Revista Colombiana de Estadística
Nº 14, Diciembre de 1986

ESTADISTICA DE PRUEBA PARA LOCALIZACION BASADA EN SECUENCIAS

Felipe Fernández H.
Asistente de Docencia
Universidad Nacional

Jorge Ortiz P.
Profesor Asociado
Universidad Nacional

Resumen. El presente trabajo contiene algunos de los principales resultados obtenidos de la aplicación de la noción de secuencias al problema de comparación de la localización de dos poblaciones con base en dos muestras independientes. Se propone modificar una estadística estudiada anteriormente, para explorar sus consecuencias, obteniéndose resultados que la hacen más favorable con respecto a propiedades tales como la convergencia, potencia y simetría.

1. Introducción.

Uno de los conceptos más estudiados en estadística es el de Localización. Un caso citado corrientemente en la literatura disponible es la comparación de los parámetros de localización de dos poblaciones de igual varianza distribuidas normalmente y varianza desconocida; si estos supuestos se satisfacen y se dispone de dos muestras aleatorias independientes, la estadística más indicada para realizar ésta comparación está dada por el cociente.

$$\frac{\bar{X}-\bar{Y}-\theta}{S_p(1/n+1/m)^{\frac{1}{2}}}$$

el cual sigue la distribución t-student no central con $n+m-2$ grados de libertad y parámetros de centralidad θ, donde $\bar{X}$, $\bar{Y}$, son las medias muestrales y S_p^2 la varianza ponderada. En muchas ocasiones se encuentra que los supuestos antes mencionados no se satisfacen y entonces surgen situaciones donde hay la necesidad de aplicar procedimientos diferentes. Como caso particular se encuentran los procedimientos basados en secuencias o rachas en la estadística no paramétrica donde pruebas tales como el número de secuencias o la longitud de la secuen-

cia más grande son las más conocidas. Por otra parte, pruebas no paramétricas tales como las de Wilcoxon y Van der Waerden aunque son aplicables en secuencias no fueran originalmente diseñadas con este propósito.

Específicamente en este campo, Ortiz (1983), propuso el estudio de la familia general de estadísticas de localización dadas por la expresión:

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

donde R es el número de secuencias, δ_x^i es la función indicadora del tipo de secuencia (de tipo X o Y) y $Q(i, L_i)$ es una función creciente en los parámetros i (posición de la secuencia) y L_i (longitud). Un caso considerado anteriormente es la estadística

$$T_o = \sum_{i=1}^{R} (-1)^{1-\delta_x^i} i L_i / (R-1)$$

la cual para muestras de igual tamaño mostró por ejemplo tener igual potencia que la estadística de Wilcoxon en los niveles de significación donde fue posible establecer comparación. Sin embargo la ampliación del estudio a muestras de diferente tamaño reveló algunas defi-

ciencias en lo que se refiere a su potencia, desventajas en cuanto a su convergencia a la distribución normal y falta de simetría.

En lo que sigue de este artículo se presenta el cambio introducido sobre la estadística y algunos de los principales resultados obtenidos.

2. Modificación Propuesta.

Teniendo en cuenta que sobre la estadística T_o no había estudios anteriores para muestras de tamaños desiguales, se procedió a calcular su distribución encontrándose algunas faltas de sensibilidad para captar diferencias extremas. Por ejemplo cuando $n=7$ y $m=3$, la estadística debería tomar sus valores máximo y mínimo en las configuraciones extremas, lo cual no se satisfizo. Para observar esto considérense las siguientes secuencias:

$$u_1 : X\ X\ X\ X\ X\ X\ X\ Y\ Y\ Y$$

$$u_2 : X\ X\ X\ X\ X\ X\ Y\ X\ Y\ Y$$

$$u_3 : X\ X\ X\ X\ X\ Y\ Y\ Y\ X\ X$$

denotando como $T_o(u_i)$ el valor que toma T_o

en la configuración u_i rápidamente se puede verificar que

$$T_o(u_1) = 1, \quad T_o(u_2) = -1/3 \quad y \quad T_o(u_3) = 5/2$$

y al ordenar estos valores tenemos que

$$T_o(u_3) > T_o(u_1) > T_o(u_3)$$

sin embargo obsérvese que u_1 es una configuración extrema. En realidad, el cálculo de la distribución de probabilidad de T_o para estos tamaños de muestras, indica que el cuantil $T_o(u_1)$ corresponde a un percentil del 9.17%, lo cual implica que T_o no detecta una diferencia máxima en localización como la que se presenta en la configuración u_1 en niveles de significación inferiores a este valor, lo cual necesariamente conlleva a un empobrecimiento en la potencia de la prueba. Por otra parte se puede verificar que el número de valores distintos que se presenta hasta el cuantil del 10%, es menor que el número que se presenta por encima del 90%, por lo cual se puede afirmar que T_o no es simétrica, propiedad que sería deseable para una mejor convergencia hacia la distribución normal.

Aspectos como los anteriormente mencionados, pueden ser atribuibles al desequilibrio que se

presenta en las ponderaciones asignadas por las longitudes L_i de las secuencias ante situaciones, donde los tamaños de las muestras son diferentes. Esta característica se hace más evidente en la medida en que el tamaño de una de las muestras es relativamente grande comparado con el de la otra. Por ejemplo en u_1 donde $L_1 = 7$ y $L_2 = 3$ el peso que asigna L_2 no es suficiente para contrarrestar el efecto de L_1. Como las ponderaciones deberían ser equilibradas independientemente del tamaño de las muestras, se consideró apropiado asignar los pesos de las longitudes en términos relativos de acuerdo con el tamaño de éstas. Por lo cual la estadística considerada fue la siguiente:

$$T_1 = \sum_{i=1}^{R} (-1)^{1-\delta_X^i} \frac{iL_i}{(R-1)n_i}$$

donde n_i toma el valor n cuando la secuencia i es de tipo X ó m cuando es de tipo Y. Debe observarse que si $n = m$ las estadísticas T_0 y T_1 son equivalentes en el sentido de que cualquiera de ellas puede obtenerse como combinación lineal de la otra. Por otra parte cuando la diferencia en el tamaño de las muestras va creciendo, la correlación entre T_0 y T_1 va disminuyendo.

3. Resultados.

En esta sección se presentan y comentan a continuación algunos de los resultados obtenidos.

3.1 Simetría.

En general puede demostrarse que para cualquier estadística de la familia T_L con función de parámetros $Q(i,L_i)$ de la forma $q_i * L_i/n_i$ donde $q_i + q_{R-i+1}$ es constante para todo $i = 1,2,...R$ es simétrica.(1) En el caso de T_1, obsérvese que $Q(i,L_i) = i * L_i/n_i$ y que

$$q_i + q_{R-i+1} = i + (R-i+1) = R+1$$

por lo cual se satisface la condición anterior.

3.2 Valores Extremos.

La estadística T_1 asume todos sus valores dentro del intervalo $[-1,1]$ tomando los valores 1 ó -1 en las configuraciones extremas, propiedad que la estadística T_0 algunas veces no la satisface como escomentado en la sección anterior.

---

(1) Este resultado se encuentra demostrado en Fernández (1988)

3.3 Estudio de Convergencia.

Las gráficas de la página siguiente muestran las diferencias máximas que ocurren para las distribuciones de probabilidad de T_0 y T_1 con respecto a la distribución normal para algunos tamaños de muestras.[2] Como puede observarse la convergencia de la estadísca T_1 mejora en comparación con el de la estadística anterior. Resultados similares se obtuvieron en otros tamaños de muestra estudiados.

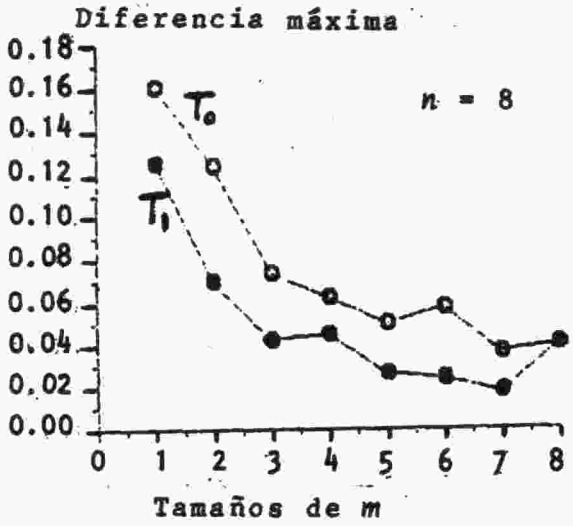
3.4 Estudio de Potencia en Poblaciones Normal, Exponencial y Uniforme para Muestras Pequeñas de Tamaños Diferentes.

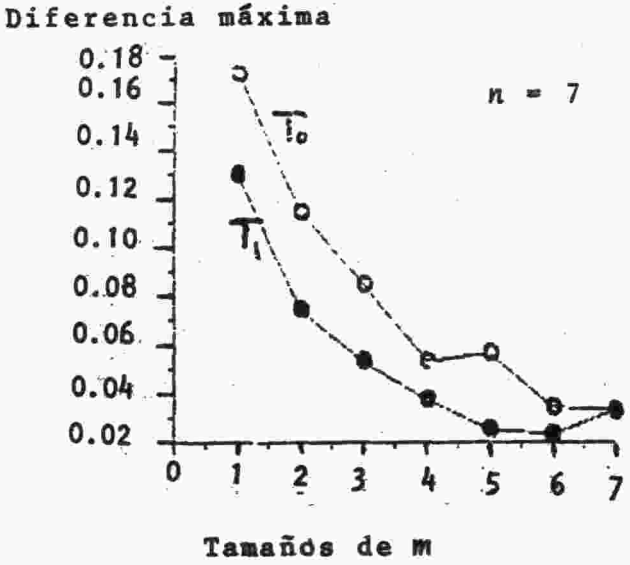
Poblacion Normal. Las gráficas 1 a 9 muestran las curvas de potencia de las estadísticas de Wulcoxon y T_1 calculadas en niveles de significación comparables, adicionalmente la tabla 1 presenta los resultados numéricos. Se observa en términos generales que la potencia de ambas estadísticas es igual, sin embargo T_1 induce un espectro de valores mucho más amplio que

(2) El estudio se realizó hasta n y m satisfaciendo la condición $n+m < 18$.

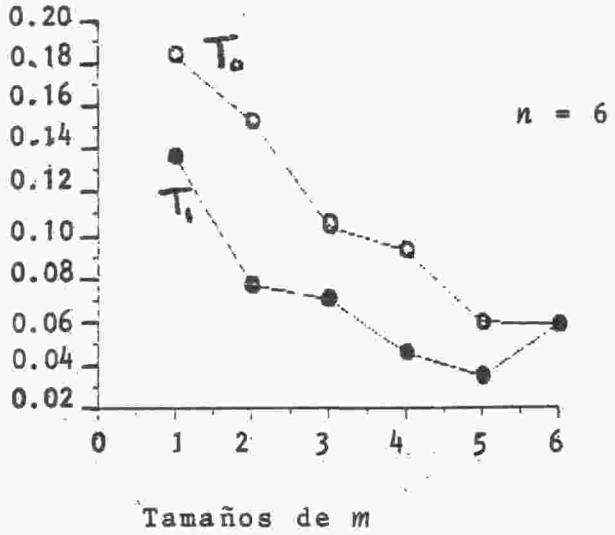
Estudio de convergencia

Máximo error de la aproximación normal de la función de distribución de T_o y T_1 para algunos tamaños de muestras.



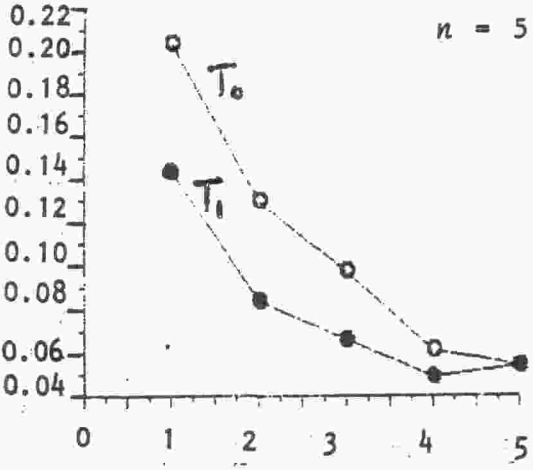


Diferencia máxima



Tamaños de m

Diferencia máxima



Tamaños de m

el de T_0 y por lo tanto con T_1 pueden realizarse pruevas de hipótesis sobre una mayor variedad de niveles de significación.

Para el cálculo de estas potencias se utilizaron las tablas de Milton (1970), con las cuales se obtiene exactitud en los resultados hasta el orden del octavo decimal.

Por otra parte debe mencionarse que la prueba t-student señalada en la introducción, es uniformemente más potente en estas poblaciones, razón por lo cual su potencia no puede ser superada.

Población Exponencial. Las gráficas 10 a 14 junto con la tabla 2 presenta los resultados de los tamaños de muestra estudiados. Como patrón general se observa un comportamiento similar en las estadísticas W y T_1 con un desempeño moderado mejor respecto a la t-student.

Los cálculos fueron obtenidos utilizando métodos de simulación para lo cual se generaron 8000 muestras para cada uno de los tamaños indicados, este tamaño garantiza que las estimaciones de las potencias cercanas a 0.5 tienen un margen de error de 0.012 mientras que las cercanas a 0.05 ó 0.95 tienen un error máximo de 0.002, ambas con un coeficiente de confianza del

95% (Burstein, 1971).

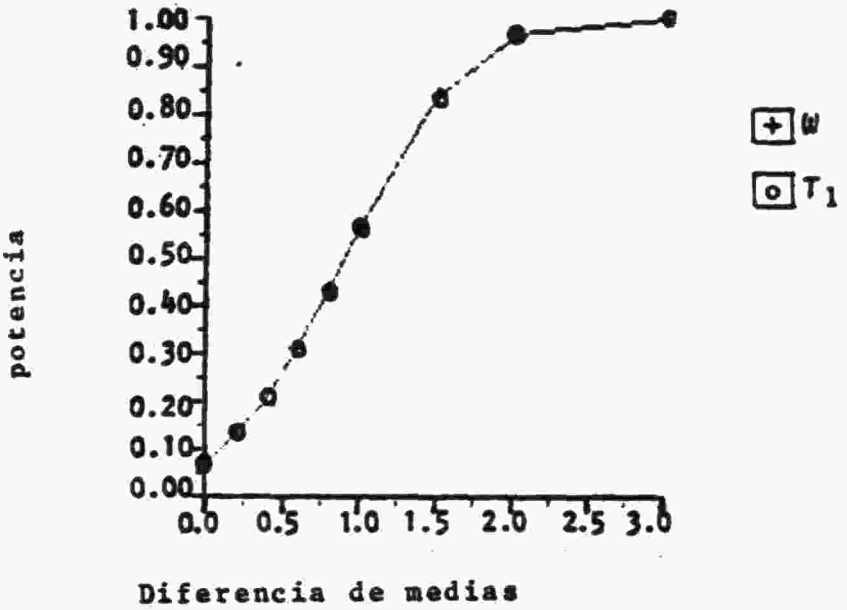
Población Uniforme. La tabla 3 junto con los gráficos 15 a 19, muestran que en general la potencia de la t-student es sensiblemente mejor que las de T_1 y W con excepción del resultado obtenido para n=7 y m=6 (gráfica 15) donde se observa un comportamiento favorable de estas dos últimas con respecto a la t-student. Como en el caso anterior, los resultados se obtuvieron utilizando métodos de simulación bajo las mismas condiciones de error.

TABLA 1

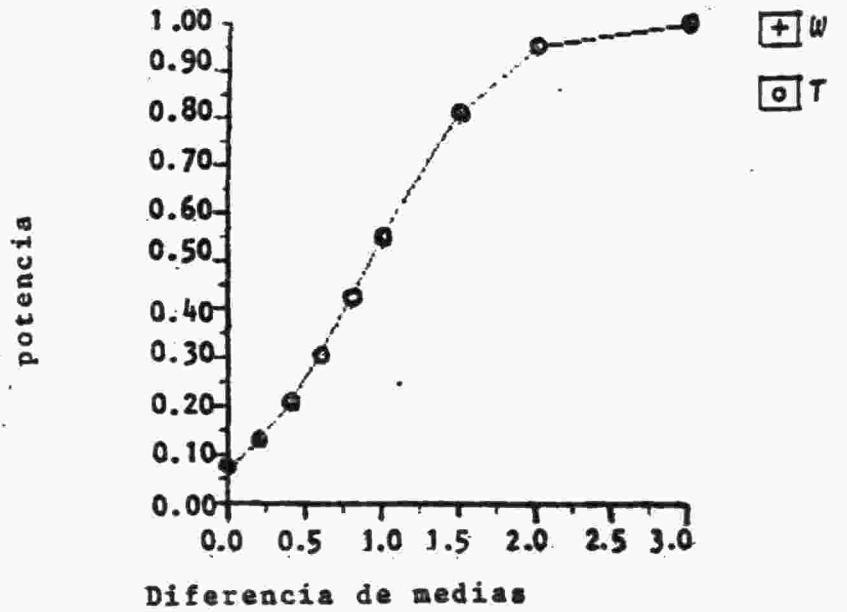
Comparación de algunas potencias exactas de W y T_1 en poblaciones normales. Prueba unilateral contra alternativas d en localización.

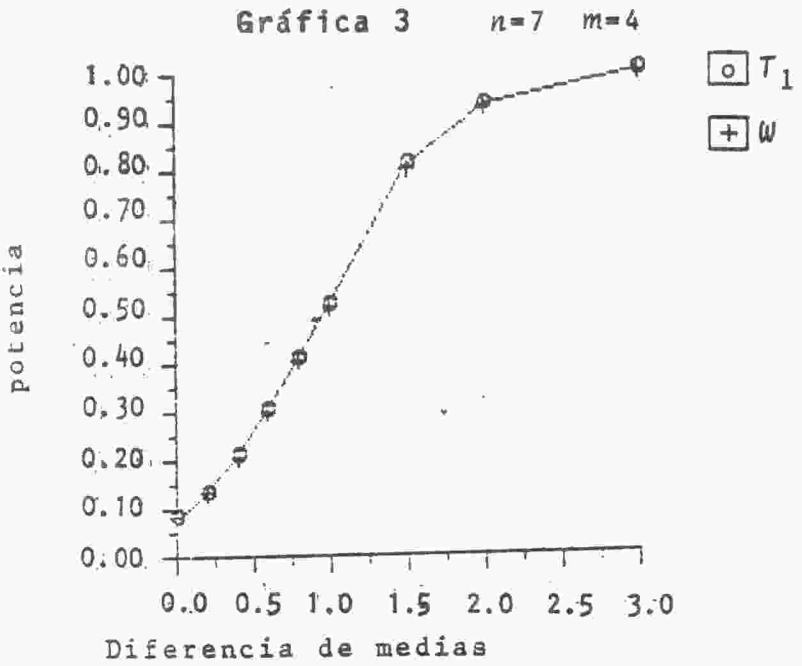
Prueba	Tamaños	d=0,0	d=0,2	d=0,4	d=0,6	d=0,8	d=1,0	d=1,5	d=2,0	d=3,0
W	n=7 m=6	0,069	0,133	0,205	0,311	0,435	0,565	0,837	0,969	1,000
T_1		0,069	0,132	0,204	0,308	0,430	0,559	0,830	0,965	0,999
W	n=7 m=5	0,074	0,129	0,207	0,307	0,424	0,547	0,814	0,951	0,999
T_1		0,074	0,129	0,205	0,303	0,418	0,539	0,805	0,946	0,999
W	n=7 m=4	0,082	0,136	0,209	0,303	0,410	0,524	0,810	0,931	0,998
T_1		0,082	0,135	0,208	0,300	0,405	0,518	0,802	0,925	0,997
W	n=7 m=3	0,058	0,095	0,146	0,213	0,294	0,387	0,633	0,831	0,984
T_1		0,058	0,095	0,147	0,214	0,296	0,390	0,637	0,834	0,985
W	n=6 m=5	0,041	0,075	0,126	0,196	0,287	0,393	0,533	0,877	0,995
T_1		0,041	0,074	0,124	0,193	0,281	0,384	0,515	0,859	0,990
W	n=6 m=4	0,057	0,097	0,153	0,227	0,318	0,421	0,684	0,874	0,993
T_1		0,057	0,096	0,152	0,225	0,315	0,416	0,677	0,868	0,991
W	n=6 m=3	0,048	0,078	0,121	0,178	0,249	0,332	0,567	0,746	0,972
T_1		0,048	0,078	0,121	0,178	0,249	0,332	0,567	0,746	0,972
W	n=5 m=4	0,056	0,092	0,144	0,212	0,295	0,391	0,643	0,842	0,987
T_1		0,056	0,092	0,144	0,212	0,295	0,391	0,643	0,842	0,987
W	n=5 m=3	0,056	0,092	0,144	0,212	0,295	0,391	0,643	0,842	0,987
T_1		0,056	0,092	0,144	0,212	0,295	0,391	0,643	0,842	0,987

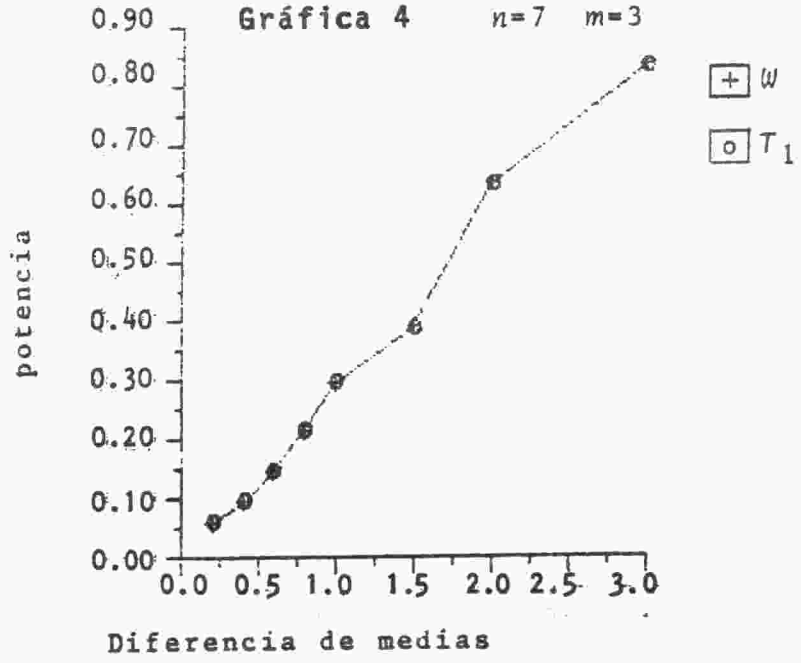
Gráfica 1 $n=7$ $n=6$



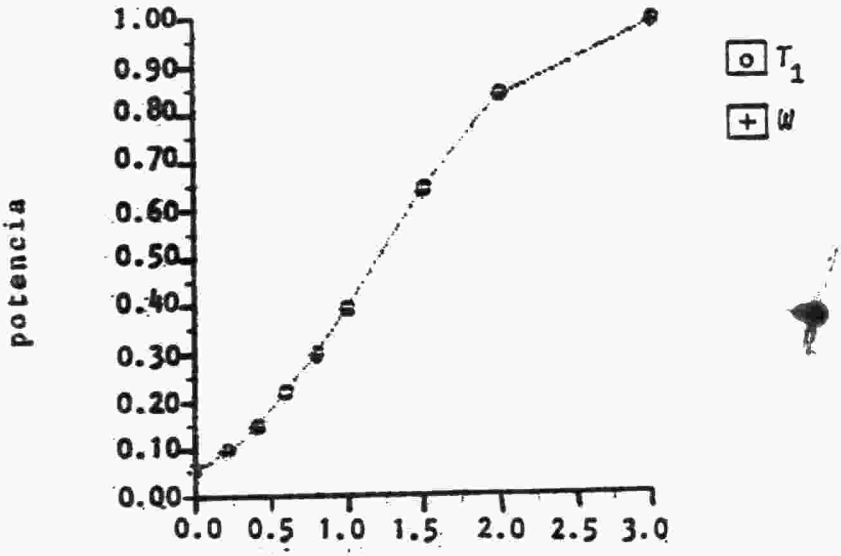
Grafica 2 $n=7$ $n=5$



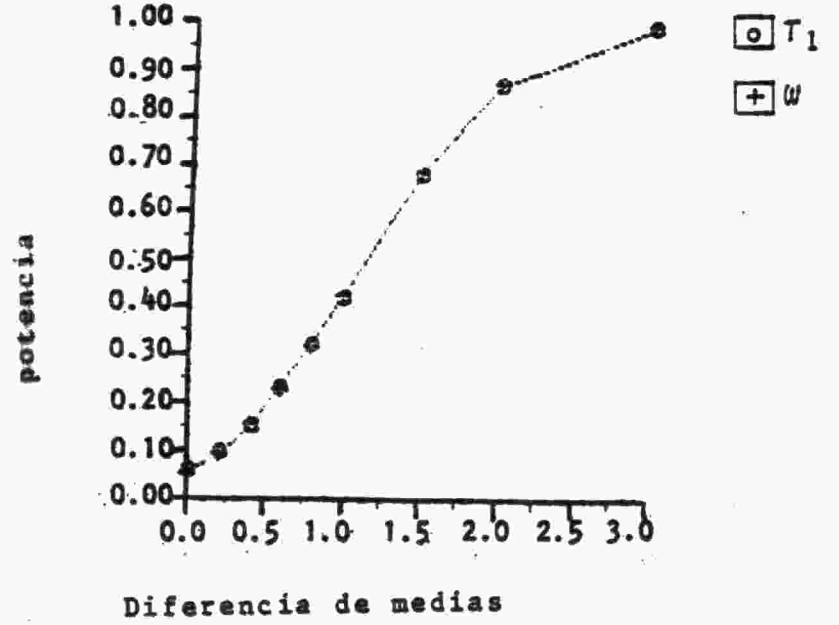

Gráfica 3 n=7 m=4
T1
W
potencia
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Diferencia de medias


Gráfica 4 n=7 m=3
W
T1
potencia
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Diferencia de medias

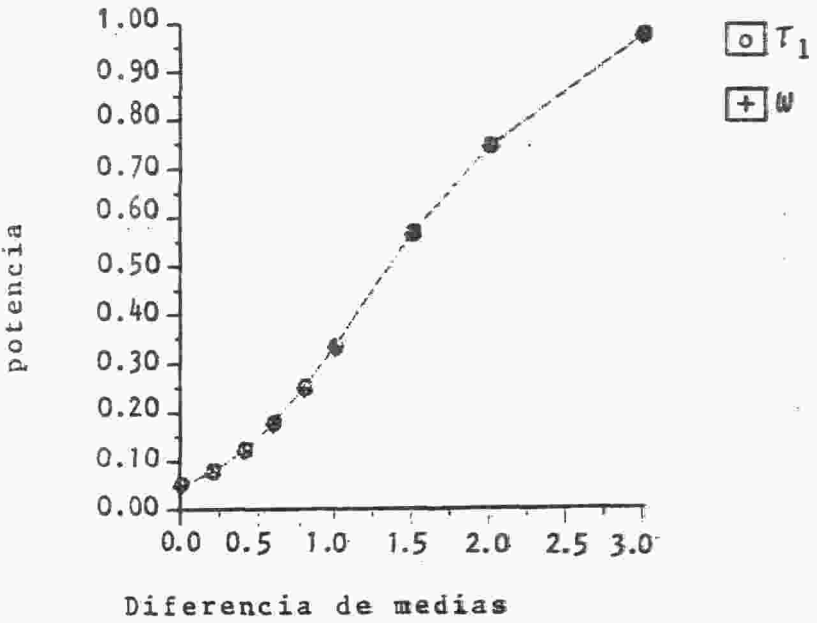
Gráfica 5 $n=6$ $m=5$



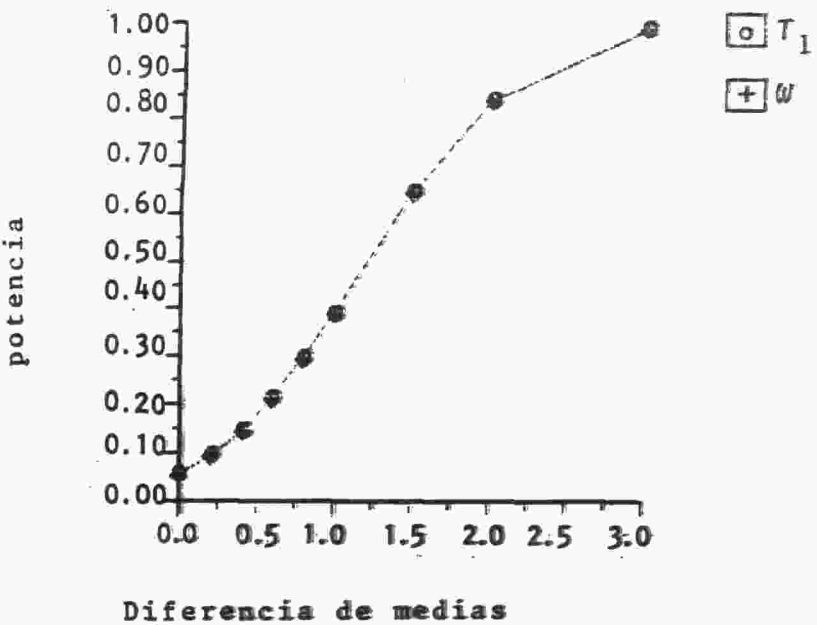
Gráfica 6 $n=6$ $m=4$



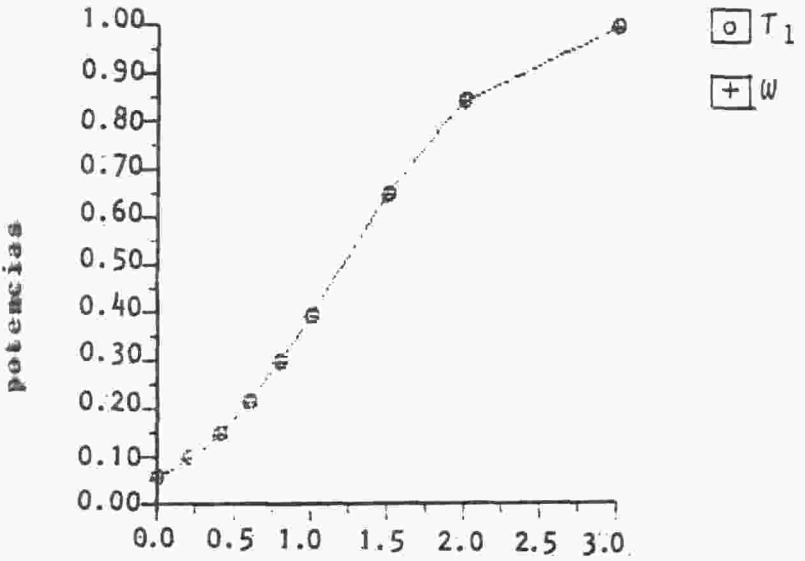
Gráfica 7 $n=6$ $m=3$



Gráfica 8 $n=5$ $m=4$



Gráfica 9 $n=5$ $m=3$



Gráfica 10 $n=7$ $m=6$ (Exp.)

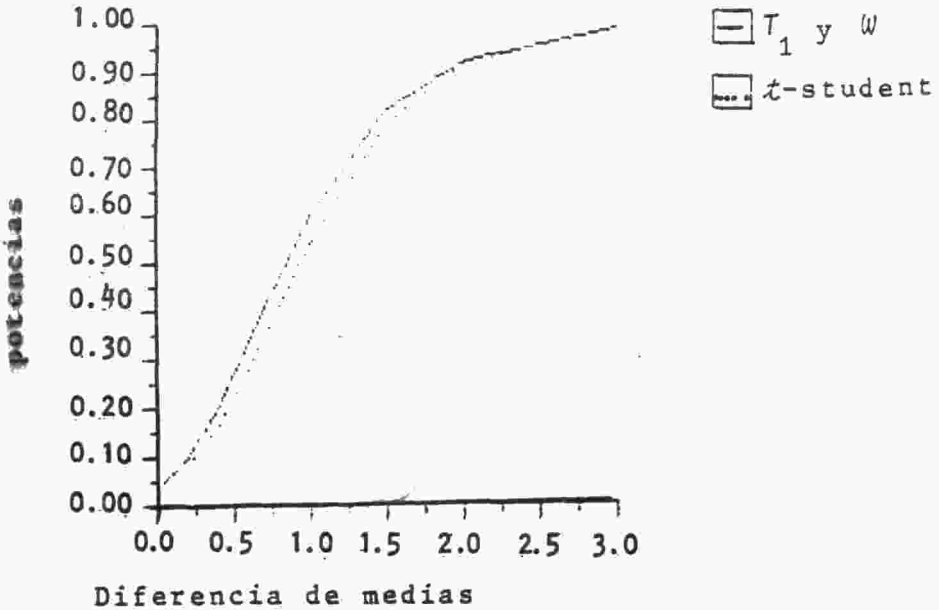


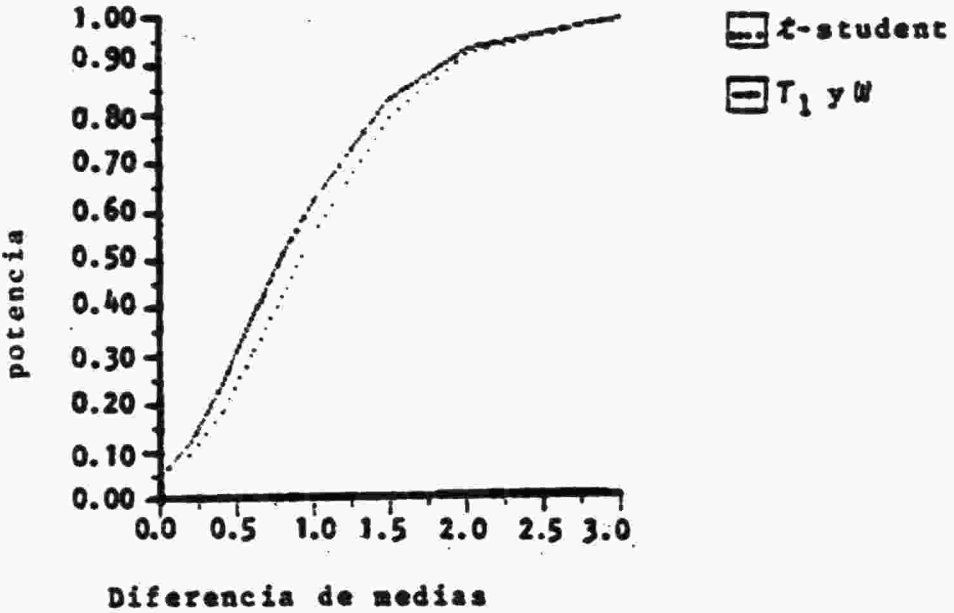
TABLA 2.

Comparación de algunas poblaciones simuladas W, T_1 y t-student en poblaciones exponenciales con parâmetro lamda = 1.

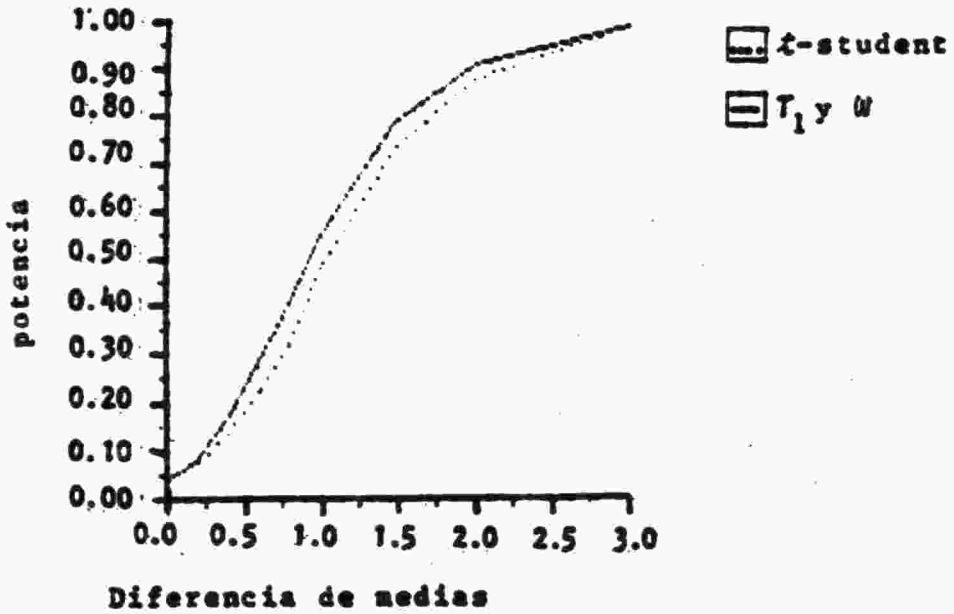
Prueba unilateral contra alternativas d en localización.

Prueba	Tamaños	d=0,0	d=0,2	d=0,4	d=0,6	d=0,8	d=1,0	d=1,5	d=2,0	d=3,0
W	n=7 m=5	0,0366	0,1008	0,2061	0,3325	0,4682	0,5923	0,8224	0,9239	0,9886
T_1		0,0351	0,1013	0,2104	0,3365	0,4689	0,5930	0,8132	0,9277	0,9873
t-student		0,0366	0,0871	0,1692	0,2755	0,4089	0,5386	0,7897	0,9162	0,9902
W	n=7 m=5	0,0490	0,1205	0,2365	0,3707	0,5077	0,6340	0,8373	0,9347	0,9885
T_1		0,0483	0,1227	0,2390	0,3730	0,5042	0,6192	0,8278	0,9295	0,9872
t -student		0,0480	0,1003	0,1835	0,2985	0,4235	0,5475	0,7883	0,9147	0,9892
W	n=7 m=4	0,0460	0,0852	0,1730	0,2835	0,3994	0,5451	0,7882	0,9043	0,9830
T_1		0,0430	0,0875	0,1629	0,2750	0,3860	0,5168	0,7845	0,9004	0,9821
t-student		0,0420	0,0762	0,1380	0,2231	0,3152	0,4764	0,7358	0,8741	0,9839
W	n=6 m=5	0,0649	0,1435	0,2565	0,3936	0,5167	0,6340	0,8382	0,9355	0,9916
T_1		0,0649	0,1442	0,2575	0,3895	0,5122	0,6264	0,8293	0,9284	0,9909
t -student		0,0685	0,1336	0,2235	0,3445	0,4718	0,6020	0,8155	0,9298	0,9904
W	n=6 m=4	0,1297	0,2514	0,3873	0,5257	0,6513	0,7544	0,9043	0,9636	0,9959
T_1		0,1326	0,2488	0,3848	0,5174	0,6330	0,7327	0,8856	0,9542	0,9937
t-student		0,1333	0,2326	0,3456	0,4638	0,5845	0,6844	0,8626	0,9498	0,9939

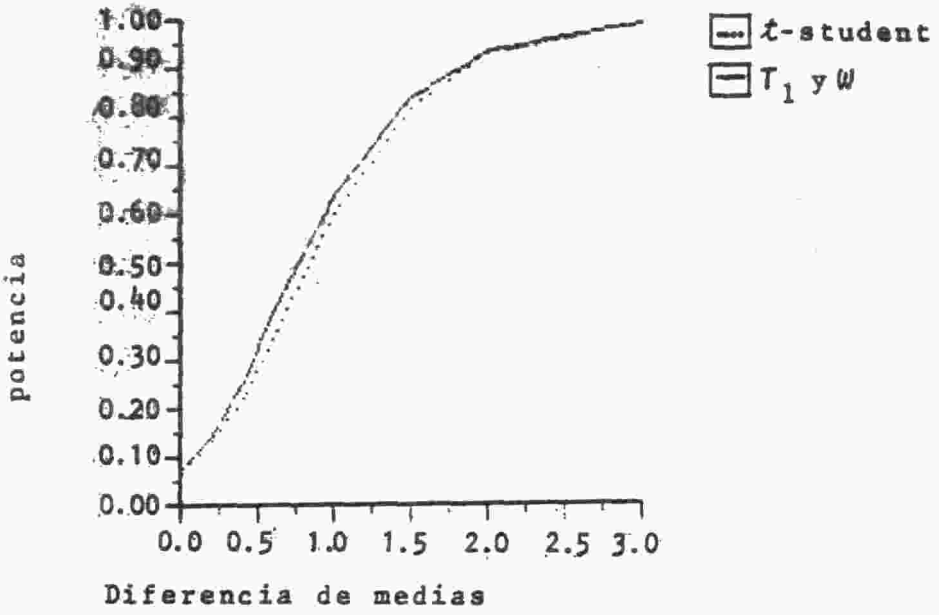
Gráfica 11 $n=7$ $m=5$ (Exp.)



Gráfica 12 $n=7$ $m=4$ (Exp.)



Gráfica 13 $n=6$ $m=5$ (Exp.)



Gráfica 14 $n=6$ $m=4$ (Exp.)

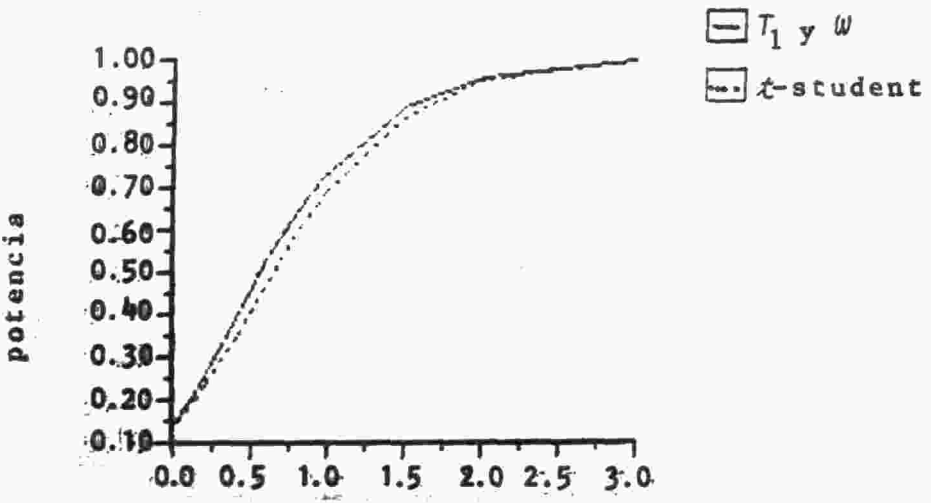
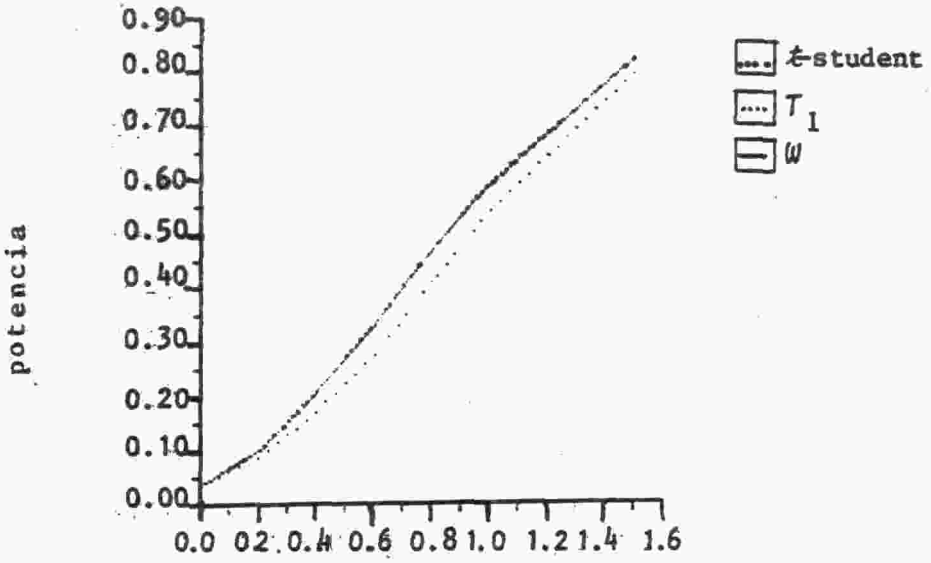


TABLA 3

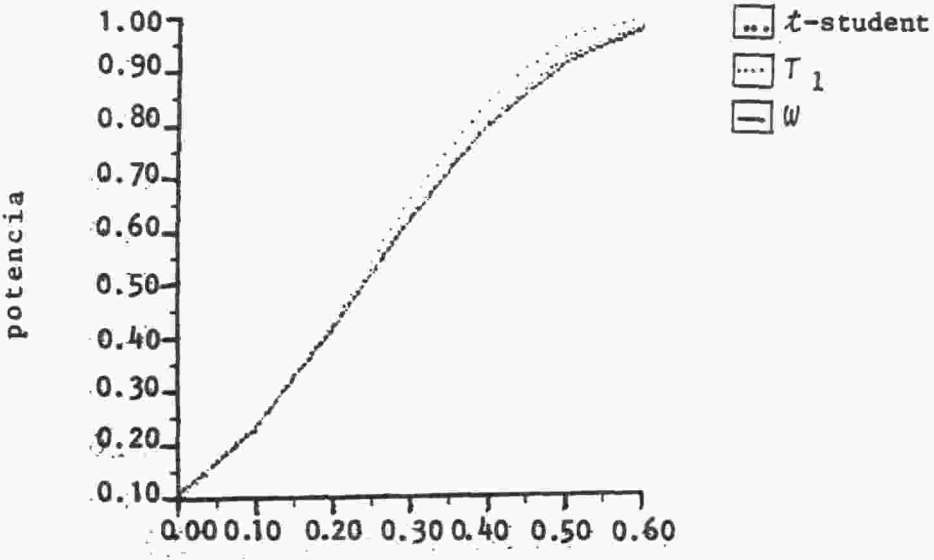
Comparación de algunas potencias simuladas de W, T_1 y t-student en poblaciones uniformes $(0,1)$. Prueba unilateral contra alternativas d en localización.

Prueba	Tamaños	$d=0,0$	$d=0,1$	$d=0,2$	$d=0,3$	$d=0,4$	$d=0,5$	$d=0,6$
W	$n=7$ $m=6$	0,1145	0,2638	0,4748	0,6853	0,8483	0,9430	0,9843
T_1		0,1091	0,2515	0,4555	0,6714	0,8381	0,9382	0,9836
t-student		0,1042	0,2405	0,4516	0,6890	0,8745	0,9636	0,9949
W	$n=7$ $m=5$	0,1104	0,2309	0,4120	0,6172	0,7960	0,9108	0,9720
T_1		0,1084	0,2340	0,4208	0,6228	0,7988	0,9208	0,9768
t-student		0,1008	0,2248	0,4224	0,6604	0,8388	0,9560	0,9908
W	$n=7$ $m=4$	0,0523	0,1312	0,2592	0,4300	0,6093	0,7710	0,8995
T_1		0,0567	0,1530	0,2940	0,4713	0,6450	0,8077	0,9160
t-student		0,0510	0,1227	0,2498	0,4268	0,6360	0,8148	0,9415
W	$n=6$ $m=5$	0,0922	0,2035	0,3642	0,5580	0,7425	0,8709	0,9482
T_1		0,0922	0,1991	0,3651	0,5622	0,7487	0,8765	0,9536
t-student		0,0953	0,2004	0,3891	0,6005	0,8007	0,9224	0,9798
W	$n=6$ $m=4$	0,0898	0,1982	0,3453	0,5163	0,6867	0,8333	0,9287
T_1		0,0852	0,1905	0,3417	0,5142	0,6925	0,8323	0,9338
t-student		0,0862	0,1948	0,3427	0,5453	0,7290	0,8805	0,9620

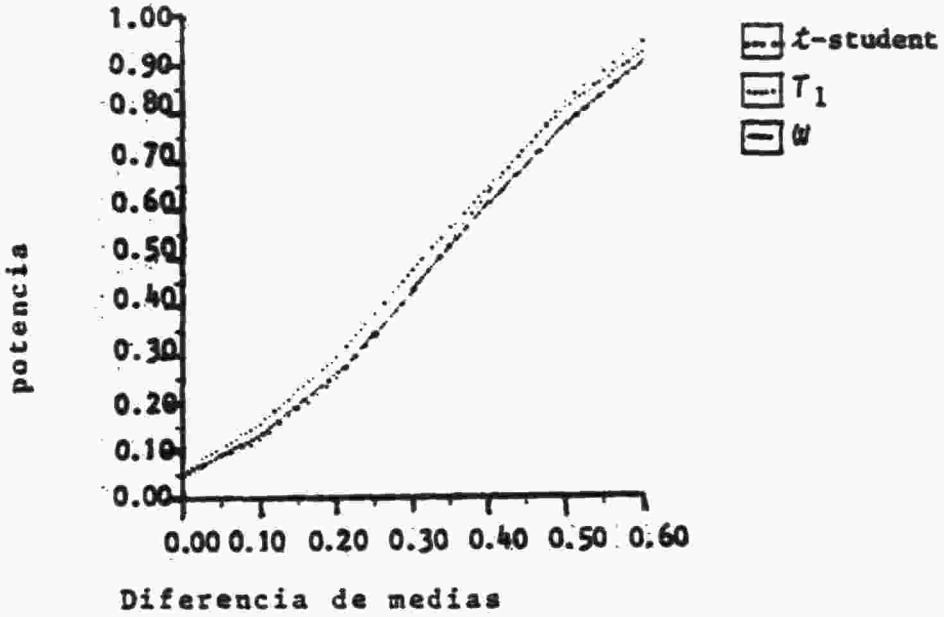
Gráfica 15 $n=7$ $m=6$



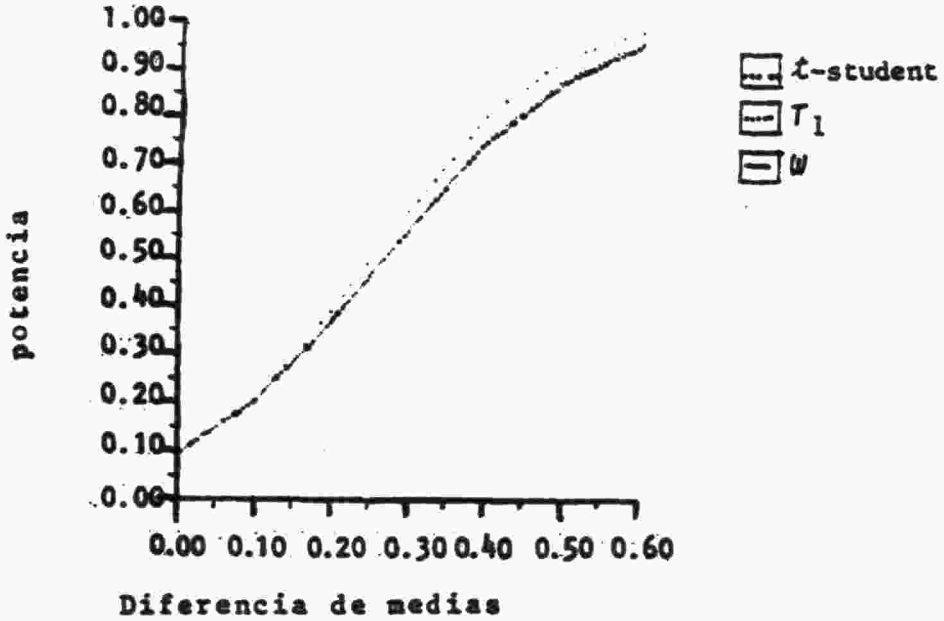
Gráfica 16 $n=7$ $m=5$



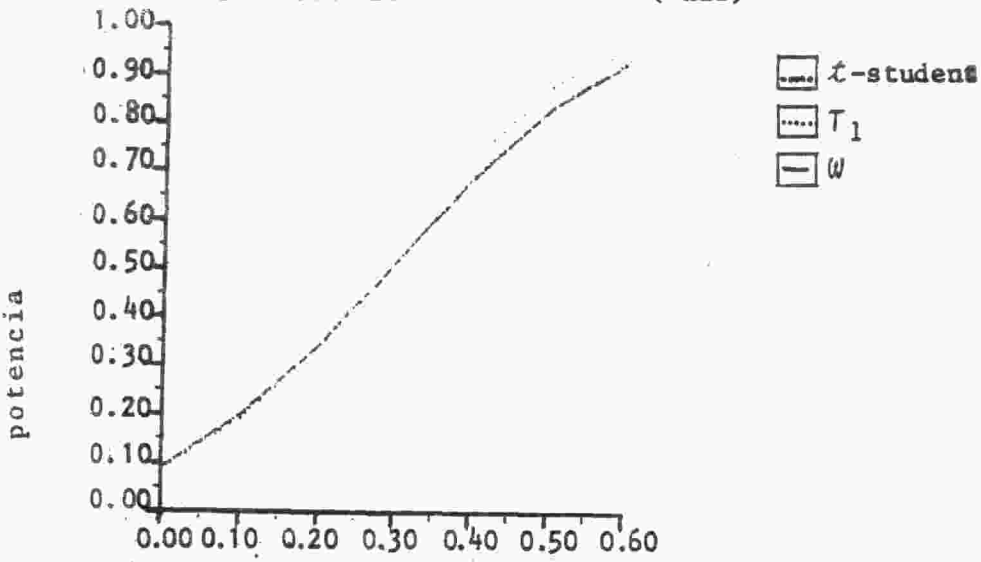
Gráfica 17 $n=7$ $m=4$ (Unif.)



Gráfica 18 $n=6$ $m=5$ (unif)



Gráfica 19 $n=6$ $m=4$ (Unif)



*

BIBLIOGRAFIA

Ortiz, J., (1983). Pruebas de hipótesis sobre parámetros de localización y dispersión basadas en secuencias. Revista Colombiana de Estadística, 8, 20-33.

Ortiz, J., y Corzo, J., (1983). Una prueba de dispersión basada en secuencias. Revista Colombiana de Estadística, 8, 34-48.

Conover, W.J., (1980). Practical Nonparametric Statistics. New York: John Wiley.

Burstein, H., (1971). Attribute Sampling. New

York: Mc-Graw-Hill.

Milton Roy, C., (1970). Rank Order Probabilities. New York: John Wiley.

Fernández, F., (1988). Transformaciones en Estadísticas Basadas en Secuencias. Tesis de magister, Universidad Nacional de Colombia, Bogotá.

* *