QÜESTIIÓ, Vol 15, 2. pp. 163-176, 1991

TESTS ESTADÍSTICOS SOBRE LA DISTRIBUCIÓN EXPONENCIAL BASADOS EN LA DISTANCIA DE RAO

AZUARA, A., VILLARROYA, A., SALDAÑA* J., RIOS, M.

Dept. Estadística
Universitat de Barcelona

En este artículo se construyen varios tests sobre el parámetro de la distribución exponencial, basados en la distancia de Rao. Los tests así desarrollados son comparados con los tests clásicos de la inferencia estadística.

Statistical tests for the exponential distribution based on Rao distance.

Keywords: Distancia de Rao, distribución exponencial, test de hipótesis.

Clasificación AMS: 62B10, 62H12.

1. INTRODUCCIÓN

El concepto de distancia entre dos distribuciones de probabilidad es fundamental en problemas de inferencia estadística. Rao (1945) propone un método para definir distancias entre distribuciones de una familia paramétrica, construyendo un campo tensorial a partir de la matriz de información de Fisher. Muchos autores han utilizado las distancias así construidas en la resolución de problemas estadísticos: Efron (1975), Ríos y Cuadras (1986), Burbea and Oller (1988), Cuadras y Arenas (1990), Villarroya y Oller (1991).

-J. Saldaña. Dep. Matemàtiques. Universitat Autònoma de Barcelona.

-Article rebut el març de 1991.

La construcción de tests de hipótesis es una de las aplicaciones estadísticas de las distancias, Matusita (1964), Mitchell and Krzanowski (1985).

En este trabajo utilizamos la distancia de Rao entre dos distribuciones exponenciales para la elaboración de contrastes de hipótesis sobre sus parámetros.

Una variable aleatoria X sigue la distribución exponencial si su función de densidad es:

$$(1) \qquad \begin{aligned} f(x) &= \alpha e^{-\alpha x} \qquad \text{si } x \geq 0 \quad \alpha > 0 \\ &= 0 \qquad\qquad\quad \text{si } x < 0 \end{aligned}$$

La media es $\frac{1}{\alpha}$ y la varianza $\frac{1}{\alpha^2}$. Se utiliza como ley de probabilidad del tiempo de espera en análisis de datos de supervivencia y fiabilidad, cuando la función de riesgo se puede considerar constante (la distribución no tiene memoria).

Dada una muestra aleatoria simple $x_1, \ldots, x_n$ de X con distribución (1), la estimación máximo verosímil de α es

$$(2) \qquad \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Sin embargo, $\hat{\alpha}$ es un estimador sesgado. La estimación con corrección de sesgo es

$$(3) \qquad \hat{\alpha}' = \frac{n-1}{\sum_{i=1}^{n} x_i}$$

que por estar basada en el estadístico suficiente $T = \sum_{i=1}^{n} x_i$, es insesgada y de varianza mínima.

Dadas dos distribuciones exponenciales de parámetros α_1 y α_2, el procedimiento clásico para comparar ambos parámetros está basado en el test de la razón de verosimilitud, que fue estudiado por Cox (1953). En este trabajo comparamos el test de la razón de verosimilitud con el que está basado en la distancia de Rao.

2. TEST ESTADÍSTICO SOBRE LA DISTRIBUCIÓN EXPONENCIAL

Dadas dos poblaciones estadísticas independientes, caracterizadas por los parámetros α_1 y α_2, la distancia de Rao entre ellas viene dada, Oller (1982), por:

$$d(\alpha_1, \alpha_2) = \left| \ln \frac{\alpha_1}{\alpha_2} \right| \tag{4}$$

Si $x_1, \ldots, x_n$ es una muestra aleatoria simple, cuyas componentes están distribuidas según una exponencial de parámetro α, vamos a considerar algunos tests estadísticos sobre este parámetro.

a) El contraste de hipótesis unilateral derecho

$$H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha > \alpha_0 \tag{5}$$

aplicando el test de la razón de verosimilitud se obtiene la región crítica

$$W_\epsilon = \left\{ (x_1, \ldots, x_n) \in \Re_+^n \mid \sum_{i=1}^{n} x_i < K_\epsilon \right\} \tag{6}$$

que es uniformemente más potente. Luego el test no puede ser mejorado, siendo fácil probar que aplicando la distancia de Rao se obtiene la misma región crítica. Análogo resultado se obtiene para el contraste unilateral izquierdo.

b) A continuación vamos a considerar el contraste bilateral

$$H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha \neq \alpha_0 \tag{7}$$

Este contraste de hipótesis lo podemos plantear a través de la distancia de Rao con las hipótesis nula y alternativa siguientes:

$$H_0 : d(\alpha, \alpha_0) = 0, \qquad H_1 : d(\alpha, \alpha_0) > 0 \tag{8}$$

La estimación de la distancia de Rao $d(\alpha_0, \alpha)$ es, según (5)

$$\hat{d}(\alpha_0, \alpha) = \left| \ln \frac{\alpha_0}{\hat{\alpha}} \right| \tag{9}$$

siendo $\hat{\alpha}$ la estimación máximo verosímil (2) del parámetro α.

La región de crítica, a nivel de significación ϵ, para realizar este contraste, viene dada por

$$W_\epsilon = \left\{(x_1, \ldots, x_n) \in \Re_+^n \mid \hat{d}(\alpha_0, \alpha) = \left|\ln \frac{\alpha_0}{\hat{\alpha}}\right| > C_\epsilon\right\} \tag{10}$$

ó de forma equivalente

$$W_\epsilon = \{(x_1, \ldots, x_n) \in \Re_+^n \mid T > c_2 \text{ ó } T < c_1\} \tag{11}$$

donde $T = \sum_{i=1}^n x_i$. Bajo la hipótesis nula, el estadístico T sigue una distribución gamma de parámetros (α_0, n), por lo tanto c_1 y c_2 cumplirán que

$$F(c_2) - F(c_1) = 1 - \epsilon \tag{12}$$

donde F es la función de distribución de una gamma de parámetros (α_0, n).

Por otra parte teniendo en cuenta (10) los valores de $\hat{\alpha}_1\, \hat{\alpha}_2$ que definen la región crítica deben cumplir

$$\left|\ln \frac{\alpha_0}{\hat{\alpha}_1}\right| = \left|\ln \frac{\alpha_0}{\hat{\alpha}_2}\right| \tag{13}$$

con $\hat{\alpha}_1 < \alpha_0$, $\hat{\alpha}_2 > \alpha_0$, por consiguiente

$$-\ln \frac{\alpha_0}{\hat{\alpha}_1} = \ln \frac{\alpha_0}{\hat{\alpha}_2} \Rightarrow \hat{\alpha}_1 \hat{\alpha}_2 = \alpha_0^2 \tag{14}$$

Teniendo en cuenta (2) y (11), los valores de c_1 y c_2 deben cumplir

$$c_1 c_2 = \frac{n^2}{\alpha_0^2} \tag{15}$$

De esta forma, con las ecuaciones (12) y (15) obtenemos los valores de c_1 y c_2 que nos determinan la región crítica (11).

Cuando tomamos el estimador máximo verosímil corregido de α, $\hat{\alpha}'(3)$, obtenemos unos valores c_1' y c_2' que definen una nueva región crítica, verificando

$$
\begin{aligned}
F(c_2') - F(c_1') &= 1 - \epsilon \\
c_1' c_2' &= \frac{(n-1)^2}{\alpha_0^2}
\end{aligned}
\tag{16}
$$

A continuación vamos a comparar el contraste antes desarrollado con los obtenidos por los métodos clásicos.

Es conocido que, en este caso, el lema de Neyman-Pearson no proporciona una región crítica óptima uniformemente. Si realizamos el contraste (7) utilizando el test de la razón de verosimilitud, la región crítica que se obtiene, dado un nivel de significación ϵ, es:

$$
W_\epsilon = \left\{(x_1, \ldots, x_n) \in \Re_+^n \mid T < k_1 \text{ ó } T > k_2\right\} \tag{17}
$$

donde

$$
T = \sum_{i=1}^{n} x_i \tag{18}
$$

y donde k_1 y k_2 son dos constantes tales que

$$
F(k_2) - F(k_1) = 1 - \epsilon \tag{19}
$$

siendo F la función de distribución de una gamma de parámetros (α_0, n).

Teniendo en cuenta que la razón de verosimilitud es:

$$
\Lambda(x) = \alpha_0^n \bar{x}^n e^{n(1-\alpha_0 \bar{x})} \tag{20}
$$

las constantes k_1 y k_2 deben cumplir

$$
k_1^n e^{-\alpha_0 k_1} = k_2^n e^{-\alpha_0 k_2} \tag{21}
$$

De las ecuaciones (12) y (15) obtenemos los valores c_1, c_2, de las ecuaciones (16) obtenemos c_1', c_2' y de las ecuaciones (19) y (21) obtenemos los valores k_1 y k_2.

Sabemos que T se distribuye según una distribución gamma de parámetros (α, n), por tanto la función de potencia para el contraste bilateral utilizando el test basado en la distancia de Rao y el estimador $\hat{\alpha}$ vendrá dada por:

$$\beta_1(\alpha) = 1 - \frac{1}{\Gamma(n)} \int_{c_1}^{c_2} \alpha^n x^{n-1} e^{-\alpha x} dx \tag{22}$$

Análogamente se obtienen las funciones de potencia $\beta_2(\alpha)$ utilizando el test de la razón de verosimilitud y $\beta_3(\alpha)$ utilizando $\hat{\alpha}'$ en el test basado en la distancia de Rao, con tal de cambiar c_1, c_2 por k_1, k_2 y por c_1', c_2' respectivamente.

La derivada de $\beta_2(\alpha)$ con respecto a α es

$$\beta_1'(\alpha) = -\frac{\alpha^{n-1}}{\Gamma(n)} \left[c_2^n e^{-c_2\alpha} - c_1^n e^{-c_1\alpha}\right] \tag{23}$$

Para $\alpha = \alpha_0$, teniendo en cuenta (21), $\beta_2'(\alpha_0) = 0$, como $\beta_2'' > 0$, la función de potencia $\beta_2(\alpha)$ alcanza un mínimo en α_0, por consiguiente tan sólo el test de la razón de verosimilitud es insesgado.

A continuación vamos a estudiar la función de potencia para los tres tests obtenidos. Sin pérdida de generalidad podemos tomar $\alpha_0 = 1$, ya que se trata de una familia invariante frente al producto por un escalar positivo. Así pues, el contraste (7) es equivalente a

$$H_0 : \alpha = 1, \qquad H_1 : \alpha \neq 1 \tag{24}$$

para este caso, tomando el nivel de significación de 0.05 y con un tamaño muestral $n = 5$, los valores críticos y las potencias medias, en un entorno de la hipótesis nula I = (0.5 , 1.5), para los tres tests, vienen dados en la tabla 1, mientras que en la figura 1 se han representado las curvas de potencia correspondientes. El cálculo de las potencias medias en un intervalo (a, b) se ha efectuado mediante la fórmula: $\bar{\beta} = \frac{1}{(b-a)} \int_a^b \beta(\alpha) d\alpha$.

TABLA 1

Test Estadístico	Valores Críticos	Potencia media
Test de la razón de verosimilitud	1.75 10.86	0.1091
Distancia de Rao con estimador corregido	1.58 10.10	0.1155
Distancia de Rao con estimador M.V.	1.92 12.98	0.0899

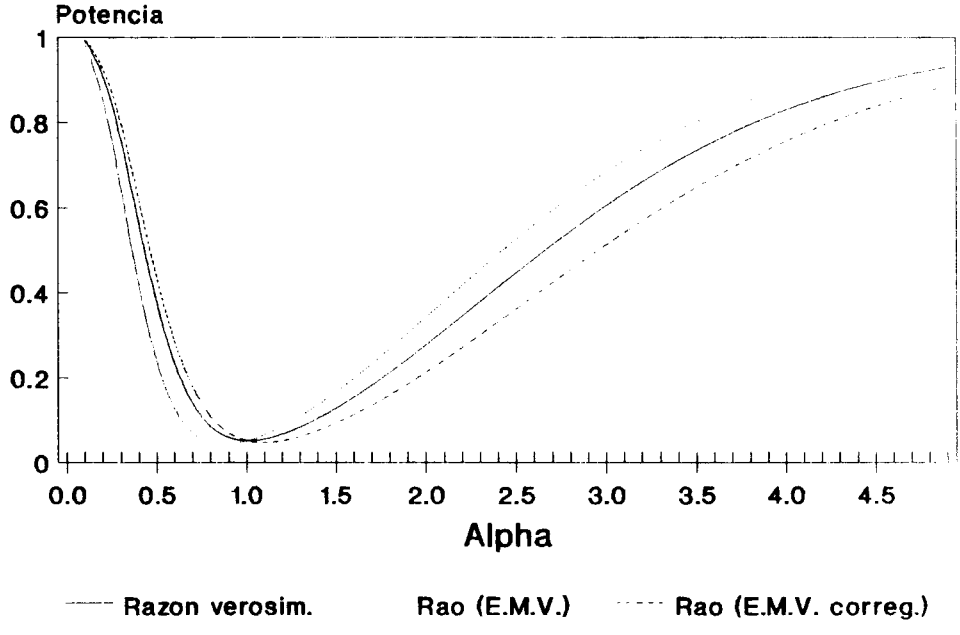

Figura 1. Curvas de potencias para los tests obtenidos a través de la distancia de Rao y de la razón de verosimilitud. $H_0 : \alpha = 1$, $H_1 : \alpha \neq 1$, muestras de tamaño 5, nivel de significación 0.05.

c) Si $x_1, \ldots, x_n$ y $y_1 \ldots y_m$ son dos muestras aleatorias simples distribuidas según dos exponenciales de parámetros α_1 y α_2 vamos a estudiar el siguiente contraste de hipótesis

(25) $$H_0 : \alpha_1 = \alpha_2 = \alpha, \qquad H_1 : \alpha_1 \neq \alpha_2$$

que también podemos plantearlo a través de la distancia de Rao:

(26) $$H_0 : d(\alpha_1, \alpha_2) = 0, \qquad H_0 : d(\alpha_1, \alpha_2) > 0$$

La estimación de la distancia de Rao, $d(\alpha_1, \alpha_2)$, es según (4)

(27) $$\hat{d}(\alpha_1, \alpha_2) = \left| \ln \frac{\hat{\alpha}_1}{\hat{\alpha}_2} \right|$$

La región crítica para realizar este contraste, dado un nivel de significación ϵ, viene dada por:

(28) $$W_\epsilon = \left\{ (x_1, \ldots, x_n, y_1, \ldots, y_m) \in \Re_+^{n+m} \mid \left| \ln \frac{\hat{\alpha}_1}{\hat{\alpha}_2} \right| > C_\epsilon \right\}$$

ó de forma equivalente

(29) $$W_\epsilon = \left\{ (x_1, \ldots, x_n, y_1, \ldots, y_m) \in \Re_+^{n+m} \mid \frac{T_2}{T_1} > c_2 \text{ ó } \frac{T_2}{T_1} < c_1 \right\}$$

donde

(30) $$T_1 = \sum_{i=1}^{n} x_i \ , \ T_2 = \sum_{i=1}^{m} y_i$$

Bajo la hipótesis nula, los estadísticos T_1 y T_2 están distribuidos según unas distribuciones gammas de parámetros (α, n) y (α, m) respectivamente, por tanto su cociente sigue una F de Fisher-Snedecor con $2n$ y $2m$ grados de libertad, con lo que c_1 y c_2 deben cumplir

(31) $$F(c_2) - F(c_1) = 1 - \epsilon$$

Siendo F la función de distribución de una F de Fisher de $2m$ y $2n$ grados de libertad.

Por otra parte teniendo en cuenta (28) los valores de los estimadores que definen la región crítica deben cumplir:

$$\left|\ln \frac{\hat{\alpha}_1}{\hat{\alpha}_2}\right| = \left|\ln \frac{\hat{\alpha}'_1}{\hat{\alpha}'_2}\right| \tag{32}$$

por consiguiente

$$\ln \frac{\hat{\alpha}_1}{\hat{\alpha}_2} = -\ln \frac{\hat{\alpha}'_1}{\hat{\alpha}'_2} \Rightarrow \hat{\alpha}_1 \hat{\alpha}'_1 = \hat{\alpha}_2 \hat{\alpha}'_2 \tag{33}$$

Según (29), los valores de c_1 y c_2 cumplirán:

$$c_1 c_2 = \frac{m^2}{n^2} \tag{34}$$

De esta forma, con las ecuaciones (31) y (34) obtenemos los valores c_1 y c_2 que nos determinan la región crítica (29).

Tomando los estimadores insesgados $\hat{\alpha}'_1 = \frac{n-1}{T_1}, \hat{\alpha}'_2 = \frac{m-1}{T_1}$, debemos determinar c'_1 y c'_2 tales que

$$\begin{aligned} F(c'_2) - F(c'_1) &= 1 - \epsilon \\ c'_2 c'_1 &= \frac{(m-1)^2}{(n-1)^2} \end{aligned} \tag{35}$$

A continuación vamos a comparar el contraste antes desarrollado con los utilizados clásicamente en Estadística. Es conocido que el lema de Neyman-Pearson no proporciona una región crítica para realizar el contraste de hipótesis (25), por lo tanto no existe un test uniformemente más potente.

El contraste (25) utilizando el test de Cox (1953) nos lleva a la misma región crítica (29), pero substituyendo c_1 y c_2 por k_1 y k_2 tales que

$$F(k_2) - F(k_1) = 1 - \epsilon \tag{36}$$

Teniendo en cuenta que la razón de verosimilitud es

$$\Lambda = \frac{\lambda^m}{(1+\lambda)^{m+n}} \tag{37}$$

con $\lambda = \frac{T_2}{T_1}$. Las constantes k_1 y k_2 deben cumplir

$$\frac{k_1^m}{(1+k_1)^{m+n}} = \frac{k_2^m}{(1+k_2)^{m+n}} \tag{38}$$

De las ecuaciones (31) y (34) obtenemos los valores c_1 y c_2, de las ecuaciones (35) c_1' y c_2' y de las ecuaciones (36) y (38) obtenemos los valores k_1 y k_2.

Se observa que cuando m=n las condiciones (34), (35) y (38) coinciden y por tanto el test de Cox coincide con los tests basados en la distancia de Rao.

Bajo la hipótesis alternativa del contraste (25) los estadísticos T_1 y T_2 están distribuidos según unas gammas de parámetros (α_1, n) y (α_2, m) respectivamente y su cociente $\frac{T_2}{T_1}$ sigue una distribución F de Fisher generalizada de parámetros (μ, m, n) cuya función de distribución es:

$$F(z) = \frac{\Gamma(m+n)}{\Gamma(n)\Gamma(m)} \int_0^{\mu z} \frac{y^{m-1}}{(1+y)^{n+m}} dy \tag{39}$$

siendo $\mu = \frac{\alpha_2}{\alpha_1}$.

Teniendo en cuenta este resultado, la función de potencia para el test basado en la distancia de Rao es

$$\beta_1(\mu) = 1 - \frac{\Gamma(m+n)}{\Gamma(n)\Gamma(m)} \int_{\mu c_1}^{\mu c_2} \frac{y^{m-1}}{(1+y)^{n+m}} dy \tag{40}$$

Análogamente se obtienen $\beta_2(\mu)$ y $\beta_3(\mu)$ considerando los otros tests y cambiando μc_i por μk_i y por $\mu c_i'$ respectivamente.

La derivada de β_2 es

$$\beta_2' = -\frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} \left[\frac{\mu^{m-1} k_2^m}{(1+\mu k_2)^{n+m}} - \frac{\mu^{m-1} k_1^m}{(1+\mu k_1)^{n+m}} \right] \tag{41}$$

Por (38), para $\mu = 1$, $\beta_2' = 0$, la función de potencia $\beta_2'(\mu)$ alcanza un mínimo en 1, por consiguiente el test es insesgado, a diferencia de lo que sucede en los tests basados en la distancia de Rao.

En la tabla 2 se muestran los valores críticos y la potencia media en el entorno de la hipótesis nula $I = (0.5, 1.5)$ para la hipótesis nula $\mu = 1$, un nivel de significación 0.05 y muestras de tamaño $n = 4$ y $m = 6$. En la figura 2 se muestran las funciones de potencias de los tres tests bajo las anteriores condiciones.

TABLA 2

Test Estadístico	Valores Críticos	Potencia media
Test de Cox	0.16 2.41	0.0739
Distancia de Rao con estimador corregido	0.15 2.30	0.0753
Distancia de Rao con estimador M.V.	0.17 2.56	0.0720

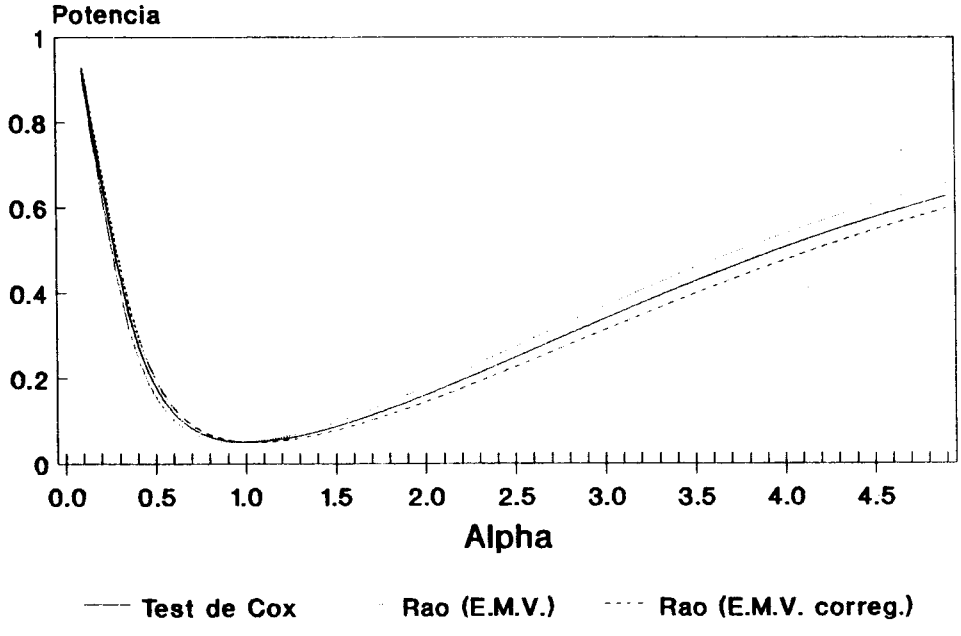

Figura 2. Curvas de potencias para los tests obtenidos a través de la distancia de Rao y de la razón de verosimilitud (test de Cox). $H_0 : \alpha_1 = \alpha_2$, $H_1 : \alpha_1 \neq \alpha_2$, muestras de tamaño $n = 4$ y $m = 6$, nivel de significación 0.05.

Con la misma hipótesis nula $\mu = 1$ y con el mismo nivel de significación 0.05, tomando muestras de tamaño $n = 6$ y $m = 4$, se obtienen los valores críticos y la potencia media en el entorno de la hipótesis nula $I = (0.5, 1.5)$ que vienen dados

en la tabla 3. Asímismo las funciones de potencia para este caso se muestran en la figura 3.

TABLA 3

Test Estadístico	Valores Críticos	Potencia media
Test de Cox	0.41 6.09	0.0731
Distancia de Rao con estimador corregido	0.43 6.40	0.0723
Distancia de Rao con estimador M.V.	0.39 5.76	0.0742

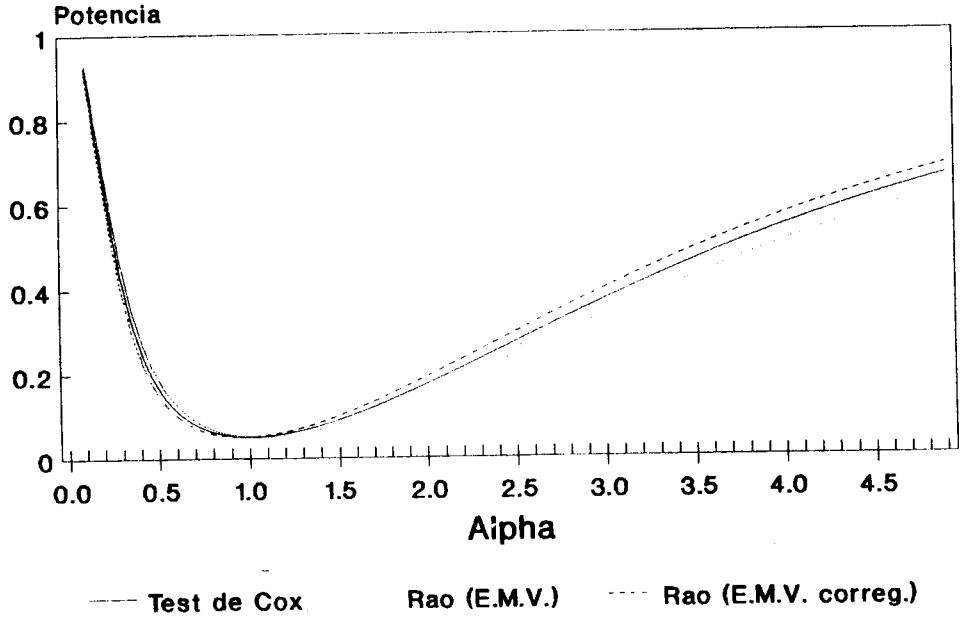

Figura 3. Curvas de potencias para los tests obtenidos a través de la distancia de Rao y de la razón de verosimilitud (test de Cox). $H_0 : \alpha_1 = \alpha_2$, $H_1 : \alpha_1 \neq \alpha_2$, muestras de tamaño $n = 6$ y $m = 4$, nivel de significación 0.05.

3. DISCUSIÓN Y CONCLUSIONES

De los test estudiados en este artículo se concluye que los tests basados en la distancia de Rao coinciden o son comparables a los obtenidos clásicamente.

Cuando realizamos el contraste de la hipótesis nula simple frente a la alternativa compuesta para el parámetro de la distribución exponencial, el test de la razón de verosimilitud es insesgado mientras que los basados en la distancia de Rao son sesgados, sin embargo la potencia media, en un entorno centrado en la hipótesis nula, del test basado en la distancia de Rao con el estimador máximo verosímil corregido es mayor que la de los otros tests.

En el caso de dos poblaciones, los tres contrastes estudiados coinciden cuando los tamaños de las muestras de las poblaciones a comparar son iguales. En el caso de que los tamaños muestrales sean diferentes, el test de la razón de verosimilitud sigue siendo insesgado, sin embargo la potencia media en un entorno de la hipótesis nula es superior en alguno de los tests basados en la distancia de Rao, dependiendo de los tamaños muestrales.

En definitiva podemos concluir que, para la distribución exponencial, los tests basados en la distancia de Rao coinciden con el resultado obtenido a través del lema de Neyman-Pearson, cuando éste último es aplicable. En caso contrario, los tests obtenidos a través de la razón de verosimilitud presentan la ventaja de ser insesgados, aunque el criterio de la potencia media en un entorno de la hipótesis nula favorece a los tests basados en la distancia de Rao. Resultados similares a los aquí descritos se han obtenido para la distribución de Wald (Villarroya y Oller, 1991).

4. BIBLIOGRAFÍA

[1] **Burbea, J. and Oller, J.M.** (1988). "The information metric for unvariate linear elliptic models. *Statist. & Decisions*, **6**, 209 - 221.

[2] **Cox, D.R.** (1953). "Some simple tests for Poisson variates". *Biometrika*, **40**, 354 - 360.

[3] **Cuadras, C. y Arenas, C.** "A distance-based regression model for prediction with mixed data". *Commun. Statist. - Theory Meth.*, **19** (6), 2261 - 2279.

[4] **Efron, B.** (1975). "Defining the curvature of a statistical problem (with applications to second oreder efficiency". *Ann. Statis.* **3**, 1189 - 1242 (with discussion).

[5] **Matusita, K.** (1964). "Distance and Decision Rules". *Ann. Inst. Stat. Math.*, **16**, 301 - 315.

[6] **Mitchell, A.F.S. and Krzanowski, W.J.** (1985). "The Mahalanobis distance and elliptics distributions". *Biometrika*, **72**, 464 - 467.

[7] **Oller, J.M.** (1982). *Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la Biología.* Publicaciones de Bioestadística, Barcelona.

[8] **Rao, C.R.** (1945). "Information and accuracy attainable in estimation of parameters". *Bull. Calcutta Math. Soc.*, **37**, 81 - 91.

[9] **Ríos, M. y Cuadras, C.** (1986). "Distancia entre modelos lineales normales". *Qüestió*, **10**, 83 - 92.

[10] **Villarroya, A. and Oller J.M.** "Statistical tests for the Inverse Gaussian distribution based on Rao distance". *Sankhya* (en prensa).