticas tuvieron el mismo comportamiento para tamaño de muestra 4, mientras que las Gráficas 2, 3, 4 y 5 que corresponden a tamaños de muestra 5,6,7 y 8 se notan ya algunas ventajas cuando de poblaciones uniformes se trata. Ahora, para el caso de poblaciones normales se ha hecho ya la comparación con la prueba F, que como era de esperarse tiene grandes ventajas respecto a las otras estadísticas estudiadas, lo cual se observa en los Gráficos 6, 7, 8 y 9. Sin embar-

go la estadística propuesta se muestra en igualdad de condiciones respecto a las estadísticas de Mood y Klotz para tamaños de muestra 4 en tanto que lleva algunas ventajas para los tamaños de muestra 5, 6 y 7.

*

BIBLIOGRAFIA

- Gibbons, J., *Nonparametric statistical inference*. McGraw Hill, Tokio, 1971.
- Lehmann, E.L., *Nonparametrics: Statistical methods Based on Ranks*, Holden Day, San Francisco, 1975.
- Milton Roy, C., *Rank Order Probabilities*, John Eilwy, New York, 1970.

* *