ESTADISTICA ESPAÑOLA
Vol. 35. Núm. 133, 1993, págs. 195 a 268

Teoría de la Información Estadística *

LEANDRO PARDO LLORENTE

Departamento de Estadística e I. O.
Facultad de Matemáticas
Universidad Complutense de Madrid

RESUMEN

En este trabajo se analizan algunas de las aplicaciones de las medidas de entropía y divergencia en Estadística poniéndose de manifiesto la creciente atención que esta rama de la Estadística, conocida con el nombre de Teoría de la Información Estadística, ha tenido en los últimos años por parte de un gran número de investigadores. Se destaca la nueva perspectiva que se abre en la construcción de contrastes de hipótesis mediante la utilización de medidas de entropía y divergencia.

Palabras clave: Medidas de entropía, Medidas de divergencia, Medidas paramétricas de información, Indice de Diversidad, Contrastes de hipótesis.

Clasificación AMS: 62B10, 94A17, 62E20.

1. INTRODUCCION

Un hecho importante que puso de manifiesto Shannon (1948) al dar las bases de lo que hoy se entiende por Teoría de la Información, fue que a los términos coloquiales «incertidumbre» e «información» era posible darles un significado matemático sobre la base de un modelo probabilístico y en términos de una cantidad medible numéricamente, que él denominó entropía y que en la actualidad recibe el nombre de Entropía de Shannon. De esta forma muchos problemas, primero de transmisión de la información y posteriormente de ecología, biología,

(*) Este trabajo ha sido parcialmente subvencionado con la ayuda de la Dirección General de Investigación Científica y Técnica (DGICYT) como parte del proyecto de referencia PB91-0387.

probabilidad, música, psicología, economía, urbanismo, tomografía computerizada y reconstrucción de imágenes, estadística, criptografía, etc., se han podido modelizar y resolver mediante la citada medida de entropía.

En general, se pueden distinguir dos líneas de actuación en el estudio de la Teoría de la Información: Una axiomática y otra de aplicaciones. La primera requiere establecer una serie de axiomas aceptables y encontrar funciones que satisfagan los axiomas impuestos, mientras que la segunda establece la aplicabilidad de estas funciones en términos de las propiedades que verifiquen. Si bien a veces han surgido medidas a partir de situaciones reales concretas, otras no han tenido ni antes ni después una justificación de su necesidad. Esto llevó en 1980 a la creación del denominado «código de ética» de Teoría de la Información: «Avoid defining new information measures without a realistic hope of applications, deriving their often not very interesting properties and then characterizing, by several of these properties, the «measures» which you have introduced in the first place» (consensus at the special Session on Information Measures, 1980).

El enfoque que se dará en este trabajo será esencialmente el de poner de manifiesto diversas aplicaciones de la Teoría de la Información en Estadística; es decir, lo que hoy en día se conoce con el nombre de **Teoría de la Información Estadística.** No obstante, en el apartado siguiente se analizarán las diversas medidas introducidas hasta la fecha y se justificará la necesidad de establecer funcionales que las unifiquen.

Sea X una variable aleatoria con valores en el espacio $(\mathfrak{X}, \beta_{\mathfrak{X}}, P)$ con $P \in \mathfrak{P}$,, dependiendo o no de un parámetro desconocido $\theta \in \Theta \subset \mathbb{R}^M$ y sea f la función de densidad de P respecto de una medida σ-finita μ. La entropía de Shannon se define mediante la expresión

$$H(X) = -\int_{\mathfrak{X}} f(x) \log f(x)\, d\mu(x).$$

Esta medida cuantitativa, propuesta por Shannon, acerca de la cantidad de información proporcionada por un experimento aleatorio está basada en la entropía clásica de Boltzmann (1896) de la física estadística. Boltzmann fue el primero en enfatizar el significado probabilístico de la entropía clásica en termodinámica y por tanto parece lógico que se le considere como el precursor de la Teoría de la Información. Aunque Bolzmann no hizo referencia explícita a la palabra información, observó que la entropía de un sistema físico se puede considerar como una medida del desorden del mismo. En un sistema físico con muchos grados libertad (por ejemplo, un gas perfecto), el número que mide el desorden del sistema mide también la incertidumbre en relación a los estados de las partículas individuales. La entropía clásica de un sistema fue primeramente definida por Clausius (1864) como una función de algunas otras macrocoordenadas que se pueden medir directamente. La entropía de Clausius es un con-

cepto no probabilístico. En termodinámica estadística, la entropía se definió como una función de posiciones y velocidades de todas las partículas incluidas en el sistema físico, y Boltzmann se dio cuenta de su significado probabilístico. Por analogía con la expresión probabilística de Boltzmann de la entropía clásica, Shannon introdujo en 1948 la entropía, en abstracto, como una medida de la información o incertidumbre de experimentos probabilísticos arbitrarios. Por supuesto, la entropía de Shannon es independiente de la entropía clásica. Entre la entropía clásica e información existe, sin embargo, una relación de naturaleza diferente. Viene del hecho, señalado por Georgescu (1971), de que no podemos obtener, transmitir o incluso almacenar información de cualquier clase sin un aumento en la entropía total del sistema aislado en el cual se actúa.

Son numerosas las caracterizaciones que de la entropía de Shannon se han dado desde que Shannon (1948) y Khintchin (1953) establecieron las dos primeras. Estas se suelen agrupar en cuatro grandes grupos dependiendo de algunas características comunes: **i) Caracterizaciones basadas en la recursividad:** Faddev (1956), Renyi (1959, 1961), Tverberg (1958), Daróczy (1969), Kendall (1964a), Lee (1964), Daróczy y Katai (1970), Diderrich (1975), Aczél y Daróczy (1975). **ii) Caracterizaciones en base a la aditividad:** Chaundy y McLeod (1960), Aczél y Daróczy (1963), Daróczy (1971). **iii) Caracterizaciones en base a la desigualdad de Shannon:** Aczél y Pfanzagl (1966), Fisher (1972), Aczél y Ostrowski (1973). **iv) Caracterización en base a la propiedad de ramificación:** Forte y Daróczy (1968a). Todas estas caracterizaciones hacen referencia al caso en el que μ es una medida contable. En el caso de que μ sea la medida de Lebesgue destacaremos únicamente la debida a Hatori (1958).

En contraposición a la entropía de Shannon y todas sus generalizaciones que se verán en el apartado siguiente, están las denominadas medidas de divergencia que expresan la cantidad de información en los datos para discriminar en favor de una distribución f contra otra g o la afinidad o distancia entre f y g. Si bien la primera medida de divergencia se debe a Bhattacharyya (1943) la que ha tenido un más amplio eco es la debida a Kullback y Leibler (1951). La idea de esta medida de divergencia la tomaron Kullback y Leibler de un trabajo de Jeffreys (1946) en el que aparecía la expresión de la divergencia al abordar el problema de encontrar una densidad invariante respecto de una probabilidad «a priori».

Sean X e Y dos variables aleatorias con valores en los espacios estadísticos $(\mathfrak{X}, \beta_{\mathfrak{X}}, P)$ e $(\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q)$ con P y Q $\in \mathfrak{P}$ y sean $f(x) = (dP/d\mu)(x)$ y $g(x) = (dQ/d\mu)(x)$ sus respectivas densidades. Se denomina información media por discriminación en favor de P frente a Q, divergencia de Kullback-Leibler entre P y Q, a la expresión

$$D(X, Y) = \int_{S_g} f(x) \log \frac{f(x)}{g(x)} \, d\mu(x), \qquad S_g = \{ x / g(x) > 0 \}.$$

Es claro que D(X,Y) es no negativa pero no es una distancia o una métrica en sentido general, ya que D(X,Y) no es una función simétrica de f y g. No obstante, D(X,Y) caracteriza en realidad desde el punto de vista estadístico la desviación de Y respecto a X.

La medida de divergencia de Kullback está estrechamente relacionada a los conceptos de entropía de Shannon e información mutua, I(X,Y), entre las variables X e Y (diferencia entre la entropía de Y y la esperanza de la entropía de la variable Y condicionada a cada valor x de la variable aleatoria X). Si en la expresión de la divergencia de Kullback, se considera como primera variable la correspondiente al par (X,Y) con la distribución conjunta asociada, y como segunda variable la misma pareja con la distribución asociada a la condición de independencia (que representaremos por X*Y) resulta, empleando las notaciones habituales, D[(X,Y),X*Y] = I(X,Y). Por otro lado si se supone que μ es una medida contable y Q es una medida de probabilidad uniforme discreta, entonces: D(X,Y) = = logn -H(X), expresión que para n fijo, es máxima cuando la entropía se minimiza, esto es, cuando la distribución de probabilidad P, está concentrada en un solo valor, y se hace mínima cuando la distribución P es uniforme discreta, lo que permite la interpretación de la divergencia como una «distancia» entre las distribuciones, idea válida aun en el caso de no elegir necesariamente la distribución uniforme como punto de referencia. Desde otro punto de vista, la igualdad anterior permite caracterizar la entropía de una distribución discreta P en función de su divergencia a una distribución uniforme discreta Q.

Finalmente se tiene H (X) + H (X ‖ Y) = D (X,Y) siendo, H (X‖ Y), la inacuracidad introducida por Kerridge (1961), al estudiar algunos problemas de inferencia estadística y cuya expresión viene dada por

$$H(X \parallel Y) = -\int_{\mathcal{X}} f(x) \log g(x)\, d\mu(x).$$

Conviene señalar que en el caso de poblaciones normales n-dimensionales, la divergencia de Kullback-Leibler, con vectores de medias μ_1 y μ_2 y matrices de varianzas covarianzas Σ_1 y Σ_2, viene dada por

$$D[(\mu_1, \Sigma_1), (\mu_2, \Sigma_2)] = \frac{1}{2}(\mu_1 - \mu_2)^t \Sigma_2^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} \operatorname{tr}(\Sigma_2^{-1} \Sigma_1 - I),$$

y en caso de que $\Sigma_1 = \Sigma_2$, esta expresión coincide con la distancia de Mahalanobis (1936).

Al igual que ocurría con la entropía de Shannon, son numerosas las caracterizaciones axiomáticas que de la divergencia de Kullback-Leibler se han dado en el caso discreto. Entre ellas mencionaremos las siguentes: Kannappan y Rathie (1973) y Mathai y Rathie (1975). En el libro de Aczél y Daróczy (1975), se recogen

otras caracterizaciones debidas a Ng (1974), Kannappan y Ng (1973), etc. Para el caso general conviene mencionar la de Ikeda (1962a, b).

2. MEDIDAS GENERALIZADAS DE ENTROPIA

Con el nombre de entropías generalizadas se recogen aquellas medidas de entropía dependientes de parámetros y tales que a partir de ellas, bien para algún valor particular de los mismos bien por paso al límite, se obtiene la entropía de Shannon. El primer intento para desarrollar una generalización de la entropía de Shannon fue llevado a cabo por Renyi (1961), el cual definió la entropía de orden r en los siguientes términos:

$$H_r(X) = (1-r)^{-1} \log \left(\int_{\mathcal{X}} f(x)^r \, d\mu(x) \right), \; r \neq 1, \; r > 0$$

Puede comprobarse que se cumple $\lim_{r \to 1} H_r(X) = H(X)$; es decir, la entropía de orden r, $H_r(X)$, contiene como caso límite la entropía de Shannon. La anterior medida, $H_r(X)$, fue caracterizada posteriormente por Daróczy (1963).

Por motivos operativos, parece más natural considerar la expresión

$$K_r(X) = \int_{\mathcal{X}} f(x)^r \, d\mu(x)$$

como una medida de información en lugar de la entropía de Renyi de orden r. Así, Havrda y Charvat (1967) propusieron la siguiente entropía de grado s

$$H^s(X) = (1-s)^{-1} \left[\int_{\mathcal{X}} f(x)^s \, d\mu(x) - 1 \right], \; s \neq 1, \; s > 0.$$

Nuevamente, la entropía de grado s, $H^s(X)$, contiene como caso límite (s = 1) a la entropía de Shannon. Daróczy estableció en 1970 una caracterización de la entropía de grado s mediante una ecuación funcional.

Arimoto (1971) presentó otra generalización de la entropía de Shannon que se llama entropía de clase t y que viene dada por

$${}_tH(X) = (1-t)^{-1} \left[\left(\int_{\mathcal{X}} f(x)^{1/t} \, d\mu(x) \right)^t - 1 \right], \; t \neq 1, \; t > 0.$$

En este caso también es cierto que

$$\lim_{t \to 1} {}_tH(X) = H(X).$$

Las tres entropías generalizadas, tienen en común la expresión $K_r(X)$, lo cual permite relacionarlas en los siguientes términos:

$$H_r(X) = (1-r)^{-1} \log\left[(1-r)\, H^r(X) + 1 \right] = r(1-r)^{-1} \log\left[\frac{r-1}{r}\, {}_{\frac{1}{r}}H(X) + 1 \right].$$

Sharma y Mittal (1975) introducen y caracterizan dos entropías generalizadas que denominan entropía de orden 1 y grado s y entropía de orden r y grado s, dadas por las expresiones:

$$H_1^s(X) = (1-s)^{-1} \left\{ \left[\exp(s-1) \int_{\mathcal{X}} f(x) \log f(x)\, d\mu(x) \right] - 1 \right\},\ s \neq 1,\ s > 0$$

y

$$H_r^s(X) = (1-s)^{-1} \left\{ \left[\int_{\mathcal{X}} f(x)^r\, d\mu(x) \right]^{\frac{s-1}{r-1}} - 1 \right\},\ r \neq 1,\ s \neq 1,\ r > 0.$$

La idea de Sharma y Mittal fue generalizar las tres entropías, $H_r(X)$, $H^s(X)$ y ${}_tH(X)$. Las relaciones entre ellas son las siguientes:

(i) Cuando $r = s$, $H_r^s(X) = H_s^s(X) = H^s(X)$

(ii) Cuando $t = r^{-1} = 2 - s$, $H_r^s(X) = H_{\frac{1}{t}}^{2-t}(X) = {}_tH(X)$

(iii) $\lim_{s \to 1} H_r^s(X) = H_r(X)$

(iv) $\lim_{r \to 1} H_r^s(X) = H_1^s(X)$

(v) $\lim_{r \to 1} H_r(X) = \lim_{s \to 1} H^s(X) = \lim_{t \to 1} {}_tH(X) = \lim_{s \to 1} H_1^s(X) = H(X)$

Existen otras muchas medidas de entropía y certidumbre que han jugado un papel más o menos relevante dentro de la Teoría de la Información. Gran parte de ellas se pueden expresar, Salicrú y otros (1993b), en la forma,

$$H_\phi^h(X) = h\left[\int_{\mathcal{X}} \phi\,(f(x))\, d\mu(X) \right]$$

donde $\phi : [0, \infty) \longrightarrow \mathbb{R}$ es cóncava y $h: \mathbb{R} \longrightarrow \mathbb{R}$ creciente ó $\phi : [0, \infty) \longrightarrow \mathbb{R}$ es convexa y $h: \mathbb{R} \longrightarrow \mathbb{R}$ decreciente. En los casos restantes; es decir, h creciente y ϕ convexa o h decreciente y ϕ cóncava $H_\phi^h(X)$ juega el papel de una medida de certidumbre (Van der Lubbe, 1981). En las tablas 1 y 2 se presentan algunas medidas de entropía y certidumbre que se obtienen como caso particular del anterior funcional.

TABLA 1
(h, φ)-Entropías

φ(x)	h(x)	(h,φ)-Entropías
$-x\log x$	x	Shannon (1948)
x^r	$(1-r)^{-1}\log x$	Renyi (1961)
x^{r-m+1}	$(m-r)^{-1}\log x$	Varma (1966)
$x^{r/m}$	$m(m-r)\log x$	Varma (1966)
$(1-s)^{-1}(x^s-x)$	x	Havrda y Charvat (1967)
$x^{1/t}$	$(2^{t-1}-1)^{-1}(x^t-1)$	Arimoto (1971)
$x\log x$	$(1-s)^{-1}\{\exp_2[(s-1)x]-1\}$	Sharma y Mittal (1975)
x^r	$(1-s)^{-1}(x^{\overset{s-1}{t-1}}-1)$	Sharma y Mittal (1975)
$x^r\log x$	$-2^{r-1}x$	Taneja (1979)
x^r-x^s	$(2^{1-r}-2^{1-s})^{-1}x$	Sharma y Taneja (1975)
$(1+\lambda x)\log(1+\lambda x)$	$\left(1+\frac{1}{\lambda}\right)\log(1+\lambda)-\frac{x}{\lambda}$	Ferreri (1980)
x	$-x\log\left(\frac{x\,\mathrm{sens}}{2\,\mathrm{sen}(s/2)}\right)$	Sant'anna y Taneja (1985)
x	$-\frac{x\,\mathrm{sen}x}{2\,\mathrm{sen}(s/2)}\log\left[\frac{x\,\mathrm{sen}x}{2\,\mathrm{sen}(s/2)}\right]$	Sant'anna y Taneja (1985)

TABLA 2
Medidas de Certidumbre

h(x)	φ(x)	Medidas de Certidumbre
x	x^2	Energía Infomacional (Onicescu, 1966)
$x^{1/r}$ $(r>1)$	x^r	r-Norma (Van der Lubbe, 1977)
$x^{1/(r-1)}(r\neq1, r>0)$	x^r	r-Media (Van der Lubbe, 1981)
x^s	x^r	Medida Generalizada de Certidumbre (Van der Lubbe, 1981)

En el cálculo efectivo de algunas entropías, juega un papel esencial la expresión

$$K_r(X)=\int_{\mathbb{R}} f(x)^r\,d\mu(x)$$

por ello, Pardo, J. A. y otros (1993) determinaron su valor para las distribuciones continuas más comunes. La tabulación se hace de forma similar a la empleada por Lazo y Rathie (1978), correspondiente a la entropía de Shannon para variables aleatorias continuas.

3. MEDIDAS DE DIVERGENCIA GENERALIZADAS

Rényi (1961) dio la primera generalización de la divergencia de Kullback-Leibler, en los siguientes términos,

$$D_r^1(X,Y) = (r-1)^{-1} \log \left\{ \int_{\mathcal{X}} f(x)^r \, g(x)^{1-r} \, d\mu(x) \right\}, \; r \neq 1, \; r > 0.$$

Sharma y Mittal (1975), estudiaron dos generalizaciones paramétricas de D(X,Y) que incluyen como casos límite a la establecida por Renyi, $D_r^1(X, Y)$

$$D_r^s(X, Y) = (r-1)^{-1} \left\{ \left[\int_{\mathcal{X}} f(x)^r \, g(x)^{1-r} \, d\mu(x) \right]^{\frac{s-1}{r-1}} - 1 \right\}, \; r \neq 1, \; s \neq 1, \; r > 0, s > 0.$$

En particular, cuando r = s, se tiene

$$D_s^s = (X, Y) = (s-1)^{-1} \left\{ \int_{\mathcal{X}} f(x)^s \, g(x)^{1-s} \, d\mu(x) - 1 \right\}, \; s \neq 1, \; s > 0.$$

La medida $D_s^s(X; Y)$ ha sido estudiada extensamente por diversos autores. Un amplio estudio de ella, para el caso discreto, puede verse en Mathai y Rathie (1975) y Taneja (1979).

Es fácil comprobar las siguientes relaciones:

$$\lim_{s \to 1} D_r^s(X; Y) = D_r^1(X; Y); \; \lim_{r \to 1} D_r^s(X; Y) = D_1^s(X; Y);$$
$$\lim_{r \to 1} D_r^1(X; Y) = \lim_{s \to 1} D_s^s(X; Y) = \lim_{s \to 1} D_1^s(X; Y) = D(X; Y),$$

donde

$$D_1^s(X, Y) = (s-1)^{-1} \left\{ \exp\left[(s-1) \, D(X; Y)\right] - 1 \right\}, \; s \neq 1.$$

Al estudiar la divergencia de Kullback-Leibler, se analizaba la relación de ésta con la entropía de Shannon en el caso de variables aleatorias discretas. La relación existente entre las divergencias generalizadas y las entropías generalizadas viene dada, para variables aleatorias discretas, mediante la expresión:

$$D_r^s(X, Y) = n^{s-1} \left[H_r^s(Y) - H_r^s(X) \right],$$

siempre que la distribución de probabilidad de la variable aleatoria Y sea la uniforme discreta. Esta relación es importante ya que una vez estudiadas las propiedades de las divergencias generalizadas se pueden obtener como caso

particular las propiedades de las entropías generalizadas. Esto, por ejemplo, puede verse en Taneja y otros (1989b).

La expresión

$$g_s^r(X, Y) = \frac{D_r^s(X, Y)}{n^{s-1} H_r^s(Y)},$$

con X e Y variables aleatorias discretas e Y uniforme, representa una generalización de las medidas de desigualdad de renta dadas por Theil (1972, 1980).

Las R-divergencias fueron introducidas y estudiadas por Burbea y Rao (1982a,b), Rao (1982). Es sabido que

$$\frac{H(X) + H(Y)}{2} \leq H\left(\frac{X + Y}{2}\right),$$

siendo H(X) la entropía de Shannon de la variable aleatoria X. La diferencia

$$R(X, Y) = H\left(\frac{X + Y}{2}\right) - \frac{H(X) + H(Y)}{2} =$$

$$= \int_{\mathcal{X}} \left[\frac{f(x) \log f(x) + g(x) \log g(x)}{2}\right] d\mu(x) - \int_{\mathcal{X}} \left(\frac{f(x) + g(x)}{2}\right) \log\left(\frac{f(x) + g(x)}{2}\right) d\mu(x)$$

se conoce con el nombre de radio de información (Sibson, 1969) o divergencia diferencia de Jensen (Burbea and Rao, 1982a,b; Rao, 1982). Es sencillo comprobar que

$$R(X, Y) = \frac{1}{2}\left[D\left(X, \frac{X + Y}{2}\right) + D\left(Y, \frac{X + Y}{2}\right)\right].$$

Por simplicidad, la medida R(X,Y) se denomina R-divergencia. Rao (1982) y Sgarro (1981) establecieron que J(X,Y), ≥ 4 R(X,Y), siendo J(X,Y) = D(X,Y) + D(Y,X).

El concepto general de φ-Divergencia fue introducido en la literatura por Csiszar (1967) en la forma siguiente:

$$D_\varphi(X, Y) = \int_{\mathcal{X}} f(x)\, \varphi[f(x)/g(x)]\, d\mu(x),$$

donde φ es una función continua y convexa en $[0,\infty)$, finita en $(0,\infty)$ y estrictamente convexa en algún punto x, $0<x<\infty$.

La φ-Divergencia fue también introducida independientemente por Alí y Silvey (1966). Las propiedades y aplicaciones de la φ-divergencia han sido estudiadas posteriormente por Pérez (1967a,b), Csiszar (1967, 1972, 1978), Vajda (1972), Liese (1975), etc. Una excelente recopilación de las propiedades y aplicaciones

de esta familia de divergencias puede verse en los libros de Liese y Vajda (1987) y Vajda (1989).

Antes de seguir adelante conviene poner de manifiesto que las φ-divergencias constituyen una forma alternativa, de generalizar la divergencia de Kullback. Obsérvese que si se toma φ(x) = xlogx, entonces la φ-Divergencia de Csiszar coincide con la divergencia de Kullback. Es interesante señalar que si bien las propiedades que en general verifican las φ-Divergencias son conocidas desde hace poco tiempo, algunas φ-Divergencias en particular se conocían y habían sido estudiadas sus propiedades y aplicaciones desde hace bastante tiempo, dada su importancia en el Análisis Funcional, Estadística Matemática, Teoría de la Información, etc. Casos particulares de φ-Divergencias se han utilizado como métricas o distancias no métricas. Así por ejemplo, si φ(x)=|x−1|, se obtiene la denominada divergencia variacional con importantes aplicaciones en Análisis Funcional y Estadística Matemática: En este caso se obtiene

$$D_{\varphi}(X, Y) = \int_{\chi} |f(x) - g(x)| \, d\mu(x)$$

que no es sino la norma variacional de la medida signada P-Q en el espacio de Banach $F^{\pm}$ (conjunto de todas las medidas signadas finitas). Referencias a esta norma pueden verse en Halmos (1964), Dunford y Schwartz (1958). A su vez la divergencia variacional tiene una gran importancia en Estadística, donde se denomina distancia variacional o estadística entre distribuciones de probabilidad, como puede verse en Le Cam (1964, 1974), Torgersen (1976, 1980), Ibragimov y Chasminskij (1981). En caso de tomar $\varphi(x) = (1 - x)^2$, se obtiene

$$D_{\varphi}(X, Y) = \int_{\{ x/g(x) > 0 \}} \frac{[f(x) - g(x)]^2}{g(x)} \, d\mu(x),$$

que generalmente se denomina χ^2-Divergencia. Esta expresión fue utilizada por Pearson (1900). La χ^2-Divergencia en el contexto de las divergencias fue introducida y estudiada en relación con generalizaciones de la desigualdad de Cramer-Rao y de la información de Fisher en Vajda (1973). En caso de ser $\varphi(x) = (1-x^a)^{1/a}$ $0<a\leq 1$, se obtiene la distancia introducida y estudiada por Matusita (1955, 1964), conocida con el nombre de distancia de Matusita. Otros importantes casos pueden verse en Vajda (1989).

Como en el caso de las medidas de entropía, ante la gran cantidad de medidas de divergencia en Menéndez y otros (1992b) y Morales y otros (1993b) se plantea la posibilidad de definir un funcional que las unificase. Tal objetivo no puede

conseguirse, aunque sí resulta factible con el empleo de dos funcionales. Así, se definieron las (h, ϕ)–divergencias por medio de la expresión

$$D_{\phi}^{h}(X, Y) = \int_{\Lambda} h_{\alpha} \left\{ \int_{\mathcal{X}} g(x)\, \phi_{\alpha} \left[f(x)/g(x) \right] d\mu(x) - \phi_{\alpha}(1) \right\} d\eta(\alpha),$$

donde $\underline{h} = (h_{\alpha})_{\alpha \in \Lambda}$, $\underline{\phi} = (\phi_{\alpha})_{\alpha \in \Lambda}$, ϕ_{α} y h_{α} son funciones C^2, con $h_{\alpha}(0) = 0$ y η es una medida positiva σ-finita definida sobre un espacio medible (Λ, B), supuesta la existencia de las integrales y la de los límites siguientes:

$$\lim_{x \to 0^{+}} \phi_{\alpha}(x) \qquad y \qquad \lim_{x \to 0^{+}} \frac{\phi_{\alpha}(x)}{x} \qquad \supset \alpha$$

y las R_h^{ϕ}–divergencias mediante la expresión

$$R_h^{\phi}(X, Y) = h\left(\int_{\mathcal{X}} \phi\left(\frac{f(x) + g(x)}{2} \right) d\mu(x) - \right.$$

$$\left. - \frac{1}{2} \left\{ h\left(\int_{\mathcal{X}} \phi\,[f(x)]\, d\mu(x) \right) + h\left(\int_{\mathcal{X}} \phi\,[g(x)]\, d\mu(x) \right) \right\} \right)$$

Eligiendo convenientemente las funciones h_{α} y ϕ_{α} se obtienen importantes medidas de divergencia como caso particular de las (h,ϕ)-divergencias; en este sentido se pueden obtener las siguientes: (1) Csiszar (1967), (2) Kullback-Leibler (1951), (3) Variacional o Estadística, (4) χ^2-Divergencia o Kagan (1963), (5) Matusita (1955), (6) Balakrishman y Sanghvi (1968), (7) Havrda y Charvat (1967), (8) Cressie y Read (1984), (9) Renyi (1961), (10) y (11) Sharma and Mittal (1975), (12) Battacharyya (1943), (13) Trigonometrica, (14), (15), (16) y (17) Taneja (1989).

La tabla siguiente indica, elegidos $\Lambda = \{1,2\}$ y $\eta(1)=\eta(2) = 1$, las funciones h_{α} y ϕ_{α} que permiten expresar como caso particular de (h,ϕ)-divergencias las anteriores medidas de divergencia. En algunas de ellas aparece la expresión genérica h(x) correspondiente a

$$h(x) = \frac{1}{s-1} \left[(x+1)^{\frac{s-1}{r-1}} - 1 \right].$$

N.	Divergencia	h_{ij}	ϕ_{ij}
(1)	$D_{\varphi}(P, Q)$	$h1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = \varphi(x) + \varphi(1)$
(2)	$D(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = x\log x$
(3)	$D^{VA}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = \lvert x - 1\rvert$
(4)	$D^{KA}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = (1 - x)^2$
(5)	$D^{MA}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = (1 - x^a)^{1/a}$, $0<a\leq 1$
(6)	$D^{BS}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = \dfrac{(x-1)^2}{x+1}$
(7)	$D^{HC}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = \dfrac{(x - x^a)}{1-a}$, $a > 0$
(8)	$D^{CR}(P, Q)$	$h_1(x) = x$, $h_2(x) = 0$	$\phi_1(x) = \dfrac{x^{a+1} - 1}{a(a+1)}$
(9)	$D_r^1(P, Q)$	$h_1(x) = \dfrac{\log(x+1)}{r-1}$, $h_2(x) = 0$	$\phi_1(x) = x^r$
(10)	$D_1^s(P, Q)$	$h_1(x) = \dfrac{e^{(s-1)x} - 1}{s-1}$, $h_2(x) = 0$	$\phi_1(x) = x\log x$
(11)	$D_r^s(P, Q)$	$h_1(x) = h(x)$, $h_2(x) = 0$	$\phi_1(x) = x^r$
(12)	$D^{BA}(P, Q)$	$h_1(x) = -\log(x+1)$, $h_2(x) = 0$	$\phi_1(x) = x^{1/2}$
(13)	$D^{Tri}(P, Q)$	$h_1(x) = tg^{-1}(x)$, $h_2(x) = 0$	$\phi_1(x) = x^{\alpha}$, $0<\alpha<1$
(14)	${}^1J_r^s(P, Q)$	$h_1(x) = h_2(x) = h(x)$	$\phi_1(x) = x^r$, $\phi_2(x) = x^{1-r}$
(15)	${}^2J_r^s(P, Q)$	$h_1(x) = 2h(x)$, $h_2(x) = 0$	$\phi_1(x) \dfrac{1}{2}(x^r + x^{1-r})$
(16)	${}^1R_r^s(P, Q)$	$h_1(x) = h_2(x) = \dfrac{1}{2}h(x)$	$\phi_1(x) = \left(\dfrac{1+x}{2}\right)^{1-r}$, $\phi_2(x) = x^r\left(\dfrac{1+x}{2}\right)^{1-r}$
(17)	${}^2R_r^s(P, Q)$	$h_1(x) = h(x)$, $h_2(x) = 0$	$\phi_1(x) = 2^{r-2}(1 + x^r)(1 + x)^{1-r}$

En relación con la R_h^{ϕ}-divergencias conviene destacar que además de englobar, como es obvio, a las que se obtienen a través de las medidas de entropías y certidumbre para $h(x) = x$, se tiene la R-divergencia de Burbea y Rao (1982 a,b) y si además $\phi(x) = -x\log x$ se tiene el Radio de Información introducida por Sibson (1969).

Ciertamente existen algunas otras medidas de divergencia que no quedan incluidas dentro de estos funcionales pero, aparte de las M-divergencias y K-divergencias definidas y estudiadas por Burbea y Rao (1982a,b) carecen de importancia.

4. PRINCIPIO DE MAXIMIZACION DE LA ENTROPIA Y MINIMIZACION DE LA DIVERGENCIA

Quizá sea el contenido de este apartado el que ha tenido un más amplio eco por parte de investigadores de diversos campos. Los principios de maximización

de la entropía de Shannon y de minimización de la divergencia de Kullback han sido aplicados con notable éxito en innumerables y distintos campos. Kapur (1989) señala, con amplia documentación, los siguientes campos en donde estos principios han alcanzado resultados importantes: (1) termodinámica y mecánica estadística; (2) estimación no paramétrica de funciones de densidad; (3) diseño de experimentos; (4) tablas de contingencia; (5) análisis de series temporales relativas a datos: geográficos, astronómicos, metereológicos, pluviométricos, etc.; (6) análisis de señales y de textos; (7) tomografía computerizada y reconstrucción de imágenes; (8) procesos de búsqueda; (9) análisis de la fiabilidad de sistemas; (10) teoría de colas; (11) teoría de juegos y de la decisión; (12) economía, comercio internacional, concentración industrial, problema de la cartera, etc.; (13) clasificación y reconstrucción de formas; (14) actividades bancarias y seguros; (15) publicidad y estudios de mercado; (16) geografía; (17) distribución de la población, planificación urbana y regional y transportes; (18) modelos biológicos, ecológicos y médicos; etc.

Antes de pasar a analizar algunos de los resultados más importantes me parece interesante recordar el teorema de Boltzmann, sobre la irreversibilidad macroscópica (la entropía de un sistema evoluciona siempre en un sentido creciente), como una justificación al principio de maximización de la entropía. Este principio, dado inicialmente por Jaynes, escoge como distribución de probabilidad en un cierto momento aquella que, siendo compatible con los conocimientos en dicho momento, maximiza la entropía.

Mediante el método de los multiplicadores de Lagrange, se obtiene que la densidad o probabilidad que maximiza la entropía de Shannon bajo restricciones del tipo $\beta_j = \int_{\mathbb{R}} b_j(x)\, f(x)\, d\mu(x)$ $j = 1,\ldots, s$, viene dada por

$$f(x) = \exp\{-\lambda_0 - \lambda_1\, b_1(x) - \ldots - \lambda_s b_s(x)\}.$$

Conviene observar que este procedimiento conduce a que $f(x) \geq 0$. Si en lugar de considerar la entropía de Shannon se considera otra medida de entropía, la obtención de la distribución de máxima entropía bajo restricciones análogas a las dadas anteriormente no queda, a veces, resuelto mediante la utilización del método de los multiplicadores de Lagrange ya que éste no cubre las restricciones de no negatividad.

El proceso para determinar los multiplicadores de Lagrange corresponde a la resolución del sistema proporcionado por las s + 1 ecuaciones de restricción, pero esto resulta muchas veces tedioso. A veces se consigue simplificar el problema introduciendo la función denominada de partición

$$Z(\lambda_1,\ldots, \lambda_s) = \int_{\mathbb{R}} \exp\{-\lambda_1\, b_1(x) - \ldots - \lambda_s\, b_s(x)\}\, d\mu(x).$$

Denotando $\lambda_0 = \log Z(\lambda_1,..,\lambda_s)$, el problema de encontrar los multiplicadores de Lagrange conduce a resolver el sistema

$$\begin{cases} \exp(\lambda_0) = Z(\lambda_1,..., \lambda_s) \\ \exp(-\lambda_0) \dfrac{\partial Z(\lambda_1,..., \lambda_s)}{\partial \lambda_j} = -\beta_j; \quad j = 1,..., s. \end{cases}$$

Además, la determinación efectiva de la entropía máxima correspondiente a la distribución hallada, resulta de la expresión

$$H_{MAX}(X) = \lambda_0 + \lambda_1 \beta_1 + ... + \lambda_s \beta_s.$$

La existencia de los multiplicadores de Lagrange queda garantizada siempre que el conjunto de funciones $\{1,b_1(x),...,b_s(x)\}$ sea un conjunto de funciones linealmente independiente ya que hay s ecuaciones implícitas para determinar los s multiplicadores de Lagrange, $\lambda_1, \lambda_2,...,\lambda_s$,

$$f_j(\lambda_1, \lambda_2,..., \lambda_s) = \exp(-\lambda_0) \frac{\partial Z(\lambda_1,..., \lambda_s)}{\partial \lambda_j} - \beta_j = 0; \qquad j = 1,..., s,$$

y esta condición implica que el determinante de la matriz $\left(\partial f_j / \partial \lambda_k \right)_{j,k}$ es distinto de cero y en consecuencia el teorema de la función implícita garantiza que es posible usar el sistema anterior para encontrar $\lambda_1, \lambda_2,...,\lambda_s$.

La función de partición tiene interesantes propiedades:

$$a) \frac{\partial \lambda_0}{\partial \lambda_j} = \beta_j, \; b) \frac{\partial^2 \lambda_0}{\partial \lambda_j^2} = \mathrm{Var}\,[b_j(X)], \; c) \frac{\partial^2 \lambda_0}{\partial \lambda_1 \partial \lambda_k} = \mathrm{Cov}\,[b_j(X), b_k(X)]$$

y como consecuencia la función λ_0 es convexa.

Son importantes los estudios encaminados a la caracterización de distribuciones, tanto continuas como discretas, utilizando el principio de máxima entropía, Kapur (1989) recoge algunos resultados interesantes dependiendo de que el conjunto de valores que toma la variable sea finito o infinito numerable. Imponiendo restricciones acerca de la media aritmética, media geométrica, varianza, etc., obtiene diversas distribuciones discretas tales como la geométrica, geométrica generalizada, zeta de Riemann, etc. En lo que hace referencia a las distribuciones continuas existe una amplia e interesante bibliografía. Así, Kagan, Linnik y Rao (1973), caracterizaron las distribuciones uniforme, exponencial, gamma, normal y Laplace; Gokhale (1975) caracterizó la normal, doble exponencial y Cauchy; Wragg y Dowson (1970), la normal truncada y la exponencial; Tribus, M. (1969) la Beta; Kapur, J. N. (1982) la normal truncada. Este mismo autor en 1983 caracterizó las distribuciones de Pareto, la Beta de tipo II, la logaritmo-Normal, la Cauchy generalizada, así como otras distribuciones de

interés. Finalmente, Preda V. C. (1980) ha caracterizado la distribución t-Student. En el caso de distribuciones k-dimensionales conviene destacar los resultados debidos a Kapur (1989), y la caracterización de las distribuciones de Wishart y Dirichlet debidas a Gokhale (1975). Otra aplicación importante del principio de máxima entropía se encuentra en la estimación no paramétrica de densidades como puede verse en Kapur (1989), Theil y Fiebig (1984), etc.

El principio de máxima entropía tiene su aplicabilidad en aquellas situaciones en las que se necesita una distribución de probabilidad e inicialmente sólo se dispone de una cierta información parcial acerca de ella, cuantificada en términos de valores esperados o restricciones sobre la o las variables. Un nuevo problema se presenta cuando además de esta información se dispone de una estimación previa de la distribución de interés. En líneas generales, el principio de mínima divergencia establece un procedimiento de inferencia acerca de una densidad o probabilidad desconocida, g(x), cuando se dispone de cierta información sobre ella, en términos de valores esperados o restricciones sobre la o las variables además de una estimación previa de la distribución de interés. Si f(x) es la estimación previa que se tiene de la densidad o probabilidad de una variable aleatoria X y supuesto un conocimiento inicial adicional en términos de $\beta_j = \int_{\mathbb{R}} b_j(x)\, g(x)\, d\mu(x)$; $j = 1,\ldots, s$, se establece que la densidad o probabilidad resultante es de la forma

$$g(x) = f(x) \exp \{- \lambda_0 - \lambda_1 b_1 (x) - \ldots - \lambda_s b_s(x)\}.$$

La función obtenida debe verificar las restricciones impuestas, lo que permite construir un sistema con el que determinar los multiplicadores, si bien en la práctica la obtención explícita de los mismos se realiza con ayuda de la «función de partición», que ahora queda determinada por

$$Z(\lambda_1,\ldots, \lambda_s) = \int_{\mathbb{R}} f(x) \exp \left\{ - \sum_{j=1}^{s} \lambda_j b_j(x) \right\} d\mu(x) ,$$

y para la que se cumplen, entre otras, propiedades análogas a las encontradas en el desarrollo del principio de maximización de la entropía. La determinación efectiva de la divergencia mínima da lugar a

$$D(^*, X) = -\log Z - \sum_{j=1}^{s} \lambda_j b_j .$$

La matriz cuyos elementos son las derivadas parciales segundas del logaritmo de la función de partición con respecto a los multiplicadores, que coincide con la matriz de varianzas-covarianzas de las variables que determinan las restricciones, se denotará en lo sucesivo mediante $\Sigma_{b_1(X),\ldots, b_s(X)}$. Nuevamente una condi-

ción suficiente para garantizar la existencia de los multiplicadores λ_j, es que el conjunto de funciones $\{1,b_1(x),..,b_s(x)\}$ sea linealmente independiente.

La aplicación del principio de mínima divergencia a la construcción de tests de hipótesis resulta interesante. A tal fin se utilizará la notación que recuerde directamente los problemas que van a ser tratados; así, la variable X de partida quedará sustituida por una muestra aleatoria simple $X_1,...,X_n$ de una cierta variable aleatoria (su función de densidad será representada por $L_f(x_1,...,x_n) = L_f(x)$, los valores medios asociados serán sustituidos por valores θ_j correspondientes a parámetros de la distribución, y la formulación de las restricciones se dará en la forma $E(T_j) = \theta_j$. En la práctica los valores θ_j son desconocidos y se estiman a través del correspondiente T_j lo que proporcionará a su vez unos estimadores para los multiplicadores λ_j mediante la resolución del sistema

$$T_j(x_1,..., x_n) = \frac{\partial \log Z(\hat{\lambda}_1,..., \hat{\lambda}_s)}{\partial \lambda_j}$$

y de este modo no se dispondrá de la divergencia mínima sino de su estimador, que se denotará por $\hat{D}[*, L_f(x_1,..., x_n)]$ y viene dado por

$$\hat{D}[*, L_f(x_1,..., x_n)] = -\log Z(\hat{\lambda}_1,..., \hat{\lambda}_s) - \sum_{j=1}^{s} \lambda_j T_j .$$

En Kullback (1959) se establece, supuesto que $\lambda \neq 0$, que

$$n^{1/2}\{\hat{D}[*, L_f(x_1,..., x_n)] - D[*, L_f(x_1,..., x_n)]\} \xrightarrow{L} N(0, \lambda^t \Sigma_{T_1,...,T_s} \lambda)$$

donde $\Sigma_{T_1,..., T_s}$ es la matriz de varianzas covarianzas de $T_1,...,T_s$. Bajo la hipótesis de que $\lambda = 0$, se tiene

$$2n\hat{D}[*, L_f(x_1,..., x_n)] \xrightarrow{L} \chi^2_s.$$

Supongamos el problema general de contrastar la hipótesis nula H_0: $f(x) \in C_0$, contra la alternativa H_1: $f(x) \in C_1$. Se trata de decidir, a partir de un estadístico T [tal que $E(T)=\theta$], si la muestra obtenida es consistente con la hipótesis nula o con la alternativa, en el sentido de disponer de una mayor información mínima de discriminación contra una de ellas. El procedimiento general de decisión consistirá en determinar los estimadores de las informaciones mínimas de discriminación contra cada hipótesis,

$$\hat{D}(*, H_0) = \underset{f \in C_0}{Min}\, \hat{D}[*, L_f(x)] \qquad \text{y} \qquad \hat{D}(*, H_0) = \underset{f \in C_1}{Min}\, \hat{D}[*, L_f(x)].$$

Estan serán tanto mayores cuanto «peor» es la similitud entre la muestra obtenida y la población considerada en la hipótesis, rechazándose entonces la hipótesis nula si la diferencia entre la primera y la segunda cantidad supera un cierto valor c que se determinará una vez fijado el nivel de significación (α) del test, esto es: $P_{H_0}\left(\hat{D}(*, H_0) - \hat{D}(*, H_1) \geq c\right) \leq \alpha$.

Si $C_0 = \{f_0\}$, $C_1 = \{f_1\}$ y se elige $T(x_1,..., x_n) = \log \frac{L_1(x_1,..., x_n)}{L_0(x_1,..., x_n)}$, que es el estimador insesgado de la diferencia de divergencias entre cualquier otra distribución muestral y las correspondientes a ambas hipótesis, se obtiene que

$$\hat{D}(*, H_0) - \hat{D}(*, H_1) = \log \frac{L_1(x_1,\ldots,x_n)}{L_0(x_1,\ldots,x_n)};$$

es decir, el procedimiento de contraste dado por Kullback coincide con el dado por Neyman-Pearson.

Sea M un conjunto de poblaciones pertenecientes a la familia exponencial, Ω el conjunto donde varía λ sobre la familia exponencial y sea ω el conjunto donde varía λ sobre M, se verifica que

$$\hat{D}_M[*, L_f(x_1,\ldots, x_n)] = \log \frac{\sup_{\lambda \in \Omega} g(x_1,\ldots, x_n)}{\sup_{\lambda \in \omega} g(x_1,\ldots, x_n)}.$$

Si ahora se desea contrastar, H_0: $\lambda \in \omega_0 \subset \Omega$ frente H_1: $\lambda \in \omega_1 \subset \Omega$, con $\omega_0 \cap \omega_1 = \varnothing$, se tiene,

$$\hat{D}(*, H_0) - \hat{D}(*, H_1) = \log \frac{\sup_{\lambda \in \omega_0} g(x_1,\ldots, x_n)}{\sup_{\lambda \in \omega_1} g(x_1,\ldots, x_n)}$$

luego nuevamente, en este caso, el test de Kullback vuelve a coincidir con el test del cociente de verosimilitudes.

Una aplicación especialmente importante del principio de minimización de la divergencia de Kullback, puede verse en Berger (1980), se presenta cuando nos situamos en el contexto bayesiano y nuestro conocimiento a priori viene dado en términos de valores esperados de funciones relacionadas con la distribución a priori. En esta situación parece especialmente interesante considerar como estimación de la distribución a priori dada, la correspondiente distribución a priori no informativa. Esta es la distribución que no contiene información acerca del parámetro desconocido; es decir, la que no favorece

posibles valores de θ sobre otros. Claramente, en el caso discreto, la distribución a priori no informativa es aquella distribución que asigna igual probabilidad a todos los valores del espacio paramétrico. Esta idea al trasladarla al caso continuo da lugar a considerar distribuciones a priori no informativas del tipo $\pi_0(\theta) = c, \supset \theta \in \Theta$. La principal crítica que tiene esta forma de proceder es la falta de invariancia por transformaciones. Propiedad que sería deseable que tuviera ya que parece lógico que dado el conocimiento de una cierta distribución a priori no informativa respecto de Θ, se conozca la correspondiente a una transformación del parámetro Θ, sin más que aplicarla, para no tener que efectuar de nuevo los cálculos. Para la resolución de este problema se han dado innumerables procedimientos, siendo quizá el más sencillo el consistente en utilizar $\pi_0(\theta) \propto I(\theta)^{1/2}$ donde $I(\theta)$ es la información de Fisher. Este procedimiento fue establecido por Jeffreys (1961).

Sea cual sea el procedimiento seguido para la determinación de la distribución a priori no informativa, el método de la divergencia mínima permite la construcción de una distribución a priori que se separe de la no informativa lo menos posible y además aproveche la información adicional dada por las restricciones conocidas, ésto es: encontrar una distribución a priori $\pi(\theta)$ que satisfaga restricciones del tipo $\beta_j = \int_\Theta b_j(\theta)\, \pi(\theta)\, d\theta;\ j = 1,\ldots, s,$ y haga mínima la expresión $\int_\Theta \pi(\theta) \log \frac{\pi(\theta)}{\pi_0(\theta)}\, d\theta.$

Obsérvese que si la distribución no informativa a priori se elige como habíamos señalado en primer lugar, valores constantes para la probabilidad o densidad, los resultados de la aplicación de los principios de mínima divergencia y máxima entropía son coincidentes, en virtud de la relación vista entre la entropía de Shannon y la divergencia de Kullback-Leibler.

Conviene señalar la importante relación existente entre la divergencia de Kullback y la función de verosimilitud. De acuerdo con el principio de mínima divergencia, dada la densidad o probabilidad f(x), se desea encontrar la densidad o probabilidad g(x) que minimice

$$\int_{\mathcal{X}} g(x) \log g(x)\, d\mu(x) - \int_{\mathcal{X}} g(x) \log f(x)\, d\mu(x)$$

y satisfaga las restricciones dadas. Alternativamente se puede dar g(x) y encontrar f(x) que minimice la divergencia de Kullback o lo que es lo mismo maximice,

$$G_f = \int_{\mathcal{X}} \log f(x)\, dG(x).$$

Sea $x_1,...,x_n$ una muestra aleatoria simple obtenida de G y sea $G_n(x)$ la correspondiente distribución empírica de la muestra. Un estimador de la expresión anterior es

$$\hat{G}_f = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i)$$

y como consecuencia la maximización de G_f conduce, aproximadamente, a la maximización de $\hat{G}_f$. Si ahora se supone que f pertenece a una familia paramétrica que denotaremos por $f_\theta(x)$, el problema de encontrar f maximizando G_f será aproximadamente el mismo que el de encontrar θ maximizando,

$$\frac{1}{n} \sum_{i=1}^{n} \log f_\theta(x_i) = L_n(\theta; x_1, \ldots, x_n)$$

es decir, maximizando la función de verosimilitud.

En Bernardo (1979a, 1979b), Guttman y Peña (1988) y Peña y Guttman (1989) pueden verse aplicaciones estadísticas importantes de la divergencia de Kullback.

5. MEDIDAS PARAMETRICAS DE INFORMACION

Las medidas de entropía dan la cantidad de información contenida en una distribución; esto es, la cantidad de incertidumbre o información concerniente con la realización de un experimento. Por otro lado, las medidas de divergencia expresan la cantidad de información que los datos proporcionan para discriminar en favor de la primera población contra la segunda o la medida de la distancia o afinidad entre las dos poblaciones. Frente a estas dos concepciones, las medidas paramétricas de información miden la cantidad de información proporcionada por los datos acerca del parámetro desconocido y son funciones del mismo.

Sea $(\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$, $\theta \in \Theta$, Θ abierto de $\mathbb{R}$, un espacio estadístico y supongamos que la familia de distribuciones de probabilidad $\{P_\theta, \theta \in \Theta\}$ está dominada por una medida σ-finita μ, siendo $f(x,\theta) = (dP_\theta/d\mu)(x)$ la densidad correspondiente. Supuesto que se verifican las habituales condiciones de regularidad, se define la cantidad de información de Fisher sobre θ contenida en P_θ, mediante la expresión

$$I_X(\theta) = E_\theta \left[\left(\frac{\partial \log f(X, \theta)}{\partial \theta} \right)^2 \right]$$

y dadas las variables aleatorias X e Y, se denomina Información de Fisher contenida en Y acerca de θ cuando X es igual a x, a la expresión:

$$I_{Y/X=x}(\theta) = E_\theta\left[\left(\frac{\partial \log f(Y/x, \theta)}{\partial \theta}\right)^2\right].$$

Verificándose

$$I_{X,Y}(\theta) = I_x(\theta) + I_{Y/X}(\theta) = I_Y(\theta) + I_{X/Y}(\theta) \qquad \supset \theta \in \Theta,$$

donde

$$I_{Y/X}(\theta) = E_X\,[I_{Y/X=x}(\theta)].$$

Si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, se tiene

$$n^{1/2}\left(I_X(\hat{\theta}) - I_X(\theta)\right) \xrightarrow[n\to\infty]{L} N\left(0, \frac{I'_X(\theta)}{I_X(\theta)^{1/2}}\right)$$

resultado que permite la comparación de cantidades de información.

Como generalizaciones de la medida de información de Fisher conviene destacar las dadas por Vajda, Mathai y Boekee. Estas se definen, en los siguientes términos:

$$I^V(\theta) = E_\theta\left[\left|\frac{\partial \log f(X, \theta)}{\partial \theta}\right|^\alpha\right], \alpha \geq 1 \qquad \text{Vajda (1973)}$$

$$I^M(\theta) = \left(E_\theta\left[\left|\frac{\partial \log f(X, \theta)}{\partial \theta}\right|^\alpha\right]\right)^{1/\alpha}, \alpha \geq 1 \quad \text{Mathai (1967)}$$

$$I^B(\theta) = \left[E_\theta\left[\left|\frac{\partial \log f(X, \theta)}{\partial \theta}\right|^{s/(s-1)}\right]\right]^{s-1}, s \geq 1 \text{ Boekee (1977)}$$

La cantidad de información de Fisher puede generalizarse al caso de que el parámetro del que depende la densidad $f(x,\theta)$ varíe en un abierto Θ de $\mathbb{R}^M$. En esta situación supuestas las habituales condiciones de regularidad, se define, como cantidad de información de Fisher, la matriz $I_F(\theta) = \| I^{ij}(\theta) \|$ cuyos elementos son

$$I^{ij}(\theta) = E_\theta\left(\frac{\partial}{\partial\theta_i} \log f(X, \theta)\, \frac{\partial}{\partial\theta_j} \log f(X, \theta)\right).$$

En el caso M-dimensional es posible también definir la Información de Fisher contenida en Y acerca de θ cuando X = x. Esta viene dada por la matriz $I^{\xi}_{Y/X=x}(\theta) = \| I^{ij}_{Y/X=x}(\theta) \|$, donde

$$I^{ij}_{Y/X=x}(\theta) = E_{\theta}\left(\frac{\partial}{\partial\theta_i} \log f(Y/x, \theta) \frac{\partial}{\partial\theta_j} \log f(Y/x, \theta) \right).$$

La relación vista anteriormente en el caso uniparamétrico entre la información conjunta, condicionada y marginal se mantiene en el caso M-paramétrico.

Las propiedades de no negatividad, aditividad, invarianza y suficiencia, convexidad y pérdida de información debida al agrupamiento ponen de manifiesto que la matriz de información de Fisher posee las principales propiedades de una medida de información.

Sea θ un parámetro real y $\theta = g(\phi_1, \phi_2, \ldots, \phi_r)$ una transformación en un espacio unidimensional Θ, desde un espacio r-dimensional Φ. El caso r=1, es quizá el más común en la práctica. Supongamos que g es una función real que posee primeras derivadas parciales con respecto a sus argumentos. Un simple cálculo da lugar a:

$$I_F(\phi) = \| I^{ij}(\phi) \| = I(\theta) \left\| \frac{\partial g(\phi_1, \ldots, \phi_r)}{\partial\phi_i} \frac{\partial g(\phi_1, \ldots, \phi_r)}{\partial\phi_j} \right\|_{rxr}.$$

Si $\theta \in \mathbb{R}^M$ se obtiene de la misma forma la expresión de la matriz de información resultante, si bien ahora es más complicado. Fisher (1925), suponiendo igualdad en las dimensiones de los espacios paramétricos Θ y Φ, dio la fórmula general de reparametrización, $I_F(\phi) = A^t I_F(\theta) A$ donde

$$A = \| \partial g_i / \partial\phi_j \|_{kxr}.$$

En definitiva se tiene que la medida de información de Fisher no permanece invariante bajo transformaciones paramétricas.

Una cuestión resulta interesante: ¿Existe alguna reparametrización del modelo estadístico que haga que la información de Fisher permanezca insensible a los valores del parámetro? Es claro que estas parametrizaciones se obtendrán mediante la resolución de las ecuaciones diferenciales determinadas por las fórmulas de reparametrización.

La cuestión que se aborda ahora es la de definir medidas paramétricas de información a partir de medidas de divergencia. El primer método de construir

medidas paramétricas de información consiste en definirlas mediante la siguiente expresión:

$$(\theta) = \lim_{t \to 0} \inf \frac{1}{t^2} \mathfrak{D} [(f(x, \theta), f(x, \theta + t)] \qquad \theta \in \Theta$$

siendo $\mathfrak{D}$ una medida cualquiera de divergencia. Este procedimiento ha sido utilizado por Kagan (1963), Papaioannou and Kempthorne (1971), Vajda (1973), Aggarwal (1974), Boekee (1977), Ferentinos y Papaioannou (1981), Taneja (1987), Pardo, L. Taneja y Morales (1993), etc. Según sea la divergencia que se utilice así será la información paramétrica resultante. Estas medidas paramétricas resultantes verifican interesantes propiedades dependiendo de la medida de divergencia utilizada.

Kale (1964), Papaioannou and Kempthorne (1971) y Ferentinos y Papaioannou (1981) han utilizado otro procedimiento para construir medidas paramétricas de información. Sea $h(\theta)$ una transformación biyectiva del espacio paramétrico en sí mismo con $h(\theta) \neq 0$. La cantidad

$$^{*}I(\theta) = I [f(x,\theta), f (x,h(\theta))]$$

se puede considerar como una medida paramétrica basada en $h(\theta)$. En cuanto a las propiedades de estas medidas, se puede decir lo mismo que se había dicho de las medidas paramétricas definidas por el procedimiento anterior.

En el caso M-dimensional existen diversos procedimientos de obtención de matrices informativas basadas en divergencias. Siendo quizá el más interesante el basado en la consideración de las matrices informativas asociadas a métricas diferenciales. Atendiendo a que toda función de densidad $f(x,\theta)$ viene caracterizada por el parámetro $\theta \in \Theta$, se puede interpretar $\{f(x,\theta), \theta \in \Theta\}$ como una superficie en donde las componentes del parámetro $\theta = (\theta_1, \ldots, \theta_k)$ juegan el papel de coordenadas. En este contexto, para un valor fijo de parámetro $\theta \in \Theta$, cuando las M funciones $\frac{\partial f(x, \theta)}{\partial \theta_i}$, $i = 1, \ldots,$ M son independientes, entonces el espacio tangente T_θ en el punto θ es el que viene generado por

$$\left\{ \frac{\partial f(x, \theta)}{\partial \theta_i} \right\}_{i = 1, \ldots, M}.$$

En la dirección de este espacio tangente, es posible definir una métrica diferencial en la que el elemento diferencial de arco venga caracterizado por

$$ds^2(\theta) = \lim_{t \to 0} \frac{\mathfrak{D} [f(x, \theta), f(x, \theta) + tdf(x, \theta)]}{t^2}$$

siendo $\mathfrak{D}$ una medida de divergencia.

En caso de utilizar como divergencia las (h, ϕ)–divergencias en Pardo, L. y otros (1993c), se estableció que

$$ds^2(\theta) = \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{\Lambda} h'_{\alpha}(0) \frac{\phi''_{\alpha}(1)}{2} I^{ij}(\theta)\, d\theta_i\, d\theta_j.$$

Es obvio que cuando la matriz

$$IM^{h}_{\phi}(\theta) = \| h'_{\alpha}(0) \frac{\theta''_{\alpha}(1)}{2} I^{ij}(\theta) \|_{M \times M}$$

es definida positiva, entonces $ds^2(\theta)$ define una métrica Riemaniana, ya que $IM^{h}_{\phi}(\theta)$ define un tensor simétrico y covariante de segundo orden para todo $\theta \in \Theta$. En el contexto de la Teoría de la Información las matrices que definen la métrica diferencial se las denomina matrices informativas.

En lo que hace referencia al funcional de las R^{h}_{ϕ}–divergencias en Morales y otros (1993b) se obtiene que la diferencial de arco viene dada por la expresión

$$ds^2(\theta) = \sum_{i=1}^{M} \sum_{j=1}^{M} g_{ij}(\theta)\, d\theta_i\, d\theta_j$$

siendo,

$$g_{ij}(\theta) = -\frac{1}{8}\left\{ h''\left(\int_{\mathcal{X}} \phi\,[f(x,\theta)]\, d\mu\right)\left[\int_{\mathcal{X}} \phi'\,[f(x,\theta)] \frac{\partial f(x,\theta)}{\partial \theta_i}\, d\mu \int_{\mathcal{X}} \phi'\,[f(x,\theta)] \frac{\partial f(x,\theta)}{\partial \theta_j}\, d\mu\right] + \right.$$

$$\left. + h'\left(\int_{\mathcal{X}} \phi\,[f(x,\theta)]\, d\mu\right)\left[\int_{\mathcal{X}} \phi''\,[f(x,\theta)] \frac{\partial f(x,\theta)}{\partial \theta_i} \frac{\partial f(x,\theta)}{\partial \theta_j}\, d\mu\right]\right\}.$$

Rao (1945) fue el primero que puso de manifiesto la importancia de la geometría diferencial en estadística, introduciendo la métrica Riemaniana en un modelo estadístico al considerar los elementos de la matriz de información de Fisher como los elementos del tensor covariante de orden 2 que define la métrica. Además calculó la distancia geodésica entre dos distribuciones para diversos modelos estadísticos. Las propiedades y aplicabilidad de la distancia de Rao en modelos estadísticos ha tenido y sigue teniendo un amplio eco y son muchos los trabajos desarrollados en esta dirección. Así Campbell (1985, 1987) hace un desarrollo de los resultados clásicos de Teoría de la Información utilizando la geometría diferencial. Amari (1984, 1985) aborda diversos problemas estadísticos usando métricas diferenciales. Burbea (1986) define una geometría informativa en espacios de probabilidad. Atkinson y Mitchell (1981) obtienen el valor de la distancia geodésica para distribuciones unidimensionales, multinomiales y

normales con el mismo vector de medias o misma matriz de varianzas-covarianzas. Oller y Cuadras (1985) y Oller (1989) calculan las distancias para distribuciones logísticas de valores extremos y multinomiales negativas y acotan la distancia para distribuciones normales multivariantes.

A la vista de las propiedades de las matrices de información, se puede concluir que poseen las principales propiedades de una medida de información y en ello radica su importancia. Sin embargo, pensemos en el problema de evaluar la posible pérdida de información que, acerca del parámetro, se produce cuando aparece una variable de censura, o aquel otro en que, entre varios experimentos con el mismo espacio paramétrico deseamos elegir el que proporciona mayor información acerca de dicho parámetro. En estos y otros muchos problemas nuestro interés será cuantificar la información, y en este sentido, la matriz de Fisher es inoperante y por consiguiente insuficiente para nuestros objetivos.

A esta insuficiencia práctica podemos añadir otra intuitiva de la matriz de información de Fisher como medida de información. Así, mientras los elementos de la diagonal principal de dicha matriz, $I^{ii}(\theta)$, son las medidas de información de Fisher (unidimensionales) acerca de cada parámetro θ_i, dados los valores del resto de parámetros θ_j, $j \neq i$; no existe, sin embargo, una interpretación teórica análoga en términos de información para los elementos que están fuera de la diagonal principal. Parece pues, difícil medir la información acerca de un parámetro M-variante por medio de una matriz MxM, algunos de cuyos elementos no tienen connotación de información.

A este respecto se han propuesto dos procedimientos para solventar estas dos dificultades. El primer procedimiento consiste en definir una medida real asociada a la matriz de Fisher con una interpretación intuitiva de información y como tal, buenas propiedades.

Turrero (1988) propone como medidas escalares de información, cuando θ es M-variante, las siguientes funciones reales:

a) $A_X(\theta) = f_1(\lambda_1, \lambda_2, \ldots, \lambda_M) = \alpha \sum_{i=1}^{M} \lambda_i$, α constante, $\alpha > 0$

b) $B_X(\theta) = f_2(\lambda_1, \lambda_2, \ldots, \lambda_M) = \left(\prod_{i=1}^{M} \lambda_i \right)^{\alpha}$, α constante, $\alpha > 0$

c) $C_X(\theta) = f_3(\lambda_1, \lambda_2, \ldots, \lambda_M) = \left(\sum_{i=1}^{M} \lambda_i^2 \right)^{1/2}$.

donde los λ_i son los autovalores de la matriz de información de Fisher.

Ferentinos y Papaioannou (1981), basándose en las propiedades que como medidas de información poseen, proponen las siguientes medidas: a) Traza

de $I_F(\theta)$; b) Determinante de $I_F(\theta)$; c) λ_i, i-ésimo autovalor de $I_F(\theta)$; d) $L = \sum_{i=1}^{M} w_i \lambda_i$, $w_i > 0$, $i = 1,\ldots, M$. Lógicamente parecen más razonables las dos primeras que las dos últimas. El autovalor i-ésimo no parece una buena medida de información pues sólo reflejará una parte de la dispersión de $I_F(\theta)$. De elegir algún autovalor parece lógico elegir el mayor. Respecto de la medida L, no es simétrica respecto de los autovalores, salvo cuando $w_i = w$, $i = 1,\ldots,M$, que coincide con $A_X(\theta)$.

El segundo procedimiento consiste en construir medidas de información asociadas a divergencias en base a la perturbación de todos los parámetros. En Pardo, L. y otros (1993c) se propone considerar la medida que resulta de perturbar el parámetro en todas sus componentes. Así, se considera la cantidad de información definida mediante la expresión

$$\mathrm{Inf}\left(f(x, \theta)\right) = \lim_{t \to 0} \inf \frac{\mathfrak{D}\left(f(x, \theta), f[x,(\theta_1 + t,..,\theta_M + t)] \right)}{t^2}.$$

En el caso de la $(\underline{h}, \underline{\phi})$–Divergencia se obtenía el siguiente resultado:

$$\mathrm{Inf}_{\phi}^{h}\left(f(x, \theta) \right) = \left\{ \int_{\Lambda} h'_{\alpha}(0) \frac{\phi''_{\alpha}(1)}{2} d\eta \right\} \sum_{i,j=1}^{M} I^{ij}(\vec{\theta}).$$

Además se encuentra la distribución asintótica del estadístico $\mathrm{Inf}_{\phi}^{h}\left(f(x, \hat{\theta}) \right)$, donde $\hat{\theta}$ es el estimador de máxima verosimilitud.

En Morales y otros (1993b), se obtiene la cantidad de información asociada a las R_{ϕ}^{h}–divergencias.

6. LAS MEDIDAS DE ENTROPIA COMO INDICE DE DIVERSIDAD. COMPORTAMIENTO ASINTOTICO

El problema de encontrar la distribución asintótica de las medidas de entropía y divergencia, para su posterior utilización en el análisis de datos categorizables, ha sido y sigue siendo un problema de interés entre los estadísticos. En lo que hace referencia a la utilización de las medidas de entropía en el análisis de datos categorizables, hay que destacar la importancia que éstas han tenido como índices de diversidad o concentración.

Considérese una población finita constituida por N elementos que, de acuerdo con un proceso X, puede clasificarse en M categorías o clases $x_1, \ldots, x_M$. Se denota por $\aleph$ el conjunto formado por las M clases o categorías; $\aleph = \{x_1,\ldots, x_M\}$.

Sea $\Delta_M = \{P = (p_1,\ldots, p_M) / p_i \geq 0, i = 1,\ldots, M, \sum_{i=1}^{M} p_i = 1\}$, el conjunto de todas las distribuciones de probabilidad definidas sobre $\mathcal{X}$. Se denomina índice de diversidad (Rao, 1982) a toda función $I: \Delta_M \longrightarrow \mathbb{R}$, verificando:

i) $I(P) \geq 0, \supset P \in \Delta_M$ con $I(P) = 0$ si y sólo si P es degenerada.

ii) I es una función cóncava.

Intuitivamente un índice de diversidad permite cuantificar la variabilidad en una población con respecto a un proceso de clasificación en términos del número de clases y de la probabilidad de cada una de esas clases.

En este sentido, Patil y Taille (1982) y Pielou (1975) utilizan la entropía de Shannon como índice de diversidad en estudios de ecología, Lewontin (1972) considera la entropía de Shannon en estudios de Biología, Agresti y Agresti (1979) utilizan las entropías de Havrda-Charvat de grado 2 en estudios sociológicos, Lieberson (1969) en estudios de Economía, Greenberg (1956) en lingüística, etc. Rao (1982) planteó la posibilidad de usar otras medidas de entropía distintas de la de Shannon como índices de diversidad. Posteriormente, Rao (1982) investigó esta posibilidad; siendo Nayak (1985) quien estudió el comportamiento asintótico del estimador analógico de diversas medidas de entropía en un muestreo aleatorio simple. Estos resultados fueron extendidos por Gil, M.A. (1989) al muestreo aleatorio estratificado con asignación proporcional en el caso de la entropía de Renyi, y por Pardo, L. y otros (1992) al caso de entropías generalizadas.

Para tener una mayor generalidad se presentarán los resultados en poblaciones generales y luego se particularizarán a poblaciones multinomiales tanto con muestreo aleatorio simple como estratificado.

Sea $(\mathcal{X}, \beta_{\mathcal{X}}, P_\theta)_{\theta \in \Theta}$ un espacio estadístico con Θ abierto de $\mathbb{R}^M$, denotemos por $f(x,\theta)$ la densidad o probabilidad respecto de una medida σ-finita μ, si $\hat{\theta}$ es un estimador CAN de θ; es decir,

$$n^{1/2} (\hat{\theta} - \theta) \xrightarrow{L} N [0, \Sigma(\theta)]$$

se verifica:

a) $$n^{1/2} \left(H_n^{\phi} (f (x, \hat{\theta})) - H_n^{\phi} (f (x, \hat{\theta})) \right) \xrightarrow[n \to \infty]{L} N (0, \sigma^2)$$

siendo $\sigma^2 = T^t \Sigma(\theta)\, T > 0$ y $T^t = (t_1, \ldots, t_M)$, con

$$t_j = h' \left(\int_{\mathcal{X}} \phi\, [f(x, \theta)]\, d\mu \right) \int_{\mathcal{X}} \phi' [f(x, \theta)] \frac{\partial f(x, \theta)}{\partial \theta_j}\, d\mu$$

b) Si $T^t \Sigma(\theta) T = 0$, entonces

$$2n \left(H_h^{\phi} (f(x, \hat{\theta})) - H_h^{\phi} (f(x, \theta)) \right) \xrightarrow[n \to \infty]{L} \sum_{i=1}^{M} \beta_i \chi^2_{1,i}$$

siendo $\chi^2_{1,i}$ distribuciones Ji-cuadrados independientes con un grado de libertad y β_i los autovalores de la matriz $C\Sigma(\theta)$ con $C = (c_{ij})_{i,j=1,\ldots,M}$ y

$$c_{ij} = h'' \left(\int_{\mathcal{X}} \phi\, [f(x, \theta)]\, d\mu \right) \int_{\mathcal{X}} \phi' [f(x, \theta)] \frac{\partial f(x, \theta)}{\partial \theta_i}\, d\mu \int_{\mathcal{X}} \phi' [f(x, \theta)] \frac{\partial f(x, \theta)}{\partial \theta_i}\, d\mu +$$

$$+ h' \left(\int_{\mathcal{X}} \phi [f(x, \theta)]\, d\mu \right) \left\{ \int_{\mathcal{X}} \phi'' [f(x, \theta)] \frac{\partial f(x, \theta)}{\partial \theta_i} \frac{\partial f(x, \theta)}{\partial \theta_j}\, d\mu + \int_{\mathcal{X}} \phi' [f(x, \theta)] \frac{\partial f(x, \theta)}{\partial \theta_j}\, d\mu \right\}$$

Estos resultados permiten resolver entre otros los siguientes contrastes:

1) **Adherencia a un valor prefijado,** H_0: $H_h^{\phi} [f(x, \theta)] = D_0$. En este caso, el estadístico a utilizar es

$$Z = n^{1/2} \left(\frac{H_h^{\phi} (f(x, \hat{\theta})) - D_0}{\hat{\sigma}} \right)$$

que se distribuye asintóticamente según una normal cero uno y $\hat{\sigma}$ es el valor obtenido en a) al sustituir θ por $\hat{\theta}$.

2) **Igualdad de r-diversidades a un valor prefijado,**

$$H_0: H_h^{\phi} [f(x, \theta^1)] = \ldots = H_h^{\phi} [f(x, \theta^r)] = D_0$$

En este caso, el estadístico a utilizar es

$$Z = \sum_{j=1}^{r} n_j \frac{\left(H_h^{\phi} (f(x, \hat{\theta}_j)) - D_0 \right)^2}{\hat{\sigma}_j^2}$$

que se distribuye asintóticamente según una Ji-cuadrado con r grados de libertad, siendo n_j el tamaño de la muestra utilizado para estimar la (h,ϕ)-entropía en la población j-ésima.

3) **Igualdad de r-diversidades,**

$$H_0: H_h^{\phi} [f(x, \theta^1)] = \ldots = H_h^{\phi} [f(x, \theta^r)]$$

En este caso, el estadístico a utilizar es

$$Z = \sum_{j=1}^{r} \frac{n_j \left(H_h^{\phi} (f(x, \hat{\theta}_j)) - \overline{H} \right)^2}{\hat{\sigma}_j^2}$$

que se distribuye asintóticamente según una Ji-cuadrado con r – 1 grados de libertad. Siendo,

$$\overline{H} = \left(\sum_{j=1}^{r} \frac{n_j H_h^{\phi} [f(x, \hat{\theta}_j)]}{\hat{\sigma}_j^2} \right) \Big/ \left(\sum_{j=1}^{r} \frac{n_j}{\hat{\sigma}_j^2} \right)$$

Conviene señalar que para aquellos casos en los que la (h,ϕ)-entropía sea función biyectiva de los parámetros, entonces, las hipótesis

$$H_0: \theta = \theta^0,\ H_0: \theta^1 = \theta^2 = \ldots = \theta^r = \theta^0 \text{ y } H_0: \theta^1 = \theta^2 = \ldots = \theta^r$$

pueden resolverse a partir de los contrastes

$$H_0 = H_h^{\phi} [f(x, \theta)] = H_h^{\phi} [f(x, \theta^0)],\ H_0: H_h^{\phi} [f(x, \theta^1)] = \ldots = H_h^{\phi} [f(x, \theta^r)] = H_h^{\phi} [f(x, \theta^0)],$$

y $H_0: H_h^{\phi} [f(x, \theta^1)] = \ldots = H_h^{\phi} [f(x, \theta^r)]$, respectivamente.

Una situación especialmente interesante se presenta al contrastar, si las varianzas de r poblaciones normales son iguales utilizando la entropía de Shannon. En este caso, se llega a rechazar las hipótesis nula si se verifica

$$\frac{1}{2} \sum_{j=1}^{r} n_j \left(\log \frac{\hat{\sigma}_j^2}{\prod_{j=1}^{r} \left(\hat{\sigma}_j^2 \right)^{n_j / N}} \right)^2 > \chi^2_{r-1,\,\alpha} \qquad \left(N = \sum_{j=1}^{r} n_j \right).$$

Este contraste está siendo objeto de estudio por nuestra parte en la actualidad. En concreto se está llevando a cabo un análisis similar al efectuado por Conover, W. J.; Johnson, M. E. and M. M. Johnson (1981) con los tests de Bartlett, Cochran, Bartlett y Kendall, Hartley, Cadwell, Box, Levene, etc.

Si $\hat{\theta}$ es el estimador de máxima verosimilitud entonces la matriz $\Sigma(\theta)$ sería la inversa de la matriz de información de Fisher. Un amplio estudio de estos resultados pueden verse en Salicrú (1993).

Sea ahora $\Theta = \left\{ P = (p_1,\ldots, p_M) / p_i > 0 \text{ y } \sum_{i=1}^{M} p_i = 1 \right\}$, $\mathcal{X} = \{ x_1,\ldots, x_M \}$, μ una medida contable dando masa uno a cada x_i y $f(x,\theta_i) = (dP_\theta/d\mu)(x_i) = p_i$, $i = 1,\ldots,M$. Dado $P \in \Theta$ la (h,ϕ)-entropía asociada a P viene definida mediante la expresión

$$H_h^\phi (P) = h\left(\sum_{i=1}^{M} \phi (p_i) \right).$$

Si se considera el estimador de máxima verosimilitud de P, $\hat{P}$, basado en una muestra aleatoria simple de tamaño n, se tiene

$$n^{1/2} (\hat{P}^t - P^t) \xrightarrow[n \to \infty]{L} N (0, \Sigma_p) .$$

Siendo $\Sigma_P=[p_i(\delta_{ij}-p_j)]_{i,j=1,\ldots,M}$ la inversa de la matriz de información de Fisher. Además,

$$T^t \Sigma(\theta) T = T^t \Sigma_p T = \left[h'\left(\sum_{i=1}^{M} \phi (p_i) \right) \right]^2 \left\{ \sum_{i=1}^{M} [\phi'(p_i)]^2 p_i - \left(\sum_{i=1}^{M} \phi' (p_i)\, p_i \right)^2 \right\}$$

siendo,

$$t_i = h'\left(\sum_{i=1}^{M} \phi (p_i) \right) \phi' (p_i),\ i = 1,\ldots, M .$$

Si $p_1 = p_2 = \ldots = p_M = \frac{1}{M}$, $T^t \Sigma_P T = 0$ y en este caso se obtiene,

$$\frac{2n\left(H_h^\phi (\hat{P}) - h\left[M\phi\left(\frac{1}{M} \right) \right] \right)}{h'\left[M\phi\left(\frac{1}{M} \right) \right] \phi''\left(\frac{1}{M} \right)} \xrightarrow[n \to \infty]{L} \chi^2_{M-1}$$

Este resultado permitirá efectuar contrastes de bondad de ajuste, a una distribución conocida, sin más que considerar una partición equiprobable del soporte de la distribución considerada.

Ahora, se supone que la población considerada en M clases se puede dividir en L estratos (subpoblaciones homogéneas y no solapadas), de forma que $W_1,\ldots,W_L$ sean los tamaños relativos de cada uno de los estratos.

Sea p_{il} la probabilidad de que un individuo seleccionado al azar de la población pertenezca a la clase i-ésima y al estrato l-ésimo, $p_{i.}$ la probabilidad de que un

individuo seleccionado al azar en la población pertenezca a la clase x_i y $p_{.l}$ la probabilidad de que un individuo seleccionado al azar en la población pertenezca al estrato l-ésimo. Entonces,

$$p_{i.} = \sum_{l=1}^{L} p_{il},\ p_{.l} = W_l = \sum_{i=1}^{M} p_{il},\ \sum_{l=1}^{L}\sum_{i=1}^{M} p_{il} = 1,\ \sum_{l=1}^{L} W_l = 1\ .$$

En estas condiciones, la (h,ϕ)-entropía poblacional puede cuantificarse mediante

$$H_h^{\phi}(P_{.}) = h\left(\sum_{i=1}^{M} \phi(p_{i.})\right).$$

Para obtener una estimación de la (h,ϕ)-entropía poblacional, procederemos a la obtención de una muestra estratificada de tamaño n, donde las submuestras correspondientes a los distintos estratos se tomarán de forma independiente y de acuerdo con una afijación predeterminada $w_1, \ldots, w_L$. En este contexto se estimará la (h,ϕ)-entropía mediante el estimador.

$$H_h^{\phi}(\hat{P}_{.}) = h\left[\sum_{i=1}^{M} \phi\,(\hat{p}_{i.})\right] = h\left[\sum_{i=1}^{M} \phi\left(\sum_{l=1}^{L} \frac{W_l}{w_l}\,\hat{p}_{il}\right)\right]$$

siendo $\hat{p}_{il}$ la frecuencia relativa de la intersección del estrato l-ésimo con la clase x_i y $\hat{p}_{i.} = \sum_{l=1}^{L} \frac{W_l}{w_l}\,\hat{p}_{il}$ el estimador insesgado de $p_{i.}$ en muestreo estratificado.

Ahora bien como se verifica,

$$n^{1/2}\,(\hat{p}_{1.} - p_{1.}, \ldots, \hat{p}_{M.} - p_{M.}) \xrightarrow[n \to \infty]{L} N\,(0, \Sigma^*)$$

siendo

$$\Sigma^* = \sum_{l=1}^{L} \frac{W_l^2}{w_l}\,\Sigma(l)$$

con

$$\Sigma(l) = \left[\frac{p_{il}}{W_l}\left(\delta_{ij} - \frac{p_{sl}}{W_l}\right)\right]_{i,\,j\,=\,1,\ldots,\,M}$$

se sigue que,

$$n^{1/2}\left(H_h^{\phi}(\hat{P}_{..}) - H_h^{\phi}(P_{..}) \right) \xrightarrow[n \to \infty]{L} N(0, {}^{st}\sigma^2)$$

donde

$${}^{st}\sigma^2 = \sum_{l=1}^{L} \frac{W_l}{w_l} \left\{ \sum_{i=1}^{M} \left[h'\left(\sum_{i=1}^{M} \phi(p_{i.}) \right) \phi'(p_{i.}) \right]^2 p_{il} - \frac{1}{W_l} \left[\sum_{i=1}^{M} h'\left(\sum_{i=1}^{M} \phi(p_i) \right) \phi'(p_{i.}) p_{il} \right]^2 \right\} > 0$$

b) Cuando $p_{1.} = p_{2.} = \ldots = p_{M.} = \frac{1}{M}$, entonces se tiene

$$\frac{2n\left(H_h^{\phi}(\hat{P}_{..}) - H_h^{\phi}(P_{..}) \right)}{h'\left[M\phi\left(\frac{1}{M} \right) \right] \phi''\left(\frac{1}{M} \right)} \xrightarrow[n \to \infty]{L} \sum_{i=1}^{M} \beta_i \chi^2_{1,i}$$

siendo los β_i los autovalores de la matriz

$$\left[\sum_{l=1}^{L} \frac{W_l}{w_1} p_{il} \left(\delta_{ij} - \frac{p_{jl}}{W_l} \right) \right]_{i,j=1,\ldots,M}$$

Una cuestión interesante es la obtención de la asignación óptima en el sentido de Neyman. En este caso se tiene que: Para un tamaño muestral fijo n, la varianza asintótica ${}^{st}\sigma^2$ se minimiza para

$$w_l = \frac{\alpha_l^{1/2}}{\sum_{l=1}^{L} \alpha_l^{1/2}} \qquad l = 1, \ldots, L$$

siendo

$$\alpha_l = W_l \left\{ \sum_{i=1}^{M} \left[h'\left(\sum_{i=1}^{M} \phi(p_{i.}) \right) \phi'(p_{i.}) \right]^2 p_{il} - \frac{1}{W_l} \left[\sum_{i=1}^{M} \left(h'\left(\sum_{i=1}^{M} \phi(p_{i.}) \right) \phi'(p_{i.}) p_{il} \right]^2 \right\}.$$

Finalmente conviene señalar que el paso a muestreo estratificado supone una ganancia en precisión. En este sentido se tiene el siguiente resultado,

$${}^{st}\sigma^2_{MINIMA} \leq {}^{st}\sigma^2_{PROPORCIONAL} \leq {}^{st}\sigma.$$

En Salicrú y otros (1993b) se hace un amplio estudio de las (h,ϕ)-entropías en muestreo estratificado.

7. MEDIDAS DE DIVERGENCIA Y CONTRASTES DE HIPOTESIS

Kupperman (1957) estableció que la distribución asintótica del estadístico

$$2nD(\hat{\theta}, \theta) = 2n \int_{\mathcal{X}} f(x, \hat{\theta}) \log \frac{f(x, \hat{\theta})}{f(x, \theta)} d\mu(x) ,$$

es una Ji-cuadrado con M grados de liberad siendo $\hat{\theta} = (\hat{\theta}_1,\ldots, \hat{\theta}_M)$ el estimador de máxima verosimilitud de $\theta=(\theta_{1,\ldots,}\theta_M)$. Este resultado permite contrastar hipótesis del tipo H_0: $\theta=\theta_0$. Para una distribución normal K-dimensional, $\theta = (\mu_i, \sigma_{ii}, \sigma_{ij}; i=1,\ldots,K, j=1,\ldots,K, j>i)$, se puede contrastar H_0: $(\mu,\Sigma) = (\mu_0, \Sigma_0)$, donde $\mu=(\mu_1,\ldots,\mu_K)t$ es el vector de medias y $\Sigma = (\sigma_{ij})$ es la matriz de varianzas covarianzas.

Para dos muestras, Kupperman (1957) estableció que si $\theta_1=\theta_2$, entonces

$$\frac{2\,mn}{m+n} D(\hat{\theta}_1, \hat{\theta}_2) \xrightarrow[n \to \infty]{L} \chi^2_{M'}$$

donde $\hat{\theta}_1 = (\hat{\theta}_{11},\ldots, \hat{\theta}_{1M})$ y $\hat{\theta}_2 = (\hat{\theta}_{21},\ldots, \hat{\theta}_{2M})$ son los estimadores de máxima verosimilitud de $\theta_1=(\theta_{11,\ldots,}\theta_{1M})$ y $\theta_2=(\theta_{21,\ldots,}\theta_{2M})$ basados en muestras independientes de tamaños n y m respectivamente. Este resultado permite contrastar hipótesis del tipo H_0: $\theta_1=\theta_2$. Para dos poblaciones normales K-dimensionales, se puede contrastar H_0: $(\mu_1,\Sigma_1)= (\mu_2, \Sigma_2)$, es decir, el contraste de homogeneidad completa. Es claro que existen otras posibilidades de contraste que no son cubiertas con las dos distribuciones asintóticas dadas por Kupperman. Así, en el caso de una muestra no se puede contrastar H_0: $\Sigma=\Sigma_0$ (cuando μ es desconocida) o H_0: $\mu=\mu_0$ (cuando Σ es desconocida) y en el caso de dos muestras H_0: $\mu_1=\mu_2$ (cuando Σ_1 y Σ_2 son iguales pero desconocidas), H_0: $\Sigma_1=\Sigma_2$ (cuando μ_1 y μ_2 son iguales pero desconocidas), H_0: $\mu_1 = \mu_2^*$ y $\Sigma_1 = \Sigma_2$ (cuando μ_2^* es un valor prefijado de μ_2) y H_0: $\mu_1 = \mu_2^*$ y $\sigma_{1ii} = \sigma_{2ii}$ (cuando σ_{1ij} y σ_{2ij} son iguales pero desconocidas para cada i<j).

En Salicrú y otros (1993a) se da solución a los problemas enunciados anteriormente, utilizando las (h, ϕ)–divergencias. Por un lado se abordan estos nuevos problemas y por otro se da una metodología general que puede utilizarse con las medidas de divergencia que se obtienen como caso particular de las (h, ϕ)–divergencias. De forma más precisa se supone que $\theta_1=(\theta_{11},\ldots,\theta_{1M})$ es desconocido y que

$$\theta_2 = (\theta_{21},\ldots, \theta_{2k}, \theta_{2(k+1)},\ldots, \theta_{2M_0}, \theta^*_{(M_0+1)},\ldots, \theta^*_M)$$

es parcialmente conocido. Así, $\theta_{2i}=\theta_{1i}$ si y sólo si $i \in I_1=\{1,2,\ldots,k\}$, θ_{2i} es desconocido y diferente de θ_{1i} cuando $i \in I_2 = \{k + 1,\ldots,M_0\}$ y θ_{2i} es conocido e igual a θ_i^* cuando $i \in I_3 = \{M_0 + 1,\ldots, M\}$. Por tanto el espacio paramétrico conjunto, Γ,

es un subconjunto abierto de $\mathbb{R}^{M+M_0-k}$, sin embargo, si se añade la hipótesis $\theta_1=\theta_2$, el espacio paramétrico conjunto, Γ_0, es un subconjunto abierto de $\mathbb{R}^{M_0}$. Además los elementos del espacio paramétrico $\gamma = (\gamma_1,\ldots, \gamma_{M+M_0-k}) \in \Gamma$, son de la forma siguiente

$$\gamma_i = \begin{cases} \theta_{1i} & \text{si } 1 \le i \le M \\ \theta_{2(i-M+k)} & \text{si } M+1 \le i \le M+M_0-k \end{cases}$$

es decir, $\gamma = (\theta_{11},\ldots, \theta_{1M}, \theta_{2(k+1)},\ldots, \theta_{2M_0})$.

De cada población se obtienen sendas muestras independientes de tamaños n y m respectivamente. Sea $\hat{\theta}_{1i}$; $\hat{\theta}_{2j}$, $i = 1,\ldots, M$, $j = k+1,\ldots, M_0$, el estimador que maximiza el logaritmo de la verosimilitud conjunta $\log L(\gamma) = \sum_{i=1}^{m} \log f(x_i, \theta_1) + \sum_{i=1}^{m} \log f(x_i, \theta_2)$ y sea $\hat{\gamma} = (\hat{\gamma}_1,\ldots, \hat{\gamma}_{M+M_0-k})$ el estimador de máxima verosimilitud de γ, es decir,

$$\hat{\gamma}_i = \begin{cases} \hat{\theta}_{1i} & \text{si } 1 \le i \le M \\ \hat{\theta}_{2(i-M+k)} & \text{si } M+1 \le i \le M+M_0-k\,. \end{cases}$$

La distribución asintótica del estadístico $D_\phi^h(\hat{\theta}_1, \hat{\theta}_2)$, donde

$$\hat{\theta}_1 = (\hat{\theta}_{11},\ldots, \hat{\theta}_{1M})^t, \qquad \hat{\theta}_2 = (\hat{\theta}_{21},\ldots, \hat{\theta}_{2k}, \hat{\theta}_{2(k+1)},\ldots, \hat{\theta}_{2M_0}, \theta^*_{(M_0+1)},\ldots,\theta^*_M)^t, \qquad \hat{\theta}_{2i} = \hat{\gamma}_i$$

si $1 \le i \le k$ y $\hat{\theta}_{2i} = \theta_i^*$ si $M_0 + 1 \le i \le M$, viene dada en los siguientes términos:

Resultado 1

Supuestas las condiciones de regularidad estandar, se verifica:

(a) Si $\frac{m}{m+n} \xrightarrow[m,n\to\infty]{} \lambda \in (0, 1)$ y $T^t \Sigma (\theta_1, \theta_2)^{-1} T > 0$, entonces

$$\left[\frac{mn}{m+n}\right]^{1/2} [D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) - D_\phi^h(\theta_1, \theta_2)] \xrightarrow[n,m\to\infty]{L} N(0, T^t \Sigma (\theta_1, \theta_2)^{-1} T)$$

(b) Si $\theta_1 = \theta_2$, entonces

$$\frac{2nm}{m+n} D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) \int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta \xrightarrow[n,m\to\infty]{L} \sum_{i=1}^{M-k} \beta_i \chi_1^2,$$

donde las χ^2_1 son independientes y los β_i son los autovalores de la matriz $A\Sigma_\beta$. además, $T^t = (t_1, \ldots, t_{M+M_0-k})$, con

$$t_i = \int_\Lambda \left[h'_\alpha \left\{ \int_{\mathcal{X}} f(x, \theta_2)\, \phi_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) d\mu - \phi_\alpha(1) \right\} \int_{\mathcal{X}} \left\{ \frac{\partial f(x, \theta_2)}{\partial \theta_{1\,i}} - \phi_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) + \right. \right.$$

$$\left. \left. + \frac{\partial f(x, \theta_1)}{\partial \theta_{1\,i}} \phi'_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) - \phi'_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) \frac{\partial f(x, \theta_2)}{\partial \theta_{1\,i}} \frac{f(x, \theta_1)}{f(x, \theta_2)} \right\} d\mu \right] d\eta$$

si $1 \le i \le k$,

$$t_i = \int_\Lambda \left[h'_\alpha \left\{ \int_{\mathcal{X}} f(x, \theta_2)\, \phi_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) d\mu - \phi_\alpha(1) \right\} \int_{\mathcal{X}} \left\{ \frac{\partial f(x, \theta_1)}{\partial \theta_{1\,i}} - \phi_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) \right\} d\mu \right] d\eta$$

si $k + 1 \le i \le M$,

$$t_i = \int_\Lambda \left[h'_\alpha \left\{ \int_{\mathcal{X}} f(x, \theta_2)\, \phi_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) d\mu - \phi_\alpha(1) \right\} \int_{\mathcal{X}} \left\{ \frac{\partial f(x, \theta_2)}{\partial \theta_{2,i-M+k}} \phi_\alpha \left(\frac{f(x, \theta_1)}{t(x, \theta_2)} \right) - \right. \right.$$

$$\left. \left. - \phi'_\alpha \left(\frac{f(x, \theta_1)}{f(x, \theta_2)} \right) \frac{\partial f(x, \theta_2)}{\partial \theta_{2,i-M+k}} \frac{f(x, \theta_1)}{f(x, \theta_2)} \right\} d\mu \right] d\eta \qquad \text{si } M + 1 \le i \le M + M_0 - k$$

Mientras que las matrices $\Sigma(\theta_1, \theta_2)$, Σ_β y A tienen por expresiones,

$$\Sigma(\theta_1, \theta_2) = \begin{pmatrix} \frac{1}{\lambda_{1,k}} {}^{1,k}_{1,k} I_F(\theta_1) + \frac{1}{1-\lambda_{1,k}} {}^{1,k}_{1,k} I_F(\theta_2) & \frac{1}{\lambda_{k+1,M}} {}^{1,k}_{k+1,M} I_F(\theta_1) & \frac{1}{1-\lambda_{k+1,M_0}} {}^{1,k}_{k+1,M_0} I_F(\theta_2) \\ \frac{1}{\lambda_{1,k}} {}^{k+1,M}_{1,k} I_F(\theta_1) & \frac{1}{\lambda_{k+1,M}} {}^{k+1,M}_{k+1,M} I_F(\theta_1) & 0 \\ \frac{1}{1-\lambda_{1,k}} {}^{k+1,M_0}_{1,k} I_F(\theta_2) & 0 & \frac{1}{1-\lambda_{k+1,M_0}} {}^{k+1,M_0}_{k+1,M_0} I_F(\theta_2) \end{pmatrix}$$

(${}^{i,j}_{r,s} I_F(\theta)$ es la submatriz de dimensión $(j - i + 1) \times (r - s + 1)$, cuyas filas son la i, i + 1, ..., j cuyas columnas son la r, r + 1, ..., s, de la matriz de Información de Fisher, $I_F(\theta)$)

$$\Sigma_\beta = \begin{pmatrix} \lambda_{k+1,M} {}^{k+1,M}_{k+1,M} I_F(\theta_2)^{-1} & 0 \\ 0 & (1-\lambda)_{k+1,M_0} {}^{k+1,M_0}_{k+1,M_0} I_F(\theta_2)^{-1} \end{pmatrix}$$

y

$$A = \begin{pmatrix} {}^{k+1,\,M_0}_{k+1,\,M_0} I_F(\theta) & {}^{k+1,\,M_0}_{M_0+1,\,M} I_F(\theta) & -\left({}^{k+1,\,M_0}_{k+1,\,M_0} I_F(\theta)\right) \\ {}^{M_0+1,\,M}_{k+1,\,M_0} I_F(\theta) & {}^{M_0+1,\,M}_{M_0+1,\,M} I_F(\theta) & -\left({}^{M_0+1,\,M}_{k+1,\,M_0} I_F(\theta)\right) \\ -\left({}^{k+1,\,M_0}_{k+1,\,M_0} I_F(\theta)\right) & -\left({}^{k+1,\,M_0}_{M_0+1,\,M} I_F(\theta)\right) & {}^{k+1,\,M_0}_{k+1,\,M_0} I_F(\theta) \end{pmatrix},$$

respectivamente.

Un caso especialmente importante del resultado 1 se presenta cuando $M_0 = M$, es decir, cuando el parámetro de la segunda distribución es $\theta_2 = (\theta_{21},\ldots,\theta_{2k},\theta_{2(k+1)},\ldots,\theta_{2M})$, donde $\theta_{21},\ldots,\theta_{2k}$ son desconocidos e iguales a $\theta_{11},\ldots,\theta_{1k}$ mientras $\theta_{2(k+1)},\ldots,\theta_{2M}$ son desconocidos y diferentes, en general, de los parámetros de la primera distribución.

i) Si las k primeras componentes de los parámetros θ_1 y θ_2 son desconocidas e iguales y las M-k restantes son desconocidas, se tiene que

(a) Si $\frac{m}{m+n} \xrightarrow[m,n\to\infty]{} \lambda \in (0, 1)$ y $T^t \Sigma(\theta_1, \theta_2)^{-1} T > 0$, entonces

$$\left(\frac{mn}{m+n}\right)^{1/2} \left(D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) - D_\phi^h(\theta_1, \theta_2)\right) \xrightarrow[m,n\to\infty]{L} N[0, T^t \Sigma(\theta_1, \theta_2)^{-1} T]$$

donde $T = (t_1,\ldots,t_{2M-k})^t$ y los t_i vienen dados en el resultado 1.

(b) Si $\frac{m}{m+n} \xrightarrow[m,n\to\infty]{} \lambda \in (0, 1)$ y $\theta_1 = \theta_2$, entonces

$$\frac{2nm}{m+n} \frac{D_\phi^h(\hat{\theta}_1, \hat{\theta}_2)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta} \xrightarrow[n,m\to\infty]{L} \chi^2_{M-k}.$$

Otro caso interesante se presenta cuando no se posee ninguna información previa sobre ambos parámetros, $M_0 = M$ y $k = 0$, en cuyo caso se tiene:

ii) Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ los estimadores de máxima verosimilitud de θ_1 y θ_2. Entonces

(a) Si $\frac{m}{m+n} \xrightarrow[m,n\to\infty]{} \lambda \in (0, 1)$ y $\lambda T^t I_F(\theta_1)^{-1} T + (1-\lambda) S^t I_F(\theta_2)^{-1} S) > 0$,

entonces

$$\left(\frac{m\,n}{m+n}\right)^{1/2}[D_\phi^h(\hat\theta_1,\hat\theta_2)-D_\phi^h(\theta_1,\theta_2)]\xrightarrow[m,n\to\infty]{L}N(0,\lambda\,T^t\,I_F(\theta_1)^{-1}T+(1-\lambda)\,S^t\,I_F(\theta_2)^{-1}\,S,$$

donde $T=(t_1,\ldots,t_M)^t$ y $S=(s_1,\ldots,s_M)^t$ con

$$t_i=\int_\Lambda\left[h'_\alpha\left\{\int_{\mathcal{X}}f(x,\theta_2)\,\phi_\alpha\left(\frac{f(x,\theta_1)}{f(x,\theta_2)}\right)d\mu-\phi_\alpha(1)\right\}\int_{\mathcal{X}}\left\{\frac{\partial f(x,\theta_1)}{\partial\theta_{1\,i}}\,\phi_\alpha\left(\frac{f(x,\theta_1)}{f(x,\theta_2)}\right)\right\}d\mu\right]d\eta$$

y

$$s_i=\int_\Lambda\left[h'_\alpha\left\{\int_{\mathcal{X}}f(x,\theta_2)\,\phi_\alpha\left(\frac{f(x,\theta_1)}{f(x,\theta_2)}\right)d\mu-\phi_\alpha(1)\right\}\int_{\mathcal{X}}\left\{\frac{\partial f(x,\theta_1)}{\partial\theta_{2\,i}}\,\phi_\alpha\left(\frac{f(x,\theta_1)}{f(x,\theta_2)}\right)\right\}d\mu\right.$$

$$\left.-\phi'_\alpha\left(\frac{f(x,\theta_1}{f(x,\theta_2)}\right)\frac{\partial f(x,\theta_2)}{\partial\theta_{2\,i}}\,\frac{f(x,\theta_1)}{f(x,\theta_2)}\right\}d\mu\Bigg]d\eta$$

(b) Si $\dfrac{m}{m+n}\xrightarrow[m,n\to\infty]{}\lambda\in(0,1)$ y $\theta_1=\theta_2$, entonces

$$\frac{2\,m\,n}{m+n}\,\frac{D_\phi^h(\hat\theta_1,\hat\theta_2)}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta}\xrightarrow[n\to\infty]{L}\chi^2_M$$

Cuando $M_0=k$, es decir, solamente se observa una muestra de tamaño n en la primera población, se obtiene un nuevo resultado

Resultado 2

Sean $\hat\theta=(\hat\theta_1,\ldots,\hat\theta_M)$ y $\hat\theta^*=(\hat\theta_1,\ldots,\hat\theta_{M_0},\theta^*_{M_0+1},\ldots,\theta^*_M)$ los estimadores de máxima verosimilitud de $\theta=(\theta_1,\ldots,\theta_M)$ y $\theta^*=(\theta_1,\ldots,\theta_{M_0},\theta^*_{M_0+1},\ldots,\theta^*_M)$ respectivamente basados en una muestra aleatoria simple de tamaño n de $f(x,\theta)$. Entonces

$$\text{(a)}\quad n^{1/2}\,(D_\phi^h(\hat\theta,\hat\theta^*)-D_\phi^h(\theta,\theta^*))\xrightarrow[n\to\infty]{L}N[0,T^tI_F(\theta)^{-1}T],$$

cuando $T^tI_F(\theta)^{-1}\,T>0$, donde $T=(t_1,\ldots,t_M)^t$, con

$$t_i=\int_\Lambda\left[h'_\alpha\left\{\int_{\mathcal{X}}f(x,\theta^*)\,\phi_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right)d\mu-\phi_\alpha(1)\right\}\int_{\mathcal{X}}\left\{\frac{\partial f(x,\theta^*)}{\partial\theta_i}\,\phi_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right)+\right.\right.$$

$$+ \frac{\partial f(x,\theta)}{\partial \theta_i}\,\phi'_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right) - \phi'_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right)\frac{\partial f(x,\theta)}{\partial \theta_i}\,\frac{f(x,\theta)}{f(x,\theta^*)}\Bigg\}\,d\mu\Bigg]d\eta$$

si $1 \le i \le M_0$,

$$t_i = \int_\Lambda \left[h'_\alpha \left\{ \int_{\mathcal{X}} f(x,\theta^*)\,\phi_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right)d\mu - \phi_\alpha(1)\right\} \int_{\mathcal{X}} \frac{\partial f(x,\theta)}{\partial \theta_i}\,\phi'_\alpha\left(\frac{f(x,\theta)}{f(x,\theta^*)}\right)d\mu\right] d\eta$$

si $M_0 + 1 \le i \le M$

(b) Si $\theta = \theta^*$, entonces

$$\frac{2n\,D_\phi^h(\hat\theta, \hat\theta^*)}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta} \xrightarrow[n\to\infty]{L} \chi^2_{M-M_0}.$$

Si $M_0=k=0$, solamente se observa una muestra aleatoria simple de tamaño n de la primera población. Obsérvese que θ_2 es completamente conocido. Se define $\theta=\theta_1$ y $\theta_0=\theta_2$ y se tiene:

iii) Si $\hat\theta$ es el estimador de máxima verosimilitud de θ y θ_0 es conocido, entonces

(a) $$n^{1/2}\left(D_\phi^h(\hat\theta, \theta_0) - D_\phi^h(\theta, \theta_0)\right) \xrightarrow[n\to\infty]{L} N\,[0,\ T^t I_F(\theta)^{-1}T],$$

donde $T = (t_1,\ldots,t_M)^t$ y los t_i están definidos en (ii) (a) y $T^t I_F(\theta)^{-1}T>0$.

(b) Si $\theta = \theta_0$, entonces

$$2n\,\frac{D_\phi^h(\hat\theta, \theta_0)}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta} \xrightarrow[n\to\infty]{L} \chi^2_M.$$

Los resultados obtenidos anteriormente se pueden utilizar para construir diversos contrastes de hipótesis. En este sentido se tiene:

(1) Para contrastar H_0: $D_\phi^h(\theta, \theta_0) = D_0$, se puede utilizar el estadístico:

$$Z_1 = \frac{n^{1/2}\,(D_\phi^h(\hat\theta, \theta_0) - D_0)}{\sigma},$$

que se distribuye asintóticamente, bajo la hipótesis nula, como una normal de media cero y varianza 1. El valor de σ se obtiene de (iii) al reemplazar θ por $\hat\theta$ en $[T^t I_F^{-1}(\theta)T]^{1/2}$.

(2) Para constrastar H_0:$\theta = \theta_0$, donde θ_0 es un valor dado del parámetro, se considera el estadístico,

$$T_1 = \frac{2n\, D_\phi^h(\theta, \theta_0)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta},$$

que asintóticamente y bajo la hipótesis nula se distribuye como una Ji-cuadrado con M grados de libertad, según (iii)(b).

(3) Sea $\theta=(\theta_1,\ldots,\theta_M)$ el valor desconocido del parámetro y sea $\theta^* = (\theta_1,\ldots, \theta_{M_0}, \theta^*_{M_0+1},\ldots, \theta^*_M)$ un valor de θ teniendo las M_0 primeras componentes como θ y con sus M-M_0 restantes componentes un valor prefijado de $\theta_{M_0+1},\ldots, \theta_M$. Para contrastar la hipótesis de que la divergencia entre θ y θ^* es de una cierta magnitud; es decir, H_0: $D_\phi^h(\theta, \theta*) = D_0$, se puede usar el estadístico

$$Z_2 = \frac{n^{1/2}\,(D_\phi^h(\hat{\theta}, \theta*) - D_0)}{\hat{\sigma}},$$

que asintóticamente y bajo la hipótesis nula, se distribuye según una normal de media cero y varianza 1 y $\hat{\sigma}$ se obtiene del resultado 2(a) al sustituir θ por su estimador de máxima verosimilitud $\hat{\theta}$.

(4) Bajo las condiciones dadas en (3), para contrastar H_0:$\theta=\theta*$, se usará el estadístico

$$T_2 = \frac{2n\, D_\phi^h(\hat{\theta}, \hat{\theta}*)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta},$$

cuya distribución asintótica está dada en 2(b).

(5) Para contrastar, H_0: $D_\phi^h(\theta_1, \theta_2) = D_0$, bajo las hipótesis del resultado ii) se usará el estadístico

$$Z_3 = \left(\frac{m\,n}{m+n}\right)^{1/2} \frac{(D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) - D_0)}{\hat{\sigma}},$$

que asintóticamente y bajo la hipótesis nula, se distribuye según una normal de media cero y varianza 1 y $\hat{\sigma}$ se obtiene de (ii)(a) al sustituir θ_1 y θ_2

por sus estimadores de máxima verosimilitud $\hat{\theta}_1$ y $\hat{\theta}_2$ en $(\lambda\, T^t\, I_F^{-1}(\theta_1)\, T + (1-\lambda)\, S^t\, I_F^{-1}(\theta_2)\, S)^{1/2}$.

(6) Bajo las condiciones dadas en (5), para contrastar $H_0{:}\theta_1{=}\theta_2$, se considera el estadístico

$$T_3 = \frac{2\,mn}{m+n} \frac{D_\phi^h(\hat{\theta}_1, \hat{\theta}_2)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta},$$

que asintóticamente y bajo la hipótesis nula se distribuye como una Ji-cuadrado con M grados de libertad, según (ii)(b).

(7) Sean $\theta_{11} = \theta_{21,\ldots,}\ \theta_{1k} = \theta_{2k}$ componentes desconocidas pero iguales de θ_1 y θ_2 respectivamente. Para contrastar, $H_0{:}\ D_\phi^h(\theta_1, \theta_2) = D_0$, se puede usar el estadístico

$$Z_4 = \left(\frac{m\,n}{m+n}\right)^{1/2} \frac{(D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) - D_0)}{\hat{\sigma}},$$

donde $\hat{\sigma}$ se obtiene de (i)(a) al sustituir θ_1 y θ_2 por $\hat{\theta}_1$ y $\hat{\theta}_2$ respectivamente en $[T^t\, \Sigma(\theta_1, \theta_2)\, T)^{1/2}$.

(8) Bajo las condiciones dadas en (7), para contrastar $H_0{:}\theta_1{=}\theta_2$, se usará el estadístico

$$T_4 = \frac{2\,mn}{m+n} \frac{D_\phi^h(\hat{\theta}_1, \hat{\theta}_2)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta},$$

cuya distribución asintótica está dada en (i)(b).

(9) Sean $\theta_{11} = \theta_{21,\ldots,}\ \theta_{1k} = \theta_{2k}$ componentes desconocidas pero iguales de θ_1 y θ_2 respectivamente. Sea $\theta^*_{2,M_0+1,\ldots,}\ \theta^*_{2,M}$ un valor dado de $\theta_{2,M_0+1,\ldots,}\ \theta_{2,M}$. Para contrastar que la divergencia entre θ_1 y θ_2 es de una cierta magnitud; es decir, $H_0{:}\ D_\phi^h(\theta_1, \theta_2) = D_0$, se usará el estadístico

$$Z_5 = \left(\frac{m\,n}{m+n}\right)^{1/2} \frac{(D_\phi^h(\hat{\theta}_1, \hat{\theta}_2) - D_0)}{\hat{\sigma}},$$

donde $\hat{\sigma}$ se obtiene del resultado 1(a) al sustituir θ_1 y θ_2 por $\hat{\theta}_1$ y $\hat{\theta}_2$ respectivamente en $[T^t\Sigma(\theta_1,\theta_2)T]^{1/2}$.

(10) Bajo las condiciones dadas en (9), para contrastar $H_0:\theta_1=\theta_2$, se puede usar el estadístico

$$T_5 = \frac{2\,mn}{m+n} \frac{D_\phi^h(\hat{\theta}_1, \hat{\theta}_2)}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta},$$

cuya distribución asintótica está dada en el resultado 1(b).

Dos cuestiones conviene destacar finalmente:

(a) Si para un α, $0\leq\alpha\leq1$, dado, existe un test no aleatorizado de Neyman-Pearson para contrastar una hipótesis nula simple frente a una alternativa simple, entonces existe un test equivalente basado en la D_ϕ^h–divergencia. Para ello no hay más que considerar como estimador de $\theta\in\Theta=\{\theta_0,\theta_1\}$,

$$\hat{\theta} = \begin{cases} \theta_1 & \text{si} \quad f(x_1,\ldots,x_n/\theta_1) \geq \lambda\, f(x_1,\ldots,x_n/\theta_0) \\ \theta_0 & \text{si} \quad f(x_1,\ldots,x_n/\theta_1) < \lambda\, f(x_1,\ldots,x_n/\theta_0) \end{cases}$$

(b) Si el test basado en la D_ϕ^h– divergencia se construye a partir del estimador de máxima verosimilitud, entonces el estadístico resultante para contrastar $\theta\in\Theta_0$ frente $\theta\in\Theta_1$, es función de cada estadístico suficiente para θ.

Ultimamente son numerosos los trabajos que han aparecido publicados en los que se abordan diversos problemas estadísticos en poblaciones multinomiales mediante las divergencias. En Zografos y otros (1990) se estudian las propiedades de la φ-divergencias estimadas. Zografos (1992) y Pardo, L. y otros (1993b) establecen un test de independencia utilizando las medidas de φ-divergencias. Con otras medidas diferentes de divergencia, Morales y otros (1993a) han estudiado el mismo problema en muestreo estratificado con asignación proporcional e independencia entre los estratos. Otros trabajos en esta línea son los de Menéndez y otros (1992a) y Salicrú y otros (1992), Zyarova, J. (1973), Cressie y Read (1984), Gil y Martínez (1992), etc.

Ahora es estudiará el comportamiento asintótico de las (h,ϕ) – divergencias en poblaciones multinomiales. Sea $(\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)_{\theta\in\Theta}$ con $\Theta = \{\theta = (p_1,\ldots,p_{M-1}) / p_i > 0, \sum_{i=1}^{M-1} p_i = 1 - p_M\}$ y $\mathfrak{X} = \{x_1,\ldots,x_M\}$. Se consideran los parámetros $\theta_1 = (p_1,\ldots,p_{M-1})$ y $\theta_2 = (q_1,\ldots,q_{M-1})$, donde $p_i\geq0$, $q_i\geq0$ $(i = 1,2,\ldots,M)$ y $\sum_{i=1}^{M} p_i = \sum_{i=1}^{M} q_i = 1$ (Obsérvese que ahora el parámetro es de dimensión $M - 1$ y no de dimensión M). En este caso par cada valor x_i de X, $f(x_i,\theta_1)=p_i$ y $f(x_i,\theta_2)=q_i$, donde se ha supuesto que μ

es una medida contable dando masa uno a cada uno de los valores x_i de X. Entonces, se tiene

$$\int_{\mathcal{X}} f(x, \theta_2)\, \phi_\alpha\left(\frac{f(x, \theta_2)}{f(x, \theta_2)}\right) d\mu(x) = \sum_{i=1}^{M} q_i\, \phi_\alpha\left(\frac{p_i}{q_i}\right)$$

y $D_\phi^h(\theta_1, \theta_2)$ se puede rescribir en la forma

$$D_\phi^h(P, Q) = D_\phi^h(\theta_1, \theta_2) = \int_\Lambda h_\alpha\left[\sum_{i=1}^{M} q_i\, \phi_\alpha\left(\frac{p_i}{q_i}\right) - \phi_\alpha(1)\right] d\eta\ .$$

La estimación de las ($\underline{h}$, $\underline{\phi}$) –divergencias en poblaciones multinomiales se puede hacer de dos formas: Estimando las dos distribuciones que aparecen en ellas o estimando una y considerando la otra como conocida. En el primer caso se tiene lo que se denomina un índice de similaridad entre P y Q. En el otro caso se tiene la divergencia entre la muestra y un modelo de probabilidad dado. La particularización de algunos de los resultados vistos anteriormente al caso de poblaciones multinomiales, permitirá construir tests de bondad de ajuste y homogeneidad así como obtener su potencia asintótica. De forma más precisa se tiene el siguiente resultado:

Resultado 3

Considérese el estimador analógico, $D_\phi^h(\hat{P}, \hat{Q})$, obtenido al sustituir los p_i y q_i por sus frecuencias observadas $\hat{p}_i$ y $\hat{q}_i$ (i = 1,...,M) respectivamente. Supóngase que $(\hat{p}_1,...\hat{p}_M)$ y $(\hat{q}_1,...\hat{q}_M)$ están basados en muestras aleatorias simples e independientes de tamaños n y m respectivamente. Entonces

(a) $$\left(\frac{m\,n}{m+n}\right)^{1/2} [D_\phi^h(\hat{P}, \hat{Q}) - D_\phi^h(P, Q)] \xrightarrow[m,n \to \infty]{L} N(0, \sigma_1^2),$$

donde $\sigma_1^2 = \lambda\, \sigma_P^2 + (1-\lambda)\, \sigma_Q^2 > 0$,

$$\sigma_P^2 = \sum_{i=1}^{M} p_i \left\{ \int_\Lambda h'_\alpha\left[\sum_{i=1}^{M} q_i\, \phi_\alpha\left(\frac{p_i}{q_i}\right) - \phi_\alpha(1)\right] \phi'_\alpha\left(\frac{p_i}{q_i}\right) d\eta \right\}^2 -$$

$$- \left\{ \sum_{i=1}^{M} p_i \int_\Lambda h'_\alpha\left[\sum_{i=1}^{M} q_i\, \phi_\alpha\left(\frac{p_i}{q_i}\right) - \phi_\alpha(1)\right] \phi'_\alpha\left(\frac{p_i}{q_i}\right) d\eta \right\}^2$$

y

$$\sigma_Q^2 = \sum_{i=1}^{M} q_i \left\{ \int_\Lambda h'_\alpha \left[\sum_{i=1}^{M} q_i \phi_\alpha \left(\frac{p_i}{q_i} \right) - \phi_\alpha(1) \right] \left[\phi_\alpha \left(\frac{p_i}{q_i} \right) - \frac{p_i}{q_i} \phi'_\alpha \left(\frac{p_i}{q_i} \right) \right] d\eta \right\}^2 -$$

$$- \left\{ \sum_{i=1}^{M} q_i \int_\Lambda h'_\alpha \left[\sum_{i=1}^{M} q_i \phi_\alpha \left(\frac{p_i}{q_i} \right) - \phi_\alpha(1) \right] \left[\phi_\alpha \left(\frac{p_i}{q_i} \right) - \frac{p_i}{q_i} \phi'_\alpha \left(\frac{p_i}{q_i} \right) \right] d\eta \right\}^2 .$$

(b) Si P = Q, entonces

$$\frac{2mn}{m+n} \frac{D_\phi^h (\hat{P}, \hat{Q})}{\int_\Lambda h'_\alpha (0)\, \phi''_\alpha (1)\, d\eta} \xrightarrow[n,m \to \infty]{L} \chi^2_{M-1}.$$

Resultado 4

Si Q es conocido, entonces

(a) $n^{1/2} [D_\phi^h (\hat{P}, Q) - D_\phi^h (P, Q)] \xrightarrow[n \to \infty]{L} N(0, \sigma_P^2)$,

donde σ_P^2 se ha definido en el resultado 3.

b) Si Q es conocido y P = Q, entonces

$$2n \frac{D_\phi^h (\hat{P}, Q)}{\int_\Lambda h'_\alpha (0)\, \phi''_\alpha (1)\, d\eta} \xrightarrow[n \to \infty]{L} \chi^2_{M-1} .$$

A partir de los dos resultados dados anteriormente se pueden construir los siguientes contrastes:

(11) Test para un valor dado de la divergencia poblacional; es decir,

$$H_0: D_\phi^h (P, Q) = D_0.$$

Ver el test dado en (1) cuando Q es conocido y (5) si Q es desconocido.

(12) Test para un valor dado de r divergencias poblacionales; es decir,

$$H_0: D_\phi^h (P_1, Q_1) = ... = D_\phi^h (P_r, Q_r) = D_0$$

con Q_i conocido o desconocido. Se procede de forma análoga a como se procedía en el caso de la igualdad de r (h,ϕ)-entropías a un valor dado.

(13) Test para la igualdad de r divergencias poblacionales; es decir,

$$H_0: D_\phi^h(P_1, Q_1) = ... = D_\phi^h(P_r, Q_r)$$

con Q_i conocido o desconocido. Se procede de forma análoga a como se procedía en el caso de la igualdad de r (h,ϕ)-entropías.

(14) Test de homogeneidad de m poblaciones a una distribución conocida Q; es decir, $H_0 : P_1 = P_2 = ... = P_m = Q$.

La distribución asintótica del estadístico

$$T_a = \frac{2\, n\, D_\phi^h(\hat{P}, Q)}{\int_\Lambda h'_\alpha(0)\, \phi''_\alpha(1)\, d\eta}$$

se puede utilizar para contrastar H_0:$F_x(x) = F_0(x)$ en la forma siguiente: Supongamos que estamos obteniendo observaciones de una población con función de distribución $F_x(x)$. Dividimos el soporte de la distribución en M intervalos disjuntos, $I_1, ..., I_M$, con $P(X \in I_i) = p_i$, i = 1,...,M. Sea $X_1,...,X_n$ una muestra de tamaño n de $F_x(x)$ y $n\hat{p}_1,...,n\hat{p}_M$ el número de valores muestrales que caen en los intervalos $I_1,...,I_M$. El vector aleatorio $(n\hat{p}_1,...,n\hat{p}_M)$ sigue una distribución multinomial con parámetros $(n;p_1,...,p_M)$. Si queremos contrastar la hipótesis $H_0 : F_x(x) = F_0(x)$ calculamos $q_i = P(X \in I_i)$ (i = 1,...,M) bajo H_0. Si H_0 es cierta, entonces P = Q e intuitivamente esperamos que $n\hat{p}_i \approx nq_i$, en cuyo caso $D_\phi^h(\hat{P}, Q)$ será pequeño y en consecuencia valores grandes de $D_\phi^h(\hat{P}, Q)$ indican poca compatibilidad de los datos con la distribución considerada bajo la hipótesis nula. Por tanto, para valores grandes de n un test de tamaño α vendrá dado por,

$$\phi(\hat{p}_1,..., \hat{p}_M) = \begin{cases} 1 & \text{si } T_a > \chi^2_{M-1,\alpha} \\ 0 & \text{en el resto.} \end{cases}.$$

El problema se complica cuando parámetros que identifican la distribución bajo la hipótesis nula son desconocidos. Supongamos que las probabilidades q_i son función de $\theta = (\theta_1,..., \theta_{M_0})$, con $\theta \in \Theta \subset \mathbb{R}^{M_0}$, $M_0 < M$. El verdadero valor θ_0 se supone que es un punto interior de Θ. En otras palabras bajo la hipótesis nula, H_0, la forma de la función de probabilidad o densidad está especificada salvo uno o varios parámetros. En este contexto en Morales y otros (1993c), se establece un resultado para el problema general de bondad de ajuste cuando las probabilidades de las clases están especificadas salvo un número finito de parámetros que son desconocidos. De forma más precisa en el trabajo citado anteriormente

se establece que si P=Q y $\hat{\theta} = [\theta_1 (\hat{p}_1,\ldots, \hat{p}_M),\ldots, \theta_{M_0} (\hat{p}_1,\ldots, \hat{p}_M)]$ es un estimador eficiente de primer orden (e.p.o), entonces

$$T_b = 2n \frac{D_\phi^h [\hat{P}, Q(\hat{\theta})]}{\int_\Lambda h'_\alpha (0)\, \phi''_\alpha (1)\, d\eta} \xrightarrow[m,n-\infty]{L} \chi^2_{M-M_0-1},$$

donde $\hat{P} = (\hat{p}_1,\ldots, \hat{p}_M)^t$ y $\hat{Q} = [p_1 (\hat{\theta}),\ldots, p_M (\hat{\theta})]^t$

Este resultado se puede utilizar para contrastar las hipótesis $H_0 : F \in \{F_\theta\}_{\theta \in \Theta}$. Para ello en primer lugar se estima el parámetro $\theta = (\theta_1,\ldots, \theta_{M_0})$, $M_0 < M$, utilizando un estimador e.p.o. $\hat{\theta} = [\theta_1 (\hat{p}_1,\ldots, \hat{p}_M),\ldots, \theta_{M_0} (\hat{p}_1,\ldots, \hat{p}_M)]$. En un segundo paso se calcula, $q_i (\hat{\theta}) = \int_{I_i} dF_{\hat{\theta}} (z)$, $i = 1,\ldots, M$. Si H_0 es cierta, entonces $P = Q$ e intuitivamente se espera que $np_i \approx nq_i (\hat{\theta})$, en cuyo caso $D_\phi^h [\hat{P}, Q (\hat{\theta})]$ es pequeño y en consecuencia un test de tamaño α para la hipótesis señalada viene dado por

$$\phi [\hat{P}, Q (\hat{\theta})] = \begin{cases} 1 & \text{si } T_b > \chi^2_{M-M_0-1,\alpha} \\ 0 & \text{resto} \end{cases}.$$

Un problema interesante que surge en este contexto es el de utilizar las (h, ϕ)divergencias como un procedimiento de estimación puntual: Elegir como estimador de θ el valor $\hat{\theta}$ que verifique,

$$D_\phi^h [\hat{P}, Q (\hat{\theta})] = \underset{\theta \in \Theta \subset \mathbb{R}^{M_0}}{M\, i\, n} \; D_\phi^h [\hat{P}, Q (\hat{\theta})].$$

En caso de utilizar como caso particular de (h, ϕ)-divergencia la divergencia de Kullback se llega a que $\hat{\theta}$ coincide con el estimador de máxima verosimilitud del modelo discretizado.

En los dos test anteriores hemos supuesto que únicamente se disponía del número de observaciones, n_i ($\hat{p}_i = n_i/n$), que caían en cada uno de los intervalos, I_i. Sin embargo si se dispone de las observaciones originales $x_1,\ldots,x_n$ parece lógico utilizar estimaciones más eficientes tales como las resultantes de considerar el estimador de máxima verosimilitud, $\tilde{p}_i$, basado en la utilización de toda la información que proporcionan las x_i. Si las X_i tienen como densidad común $g(x,\theta)$ en lugar de estimar θ_0 utilizando la información que proporciona la distribución multinomial mediante $\hat{P} = (\hat{p}_1,\ldots,\hat{p}_M)$, parece más natural y lógicamente más eficiente maximizar la función de verosimilitud original; es decir $\prod_{i=1}^{n} g(x_i, \theta)$. Denotemos por $\hat{\theta}$ el estimador resultante y sea $\hat{\pi}_i = \int_{I_i} g(x, \hat{\theta})\, dx$.

En este contexto, en Morales y Pardo, L. (1993) se ha obtenido el siguiente resultado:

$$T_c = 2n \frac{D_\phi^h (\hat{P}, \Pi)}{\int_\Lambda h'_\alpha (0)\, \phi''_\alpha (1)\, d\eta} \xrightarrow[n \to \infty]{L} \chi^2_{M-M_0-1} + \sum_{i=1}^{M_0} \lambda_i \chi^2_1$$

donde la $\chi^2_{M-M_0-1}$ y las χ^2_1 son independientes, los λ_i son tales que $0 \leq \lambda_i \leq 1$ y además son las soluciones de la ecuación, $| \tilde{I}_F (\theta) - \lambda I_F (\theta) | = 0$ siendo $\tilde{I}_F (\theta) = [i_{r,s} (\theta)]_{r,s = 1,\ldots,M_0}$ la matriz de información de Fisher para los datos discretizados; es decir,

$$i_{r,s} (\theta) = \sum_{j=1}^{M} \frac{1}{q_j (\theta)} \frac{\partial q_j (\theta)}{\partial \theta_r} \frac{\partial q_j (\theta)}{\partial \theta_s} .$$

Si los extremos de las M clases consideradas dependen de parámetros desconocidos; es decir $\pi_i = \int_{z_{i-1}(\theta)}^{z_i (\theta)} f(x, \theta)\, dx$ y se eligen los $z_i (\hat{\theta})$, con $\hat{\theta}$ estimador de máxima verosimilitud de θ, de forma que $\pi_i = 1/M$, entonces

$$T_d = 2n \frac{D_\phi^h (\hat{P}, \Pi)}{\int_\Lambda h'_\alpha (0)\, \phi''_\alpha (1)\, d\eta} \xrightarrow[n \to \infty]{L} \chi^2_{M-M_0-1} + \sum_{i=1}^{M_0} \mu_i \chi^2_1$$

donde la $\chi^2_{M-M_0-1}$ y las χ^2_1 son independientes y las $\mu_i = 1 - a_i$ son tales que $0 \leq a_i \leq 1$ y además son las soluciones de la ecuación $| \tilde{\tilde{I}}_F (\theta) - a I_F (\theta) | = 0$, siendo

$$i^*_{r,s} (\theta) = \sum_{l=1}^{M} \frac{1}{\pi_l} \left(\int_{z_{l-1}(\theta)}^{z_l (\theta)} \frac{\partial f(x, \theta)}{\partial \theta_r} dx \right) \left(\int_{z_{l-1}(\theta)}^{z_l (\theta)} \frac{\partial f(x, \theta)}{\partial \theta_s} dx \right)$$

el elemento (r,s) de la matriz $\tilde{\tilde{I}}_F (\theta)$.

Tiene interés conocer la distribución asintótica de $D_\phi^h [\hat{P}, Q (\theta_0)]$ cuando θ_0 se estima utilizando información procedente de una segunda población multimonial que es independiente de X. Sea X una distribución multinomial con parámetros $[n, p_1(\theta), \ldots, p_M(\theta)]$ y sea X* otra distribución multinomial e independiente de X con parámetros $[n^*, p_1(\theta), \ldots, p_M(\theta)]$. Entonces X y X* tienen la misma dimensión y el mismo modelo $p(\theta)$, pero en general n y n* serán distintos. Sean $\tau^* = n/n^*$ y $\tau^{**} = n/(n + n^*)$, y se supone que al crecer n y n* se mantienen las tasas τ^* y τ^{**}.

Sean $\hat{P} = (\hat{p}_1,\ldots,\hat{p}_M)$ y $\hat{P}^* = (\hat{p}_1^*,\ldots,\hat{p}_M^*)$ las frecuencias relativas en las dos muestras y $\hat{\theta}^*$ el estimador de máxima verosimilitud de θ_0 basado en la primera población y $\hat{\theta}^{**}$ el estimador de máxima verosimilitud de θ_0 basado en la muestra conjunta X+X*. En Morales y Pardo, L. (1993) se establece el siguiente resultado:

$$\text{(i)} \quad T_e = 2n \frac{D_\phi^h(\hat{P}, P(\hat{\theta}^*))}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta} \xrightarrow[n\to\infty]{L} \chi^2_{M-M_0-1} + (1+\tau^*)\,\chi^2_M$$

$$\text{(ii)} \quad T_f = 2n \frac{D_\phi^h(\hat{P}, P(\hat{\sigma}^*))}{\int_\Lambda h'_\alpha(0)\,\phi''_\alpha(1)\,d\eta} \xrightarrow[n\to\infty]{L} \chi^2_{M-M_0-1} + \left(1 + \frac{\tau^{**}}{1+\tau^{**}}\right)\chi^2_M$$

En Morales y otros (1993a) se aborda el problema de obtener la distribución asintótica de las (h, ϕ)-divergencias en poblaciones multinomiales pero en lugar de considerar muestreo aleatorio simple se considera muestreo aleatorio estratificado con cualquier tipo de asignación. Un estudio similar al desarrollado en esta sección con las (h, ϕ)-divergencias se ha llevado a cabo con las R_ϕ^h -divergencias, pudiéndose ver los resultados obtenidos en Salicrú y otros (1993c) y Pardo, L. y otros (1993a).

8. COMPARACION Y DISEÑO SECUENCIAL DE EXPERIMENTOS

Supongamos que se consideran varios experimentos, cada uno de los cuales implica la observación de una variable o vector aleatorio sobre cierta población y tales que sus distribuciones dependen de cierto estado de la naturaleza o valor de un parámetro desconocido. La idea de establecer un orden entre estos experimentos, con el fin de seleccionar el «más informativo» sobre el estado o parámetro desconocido, aparece desarrollada por primera vez en una comunicación privada de Bohnenblust, Shapley y Sherman. Blackwell (1951) recogió los resultados básicos de esa comunicación, en la que se proponía un criterio de comparación de experimentos, formulado en el contexto de un problema de decisión y basado en el concepto de «conjunto de vectores de pérdida accesibles». Si el experimento X es preferido al experimento Y según este criterio se escribirá: $X \overset{B}{\geq} Y$.

Por otro lado, Blackwell sugirió además un nuevo criterio, formulado en un contexto más general y basado en la noción de suficiencia estadística. A lo largo de distintos trabajos, Blackwell, 1951, 1953, Le Cam, 1964, etc., se ha probado que el criterio de suficiencia y el de Bohnenblust, Shapley y Sherman son

equivalentes en la mayoría de los casos y que la condición que exige el primero es más fuerte que la que exige el segundo. Si el experimento X es preferido al experimento Y según este criterio se escribirá: $X \overset{S}{\geq} Y$.

Lehmann establece que el experimento X es preferido al experimento Y, si existe una variable aleatoria U, independiente de X y de distribución conocida, y una función medible, h tal que la variable aleatoria H = h(X,U), tiene la misma distribución que Y para todo $\theta \in \Theta$. Si el experimento X es preferido al experimento Y según este criterio se escribirá: $X \overset{L}{\geq} Y$. Es fácil comprobar que si $X \overset{L}{\geq} Y$ entonces $X \overset{S}{\geq} Y$.

A partir de los estudios realizados por Blackwell, la comparación de experimentos ha ido evolucionando como una teoría estadística, y el criterio de suficiencia, se ha erigido en criterio «patrón» dentro de esta teoría, en el sentido que indicamos a continuación. Aunque es natural que, debido a su interpretación, el criterio de suficiencia de Blackwell esté universalmente aceptado, presenta el inconveniente de no ser siempre aplicable, ya que existen experimentos no comparables a través de ese criterio, Hansen and Torgersen (1974). Por esta razón, han sido muchos los autores que han establecido diferentes criterios de comparación que se aplican en condiciones más generales. Para confirmar la adecuación de cada nuevo criterio propuesto, se examinan en los trabajos que los introducen una serie de propiedades deseables en una comparación razonable, siguiendo las ideas de Lindley (1956) y se conectan con el de suficiencia de Blackwell. Esta conexión consiste en comprobar si para cualquier par de experimentos en los que es aplicable ese último, el criterio que se introduce determina entre ellos el mismo orden que el de suficiencia. De no resultar de este modo, el nuevo criterio no se considera apropiado.

En la bibliografía sobre esta teoría, se observa que los criterios de comparación que se han definido se apoyan en medidas de información Lindley, 1956, Stone, 1959, Goel y DeGroot, 1979, García-Carrasco, 1978, 1982, Gil, M. A., 1982, Pardo, J. A. y otros 1993a, Pardo, L. y otros 1991a, Taneja y otros 1989, Vicente 1991, etc., en medidas del «valor» de la información Raiffa y Schlaifer, 1961, García-Carrasco, 1978, etc. o en otros conceptos o medidas Heyer, 1982, Le Cam, 1974, etc... El objetivo básico de este apartado es señalar diversos criterios de comparación de experimentos en base a medidas de entropía y divergencia.

Las interesantes propiedades de la cantidad de información de Fisher ponen de manifiesto su conveniencia para definir a partir de ella un criterio de comparación de experimentos. Esta cantidad de información tiene además la ventaja de venir definida exclusivamente a partir del experimento, comparando así éstos, de una manera absoluta y no en función de los términos concretos del problema de partida.

Dados los experimentos $X = (\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$ e $Y = (\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q_\theta)$, $\theta \in \Theta$, y Θ abierto de $\mathbb{R}$; se dice que X es más informativo que Y ($X \overset{F}{\geq} Y$) si $I_X(\theta) \geq I_Y(\theta)$, $\theta \in \Theta$. Se dirá que $X \overset{F}{\approx} Y$ si y sólo si $X \overset{F}{\geq} Y$ e $Y \overset{F}{\geq} X$. Sus propiedades son:

1) Para todo experimento X, $X \overset{F}{\geq} N$ donde $N = (\mathfrak{X}, \beta_{\mathfrak{X}}, Q)$es el experimento nulo; es decir, aquél en que Q no depende de θ.

2) La relación $\overset{F}{\geq}$ es un preorden parcial.

3) $(X_1; X_2) \overset{F}{\geq} X_1$ con $\overset{F}{\approx}$ si y sólo si $f(x_2/x_1,\theta)$ es independiente de θ c.s.

4) $X^{(n+1)} \overset{F}{\geq} X^{(n)} \supset n \geq 1$

5) Si X_1, X_2 y X_3 son tres experimentos sobre el mismo Θ y X_3 es independiente de X_1 y de X_2, entonces $X_1 \overset{F}{\geq} X_2$ implica $(X_1, X_3) \overset{F}{\geq} (X_2, X_3)$.

6) Si X_1, X_2, X_3 y X_4 son cuatro experimentos sobre el mismo Θ tales que $X_1 \overset{F}{\geq} X_2$, $X_3 \overset{F}{\geq} X_4$, X_1 es independiente de X_3 y X_2 de X_4, entonces

$$(X_1, X_3) \overset{F}{\geq} (X_2, X_4).$$

7) Sea $X = (\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$ un experimento donde P_θ, $\theta \in \Theta$, viene descrito por la densidad $f(x/\theta)$. Sea $\{E_i\}_{i \in N}$ una partición de $\mathfrak{X}$ por elementos del σ-álgebra $\beta_{\mathfrak{X}}$. Consideremos el nuevo experimento $Y = (\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q_\theta)$ donde $\mathfrak{Y} = \{E_i\}_{i \in N}$, $\beta_{\mathfrak{Y}}$ es la σ-álgebra engendrada por los E_i y Q_θ es tal que $Q_\theta(E_i) = P_\theta(E_i)$. Entonces $X \overset{F}{\geq} Y$.

8) Para todo estadístico $T = T(X^{(n)})$ de la muestra de tamaño n, $X^{(n)} \overset{F}{\geq} T$ con $\overset{F}{\approx}$ si y sólo si T es un estadístico suficiente.

9) Sean S y T dos estadísticos suficientes de $X^{(n)}$ e $Y^{(m)}$ respectivamente, entonces $X^{(n)} \overset{F}{\geq} Y^{(m)}$ si y sólo si $S \overset{F}{\geq} T$.

10) Sea $X = (\mathbb{R}, B(\mathbb{R}), P_\theta)$, $\theta \in \Theta$. Si T es una transformación de $\mathbb{R}$ en $\mathbb{R}$ estrictamente monótona y derivable, entonces $X \overset{F}{\approx} T$.

11) Sean $X_1,...,X_n$ experimentos sobre $(\mathbb{R}, B(\mathbb{R}))$, y consideremos una transformación $h = (h_1,...,h_m)$ de $\mathbb{R}^n$ en $\mathbb{R}^m$ tal que para cada $(x_1,...,x_{n-m})$ fijo la función $g_{x_1,...,x_{n-m}}(x_{n-m+1},..., x_n) = h(x_1,..., x_n)$ es biyectiva con primeras derivadas parciales continuas. Si definimos $Y_1 = h_1(X_1,...,X_n),...,$ $Y_m = h_m(X_1,...,X_n)$, entonces los experimentos compuestos $X^{(n)} = (X_1,...,X_n)$ y $Z^{(m)} = (X_1,...,X_{n-m}, Y_1,...,Y_m)$ son tales que $X^{(n)} \overset{F}{\approx} Z^{(m)}$.

Además si $X \overset{L}{\geq} Y$ entonces $X \overset{F}{\geq} Y$ y si $X \overset{S}{\geq} Y$ entonces $X \overset{F}{\geq} Y$.

Sea $\theta = (\theta_1,..,\theta_M)$ y supóngase que Θ es un subconjunto abierto de $\mathbb{R}^M$. Sean $I_F^X(\theta)$ e $I_F^Y(\theta)$ las matrices de Información de Fisher asociadas a los experimentos $X = (\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$ e $Y = (\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q_\theta)$ con $\theta = (\theta_1,..,\theta_M)$. Se dice que el experimento X es preferido al experimento Y si la matriz $I_F^X(\theta) - I_F^Y(\theta)$ es definida no negativa. Las propiedades de este criterio de comparación de experimentos así como su relación con el criterio de Fisher pueden verse en Goel y DeGroot (1979).

Es posible también definir un criterio bayesiano de comparación de experimentos en base a la adaptación de la medida de información de Fisher a este contexto. Las propiedades enunciadas anteriormente se mantienen.

Lindley (1956) estableció un criterio de comparación de experimentos en base a la entropía de Shannon. Es claro que la incertidumbre contenida en la distribución a priori, viene dada por

$$-\int_\Theta p(\theta) \log p(\theta)\, d\lambda(\theta).$$

Por tanto, la cantidad de información que la observación de un valor x del experimento realizado proporciona sobre θ, viene dada a través de la expresión siguiente

$$-\int_\Theta p(\theta) \log p(\theta)\, d\lambda(\theta) + \int_\Theta p(\theta/x) \log p(\theta/x)\, d\lambda(\theta).$$

Así, la información que se espera proporcione el experimento X, antes de su realización, viene dado por la expresión

$$I[X, p(.)] = \int_{\mathfrak{X}} \left(\int_\Theta p(\theta/x) \log p(\theta/x)\, d\lambda(\theta) \right) f(x)\, d\mu(x) - \int_\Theta p(\theta) \log p(\theta)\, d\lambda(\theta).$$

En base a esta cuantificación de la información, el experimento X es más informativo en el sentido de la entropía de Shannon que el experimento Y, respecto de la distribución $p(\theta)$ y se denotará por $X \overset{Sh}{\geq} Y$, si se verifica

$$I[X, p(\theta)] \geq I[Y, p(\theta)]$$

y se dirá que $X \overset{Sh}{\approx} Y$ respecto de $p(\theta)$ si y sólo si $X \overset{Sh}{\geq} Y$ e $Y \overset{Sh}{\geq} X$.

La relación $\overset{Sh}{\geq}$ es un preorden completo. En caso de que en la definición anterior no estuviese fija $p(\theta)$ y se dijese que $X \overset{Sh}{\geq} Y$ si y sólo si $I[(X, p(\theta)] \geq I[Y, p(\theta)] \supset p(\theta)$, se tendría un criterio no bayesiano ya que no tendría en cuenta el particular conocimiento que se tiene acerca de θ. Además es sencillo ver que en este caso la relación $X \overset{Sh}{\geq} Y$ es un preorden parcial.

El estudio de las propiedades de este criterio conduce a propiedades análogas a las establecidas con el criterio basado en la información de Fisher. Este criterio fue posteriormente generalizado en Pardo, J. A. y otros (1993a) al considerar entropías generalizadas en lugar de la entropía de Shannon.

Stone (1959) presentó una interesante aplicación del criterio dado por Lindley a la comparación de experimentos de regresión. Resultados análogos con otras medidas se han obtenido en Pardo, L. y Menéndez (1989), Vicente (1990), Pardo, J. A. y otros (1993b).

Seguidamente se pasa a analizar sendos criterios de comparación de experimentos en base a las medidas de divergencia. La diferencia esencial entre ambos criterios radica en que mientras el primero es un criterio no bayesiano introducido por Goel y DeGroot (1979), el segundo es un criterio bayesiano y ha sido ampliamente estudiado por Mallows (1959), Csiszar (1972), Goel y DeGroot (1980), etc. La elección de la divergencia de Csiszar para dar estos dos criterios de comparación de experimentos se basa en el hecho de que una vez establecido el criterio en base a esta medida se tendrá de forma inmediata el criterio y los resultados para las divergencias generalizadas.

Dados los experimentos $X = (\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$ e $Y = (\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q_\theta)$ se dice que el experimento X es preferido al experimento Y, $X \overset{\varphi}{\geq} Y$, si y sólo si

$$\supset \theta_1, \theta_2, \qquad D_\varphi (P_{\theta_1}, P_{\theta_2}) \geq D_\varphi (Q_{\theta_1}, Q_{\theta_2}),$$

donde

$$D_\varphi (P_{\theta_1}, P_{\theta_2}) = \int_{\mathfrak{X}} f(x/\theta_2)\, \varphi\left(\frac{f(x/\theta_1)}{f(x/\theta_2)} \right) d\mu(x) .$$

Entre otras las propiedades más importantes de este criterio de comparación de experimentos son las siguientes:

1) Para todo experimento X, $X \overset{\varphi}{\geq} N$ donde $N = (\mathfrak{N}, \beta_{\mathfrak{N}}, Q)$ es el experimento nulo.

2) $(X_1, X_2) \overset{\varphi}{\geq} X_1$ con $\overset{\varphi}{\approx}$ si y sólo si $f(x_2/x_1, \theta)$ es independiente de θ c.s.

3) $X^{(n+1)} \overset{\varphi}{\geq} X^{(n)} \supset n \geq 1$.

4) Sea $X = (\mathfrak{X}, \beta_{\mathfrak{X}}, P_\theta)$ un experimento donde P_θ, $\theta \in \Theta$, viene descrito por la densidad $f(x/\theta)$. Sea $\{E_i\}_{i \in N}$ una partición de $\mathfrak{X}$ por elementos del σ-algébra $\beta_{\mathfrak{X}}$. Consideremos el nuevo experimento $Y = (\mathfrak{Y}, \beta_{\mathfrak{Y}}, Q_\theta)$ donde $\mathfrak{Y} = \{E_i\}_{i \in N}$, $\beta_{\mathfrak{Y}}$ es la σ-álgebra engendrada por los E_i y Q_θ es la que $Q_\theta(E_i) = P_\theta(E_i)$. Entonces $X \overset{\varphi}{\geq} Y$.

5) Para todo estadístico $T = T(X^{(n)})$ de la muestra de tamaño n, $X^{(n)} \overset{\Phi}{\geq} T$ con $\overset{\Phi}{\simeq}$ si y sólo si T es un estadístico suficiente.

Al igual que ocurría con los criterios basados en entropías se establece que este criterio es más débil que los criterios de Lehmann y suficiencia. La validez de estos resultados para otras medidas de divergencia se mantiene como puede verse en Pardo, L. y otros (1991d).

La adaptación de las medidas de divergencia a la comparación de experimentos en el contexto bayesiano se establece en los siguientes términos: Se dice que el experimento X es preferido al experimento Y respecto a la distribución a priori $p(\theta)$, $X \overset{\Phi}{\geq} Y$, si y solamente si $D_\Phi(X,p(.)) \geq D_\Phi(Y,p(.))$. También, se dice que los experimentos X e Y son equivalentes con respecto a $p(\theta)$, $\overset{\Phi}{\simeq}$, si y solamente si $X \overset{\Phi}{\geq} Y$, e $Y \overset{\Phi}{\geq} X$,. Donde,

$$D_\Phi(X, p(.)) = E_X\left(D_\Phi(p(\theta), p(\theta/x)) \right) = \int_{\mathcal{X}} \left[\int_\Theta \Phi\left(\frac{p(\theta/x)}{p(\theta)} \right) p(\theta)\, d\lambda(\theta) \right] f(x)\, d\mu(x) =$$

$$= \int_\Theta \left[\int_{\mathcal{X}} \Phi\left(\frac{f(x/\theta)}{p(x)} \right) p(x)\, d\mu(x) \right] p(\theta)\, d\lambda(\theta) .$$

Las propiedades de este criterio de comparación de experimentos son las mismas que verificaba el criterio de comparación de experimentos basado en la entropía de Shannon y pueden verse en Pardo, J. A. y otros (1992).

La cuantificación de la información en modelos con censura para su posterior utilización en la comparación de los mismos ha ocupado un lugar importante dentro de la Teoría de la Información. Entre otros cabe destacar los resultados debidos a Hollander, Proschan y Sconing (1985, a y b, 1987), Goel (1986), Baxter (1989), Turrero (1989), Barlow y Hsiung (1983), y Brooks (1980, 1982).

Mediante un criterio de comparación de experimentos basado en una medida de entropía se elige, entre varios experimentos asociados a θ, el que proporciona más información acerca de él. Puesto que a medida que aumenta el número de observaciones de un experimento, X, que contiene información acerca del parámetro desconocido, varía la información acerca de él, parece lógico preguntarse, una vez elegido el experimento más informativo, hasta qué momento se debe continuar la experimentación. Si se determina de antemano un nivel de incertidumbre, c, una posible regla de experimentación consiste en repetir el experimento más informativo hasta que el valor de la incertidumbre que se obtiene sea menor o igual que el valor prefijado, c. Esta claro que este método secuencial no hace consideraciones de riesgo o de coste de experimentación, sin embargo supone un conocimiento a priori. El primer trabajo en esta línea se debe a Lindley (1956) quien introdujo, desde la perspectiva bayesiana, un plan de muestreo

secuencial basado en la entropía de Shannon. Posteriormente DeGroot (1962) estudió algunas de sus propiedades y Lindley (1957) y El Sayad (1969) estudiaron el comportamiento de la misma en problemas binomiales y exponenciales, respectivamente. La utilización de otras medidas de entropía e información conduce a resultados análogos como puede verse en Pardo, L. (1984), Pardo, L. y otros (1985), Pardo y Taneja (1991), Vicente (1990), Pardo, J. A. y Vicente (1993a), etc.

Si se supone que el estadístico dispone de un conjunto de experimentos, $\mathfrak{E}$, y antes de tomar una decisión, puede realizar exactamente n experimentos de forma secuencial; es decir, puede elegir un experimento $X_1 = X \in \mathfrak{E}$, y observar $x \in X$, a continuación, puede elegir $X_2 = Y \in \mathfrak{E}$ y observar $y \in Y$,... El objetivo será encontrar la secuencia de experimentos $X_1,...,X_n$ de $\mathfrak{E}$, que minimice la Entropía de Shannon Terminal. En relación con este problema cabe destacar el trabajo de DeGroot (1962), Morales y otros (1986), Vicente (1990), Pardo, J. A. y Vicente (1993a), etc.

REFERENCIAS

ACZÉL, J. (1966). *Lectures on Functional Equations and Their Applications.* Academic Press. New York.

ACZÉL, J. (1969). «On different characterizations of entropies». *Lecture Notes in Mathematics* **89,** 1-11. Springer Verlag. Berlín.

ACZÉL, J. y DARÓCZY, Z. (1963). «Charakterisierung der entropien positiver ordnung und der Shannonschen entropie». *Act. Math. Acad. Sci. Hunger* **14,** 95-121.

ACZÉL, J. y DARÓCZY, Z. (1975). *On Measures of Information and their Characterization.* Academic Press. New York.

ACZÉL, J. y OSTROWSKI (1973). «On the characterization of Shannon's entropy by Shannon's inequality». *Australian Math. Soc.* **16,** 368-374.

ACZÉL, J. y PFANZAGL, J. (1966). «Remarks on the Measurement of Subjective Probability and Information». *Metrika* **11,** 91-105.

ACZÉL, J., B. FORTE Y C.T. NG (1974): «Why the Shannon and Hartley Entropies are Natural?. *Adv. Appl. Prob.* **6,** 131-146

ACZÉL, J. y KANNAPPAN, P. L. (1978). «A mixed Theory of Information III. Inset Entropy of degree β». *(Information and Control* **39,** 315-322.

ACZÉL, J. y NATH, P. (1972). «Axiomatic characterization of some measures of divergence in information». *Z. Wahrsch. verw. Geb.* **21,** 215-224.

AGGARWAL, N.L. (1974). «Sur l'information de Fisher», In: *J. Kampe de Feriet, Ed., Theories de l'information,* Springer-Verlag, Berlín, 111-117.

AGRESTI, A. y AGRESTI, B. (1979). *Statistical Methods for the Social Sciences.* San Francisco. Dellen.

ATKINSON, C. y MITCHELL, A. F. S. (1981). «Rao's distance measure». *Sankhya,* **43A,** 345-465.

ALI, S. M. y SILVEY, S. D. (1966). «A general class of coefficient of divergence of one distribution from another». *J. Royal Stat. Soc.* **286,** 131-142.

AMARI, S. I. (1984). «Differential Geometry of Statistics: Towards new Developments», In: *NATO Workshop on Differential Geometry in Statistical Inference,* London, 9-11 April.

AMARI, S. I. (1985). *Differential-Geometric Methods in Statistics.* Lecture Notes in Statistics, Springer Verlag, Berlín.

ARIMOTO, S. (1971). «Information-theoretical considerations on estimation problems». *Information and Control* **19,** 181-194.

ASH, R. B. (1965). *Information Theory.* Wiley Inter Science. New York.

ATKINSON, C. y MITCHELL, A. F. S. (1981). «Rao's distance measure». *Sankhyã,* **43, A,** 345-65.

BALAKRISHNAN, V. y SANGHVI, L. D. (1968). «Distance between populations on the basis of attribute». *Biometrics* **24,** 859-865.

BARLOW, R. y HSIUNG, J. (1983). «Information in a life testing experiment». *Statistician,* **32,** 35-45.

BASHARIN, G. P. (1959). «On a statistical estimate for the entropy of a sequence of independent random variables». *Theory of Probability and its Applications* **4,** 333-336.

BAXTER, L. A. (1989). «A note on information and censored absolutely continuous random variables». Stat. and decisión, **7,** 193-198.

BECTOR, C. R. y BHATIA, B. L. (1986). «Nature of Renyi's Entropy and Associated Divergence functions». *Naval Res. Logistics,* **33,** 742-746.

BEHARA, M. (1990). *Additive and Non-Additive Measures of Entropy.* Wiley Eastern. New Delhi.

BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis (2.ª ed.).* Springer Verlag.

BERNARDO, J. M. (1979a). «Expected utility as expected information». Annals of Statistics **7,** 686-690.

BERNARDO, J. M. (1979b). «Reference posterior distributions for Bayesian Inference». *Journal of Royal Statistical Society B,* **41,** 113-147.

BHARGAVA, T. N. y DOYLE, P. H. (1974). «A geometric study of diversity». *Metron* **30,** 1-13.

BHATTACHARYYA, A. (1943). «On a measure of divergence between two statistical populations defined by their probability distributions». *Bull Cal. Math. Soc.* **35,** 99-109.

BHATTACHARYYA, A. (1946). «On a measure of divergence between two multinomial populations». *Sankyya* **7,** 401-406.

BLACKWELL, D. (1951). «Comparison of Experiments». In: *Proc. 2nd Berkeley Symp. Math. Statist. Probl.* Univ. of Californ. Press, Berkeley, California, 93-103.

BLACKWELL, D. (1953). «Equivalent comparison of experiments». *Ann. Math Statis.***24,** 265-272.

BLACKWELL, D. y GIRSHICK, M. A. (1954). *Theory of Games and Statistical Decisions.* Wiley, New York.

BLAUT, R. E. (1987). *Principles and practice of Information Theory.* Addison-Wesley Publishing Company.

BOEKEE, D. E. (1977). «A Generalization of the Fisher Information Measures». *Delft Univ. Press, Delft.*

BOEKEE, D. E. (1978). «The D_f-Information of Order s». *Trans. 8th Prague Conf. on Inform. Th.* **Vol C,** N., 55-66.

BOLTZMANN, L. (1896). *Vorlesungen über Gastheorie.* J. A. Barth. Leipzig.

BORGES, R. (1967). «Zur Herleitung der Shannonschen Information». *Math. Z.* **96,** 282-287.

BRILLOUIN, L. (1956). *Science and Information Theory.* Academic Press, New York.

BRILLOUIN, L. (1964). *Scientific Uncertainty and Information.* Academic Press, New York.

BROOKS, R. J. (1982). «On the loss of information through censoring». *Biometrika* **69,** 137-144.

BROOKS, R. J. (1980). «On the relative efficiency of two paired-data experiments». *J. R. Statist. Soc. B.* **42,** 186-192.

BURBEA, J. (1986): «Information Geometry of Probability Spaces» *Exposit. Math.,* **4,** 347-378.

BURBEA, J. (1984). «The Convexity with Respect to Gaussian Distribution of Divergences of Order α». *Utilitas Mathematica* **26,** 171-192.

BURBEA, J. y OLLER, J. M. (1988). «The Information Metric for Univariate Linear Elliptic Models». *Statistics & Decision,* **6,** 209-221.

BURBEA, J. y RAO, C. R. (1982a). «Entropy Differential Metric, Distance and Divergence Measures in Probability Spaces: A Unified Approach». *J. Multi. Analy.,* **12,** 575-596.

BURBEA, J. y RAO, C. R. (1982b). «On the Convexity of Some Divergence Measures Based on Entropy Functions», IEEE *Trans. on Information Theory,* **IT-28,** 489-495.

CAMPBELL, L. L. (1965a). «Entropy as a measure». *IEEE Trans. Inform. Theory* **IT-11,** 112-114.

CAMPBELL, L. L. (1970). «Equivalence of Gauss's principle and Minimum discrimination information estimation of probabilities». *Ann. Math. Statist.* **41,** 1011-1015.

CAMPBELL, L. L. (1972). «Characterization of entropy of probability distributions on the real line». *Inf. Contr.* **21**, 329-343.

CAMPBELL, L. L. (1985). «The Relation Between Information Theory and the Differential Geometry Approach to Statistics». *Information Sciences,* **35,** 199-210.

CAMPBELL, L. L. (1987). «Information Theory and Differential Geometry.» Department of Math. & Statis, Queen's University Prepint # 1987-12.

CENCOV, N. N. (1982). «Statistical Decision Rules and Optimal Inference». *Trans. Of Math. Monographs* **53,** American Math. Soc. Providence R.I.

CLAUSIUS, R. (1864). *Abhandlungen über die mechanische Wärmetheorie.* Braunschweig.

COHEN, M. L. (1968). «The Fisher information and convexity». *IEEE Trans. Inform. Theory* **IT-14,** 591-592.

CONOVER, W. J.; JOHNSON, M. E. y JOHNSON, M. M. (1981). «A comparison study of test homogeneity of variances, with applications to the Outer Continental Shelf bidding data». *Technometrics* **23,** 351-361.

CONSUL, P. C. y SHENTON, L. R. (1972). «Use of Lagrange Expansion for generating discrete generalized probability distributions». *SIAM J. Appl. Math.* **23,** 239-248.

CRESSIE, N. y READ, T. R. C. (1984). «Multinomial goodness of fit tests». *J. R. Statistic. Soc. B,* **46,** 440-464.

CSISZÁR, I. (1962). «On the dimension and entropy of order α of the mixture of probatility distributions». *Acta Math. Acad. Sci. Hung.* **13,** 245-255.

CSISZÁR, I. (1967). «Information-type measures of difference of probability distribution and indirect observations». *Studia Sci. Math. Hungar.* **2,** 299-318.

CSISZÁR, I. (1969). «On generalized entropy». *Stud. Sci. Math. Hungary* **4,** 401-410.

CSISZÁR, I. (1972). «A class of measures of informativity of observation Channels». *Periodica Math. Hungar.* **2,** 191-213.

CSISZÁR, I. (1975). «I-divergence geometry of probability distributions and minimization problems». *Ann. Probab.* **3,** 146-158.

CSISZÁR, I. (1978). «Information measures: a critical survey». *Trans. 7th. Prague Conf. on Inform. Theory* Vol. B, 73-86. Academia, Praha.

CSISZÁR, I. y ELIAS, P. (1977). *Topics in Information Theory.* Nort-Holland.

CUADRAS, C. M. (1989). «Distancias Estadísticas», *Estadística Española,* **30 (111),** 295-378.

CHAUNDY, T. W. y MCLEOD, J. B. (1960). «On a Functional Equation». *Edinburgh Math. Notes* **43,** 7-8.

CHIANG, T. P. (1958). «A note on the definition of the amount of information». *Theory of Prob. and Appl.* **3,** 93-97.

DACEY, M. P. y NORELIFFE, A. (1976). «New entropy models in social sciences». *Env. and Plann.* **8A,** 299-310.

DARÓCZY, Z. (1963). «Uber die gemeinsame charakterisierung der zu dennicht vollstandigen vertilungen gehörigen entropien von Shannon und Rényi». *Z. Wahr. Verw. Geb.,* **1,** 381-388.

DARÓCZY, Z. (1969). «Uber ein funktionalgleichungssystem der infromationstheorie». *Aeq. math.,* **2,** 144-149.

DARÓCZY, Z. (1970). «Generalized Information functions». *Information and Control* **16,** 36-51.

DARÓCZY, Z. (1971). «On the measurable solution of a functional equation». *Acta Math. Acad. Sci. Hungar.,* **22,** 11-18.

DARÓCZY, Z. y KATAI (1970). «Additive zahlentheoretische funktionen und das mass der information». *Ann. Univ. Sci. Budapest. Eötvös Sect. Math.,* **13,** 83-88.

DE GROOT (1962). «Uncertainty, information and sequential experiments». *Ann. Math. Statist.* **33,** 404-419

DE GROOT (1970). *Optimal Statistical Decisions.* McGraw Hill, New York.

DIDERRICH, G. (1975). «The Role of Boundedness in Characterizing Shannon Entropy». *Information and Control.*

DOBRUSHIN, R. L. (1972). «Survey of Soviet research in information theory». *IEEE Trans. Inform. Theory* **IT-18,** 703-724.

DUDEWICZ, E. J. y VAN DER MEULEN, E. C. (1981). «Entropy-based tests of uniformity». *J. Amer. Statist. Assoc.* **76,** 967-974.

DUNFORD, N. y SCHWARTZ, J. T. (1958). *Linear Operations. Part I: General Theory.* Interscience Publishers, New York.

DUTTA, M. (1968). «A hundred years of Entropy». *Physics Today* **21,** 75-79.

DUTTA, M. (1966). «On maximum information-theoretic entropy estimation». *Sankhya* **28A,** 319-328.

EBRAHIMI, N. y SOOFI, E. S. (1990). «Relative information loss under Type II censored exponential data». *Biometrika 77,* 429-435.

EL-SAYYAD, G. M. (1969). «Information and Sampling from the Exponential Distribution». *Technometrics* **11,** 41-50.

EMPTOZ, E. (1977). «L'Energie Informationnelle». *Seminaire Questionnaires.* Université Paris VI.

EVANS, R. A. (1969). «The principle of minimum information». *IEEE Trans. on Reliability* **R-18,** 87-90.

FADDEV, D. K. (1956). «On the concept of entropy of finite probability». *Uspeki Mat. Naut. (NS)* **11,** 227-231.

FANO, R. M. (1959). «The statistical theory of information». *Nuovo Cimento* **13,** 353-372.

FEINSTEIN, A. (1958). *Foundations of Information Theory.* McGraw Hill. New York.

FERENTINOS, K. y PAPAIOANNOU, T. (1981). «New Parametric Measures Information», *Inform. and Cont.,* **51,** 193-208.

FERENTINOS, K. y PAPAIOANNOU, T. (1982). «Information in Experiments and Sufficiency», *J. Statist. Plann. & Inferen.,* **6,** 309-317.

FERENTINOS, K. y PAPAIOANNOU, T. (1983). «Convexity of Measures of Information and Loss of Information due to Grouping of Observations», *J. Comb. Inform. & Syst. Sci.,* **4,** 286-294.

FERRERI, C. (1980). «Hypoentropy and related heterogeneity divergence measures». *Statistica* **40,** 55-118.

FISHER, R. A. (1925). «Theory of statistical estimation». *Proc. Cambridge Philos. Soc.,* **22,** 700-725

FISHER, P. (1972). «On the Inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$». *Metrika* **18,** 199-208.

FISK, C. (1985). «Entropy and Information Theory». *Env. and Planning* **14 A,** 679-710.

FISK, C. y BROWN, G. R. (1975). «A noter on the entropy formulation of distribution models». *Op. Res. Quart.* **26**(4), 755-758.

FORTE, B. (1969). «Measures of information: the general axiomatic theory». *Rev. Francaise d'Informatique et de Rech.* **3R-2,** 63-89.

FORTE, B. y DARÓCZY, Z. (1968a). «A Characterization of Shannon's Entropy». *Boll. Un. Mat. Ital.* **1,** 631-635.

FORTE, B. y DARÓCZY, Z. (1968b). «Sopra un sistema di equazioni funzionali nella teoria dell'informazione». *Ann. Univ. Ferrara Sez.* **13,** 67-75.

GALLAGER, R. G. (1968). *Information Theory and Reliable Communication.* Wiley, New York.

GARCÍA-CARRASCO, M. P. (1978). «Criterios para la comparación de experimentos». *Trabajos de Est. e Inv. Oper.* **29,** 28-51.

GARCÍA-CARRASCO, M. P. (1982). «Criterio bayesiano para la comparación de experimentos basado en la maximización de la ganancia de energía informacional». *Actas XIII Reunión Nacional de Estadística* **2,** 65-72.

GEORGESCU, N. (1971). «The entropy law and the economic process». Harvard Univer. Press, Cambridge, Mas.

GEORGESCU, N. (1975). «The Measures of Information: A critique». In: *Modern Trends in Cybernetics and Systems.* (edited by J. Ross and C. Bill). Vol. II 187-217. Springer Verlag.

GIL, P. (1980). *Teoría Matemática de la Información.* Editorial I.C.E.

GIL, M. A. (1989). «A note on stratification and Gain in precision in estimating diversity from large Samples». *Commun. Statist. (Theory and Methods)* **18 (4),** 1521-1526.

GIL, M. A. (1982). «Criterion of maximizing the expected quietness (invariant by homotheties with respect to the utilities)». *R.A.I.R.O. Rech. Oper.,* **16,** 319-331.

GIL, M. A. y MARTÍNEZ, I. (1992). «A note on the asymptotic optimum allocation in estimating inequality from complete data». *Kybernetika,* **28,** 325-332.

GOEL, P. K. y DEGROOT, M. H. (1979). «Comparison of Experiments and Information measures», *Ann. Statist.,* **7,** 1066-1077.

GOEL, P. K. (1986). «Comparison of experiments and information in censored data». *Techn. Report. Department of Statistics.* The Ohio State University.

GOEL, P. K. y DEGROOT, M. H. (1980). «Information about hyperparameters in hierarchical models», *Tech. Report. No.* **160,** Department of Statistics, Carnegie-Mellon University.

GOOD, I. J. y SMITH, E. P. (1985). «The variance and covariance of a generalized index of similarity especially for a generalization of an index of Hellinger and Bhattacharyya». *Commun. Stat., Theory and Methods* **14,** 3053-3061.

GOKHALE, D. V. (1975). «Maximum entropy characterization of some distributions». In: *Statistical Distributions in Scientific Work.* Vol. III edited by Patel, Golz and Ord., 209-304, M.A. Ridel Boston.

GOKHALE, D. V. y KULLBACK, S. (1978). *The Information on Contingency Tables.* Marcel Dekker, New York.

GOOD, I. J. (1963). «Maximum entropy for hypothesis formulation especially for multidimensional contingency tables». *Ann. Math. Stat.* **34,** 911-934.

GOOD, I. J. (1965). «Maximum entropy for hypothesis formulation especially in multi-dimensional contingency tables». *Ann. Math. Statist.* **34,** 911-934.

GRAY, R. M. (1990). *Entropy and Information Theory.* Springer-Verlag.

GREENBERG, J. H. (1956). «The measurement of Linguistic diversity». *Language* **32,** 109-115.

GYÖRFI, L. y NEMETZ, T. (1978). «f-dissimilarity: a general class of separation measures of probability distributions». *Topics in Information Theory* (Ed. by I. Csiszar and P. Elias), 309-321. North Holland, Amsterdam.

GUIASU, S. (1977). *Information Theory with Applications.* McGraw Hill, New York.

GUIASU, S. (1986). «Maximum Entropy Conditions in Queueing Theory». *J. Ope. Res. Soc.* **37,** 293-301.

GUIASU, S. y THEODORESCU, R. (1968). *La Thérorie Mathématique de l'information.* Dunod. Paris.

GUIASU, S. y THEODORESCU, R. (1971). *Incertitude et Information.* Les Presses de l'Université Laval, Québec.

GUTTMAN, I. y PEÑA, D. (1988). «Outliers and influence: Evaluation by posteriors of Parameters in the Linear model». *Bayesian Statistics* **3,** Bernardo, J. M. y otros (editores). Oxford University Press.

HALMOS, P. R. (1964). *Measure Theory.* Academic Press, New York.

HAMDAN, M. A. y TSOKOS, C. P. (1971). «An information measure of association in contingency tables». *Inf. Contr.* **19,** 174-180.

HANSEN, O. H. y TORGERSEN, E. N. (1974). «Comparison of linear normal experiments». *Ann. Math. Statistics.,* **2,** 367-373.

HARRIS, B. (1976). «The statistical estimation of entropy in the nonparametric case». *In Topics in Information Theory.* Edited by Math. Res. Inst. Univ. of Wisconsin. USA. 323-355.

HART, P. E. (1971). «Entropy and other measures of concentration». *Journ. Roy. Stat. Soc.,* **134,** 73-85.

HATORI, H. (1958). «A note on the entropy of a continuous distribution». *Kodai Math. Sem. Rep.,* **10,** 172-176.

HAVRDA, J. y CHARVAT, F. (1967). «Concept of structural α-entropy». *Kybernetika,* **3,** 30-35.

HELLINGER, E. (1909). «Neue Begründung der Theorie quadrtischen Formen von uunendlichen vielen Veränderlichen». *J. Reine Angew. mathe.*, **36,** 210-271.

HEYER, H. (1982). *Theory of Statistical Experiments.* Springer, Berlin.

HERNITER, J. D. (1973). «An entropy model of brand purchase behaviour». *Journ. Market Res.*, **10,** 361-373.

HOBSON, A. (1969): «A new Theorem of Information Theory». *J. Stat. Phy,* **1,** 383-391.

HOLLANDER, M. PROSCHAN, F. y J. SCONING (1985a). «Information in censored models». *Report M701, Department of Statistics,* Tallahassee: Florida State Univ.

HOLLANDER, M. PROSCHAN, F. y J. SCONING (1985b). «Meaures of dependence for evaluating information in censored models. *Tech. Report M706, Department of Statistics,* Tallahassee: Florida State Univ.

HOLLANDER, M. PROSCHAN, F. y J. SCONING (1987). «Measuring information in right-censored models». *Naval Res. Logist.*, **34,** 669-681.

IBRAGIMOV, I. A. y CHASMINSKIJ, R. Z. (1981). *Asymptotic Theory of Estimation.* Springer Verlag. Berlin.

IKEDA, S. (1962a). «On characterization of the Kullback-Leibler mean information for continuous probability distributions». *Ann. Inst. Statist. Math.*, **14,** 73-79.

IKEDA, S. (1962b). «Necessary conditions for the convergence of Kullback-Leibler's mean information». *Ann. Inst. Statist. Math.*, **14,** 107-118.

JAYNES, E. T. (1957). «Information theory and statistical Mechanics». *Physical Reviews,* **106,** 620-630, **108,** 171-197.

JAYNES, E. T. (1963a). «New engineering application of information theory». In: *proceedings of the First Symposium in Engineering Applications of random function theory and probability,* edited by J. A. Bogdanoff and F. Kozen 163-203. John Wiley, New York.

JAYNES, E. T. (1963b). «Information Theory and Statistical Mechanics». *Statistical Physics* 1962. *brandeis* lectures in Math. Physics Vol. 3, edi. K. W. Ford, Benjamin, New York, 181-216.

JAYNES, E. T. (1967). «Foundations of probability theory and statistical mechanics». In: *Delaware Seminar on the Foundations of Science,* Vol. I, M. Munge. Editor, Springer Verlag 77-101.

JAYNES, E. T. (1968). «Prior probabilities». *IEEE Trans. S.S.C.*, **4,** 227-248.

JEFFREYS, H. (1946). «An Invariant form for the Prior Probability in Estimation Problems». *Proc. Royal Soc.*, Ser A, 186, 453-561.

JEFFREYS, H. (1961). *Theory of Probability.* Oxford University Press. London.

JELINEK, F. (1968). *Probabilistic Information Theory.* McGraw-Hill, New York.

JONES, D. S. (1979). *Elementary Information Theory.* Oxford Applied Mathematics and Computing Science Series.

JUSTICE, J. H. (1986). *Maximum Entropy and Bayesian Method in Applied Statistics.* Cambridge University Press.

KAGAN, A. M. (1963). «On the Theory of Fisher's Amount of Information». *Dokl. Acad. Nauk SSSR,* **151,** 277-278.

KAGAN, A. M., LINNIK, J. V. y RAO, C. R. (1973). *Characterization Problems in Mathematical Statistics.* Capítulo XIII. Wiley, New York.

KALE, B. K. (1964). «A note on the loss of information due to grouping of observations». *Biometrika,* **51 (314),** 495-497.

KAMPÉ DE FERIET, J. (1963). *Théorie de l'information. Principe du Maximum de l'Entropie et ses Applications á la Statistique et á la Mécanique.* Publications du Laboratoire de Calcul de la Faculté des Sciences de l'Université de Lille, Lille.

KAMPÉ DE FERIET, J. (1974). «La Théorie generalisée de l'information et la mesure subjective d'information». *Lecture Notes in Mathematics,* **398,** 1-35, Springer, Berlin.

KANNAPPAN, P. (1972a). «On Shannon's entropy, directed divergence and inaccuracy». *Z. Wahrscheinlichkeitstheorie,* **22,** 95-100.

KANNAPPAN, P. (1972b). «On directed divergence and inaccuracy». *Wahrscheinlichkeitstheorie,* **25,** 49-56.

KANNAPPAN, Pl. y Ng. C. T. (1973). «Measurable solutions of functional equations related to information theory». *Proc. Amer. Math. Soc.* **38,** 303-310

KANNAPPAN, P. y RATHIE, P. N. (1973). «On a characterization of directed divergence». *Inf. Contr.,* **22,** 163-171. *Polon. Math.,* **26,** 95-101.

KAPUR, J. N. (1967). «Generalized Entropy of Order α and Type β». *The Math. Seminar,* **4,** 78-82.

KAPUR, J. N. (1982). «Maximum-Entropy Probability Distribution for a Continuous Random Variable over a finite interval». *Journal Math. Phy. Sci.,* **16,** 4, 97-103.

KAPUR, J. N. (1986). «Entropy Measures of Economic Inequality». *Indian J. Pure & Appl. Maths.,* **17 (3),** 273-285.

KAPUR, J. N. (1988). «On Measures of Divergence based on Jensen Difference». *Nat. Acad. Sci. Letters,* **11** (1) (1988), 23-27.

KAPUR, J. N. (1989). *Maximum-Entropy Models in Science and Engineering.* John Wiley & Sons, New York.

KATZ, A. (1967). *Principles of Statistical Mechanics: The Information Theory Approach.* Freeman. New York.

KENDALL, D. G. (1964a). «Functional equations in information theory». *Wahrscheinlichkeitstheorie,* **4,** 225-229.

KENDALL, D. G. (1964b). «Information theory and the limit theorem for Markov chains and processes with a countable infinity of state». *Ann. Inst. Statist. Math.,* **15,** 137-143.

KERRIDGE, D. F. (1961). «Inaccuracy and Inference». *J. Royal Stat. Soc. Ser. B,* **23 (1),** 184-194.

KHINCHIN, A. I. (1953). «The concept of entropy in probability theory» (in Russian). *Uspekhi Matem. Nauk,* **8,** N. 3, 3-20.

KHINCHIN, A. I. (1956). «On basis theorems of information theory (in Russian). *Matem. Nauk,* **11,** N. 1, 17-75.

KHINCHIN, A. I. (1957). *Mathematical Foundations of Information Theory.* Dover Publications. New York.

KOLMOGOROV, A. N. (1956). «On the Shannon theory of information in the case of continuous signals». *Trans. IRE,* **IT-2,** 102-108.

KOLMOGOROV, A. N. (1959). «Entropy per unit time as a metric invariant of automorphisms» (in Russian). *Dokl. Akad. Nauk SSSR,* **124,** 754-755.

KOLMOGOROV, A. N. (1968). «Logical basis for information theory and probability theory». *IEEE Trans. Inform. Theory,* **IT-14,** 662-664.

KOTZ, S. (1966). «Recent results in information theory». *J. Appl. Prob.,* **3,** 1-93.

KOTZ, S., JOHNSON, N. M. y BOID, D. M. (1967). «Series representation of distribution and quadratic forms in normal variables». *I. Central Case,* AMS, 823-837.

KULLBACK, S. (1952). «An application of information theory to multivariate analysis». *Ann. Math. Statist.,* **23,** 88-102.

KULLBACK, S. (1954). «Certain inequalities in information theory and the Cramer-Rao inequality». *Ann. Math. Statist.,* **25,** 745-751.

KULLBACK, S. (1959). *Information Theory and Statistics.* Wiley, New York.

KULLBACK, S. y KHAIRAT, M. A. (1966): «A note on minimum discrimination information». *Ann. Math. Statist.,* **37,** 279-280.

KULLBACK, S. y LEIBLER, A. (1951). «On the Information and Sufficiency». *Ann. Math. Statist.*, **22,** 79-86.

KUPPERMAN, M. (1957). «Further application of information theory to multivariate analysis and statistical inference». *Ph. D. Dissertation,* George Washington University.

KUPPERMAN, M. (1958). «Probability hypothesis and information Statistics in sampling exponential-class populations». *Ann. Math. Statist.*, **29,** 571-574.

LATTER, B. D. H. (1973). «Measures of Genetic Distance Between Individuals and Populations». *Genetic Structure of Populations,* 27-39, Publicat. Univ. Hawaii Press.

LAZO, A. C. G. V. y RATHIE, P. N. (1978): «On the Entropy of Continuous Distributions» *IEEE Transactions on Informations Theory,* Vol., **IT-24,** n.° 1, 120-121.

LE CAM, L. (1964). «Sufficiency and approximate sufficiency». *Ann. Math. Statist.*, **35,** 1419-1455.

LE CAM, L. (1974). «On the information contained in additional observations». *Ann. Statist.*, **2,** 630-649.

LEE, P. M. (1964). «On the axioms of Information Theory». *Ann. Math. Statist.*, **35,** 415-418.

LEE, R. (1974). «A Markovian entropy maximizing model of population distribution». *Env. and Plann.*, **6A,** 693-702.

LEVINE, R. D. y TRIBUS, M. (1979). *The Maximum Entropy Formalism.* M.I.T. Press, Camb. Mass.

LEWONTIN, R. C. (1972). «The apportionment of human diversity». *Evolutionary Biology,* **5,** 381-398.

LIEBERSON, S. (1969). «Measuring population diversity». *Amer. Soc. Rev.*, **34,** 850-862.

LIESE, F. (1975). «On the existence of f-projections». *Colloq. Math. Soc. J. Bolyai,* **16,** Budapest, 431-446.

LIESE, F. y VAJDA, I. (1987). *Convex Statistical Distances.* Teubner, Leipzig.

LINDLEY, D. V. (1956). «On a Measure of Information provided by an Experiment». *Ann. Math. Statist.*, **27,** 986-1005.

LINDLEY, D. V. (1957). «Binomial sampling and the concept of information». *Biometrika,* **44,** 179-186.

MAHALANOBIS, P. C. (1936). «On the generalized distance in statistics». *Proc. Nat. Inst. Sci. India,* **2 (1),** 49-55.

MALLOWS, C. (1959). «The Information in an experiment». *J. Roy. Statist. Soc. Ser. B*, **21**, 67-72.

MANSURIPIR, M. (1989). *Introduction to information Theory*. Prentice-Hall, Inc.

MATHAI, A. M. (1967). Dispersion and Information. *Metron*, **26**, 1-12.

MATHAI, A. M. y RATHIE, P. N. (1975). *Basic Concepts of Information Theory and Statistics*, Wiley, New York.

MATUSITA, K. (1954). «On the estimation by minimum distance method». *Ann. Inst. Statist. Math.*, **5**, 59-65.

MATUSITA, K. (1955). «Decision rules, based on the distance for problems of fit, two samples, and estimation». *Ann. Math. Statist.*, **26**, 631-640.

MATUSITA, K. (1964). «Distance and Decision rules». *Ann. Inst. Statist. Math.*, **16**, 305-320.

MENÉNDEZ, M. L., TANEJA, I. J. y PARDO, L. (1991). «On Generalized Information Radii and Their Properties». *Soochow Journal of Mathematics*, **17**, 131-150.

MENÉNDEZ, M. L., TANEJA, I. J. y PARDO, L. (1990). «On Unified (r.s)-Jensen Difference Divergence Measures». *Journal of Computing and Information*, **1**, 111-128.

MENÉNDEZ, M. L., MORALES, D., PARDO, L. y SALICRÚ, M. (1992a). «Some Applications of (r,s)-Directed divergences». *Utilitas Mathematica*, **42**, 115-127.

MENÉNDEZ, M. L., MORALES, D., PARDO, L. y SALICRÚ, M. (1992b). «Asymptotic distribution of the generalized distance measure in a random sampling». Distance'92. Rennes. Francia, 337-340.

MITTAL, D. P. (1975). «On additive and non-additive entropies». *Kybernetika*, **11**, 271-280.

MORALES, D., PARDO, L. y QUESADA, V. (1987). «La Energía Informacional de Proceso Puntual». *Estadística Española*, **107**, 145-156.

MORALES, D., PARDO, L. y QUESADA, V. (1986). «Las medidas de f*-divergencia en el diseño Secuencial de Experimentos en un contexto Bayesiano». *Trabajos de Estadística*, **1**, 95-109.

MORALES, D. y PARDO, L. (1991). «Statistical Application of Hypoentropy Measure to Populations Diversity Analisis». *Trabajos de Estadística*, **6**, 55-62.

MORALES, D. y PARDO, L. (1993). «Algunos resultados en relación a los tests de bondad de ajuste basados en divergencias». Departamento de Estadística e I. O. Facultad de Matemáticas. Universidad Complutense de Madrid.

MORALES, D.; PARDO, L. y QUESADA, V. (1991). «The Chi-Square Divergence Measure in Random Sampling with Dirichlet Process Priors». *Information Sciences,* **25,** 239-249.

MORALES, D.; PARDO, L. y TANEJA, I. J. (1993). «Hipoentropy as an Index of Diversity». *Theory of Probability and its Applications,* **34,** 155-158.

MORALES, D.; PARDO, L.; SALICRÚ, M. y MENÉNDEZ, M. L. (1993a). «Asymptotic properties of divergence statistics in stratified random sampling and its applications to test statistical hipotheses». *Journal of Statistical Planning and Inference,* **34,** .

MORALES, D.; PARDO, L.; SALICRÚ, M. y MENÉNDEZ, M. L. (1993b). «New parametric measures of information based on R_h^ϕ-divergences». En *Multivariate Analysis: Future Directions 2, (C. M. Cuadras, C.R. Rao, eds.).* Elsevier-North-Holland, pp. 473-488.

MORALES, D.; PARDO, L.; SALICRÚ, M. y MENÉNDEZ, M. L. (1993c). «Parametric Multinomial Goodness of fit tests based on Divergence type statistics». Enviado.

NAYAK, T. P. (1983). «Applications of entropy functions in measurement and analysis of diversity». *Ph. D. thesis. Un. Pittsburgh.* Ed. Un. Microfilm International.

NAYAK, T. K. (1985). «On Diversity Measures based on Entropy Functions». *Commun. Statist. Theor. Meth.,* **14,** 1, 203-215.

NAYAK, T. K. y GASTWIRTH, J. L. (1989). «The use of Diversity Analysis to Assess the Relative Influence of Factors Affecting the Income Distributions. *J. Bussiness & Economic Statis.* **7 (4),** 453-460.

NEMETZ, T. (1967). «Information theory and testing of a hypothesis». *Proc. Colloq. Inform.* Th., Debrecen, Vol. 2.

NEMETZ, T. (1970). «Notes on the rate of convergence of the information provided by an experiment». *Studia Sc. Math. Hungar.,* **5,** 19-30.

Ng, C. T. (1974). «Representation of measures of information with a branching property». *Information and Control,* **25,** 45-56.

OLLER, J. M. (1989). «Some Geometrical Aspects of Data Analysis and Statistics», *Statistical Data Analysis and Inference,* Ed. Y. Dodge, Elsevier Science Publishers B.V. (North-Holland), 41-58.

OLLER, J. M. y CUADRAS, C. M. (1985). «Rao's distance for negative multinomial distributions». *Sankhya,* **47A,** 75-83.

ONICESCU, O. (1966). «Energie Informationelle». *C. R. Acad. Sci. Paris.*, **Ser. A,** 841-842.

PAPAIOANNOU, T. (1985). «Measures of Information». *Encyclopaedia of Statistical Sciences.* **Vol. 5,** 391-397, Kotz, S.y Johnson, N. L. ed.s, New York. Wiley.

PAPAIOANNOU, T. y KEMPTHORNE, O. (1971). «On statistical information theory and related measures of information». Aerospace Research Laboratories Report, ARL 71-0059, Wright-Paterson A.F.B., Ohio.

PARDO, J. A. y VICENTE, M. L. (1993a). «The ϕ-entropy in the selection of a fixed number of experiments». Aparecerá en *Kybernetes.*

PARDO, J. A.; MENÉNDEZ, M. L. y PARDO, L. (1992). «Las f*-divergencias como criterio bayesiano de comparación de experimentos». *Stochastica,* **XII-1,** 75-78.

PARDO, J. A.; MENÉNDEZ, M. L., TANEJA, I. J. y PARDO, L. (1993). «A slar measure of information for different probability distributions». Aparecerá en *Journal of Combinatorics Information and Systems Sciences.*

PARDO, J. A., MENÉNDEZ, M. L., TANEJA, I. J. y PARDO, L. (1993a). «Comparison of Experiments based on generalized entropy mesures». *Communications In Statistics: Theory and Methods.*, **22, 4,** 1113-1132.

PARDO, J. A., MENÉNDEZ, M. L., TANEJA, I. J. y PARDO, L. (1993b). «The Generalized entropy measures to the design and comparison of regression experiments in a bayesian context». *Information Sciences,* **73,** 93-105.

PARDO, L. (1981). «Energía Informacional Util». *Trabajos de Estadística e Investigación Operativa,* **32,** 85-94.

PARDO, L. (1984). «Plan de muestreo secuencial basado en la Energía Informacional en el modelo de Bernouilli». *Estadística Española,* **104,** 27-49.

PARDO, L. (1986). «Order α Useful Information Energy». *Information Science,* **40,** 155-164.

PARDO, L. (1986). «The measure of f-Divergence as a Stopping rule in the sequential random sampling in a bayesian context». *Statistica,* **2,** 243-251.

PARDO, L. (1987). «Plan de muestreo Secuencial basado en la Energía Informacional». *Real Academia de Ciencias Exactas, Físicas y Naturales de Madrid,* **LXXXI 1,** 102-115.

PARDO, L., MORALES, D. y QUESADA, V. (1985). «Plan de Muestreo Secuencial basado en la Energía Informacional para una población Exponencial». *Trabajos de Estadística y de Investigación Operativa,* **36,** 233-242.

PARDO, L. y MENÉNDEZ, M. L. (1989). «Application of the Informational Energy to the Design and Comparison of Regression Experiment in a Bayesian Context». *Jour. of Combinatorics, Information & Systems Sciences,* **31,** 237-252.

PARDO, L. y MORALES, D. (1991a). «An Index of Diversity in Stratified Random Sampling based on the Hypoentropy Measure». *Qüestiio,* **14,** 11-25.

PARDO, L. y TANEJA, I. J. (1991b). «Information Energy and its Applications». In: *Advances in Electronic and Electron Physics,* **90,** 165-241. Academic Press.

PARDO, L., MORALES, D. y TANEJA, I. J. (1991c). «λ-Measures of Hypoentropy and Comparison of Experiments: Bayesian Approach». *Statistica* **LI (2),** 173-184.

PARDO, L., MORALES, D. y TANEJA, I. J. (1991d). «Generalized Jensen diference measures and comparison of experiments». *Applications Mathematics,* **6,** 440-455.

PARDO, L., PARDO, J. A. y MENÉNDEZ, M. L. (1992). «Unified (r,s)-Entropy as an index of Diversity». *Jour. of the Franklin Institute,* **329,** 907-921.

PARDO, L., TANEJA, I. J. y MORALES, D. (1993). «Generalized Jensen Difference measures and Fisher measures of information». Aparecerá en la revista *Kybernetes.*

PARDO, L., SALICRÚ, M., MENÉNDEZ, M. L. y MORALES, D. (1993a). «R_h^ϕ-divergence statistics in applied categorical data analysis with stratified sampling». *Utilitas Mathematica,* **40,** 223-236

PARDO, L., SALICRÚ, M., MENÉNDEZ, M. L. y MORALES, D. (1993b). «Generalized divergences measures: Information matrices, amount of information and asymptotic distribution». Enviado.

PARDO, L., SALICRÚ, M., MENÉNDEZ, M. L. y MORALES, D. (1993c). «The φ-divergence statistic in Bivariate multinomial populations including stratification». *Metrika,* **40,** 223-236.

PARDO, L., SALICRÚ, M., MENÉNDEZ, M. L. y MORALES, D. (1993d). «Information Measures Associated to K-Divergences». In Uncertainty in Intelligent Systems. 375-382, North-Holland.

PATIL, G. P. y TAILLE, C. (1982). «Diversity as a concept and its measurement». *J. Amer. Stat. Assoc.,* **77,** 548-567.

PEARSON, K. (1900). «On the criterion that a given system of deviations from the probable in the case of correlated system of variables in such that it can be reasonably supposed to have arisen from random sampling». *Phil. Mag.,* **50,** 157-172.

PEÑA, D. y GUTTMAN, I. (1989). «Optimal Collapsing of mixture distributions in robust recursive estimation». *Communication in Statistics, Theory and Methods.*, **18,** 817-833.

PÉREZ, A. (1957). «Notations generalisees dincertitude, d'entropie et d'information du point de vue de la theorie de martingales». *Trans. Ist Prague Conf. on Information Theory.* Academia, Praha, 183-208.

PERÉZ, A. (1967a). «Information-theoretic risk estimates in statistical decision». *Kybernetika,* **3,** 1-21.

PÉREZ, A. (1967b). «Sur l'Energie Informationnelle de M. Octav Onicescu». *Roumaine Math. Pures Appl.*, **12,** 1341-1347.

PÉREZ, A. (1970). «Information-theoretic approach to measurment reduction problems». *Kybernetika,* **6,** 90-110.

PIELOU, E. C. (1975). «The use of information theory in the study of the diversity of biological populations». *Proc. Fifth Berk. Symp.* IV, 163-177.

PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes.* Holden-Day. San Francisco.

PINTACUDA, N. (1966). «Shannon entropy, a more general derivation». *Statistica,* **39,** 1310-1315.

PREDA, V. C. (1980). «The Student distribution and the principle of maximum entropy». *Ann. Inst. Statist. Math.*, **34,** 335-338.

PRESCOTT, P. R. (1976). «On a test for normality based on sample entropy». *Jour. Roy. Stat. Soc.*, **38-B,** 254-256.

RAIFFA, H. y SCHLAIFER, R. (1961). *Applied Statistical Decision Theory.* Mit Press.

RAO, C. R. (1945). «Information and accuracy attaintable in the estimation of statistical parameters». *Bull. Calcuta Math. Soc.,* **37,** 81-91.

RAO, C. R. (1982). «Diversity and Dissimilarity Coefficients: A Unified Approach». *J. Theoret. Popul. Biology,* **21,** 24-43.

RAO, C. R. (1987). «Differential Metrics in Probability Spaces». In: S. S. Gupta (Ed.). *Differential Geometry in Statistical Inference, IMS Lecture Notes-Monograph Series 10,* Hayward, California, 217-240.

RAO, C. R. y NAYAK, P. K. (1985). «Cross Entropy, Dissimilarity Measures, and Characterizations of Quadratic Entropy». *IEEE Trans. on Inform. Theory,* **IT-31** (5), 589-593.

RATHIE, P. N. (1970). «On a generalized entropy and a coding theory». *J. App. Prob.*, **7,** 110-133.

RATHIE, P. N. (1972). «Generalized entropy in coding theory». *Metrika*, **18,** 216-219.

RATHIE, P. N. y KANNAPPAN, P. (1971). «On a Functional Equations Connected with Shannon's Entropy». *Funkcial. Ekvac.*, **14,** 153-159.

RATHIE, P. N. y SHENG, L T. (1981). «The J-Divergence of Order α», *J. Comb. Inform. & Syst. Sci.*, **6,** 197-205.

RENYI, A. (1959). «On the dimension and entropy of probability distribution». *Acta Math. Acad. Sci. Hung.*, **10,** 193-215.

RENYI, A. (1961). «On Measures of Entropy and Information». *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, **1,** 547-561.

RENYI, A. (1965). «On the fundations of information theory». *Review of Int. Statist. Inst.*, **33,** 1-14.

RENYI, A. (1969). «On some basic problems of statistics from the point of view of information theory». *Proc. Collq. Inform. Theory*, Debrecen, Vol. II.

ROBERSTSON, C. A. (1972). «On minimum discrepancy measures». *Shankhya,* **14A,** 133-144.

RÜDSCHENDORF, L. (1984). «On the minimum discrimination information theorem». *Results in Estimation Theory.* Ed. E. J. Dudewicz, D. Plachky, P. K. Sen, Oldenbourg, München.

SAKAGUCHI, M. (1964). «Information Theory and Decision Making». Unpublished Lecture Notes, Statist. Dept., George Washington Univ., Washington DC.

SAKAMOTO, Y., ISHIGURO, M. y KITAGAWA, G. (1986). *Akaike Information Criterion Statistics.* D. Reidel Publishing Company.

SALICRÚ, M. (1993). «Las (h,ϕ)-entropías en poblaciones generales». Departamento de Estadística. Universidad de Barcelona.

SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D. y PARDO, L. (1992). «A Test of Independence based on the (r,s)-Directed Divergence». *Tamkang Journal of Mathematics,* **23,** 95-107.

SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D. y PARDO, L. (1993a). «On the applications of divergence type measures in testing statistical hypothesis». Enviado.

SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D. y PARDO, L. (1993b). «Asymptotic distribution of (h,ϕ)-entropies». *Communications in Statistics: Theory and Methods,* **22, 7,** 2015-2031.

SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D. y PARDO, L. (1993c). «Divergence measure based on entropy functions and statistical inference». Enviado.

SANT'ANNA, A. P. y TANEJA, I. J. (1985). «Trigonometric Entropies, Jensen Difference Divergences and Error Bounds». *Infor. Scien.*, **35**, 145-156.

SCONING, J. (1985). «Information in censored models». *Florida State University Ph. D. Disertation.*

SGARRO, A. (1981). «Information Divergence and the Dissimilarity of Probability Distributions». *Estratto di Calcolo,* **XVIII,** 293-302.

SHANNON, C. E. (1948). «A Mathematical Theory of Communication». *Bell. Syst. Tech. J.,* **27,** 379-423.

SHANNON, C. E. y WEAVER, W. (1949). *The Mathematical Theory of Communication.* Univ., **I11.** Press, Urbana Chicago.

SHARMA, B. D. y MITTAL, D. P. (1975). «New Non-additive Measures of Relative Information». *J. Comb. Inform. & Syst. Sci.,* **2,** 122-133.

SHORE, J. E. (1979). «Minimum cross-entropy spectral analysis». *Naval Res. Lab., Washington, DC 20375,* NRL Memo. Rep. 3921, Jan.

SIBSON, R. (1969). «Information Radius», *Z. Wahrs und verw Geb.,* **14,** 149-160.

SIMPSON, E. H. (1949). «Measurement of diversity». *Nature,* **163,** 688.

SMITH, C. R. y ERICKSON, G. J. (1986). *Maximum Entropy and Bayesian Analysis and Estimation. D. Reidell.*

STONE, M. (1959). «Application of a measure of information to the design and comparison of regression experiments». *Ann. Math. Statist.,* **30,** 55-79.

SZILARD, L. (1964). «Uber die Entropieverminderung in einem Thermodynamischen System bei Einsgriff intelligent beings». *Behavioral Science,* **9,** 301-310.

TANEJA, I. J. (1979). «Some Contributions to Information Theory I (A Survey): On Measures of Information». *J. Comb., Inform. & Syst. Sci.,* **4 (4),** 253-274.

TANEJA, I. J. (1983). «On a Characterization of J-Divergence and its Generalizations». *J. Comb. Inform & Syst. Sci.,* **8 (3),** 206-212.

TANEJA, I. J. (1986a). «On the Convexity and Schur-Convexity of Burbea and Rao's Divergence Measures and Their Generalizations», 7th *National Symposium on Prob. and Statist.,* Campinas, SP, Brazil.

TANEJA, I. J. (1986b). «λ-Measures of Hypoentropy and Their Applications». *Statistica,* **XLVI,** 465-478.

TANEJA, I. J. (1986c). «Unified Measures of Information Applied to Markov Chains and Sufficiency». *J. Comb., Inform. & Syst. Sci.,* **11,** 99-109.

TANEJA, I. J. (1987). «Statistical Aspects of Divergence Measures». *J. Statist. Planning & Inference,* **16,** 136-145.

TANEJA, I. J. (1989). «On Generalized Information Mesures and their Applications». *Ad. Electronics and Electron Physics,* **76,** 327-413.

TANEJA, I. J. (1990). «Bounds on the Probability of Error in Terms of Generalized Information Radii», *Information Science,* **46.**

TANEJA, I. J., PARDO, L. y MORALES, D. (1989a). «λ-Measures of Hypoentropy and Comparison of Experiments: Balckwell and Lhemann Approach». *Kybernetika,* **27,** 413-420.

TANEJA, I. J., PARDO, L. y MENÉNDEZ, M. L. (1991). «Some Inequalities Among Generalized Divergence Measures». *Tamkang Journal of Mathematics,* **22,** 175-185.

TANEJA, I. J., PARDO, L., MORALES, D. y MENÉNDEZ, M. L. (1989b). «On Generalized Information and Divergence Measures and their Applications: A Brief review». *Qüestiió,* **13,** 47-75.

THEIL, H. (1972). *Statistical Descomposition Analysis.* North-Holland Pub. Co. Amsterdam.

THEIL, H. (1980). «The symmetric maximum entropy distribution». *Economic Letters,* **6,** 53-57.

THEIL y FIEBIG (1984): «Exploiting continuity: Maximum entropy estimation of continuons distribution». Bollinger, Cambridge

THEODORESCU, A. (1977). «Energie Informationnelle et notions apparentees». *Trabajos de Estadística e I. O.,* **27,** 276-298.

TORGERSEN, E. N. (1970). «Comparison of expriments when the prarameter space is finite». *Zeitschr. Wahrscheinlichkeitstheorie geb.,* **16,** 219-249.

TORGERSEN, E. N. (1976). «Comparison of statistical experiments». *Scand. J. Statist.,* **3,** 186-208.

TORGERSEN, E. N. (1977). «Mixtures and products of dominated experiments». *Ann. Statist.,* **5,** 44-64.

TORGERSEN, E. N. (1980). «Deviations from total information and from total ignorance as measures of information». *Banach Center Publications,* Wrocalaw, Vol. 6, 315-322.

TRIBUS, M. (1969). *Rational Descriptions Decisions and Designs.* Pergamon Press, New York.

TURRERO, A. (1988). *Pérdida de información a causa de la censura.* Tesis Doctoral, Universidad Complutense de Madrid.

TURRERO, A. (1989). «On the relative efficiency of grouped and censored data». *Biometrika,* **76,** 125-131.

TVERBERG, H. (1958). «A new derivation of the information function». *Math. Scand.,* **6,** 297-298.

TZANNES, N. S. y NOONAN, J. P. (1973). «The mutual information principle and applications». *Information and Control,* **22,** 1-12.

VAJDA, I. (1968). «Axioms of α-entropy of a generalized probability scheme». *Kybernetika,* **4,** 105-112.

VAJDA, I. (1970a). «Note on discrimination information and variation». *Trans. IEEE,* **IT-16,** 771-773.

VAJDA, I. (1970b). «On preservation and maximization of information in data reduction process». *Inform. Transm. Problems,* **6,** 31-40.

VAJDA, I. (1970c). «On the amount of information contained in a sequence of independent observations». *Kybernetika,* **6,** 306-324.

VAJDA, I. (1972). «On the f-divergence and singularity of probability measures». *Period. Math. Hungar.,* **2,** 223-234.

VAJDA, I. (1973). «χ^aDivergence and Generalized Fisher's Information». *Trans. 6th Prague Conf. on Inform. Th.,* 873-886.

VAJDA, I. (1984). «Minimum divergence principle in statistical estimation». *Statistics and Decisions,* **1,** 239-262.

VAJDA, I. (1989). *Theory of Statistical Inference and Information.* Kluvwer Academic Press, Dordrecht, The Netherlands.

VAJDA, I. y VASEK, K. (1985). «Majorization, concave entropies, and comparison of experiments». *Prob. of control and Information Theory,* **14,** 105-115.

VAN DER LUBBE, J. C. A. (1981). «A generalized probabilistic theory of the measurement of certainty and information». Ph. D. Thesis, Dept. of Electrical Engineering, Delf Univ. of Technology. Delft. The Netherlands.

VARMA, R. S. (1966). «Generalizations of Rényi's Entropy of Order α». *J. Math. Sci.,* **1,** 34-48.

VASICEK, O. (1976). «A test of normality based on sample entropy». *J. Roy. Statist. Soc.,* **B 38,** 54-59.

VICENTE, M. L. (1990). *Aplicaciones estadísticas de la entropía no aditiva de grado β de Havrda y Charvat.* Tesis doctoral, Universidad Complutense de Madrid.

VICENTE, M. L. (1991). «La información de grado β como criterio de comparación de experimentos». *Trabajos de Estadística.* **Vol. 6,** 87-109.

WRAGG, A. y DOWSON, D. C. (1970). «Fitting Continuous Probability Density Functions over [0,∞) Using Information Theory». *IEEE Trans.*, **IT-16,** 226-230.

WONG, K. M. y CHEN, S. (1990). «The entropy of ordered sequences and order statistics». *IEEE Transactions on Information Theory,* vol. **36, 2,** 276-284.

YAGLOM, A. M. y YAGLOM, I. M. (1969). *Probabilité et Information. Dunod.*

ZOGRAFOS, K. (1991). «Asymptotic distributions of estimated f-dissimilarity between populations in a stratified random sampling». Tech. Report#191, Math. Dept. Univ. of Ioannina.

ZOGRAFOS, K. (1992). «Asymptotic properties of φ-divergence statistic and its application in contingency tables». Tech. Report#185, Math. Dept., Univ. of Ioannina, to appear in the Int'l J. of Mathematical and Statistical Sciences.

ZOGRAFOS, K., FERENTINOS, K. y PAPAIOANNOU, T. (1990): «φ-Divergence Statistics: Sampling Properties and Multinomial Goodness of fit and Divergence Tests». *Commun. Statist. (Theory and Meth.),* **19 (5),** 1785-1802.

ZYAROVA, J. (1973). «On asymptotic behaviour of a sample estimator of Renyi's information of order α». *Trans. 6th Prague Conf. on Inf. Theory Stat. Funct. and Proc.,* Prague, Czech. Acad. of Sci. 914-924.

THEORY OF STATISTICAL INFORMATION

SUMMARY

In this paper, an analysis is made of some applications of the measures of entropy and divergence in Statistics, enhancing the growing interest lately shown by a great many researchers for this branch of Statistics, known as the Theory of Statistical Information. Emphasis is laid on the new prospects for the construction of hypotheses tests through the use of measures of entropy and divergence.

Key Words: Measures of entropy, Measures of divergence, Parametric information measures, Index of diversity, Hypotheses tests.

AMS Clasification: 62B10, 94A17, 62E20.

ESTADISTICA ESPAÑOLA
Vol. 35, Núm. 133, 1993, págs. 269 a 316

COMENTARIOS

RAMON ARDANUY ALBAJAR

Departamento de Matemática Pura y Aplicada
Universidad de Salamanca

1. INTRODUCCION

El artículo del Profesor Leandro Pardo Llorente presenta de forma unificada y sistematizada un conjunto de resultados sobre las medidas de entropía y divergencia que son de gran utilidad en lo que actualmente se conoce como Teoría de la Información Estadística. En relación con su artículo voy a destacar dos puntos que quizás no estén suficientemente tratados: la aplicación de la divergencia funcional en problemas de estimación paramétrica y la clasificación de distribuciones con máxima entropía y momentos dados.

2. DIVERGENCIA FUNCIONAL Y ESTIMACION PARAMETRICA

Un problema que con frecuencia se suele presentar en la teoría de la estimación es el de elegir la función de pérdida; unas veces se toma la pérdida cuadrática por la comodidad matemática, en otras ocasiones se elige una pérdida no cuadrática si no resulta satisfactorio el estimador obtenido bajo una función de pérdida cuadrática, tal es el caso de la estimación de un parámetro de escala θ al considerar, por razones de invariancia, funciones de pérdida de la forma:

$$L(\theta, \hat{\theta}) = l\left(\frac{\theta}{\hat{\theta}}\right) \tag{1}$$

Resulta razonable considerar que la estimación de un parámetro es un paso en el objetivo de estimar la correspondiente distribución de probabilidad, por lo cual, la función de pérdida debe medir la discrepancia entre la distribución de probabilidad cuando el parámetro desconocido es θ y la que se obtiene cuando

se sustituye el parámetro por su estimación $\hat{\theta}$. Así pues, si X es una variable aleatoria cuya distribución de probabilidad $P(x|\theta)$ depende de uno o varios parámetros $\theta \in \Theta \subset \mathbf{R}^n$ y $\hat{\theta}$ es un estimador de θ, la función de pérdida debe ser un cierto funcional entre las medidas de probabilidad $P(\cdot|\theta)$ y $P(\cdot|\hat{\theta})$, es decir:

$$L(\theta,\hat{\theta}) = \Psi\,[P(\cdot|\theta), P(\cdot|\hat{\theta})] \tag{2}$$

pero la discrepancia entre ambas distribuciones de probabilidad podemos medirla por medio de la divergencia funcional, Kullback (1959), definida en el caso discreto por:

$$L(\theta, \hat{\theta}) = \sum_k P(x_k|\theta) \ln \frac{P(x_k|\theta)}{P(x_k|\hat{\theta})} \tag{3}$$

y en el caso absolutamente continuo por:

$$L(\theta, \hat{\theta}) = \int f(x|\theta) \ln \frac{f(x|\theta)}{f(x|\hat{\theta})}\, dx \tag{4}$$

Kashyap (1974) utiliza la divergencia funcional para la estimación minimax del parámetro de una distribución Binomial, extendiendo algunos resultados al caso multinomial. En Ardanuy (1978) puede encontrarse una aplicación de la divergencia funcional para estimar secuencialmente el parámetro p de una distribución binomial. Teniendo en cuenta que $\ln x \geq 1 - 1/x$, resulta inmediato probar que la función de pérdida L definida en (3) y (4) verifica:

$$L(\theta, \hat{\theta}) = \begin{cases} 0 & \text{si } \theta = \hat{\theta} \\ \geq 0 & \text{si } \theta \neq \hat{\theta} \end{cases} \tag{5}$$

así, para la distribución binomial $\mathcal{B}(n,p)$, $p \in (0,1)$ y n conocido, se tiene:

$$L(p, \hat{p}) = n\left[p \ln \frac{p}{\hat{p}} + (1-p) \ln \frac{1-p}{1-\hat{p}}\right] \tag{6}$$

función que es convexa en $\hat{p}$ y estrictamente positiva para $p \neq \hat{p}$, propiedades éstas que parecen deseables para una función de pérdida; en ocasiones la propiedad de convexidad no se satisface, pero con frecuencia se cumple la de ser L direccionalmente creciente, esto es, L es creciente cuando $\hat{\theta}$ se aleja de θ en cualquier dirección fijada de antemano. Para otras distribuciones notables obtenemos las siguientes funciones de pérdida basadas en la divergencia funcional:

Poisson $\mathcal{P}(\lambda)$ $\qquad L(\lambda\,\hat{\lambda}) = (\hat{\lambda} - \lambda) + \lambda \ln \frac{\lambda}{\hat{\lambda}}$ (7)

Binomial Negativa $\mathcal{BN}(r, p)$ $L(p, \hat{p}) = r\left[\ln \frac{p}{\hat{p}} + \frac{1-p}{p} \ln \frac{1-p}{1-\hat{p}}\right]$ (r conocido) (8)

Normal $\mathcal{N}(\mu, \sigma^2)$ $L(\mu, \sigma^2, \hat{\mu}, \hat{\sigma}^2) = \frac{1}{2}\left[\frac{(\mu - \hat{\mu})^2}{\hat{\sigma}^2} + \ln \frac{\hat{\sigma}^2}{\sigma_2} - 1\right]$ (9)

Exponencial $\mathcal{E}(\mu)$ $L(\mu, \hat{\mu}) = \frac{\mu}{\hat{\mu}} + \ln \frac{\hat{\mu}}{\mu} - 1$ (10)

Es fácil comprobar, por ejemplo, que para la distribución exponencial $\mathcal{E}(\mu)$ la función de pérdida $L(\mu, \hat{\mu})$ asociada a la divergencia funcional no es convexa en $\hat{\mu}$ pero, en cambio, sí es direccionalmente creciente.

3. DISTRIBUCIONES DE MAXIMA ENTROPIA EN [a,b] CON MOMENTOS DADOS DE SEGUNDO ORDEN

Sea X una variable aleatoria contínua que varía sobre un intervalo cerrado [a,b] y sea f(x) su función de densidad, la entropía de Shannon H(f) está dada por:

$$H(f) = -\int_a^b f(x) \ln f(x)\, dx \tag{11}$$

y como muy bien indica el Profesor Pardo Llorente en su trabajo, mediante el método de los multiplicadores de Lagrange se obtiene que la densidad de probabilidad que maximiza la entropía de Shannon bajo restricciones del tipo:

$$\int_a^b g_i(x) f(x)\, dx = \alpha_i \qquad (i = 1, 2, \ldots, n) \tag{12}$$

está dada por:

$$f(x) = \begin{cases} K \exp\left[\sum \lambda_i g_i(x)\right] & \text{si } a \le x \le b \\ 0 & \text{en otro caso} \end{cases} \tag{13}$$

con entropía máxima $H_{max} = -\ln K - \sum \lambda_i \alpha_i$. Este resultado también es conocido como Principio de Jaynes de Máxima Entropía (vea por ejemplo Jaynes (1957), Dowson y Wragg (1973), Kagan, Linnik y Rao (1973), Guiasu (1977), Kapur (1983), etc.).

Si denotamos con $\varphi(\lambda) = \dfrac{\exp[\lambda(b-a)]-1}{b\exp[\lambda(b-a)]-a}$, con $-\infty < a < 0 < b < +\infty$, entonces es fácil comprobar que esta función satisface lo siguiente:

i) $\varphi(-\infty) = 1/a$ y $\varphi(+\infty) = 1/b$

ii) $\varphi(0) = 0$ y $\varphi'(0) = 1$

iii) φ es estrictamente creciente.

iv) Tomando $\lambda_0 = \dfrac{\ln(|a|/b)}{b-a}$, φ es estrictamente convexa para $\lambda \leq \lambda_0$ y estrictamente cóncava para $\lambda \geq \lambda_0$.

v) La ecuación $\varphi(\lambda) = \lambda$ tiene sólo la solución $\lambda = 0$ si $a + b = 0$, o tiene una raíz nula y otra que es no nula si $a + b \neq 0$, en cuyo caso la raíz no nula satisface:

$$\begin{cases} \dfrac{\ln(|a|/b)}{b-a} < \lambda < 1/b & \text{si } |a| > b \\ 1/a < \lambda < \dfrac{\ln(|a|/b)}{b-a} & \text{si } |a| < b \end{cases}$$

pues bien, utilizando distribuciones centradas podemos establecer el siguiente resultado.

Proposición 1: Dado un intervalo finito [a,b], con $a < 0 < b$, la función de densidad con máxima entropía y media cero es la distribución uniforme si $a + b = 0$, o la distribución exponencial truncada:

$$f(x) = \begin{cases} K\exp(\lambda x) & \text{si } a \leq x \leq b \\ 0 & \text{en otro caso} \end{cases}$$

si $a + b \neq 0$, donde $K = \dfrac{\lambda}{\exp(\lambda b) - \exp(\lambda a)} > 0$ y λ es la raíz no nula de la ecuación $\varphi(\lambda) = \lambda$. El máximo valor de la entropía es $H_{máx} = -\ln K$, con $K = 1/(b-a)$ en el caso uniforme.

Utilizando el principio de Jayne de máxima entropía podemos obtener el siguiente resultado para el caso de fijar los dos primeros momentos.

Proposición 2: Dado un intervalo finito [a,b], con $a < 0 < b$, la función de densidad con máxima entropía, media cero y varianza uno está dada por:

$$f(x) = \begin{cases} K \exp\left[\alpha x + \frac{1}{2}\beta x^2\right] & \text{si } a \le x \le b \\ 0 & \text{en otro caso} \end{cases}$$

con $K > 0$, y se satisface además que:

i) $b\,f(b) - af(a) = \beta + 1$

ii) $f(b) - f(a) = \alpha$

Así pues, las situaciones que pueden presentarse bajo las condiciones de la proposición 2 son los cinco casos siguientes (vea también la Fig. 1):

a) $\beta > 0$: Distribución exponencial cuadrática truncada.

b) $\beta < 0$: Distribución normal truncada.

c) $\beta = 0$ y $\alpha > 0$: Distribución exponencial positiva truncada.

d) $\beta = 0$ y $\alpha < 0$: Distribución exponencial negativa truncada.

e) $\beta = 0$ y $\alpha = 0$: Distribución uniforme.

pudiendo caracterizarse las distribuciones uniforme y exponencial truncada del siguiente modo:

Proposición 3: Consideremos densidades de probabilidad con media cero y varianza 1 en el intervalo cerrado [a,b], entonces:

i) La distribución de máxima entropía es la uniforme precisamente si $b = -a = \sqrt{3}$.

ii) La distribución de máxima entropía es una exponencial truncada precisamente si $\frac{b^2 - a^2}{ab + 1} = \ln\left(\frac{a^2 - 1}{b^2 - 1}\right)$ cuando $b \neq -a$, o bien si es $b = -a = \sqrt{3}$ (uniforme). En estas condiciones se tiene además $b > 1$, $a < -1$ y $\alpha = \frac{a + b}{ab + 1}$.

Algunos resultados interesantes sobre distribuciones de máxima entropía con momentos dados son los que pueden encontrarse en Dowson y Wragg (1973), Goldman (1955), Kagan, Linnik y Rao (1973, Cap. 13), Kapur (1982, 82a, 82b, 83 y 89), Lisman y van Zuylen (1972), Wilson y Wragg (1973) y Wragg y Dowson (1970) principalmente.

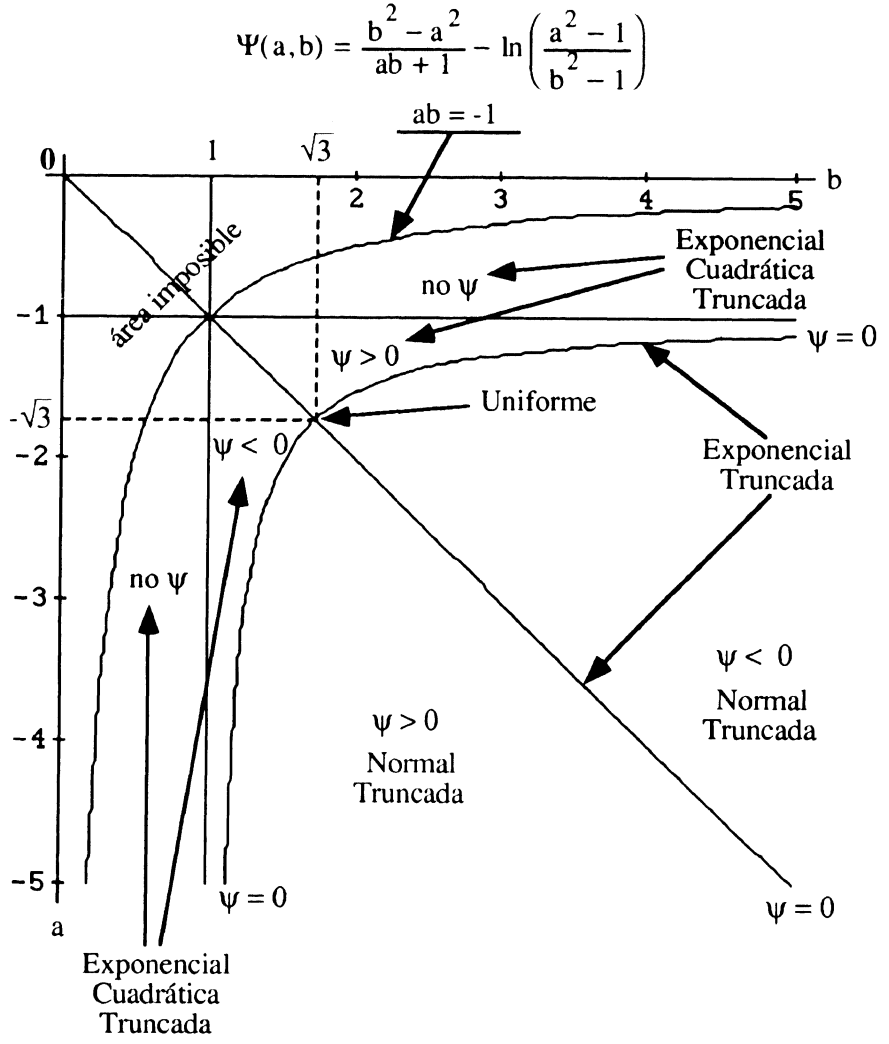

Fig. 1: *Distribuciones de máxima entropía en [a,b] con media cero y varianza uno.*

REFERENCIAS

ARDANUY, R. (1978), «Estimación Secuencial Optima de una Distribución Binomial tomando como Pérdida la Divergencia Funcional», *Trabajos de Estadística y de Investigación Operativa,* **29, (1),** 22-33.

DOWSON, D. C. and WRAGG, A. (1973), «Maximum-Entropy Distributions Having Prescribed First and Second Moments». *IEEE Transactions on Information Theory,* **IT-19,** 689-693.

GOLDMAN, S. (1955), «Information Theory», Prentice Hall, New York.

GUIASU, S. (1977), «Information Theory with Applications», McGraw-Hill International Book Company, New York.

JAYNES, E. T. (1957), «Information Theory and Statistical Mechanics», *Physical Review,* **106, (4),** 620-630, y **108, (2),** 171-190.

KAGAN, A. M., LINNIK, J. V. and RAO, C. R. (1973), «Characterisation Problems in Mathematical Statistics», Wiley, New York.

KAPUR, J. N. (1982), «Maximum-Entropy Probability Distributions for a Continuous Random Variate Over a Finite Interval». *Jour. Maths. Phy. Sci.*, **16, (1),** 97-109.

KAPUR, J. N. (1982a), «Maximum-Entropy Formalism for some Univariate and Multivariate Lagrangian Distributions», *Aligarh Journal of Statistics,* **2,** 1-16.

KAPUR, J. N. (1982b), «The Maximum-Entropy Principle and Its Applications to Science and Engineering». *Proc. National Symposium on Mathematical Modelling,* M.R.I. Allahabad, July 19-20, pp. 75-98.

KAPUR, J. N. (1983), «Maximum-Entropy Probability Distributions for Continuous Random Variates», *Journal of the Indian Society of Agricultural Statistics,* **35, (3),** 91-103.

KAPUR, J. N. (1989), «Maximum-Entropy Models in Science and Engineering», John Wiley & Sons, Inc., New Delhi, India.

KASHYAP, R. L. (1974), «Minimax Estimation with Divergence Loss Function», en *Estimation Theory,* editado por D. G. Lainiotis, American Elsevier Pub. Co., New York.

KULLBACK, S. (1959), «Information Theory and Statistics», Dover, New York.

LISMAN, J. H. C. and VAN ZUYLEN, M. C. A. (1972), «Note on the Generation of Most Probable Frequency Distributions», *Statistica Neerlandica,* **26, (1),** 19-23.

WILSON, G. A. and WRAGG, A. (1973), «Numerical Methods for Aproximating Continuous Probability Density Functions, Over [0,∞), Using Moments». *J. Inst. Maths. Applics.,* **12,** 165-173.

WRAGG, A. and DOWSON, D. C. (1970), «Fitting Continuous Probability Density Functions Over [0,∞) Using Information Theory Ideas», *IEEE Transactions on Information Theory,* **IT-16,** 226-230.

CARLES M. CUADRAS

Departament d'Estadística
Universitat de Barcelona

Agradezco a Estadística Española la oportunidad que me ofrece de comentar el artículo del Profesor Leandro Pardo Llorente, digno representante del grupo de investigadores españoles que trabajan en el tema de la medida de la información en estadística.

Cuando en una población se toman datos sobre una o varias variables estadísticas, surge de forma natural, en muchas aplicaciones, la necesidad de cuantificar la medida de la incertidumbre o cantidad de información de un sistema.

1. SOBRE LA ENTROPIA

Entiendo que la definición de entropía de una distribución

$$H = -\int p(x) \log p(x)\, dx \tag{1}$$

como fue inicialmente dada por Shannon, sigue siendo la que posee mayor justificación. En primer lugar, por su origen físico y como medida de la información en teoría de la comunicación, como Pardo explica. Otras poderosas razones son las siguientes.

1.1. Distribuciones continuas con máxima entropía

Fijada la media μ y la varianza σ^2, la distribución absolutamente contínua con soporte $(-\infty,+\infty)$, es la normal $N(\mu,\sigma^2)$ (Rao, 1973, p. 162). Resultado razonable, pues una variable normal se puede interpretar como suma límite de infinitas cantidades aleatorias independientes. Si el soporte es $(0,\infty)$ y la media es dada, entonces aparece la distribución exponencial. Si el soporte es un intervalo (a,b), la distribución de máxima entropía es la uniforme. Todos los puntos en (a,b) tienen igual densidad de probabilidad y la incertidumbre para el observador es máxima.

1.2. Distribuciones discretas con máxima entropía

Si un sistema tiene k estados y distribución de probabilidad $p = (p_1,...,p_k)$, entonces

$$H_p = -\sum_{j=1}^{k} p_j \log_2 p_j \tag{2}$$

es máximo para $p_j = 1/k$, $j=1,\ldots,k$. Por ejemplo, si $k=8$,

$$H = -\Sigma \frac{1}{8} \log_2 \left(\frac{1}{8}\right) = 3$$

tiene una sencilla interpretación: para adivinar una bola de una urna que contiene 8 bolas equiprobables, tenemos que hacer 3 preguntas (existen juegos de adivinación de cartas que se basan en expresiones análogas). La entropía es una cota inferior del número medio de bits (respuestas sí o no a preguntas binarias) y como tal posee una perfecta justificación dentro de la teoría de la información.

La entropía (2) se extiende a variables discretas con soporte numerable, pero entonces es fácil construir ejemplos de variables cuya entropía es infinita.

1.3. Aumento natural de la entropía

Existe una tendencia natural en el universo físico a aumentar la entropía con el tiempo. Sea p_0 la distribución del sistema en un momento dado y supongamos que la distribución p_1 en el paso siguiente en el tiempo, se rige según una cadena de Markov homogénea, $p_1 = P\, p_0$, siendo P la matriz de transición. Una propiedad interesante es que

$$H^*_{p_1} \geq H_{p_0} \tag{3}$$

donde H_{p_0} se calcula como en (2), mientras $H^*_{p_1}$ viene dada por

$$H^*_{p_1} = -\Sigma\, p_j^1 \log(p_j^1 / p_{.j})$$

donde $p_{.j} = \Sigma_i\, p_{ij}$ y p_j^1, $j = 1, \ldots k$, las componentes de p_1. La interpretación de (3) es que, en cada paso, la entropía o desorden natural aumenta. La tecnología sería entonces un intento humano de disminuir este crecimiento natural de la entropía mediante la imposición de restricciones sobre el entorno, aunque la entropía total no llega a disminuir (Cressie, 1991).

1.4. Como medida de divergencia en estimación

Pardo nos indica que la medida de divergencia de Kullback-Leibler (K-L), totalmente ligada a la medida de la entropía (1), proporciona una buena justificación del método de la máxima verosimilitud. La estimación ML $\hat{\theta}$ de un parámetro θ proporciona, en cierto modo, la densidad $p(x, \hat{\theta})$ que más se acerca a la verdadera $p(x,\theta_0)$. Conviene tener presente que K-L nos dice algo más. Si la verdadera densidad es $q(x)$, pero no pertenece al modelo estadístico, es decir, $q(x) \neq p(x,\theta)$, para todo θ, entonces el significado de $\hat{\theta}$ es como sigue. La

estimación ML de θ bajo $p(x,\theta)$, proporciona una estimación $\hat{\theta}$ de θ_0 tal que $p(x,\theta_0)$ es la densidad K-L más próxima a $q(x)$. Además $\hat{\theta}$ es estimador consistente y asintóticamente normal de θ_0,

$$\sqrt{n}\ (\hat{\theta} - \theta_0) \longrightarrow N(0,\ H^{-1}(\theta_0)\ J(\theta_0)\ H^{-1}(\theta_0)),$$

donde H y J son matrices apropiadas (ver Cuadras y Fortiana, 1993).

1.5. Como medida de dependencia multivariante

Joe (1987) introduce una medida de dependencia entre los vectores aleatorios $X_1,...,X_m$ como la divergencia K-L entre la densidad $p_{x_1 \ldots x_m}$ y la densidad $p_{x_1} \ldots p_{x_m}$ correspondiente a independencia estocástica. Dados tres vectores aleatorios X_1,X_2,X_3, Joe (1989) estudia la medida de dependencia condicional entre X_1 y X_2 dado X_3 como la divergencia K-L entre $p_{x_1x_2x_3}$ y la dependencia condicional entre X_1 y X_2 dado X_3. Tales divergencias, previamente normalizadas por una expresión dependiente de la entropía de Shannon, proporcionan medidas relativas de entropía que conducen a medidas de dependencia multivariante, absoluta y condicional, con propiedades bastante satisfactorias.

La divergencia K-L se utiliza también regularmente como medida de discrepancia entre una densidad p y una estimación no paramétrica de p (Abrahamowicz y Ciampi, 1991).

Un estudio reciente sobre la robustez de la estimación de la divergencia K-L, así como las distancias de Mahalanobis, Hellinger y Bhattacharyya, mediante simulación partiendo de normalidad multivariante, se debe a Rossa (1993).

1.5. Otras entropías

Sobre el conjunto de los $p = (p_1,...,p_k)$, partiendo únicamente de los axiomas

$$H(p) = 0 \qquad \text{si } p_i = 1, \text{ los demás valores } 0,$$

$$H(\lambda p + (1-\lambda)q) \geq \lambda H(p) + (1-\lambda)\ H(q), \qquad 0 \leq \lambda \leq 1,$$

se pueden construir otras entropías. Por ejemplo, la entropía de Gini-Simpson (G-S)

$$GS(p) = 1 - \sum_{j=1}^{k} p_j^2$$

que está directamente relacionada con la varianza

$$s_p^2 = \frac{1}{k} \sum_{j=1}^{k} p_j^2 - (1/k^2).$$

El uso de $H_\phi(p) = \Sigma p_i \phi(p_i)$ da lugar también a otras entropías. Pero en general, como medidas de incertidumbre, todas vienen a indicar lo mismo (véase la tabla 1, en la que se aprecia una relación monótona entre las diferentes medidas).

Tabla 1
Proporciones de grupos sanguíneos para diferentes razas humanas (datos de Mardia *et al.*, 1979) y algunas medidas de diversidad

	A	AB	B	O	G-S	H	S_p
Esquimal	0.2914	0.0000	0.0316	0.6770	0.4558	1.0569	0.2712
Bantú	0.1034	0.0866	0.1200	0.6900	0.4913	1.3806	0.2543
Inglés	0.2090	0.0696	0.0612	0.6602	0.5119	1.3817	0.2440
Coreano	0.2208	0.0000	0.2069	0.5723	0.5809	1.4122	0.2056

La entropía no tiene un significado estadístico para medir la dispersión muy diferente de la desviación típica s_p. Pero al no depender de los valores que toman las variables, es una medida de heterogeneidad especialmente apropiada para datos multinomiales sobre clases no ordenadas (lingüística, genética, sociología, concentración industrial). Del mismo modo que en los modelos líneales la dispersión se estudia mediante ANOVA, el uso de la entropía es de utilidad en el estudio de los modelos log-líneales (Haberman, 1982).

En la misma línea de ideas, cuando el estadístico debe trabajar con varias poblaciones, las medidas de entropía citadas no permiten estudiar la diversidad dentro y entre las poblaciones, como se hace con los modelos líneales multivariantes a través de MANOVA. Rao (1982, 1984) propone la llamada entropía cuadrática

$$H_\Delta(p) = p\Delta p'$$

donde $\Delta = (d_{ij})$ es una matriz tal que $d_{ii} = 0$, d_{ij} es el aumento de disimilaridad al pasar de i a j. Entonces, como en MANOVA, se puede definir un cociente G=B/(W+B), donde W y B representan diversidades dentro y entre poblaciones. G refleja un índice de diversidad relativo entre poblaciones.

1.6. Las entropías generalizadas

El hecho de que existan muchas medidas de entropía justifica la introducción de las (h,ϕ)-entropías, funcionales $H^h_\phi(p)$ de la densidad p, que contienen las ϕ-entropías, Havrda-Charvat, Rényi y otras (Shannon, Taneja, etc.). Esta es la motivación de Salicrú, Menéndez, Morales y Pardo (1993). Entonces, la declaración ética que menciona Pardo (evitar definir nuevas medidas de información sin

un fundamento aplicado), se cumpliría, si convenimos que sólo existe una, la medida $H_{\phi}^{h}(p)$, mientras que las demás serían casos particulares. Pero también se puede interpretar que, cambiando las funciones ϕ y h, podemos construir tantas medidas de información como queramos.

1.7. La medida de la diversidad y sus aplicaciones

Las aplicaciones que poseen las medidas de diversidad en una población están bien comentadas por Pardo. Me limitaré únicamente a poner énfasis en la importancia en Biología de la medida de la diversidad de las especies en una comunidad. La entropía de Shannon y de Gini-Simpson son las más utilizadas por los ecólogos. Cuando se mide por el número medio de bits (es decir, se basa en Shannon), es una cantidad indicativa de la comunidad. Raramente supera los 5 bits, valores bajos indican comunidades transitorias, aumenta en el océano con respecto a las costas (plancton) y de los polos al ecuador, existe una fuerte correlación negativa entre dominancia de algunas especies y diversidad, y su medida puede tener aplicaciones prácticas, como indicio de presencia de petróleo, grado de explotación piscícola, etc. Muchas de estas ideas fueron avanzadas por R. Margalef en una memoria publicada en 1951 (véase Margalef, 1974). Para una definición general, construcción e interpretación de la diversidad, véase Patil y Taillie (1982).

2. SOBRE LAS DIVERGENCIAS

2.1. Divergencias generalizadas

Las medidas de divergencia entre dos variables o vectores aleatorios X, Y, o entre sus funciones de densidad, es un tema de gran interés en Estadística. Las divergencias se pueden generar siguiendo tres caminos: D(X,Y) como generalización de la divergencia de Rényi, R(X,Y) como una diferencia de Jensen sobre una medida de entropía (Burbea y Rao) y a partir de las φ-divergencias de Csiszar-Alí-Silvey. Pardo expone cómo se obtienen muchas y bien conocidas medidas de divergencia (incluyendo la célebre χ^2 de K. Pearson), mediante elecciones adecuadas de las funciones. La principal contribución de Pardo y colaboradores reside en aprovechar las buenas propiedades de las R y las φ-divergencias para construir las (h,ϕ) y R_{ϕ}^{h}-divergencias, que engloban 17 divergencias conocidas.

Como potencial usuario de tales divergencias, se me presenta la razonable duda de no saber cual elegir. Cuando se trabaja con datos y se asume un determinado modelo estadístico, como el normal, existen procedimientos para

aceptar o rechazar el modelo, con un cierto grado de incertidumbre. Pero la construcción de una distancia o una divergencia se puede realizar de muchas maneras, y no parece posible decidir, mediante un test estadístico, si la elegida es correcta. Sólo podemos recurrir a criterios lógicos o simplemente pragmáticos. Como decía R.R. Sokal en DISTANCIA'92, Rennes, 1992: «Hay tantas distancias genéticas que utilizo la que me resulta más sencilla, la distancia de A. Prevosti».

Siguiendo un criterio puramente lógico, cuando se define una distancia entre los parámetros de un modelo estadístico (en vez de entre densidades), cumpliendo ciertas condiciones razonables, aparece de forma natural la entropía de Shannon y la distancia geodésica de Rao, ligada al tensor métrico que define la matriz de información de Fisher. Las métricas diferenciales construidas sobre una amplia gama de divergencias coinciden, salvo constantes, con la distancia de Rao (Cuadras *et al,* 1985, Salicrú, 1987, Cuadras, 1988). En este sentido, la generalización de Pardo y colaboradores (Morales, Pardo, Salicrú y Menéndez, 1993), consistente en perturbar el parámetro en dos direcciones, según un cálculo análogo a una velocidad, permite definir matrices informativas representando una interesante generalización de la clásica de Fisher.

2.2. Distribución asintótica y contraste de hipótesis

Una parte importante de la exposición de Pardo está dedicado al tema de la distribución asintótica de las medidas de divergencia bajo estimación. Mediante las (h,ϕ) y R_ϕ^h-divergencias se resuelven con éxito algunos problemas que no tenían solución con planteamientos clásicos. Las distribuciones asintóticas se reducen a las siguientes: $N(0, \Sigma(\theta))$, χ^2_m y una combinación líneal de χ^2_1 independientes.

Se llega a los mismos resultados asintóticos por otras vías: en estimación (Rao, 1973), distribución asintótica de la razón de verosimilitud, caso regular, y en la también distribución de $-2\log(\lambda)$ cuando la verdadera densidad no pertenece al modelo estadístico, que es una combinación líneal de χ^2_1 independientes, con los coeficientes valores propias de una cierta matriz (Kent, 1982) (por cierto, $-2\log(\lambda)$ se puede interpretar como una distancia de Mahalanobis; véase Cuadras y Fortiana, 1993). Surge entonces la pregunta de si tales contrastes ofrecen ventajas sustanciales con respecto a los clásicos, aunque en algunas situaciones (como en muestreo estratificado) el uso de las divergencias generalizadas aporta novedades.

Finalmente, mi felicitación al Profesor Leandro Pardo por su artículo tan completo, que nos proporciona una amplia panorámica sobre este importante tema.

REFERENCIAS

ABRAHAMOWICZ, M., CIAMPI, A. (1991), «Information theoretic criteria in non-parametric density estimation», *Comp. Stat. & Data Anal.*, **12,** 239-247.

CRESSIE, N. (1991), «Statistics for Spatial Data». Wiley, N. York.

CUADRAS, C. M., OLLER, J. M., ARCAS, A., RÍOS, M. (1985), «Métodos geométricos de la Estadística», *Qüestió,* **9(4),** 219-250.

CUADRAS, C. M. (1988). «Distancias estadísticas», *Estadística Española,* **30,** 295-378.

CUADRAS, C. M., FORTIANA, J. (1993), «Aplicación de las distancias en Estadística», *Qüestió,* **17,** 39-74.

HABERMAN, S. J. (1982), «Analysis of dispersion of multinomial responses», *JASA,* **77,** 568-580.

JOE, H. (1987), «Majorization, randomness and dependence for multivariate distributions», *The Ann. of Prob.,* **15,** 1217-1225.

JOE, H. (1989), «Relative entropy measures of multivariate dependence», *JASA,* **84,** 157-164.

KENT, J. T. (1982), «Robust properties of likelihood ratio tests», *Biometrika,* **69,** 19-27.

MARDIA, K. V., KENT, J. T., BIBBY, J. M. (1979), «Multivariate Analysis», Academic Press, London.

MARGALEF, R. (1974), «Ecología», Omega, Barcelona.

MORALES, D., PARDO, L., SALICRÚ, M., MENÉNDEZ, M. L. (1993), «New parametric measures of information based on generalized R-divergences». En: *Multivariate Analysis: Future Directions* 2 (C. M. Cuadras, C. R. Rao, eds.). Elsevier-North-Holland, pp. 473-488.

PATIL, G. P., TAILLIE, C. (1982). «Diversity as a concept and its measurement», *JASA,* **77,** 548-567.

RAO, C. R. (1973), «Linear Statistical Inference and its applications», Wiley, N. York.

RAO, C. R. (1982), «Diversity and dissimilarity coefficients: a unified approach», *Theoret. Popul. Biol.,* **21,** 24-43.

RAO, C. R. (1984), «Use of diversity and distance measures in the analysis of qualitative data». En: *Multivariate Statistical Methods in Physical Antropology* (G. N. Van Vark, W. W. Howells, eds.), D. Reidel Publishging Company.

ROSSA, A. (1993), «Distance measures between statistical populations-a Monte Carlo study of their robustness», *Biometrical Letters*, **30**, 39-52.

SALICRÚ, M. (1987), «Medidas de divergencia en análisis de datos», Tesis Doctoral. Fac. de Matem., Univ. de Barcelona.

SALICRÚ, M., MENÉNDEZ, M. L., MORALES, D., PARDO, L. (1993), «Asymptotic distribution of (h,ϕ)-entropies». *Commun. Sta.-Theory.*, **22**, 2015-2031.

¿Por qué no una teoría no estadística de la información?

PEDRO GIL ALVAREZ

Universidad de Oviedo

Me satisface por muchas razones participar en esta «mesa redonda» en torno al trabajo del Prof. Pardo; sin duda la que considero más entrañable es la de haber sido el encargado de enseñarle el abecé de la disciplina hace algunos años, pudiendo comprobar ahora el nivel que ha alcanzado con sus estudios e investigaciones, superando con creces a quien esto escribe.

La descripción del gran abanico de medidas de entropía y divergencia existente parece demostrar que el tema sigue siendo objeto de interés, al menos en lo que a producción de «papers» y monografías se refiere. Echo de menos alguna referencia a las medidas ponderadas (weighted entropies) que permitiría complementar el estudio de los índices de diversidad con el correspondiente a índices de desigualdad (distribuciones de renta, medidas de pobreza, etc., tan de actualidad en esta época que vivimos).

Sin duda es el apartado 7 el que aporta mayores novedades desde el punto de vista científico. El tratamiento de los distintos problemas de contraste desde una perspectiva unificada supone un gran avance en el desarrollo de modelos eficaces, aunque los resultados sean de carácter asintótico. En mi opinión, es quizá más importante el tratamiento unificado que el empleo de las medidas unificadas, sin quitar a éstas el mérito de haber condensado (y, en muchos casos, mejorado) la gran disparidad de medidas existentes.

En cualquier caso no creo ser el más indicado para dar el visto bueno o malo a los resultados obtenidos: doctores tiene la comunidad científica internacional que ya han juzgado los mismos, considerando sus merecimientos para ser publicados en revistas de gran prestigio.

Quisiera, por el contrario, extenderme (aunque en el trabajo se hable de ello al comienzo) un poco en el tema del código ético de la Teoría de la Información. Son demasiadas las medidas que han sido creadas de forma absolutamente artificial, sin mejorar para nada las propiedades que presentaban las más clásicas ni conseguir aplicaciones en campos distintos de los ya trillados. Posiblemente, salvo la entropía de Shannon y la entropía cuadrática, todas las demás sean producto de ese afán de generalización absurdo que sólo conduce a resultados cada vez más generales en los que la intuición ha desaparecido totalmente. En cuanto a las divergencias, ¿se ha ganado mucho con respecto a los modelos de Kullback?

Y sigo considerando la intuición como una base muy importante sobre la que sustentar los desarrollos de modelos de la realidad. Las «florituras» matemáticas son, evidentemente, valiosas como aportación al conocimiento general, pero poco rentables para el conocimiento del mundo que nos rodea.

Creo que estamos perdiendo las ideas esenciales de nuestra investigación, que llevaban a construir un modelo, desarrollarlo matemáticamente y devolverlo a la realidad: nos quedamos, cada vez más, encerrados en la segunda fase del proceso, sin intenciones de volver al mundo real, generalizando cada vez más resultados cuya potencia era más que suficiente para atacar los objetivos propuestos.

Y puestos a generalizar, ¿por qué no hacerlo aún más? Aunque, de acuerdo con el título, *Teoría de la Información estadística,* el Prof. Pardo no hace ninguna mención a otras «teorías de la información», conoce perfectamente las relaciones entre los componentes del binomio incertidumbre-información y sabe de situaciones en las que la incertidumbre no se plantea ante una distribución de probabilidad sino ante otro tipo de modelos (ese «¿qué habrá querido decir?» en ciertos equívocos semánticos, o la imprecisión que caracteriza nuestro lenguaje o nuestro propio conocimiento modelada con más o menos éxito por los Conjuntos Borrosos, por ejemplo).

La *Teoría Axiomática de la Información* se permite el lujo de prescindir de una distribución probabilística (aunque no la desprecie cuando está presente) para objetivar la medición de la información. La consideración del concepto de información como algo mucho más primario que el concepto de probabilidad, subyacente en una gama de situaciones mucho más amplia que la considerada por la teoría estadística, es la base de esta teoría que para muchos ha pasado y sigue pasando desapercibida.

Sobre la base de un espacio medible (Ω, S) basta definir una clase κ de pares de subconjuntos algebraicamente independientes, eventualmente vacía, para conseguir el denominado *espacio de información medible* (Ω, S, κ).

Se denomina *medida de información* a una aplicación J definida sobre un espacio de información medible que a cada resultado observable le asigna un número real positivo ($J : S \to \mathbb{R}^+$) que verifica los siguientes axiomas:

4.4.1. $J(\Omega) = 0$

4.4.2. $J(\varnothing) = +\infty$.

4.4.3. $A, B \in S, B \supset A \to J(A) \geq J(B)$

4.4.4. $(A, B) \in \kappa, A \in A, B \in B \to J(A \cap B) = J(A) + J(B)$

Con este concepto todas las medidas «clásicas» (Shannon, Rényi, Havrda-Charvat, etc.) pueden construirse mediante reglas muy simples cuyo estudio puede hacerse de forma general. Y los mismos resultados pueden aplicarse a la búsqueda de medidas adecuadas a situaciones en que no se halle presente la probabilidad.

¿Por qué no seguir este camino aunque solo fuera para poder caracterizar *todas* las medidas de incertidumbre de distribuciones probabilísticas?

Nada más. Felicitar al autor por la recopilación efectuada y, muy especialmente, por sus contribuciones al desarrollo de esta Teoría.

REFERENCIAS

BERTOLUZZA, C. (1968), «Sull'informazione condizionale», *Statistica,* XXVIII, 242-245.

BREZMES, T. & GIL, P. (1985), «Incertidumbre e información condicionada», *Trab. Est. I.O.,* **36 (2),** 39-55.

KAMPÉ DE FÉRIET, J. & FORTE, B. (1967), «Information et probabilité», *C.R.A.S.* París, **265** 110-114, 142-146, 350-353.

RAMON GUTIERREZ JAIMEZ

Universidad de Granada

1. INTRODUCCION

Quiero comenzar agradeciendo a «Estadística Española» la oportunidad de poder participar en la discusión de este excelente trabajo de L. Pardo, cuya principal cualidad para mí, aparte del completísimo panorama que presenta sobre la Información Estadística, ha sido su capacidad sugeridora de problemas abier-

tos en la línea de los ya abordados con éxito por el autor y otros relacionados con él y publicados en los últimos años.

El trabajo está además hábilmente estructurado y secuenciado para ir incluyendo en su justo lugar aquellas aportaciones originales del autor a diversos problemas estádistico-matemáticos (en especial sobre Contrastes de Hipótesis), arrancando de la consideración sistemática de las familias de (h, ϕ)-divergencias en cuya explotación posterior, en efecto, el autor ha participado en el logro de resultados técnicos muy interesantes en el contexto general de la Información en Estadística Matemática.

Así pues sobre el trabajo en sí, autolimitado por razones obvias de espacio a la Información Estadística y a algunos de sus problemas teóricos más importantes hasta el momento (Contrastes, Comparación de Experimentos, Principio de optimización en Entropía y Divergencia, etc.) poco tengo que añadir. La gran cantidad de información bibliográfica recogida es prácticamente exhaustiva incluyéndose referencias a importantes aplicaciones actuales de los temas tocados sobre Información, en la Estadística aplicada a diversos campos científicos.

Por tanto, en lo que sigue me voy a limitar a exponer algunos comentarios que ciertos aspectos del trabajo me han sugerido, más reflexiones propias a la luz del texto del trabajo, que comentarios técnicos sobre sus contenidos.

2. SOBRE LAS MEDIDAS DE ENTROPIA Y DE DIVERGENCIA

Las situaciones proliferativas, arborescentes-fractales que diríamos hoy, se han dado muchas veces en todos los campos estadísticos-probabilísticos. Y siempre se han recorrido las etapas de Globalización por un Modelo; caracterización de los casos en el modelo Global; y extensión de lo conocido en casos particulares globalizados, al Modelo síntesis global.

El panorama de las medidas de distancias y proximidades en Clasificación Automática, por ejemplo, llega a ser estremecedor. Distancias para Tablas Lógicas, de rangos, desdobladas, etc., y las distancias posteriores entre clases en los algoritmos de clasificación ascendente jerárquica, conforman, en efecto, un panorama abrumador que provoca la idea sistematizadora de la posible obtención de una fórmula de recurrencia en términos además de los criterios distintos de agregación que interesan. Y en efecto, Cormack (1971), seguido de Bock (1973), introdujo y estudió una fórmula general de recurrencia que contiene a la mayor parte de las fórmulas recurrentes que previamente habían sido introducidas.

En el abrumador panorama de las medidas de divergencia y de información se ha intentado poner orden siguiendo varias ideas confluyentes. Las (h, ϕ)-di-

vergencias son un claro exponente de ello, y L. Pardo nos ofrece una síntesis muy completa de este método globalizador. Vías distintas, que hacen más hincapié en aspectos «generativos» basados en caracterizaciones estructurales o en hipótesis como la recursividad de la Entropía, por ejemplo, son también recogidas en este trabajo de manera muy completa. La impresión que se adquiere, es que, ayudados desde luego por el «principio ético» al que L. Pardo alude, la situación está ya en una etapa madura, en la que la simple y despreocupada proposición de nuevas medidas han dado paso a una etapa de análisis sistemático de familias de medidas que juegan un relevante papel en la Estadística Matemática o en campos tan importantes como el Reconocimiento de Patrones o Clasificación general estadística.

3. SOBRE LA UTILIZACION DE LA INFORMACION EN LA GEOMETRIA DIFERENCIAL DE LA INFERENCIA ESTADISTICA; EN LOS PRINCIPIOS DE ENTROPIA Y DIVERGENCIA Y EN ANALISIS ESTADISTICO DE DATOS

En la panorámica tan completa que L. Pardo nos ofrece sobre la Información Estadística, tienen cabida múltiples temas, incluso algunos que después de cuatro décadas, en mi opinión, se han mostrado poco útiles para aportar sustanciales progresos en la Estadística en general. Por ejemplo, el autor recoge algunas consideraciones sobre la Información, la Estadística Matemática y la Geometría Diferencial sobre variedades. Fiel al enfoque técnico del autor al redactar este trabajo, prescindiendo en general de consideraciones críticas sobre el material expuesto (lo que sin duda, al menos a mí me dibuja, en principio, un panorama idílico del tema, que evidentemente no corresponde con la realidad científica del papel de la Información en la Estadística Matemática), parece deducirse, asépticamente, que el utilizar la Geometría Diferencial asociada a los modelos Estadísticos ha aportado ideas importantes e innovadoras para la Estadística. Yo más bien creo que no ha sido así. El patético prólogo del trabajo, prototipo en este campo, de Amari (1987), es bastante revelador al respecto. Curiosidades tales como relacionar las α-divergencias con las α-proyecciones en el contexto de las «α-Πat» variedades, es sin duda interesante técnicamente, pero quizás no conduzca más que a un ejercicio de virtuosismo estadístico-matemático. La idea de Rao (1945) era sin duda interesante y atrayente: Introducir una métrica Riemanniana sobre una familia paramétrica de distribuciones de probabilidad proponiendo una distancia geodésica inducida por la métrica citada y basada en la matriz de Información de Fisher, para medir la disimilaridad entre distribuciones, de tal manera que dicha métrica es invariante por transformaciones de variables y parámetros, era, en efecto, prometedor. Pero más de 45 años después cabe preguntarse si el voluminoso desarrollo asociado a la explotación

estadística, a nivel de Inferencia Asintótica por ejemplo, de las propiedades locales de la Geometría Diferencial de los Modelos Estadísticos, ha servido para aportar aspectos innovadores en la Estadística Aplicada o incluso Teórica.

El papel de los Principios de Entropía Máxima y de Divergencia Mínima en la Estadística actual, son también muy acertadamente incluidos en este trabajo de L. Pardo.

Al centrarse este trabajo en Información en Estadística (con Modelo Probabilístico subyacente), obviamente quedan al margen muchas cuestiones teóricas y prácticas relativas al papel de la moderna Teoría de la Información en la Estadística sin modelo. Particularmente interesante es su papel actual en el análisis de datos, sobre todo en las técnicas aglomerativas del Análisis Cluster jerárquico en donde se han propuesto métodos basados en la medida de Información tipo Shannon para el caso de variables dicotómicas que posteriormente se han extendido a variables categóricas y continuas (Williams y Lance, 1971). También se han utilizado otras medidas de Información, por ejemplo la de Brillouin que como es sabido es monótona respecto de la de Shannon. También Buser-Baroni-Urbani (1982) han propuesto otra entropía distinta para datos binarios que ha sido utilizada ampliamente en las mencionadas técnicas jerárquicas.

Me atrevo a sugerir que un tema pendiente en el papel de la información en Estadística, es el construir una visión integradora entre lo que es Información en Estadística Matemática y la utilización de la Información en el Análisis de datos o Estadística sin Modelo probabilístico.

4. INFORMACION Y DIVERGENCIA EN EL ANALISIS MULTIVARIANTE

Una vía poco estudiada aún en el amplísimo mundo de la Información y de la Información en Estadística es, en mi opinión, la de obtener resultados generales válidos para «clases o familias de distribuciones». Dejando a un lado el caso de la Familia Exponencial, cuyo papel en la Inferencia y también en la Información Estadística tiene su lugar propio, desde luego más bien de carácter teórico, no se ha abordado en la Bibliografía hasta el momento presente, por ejemplo, la obtención de resultados para familias de distribuciones continuas o discretas univariantes (Pearson; Ord; Kaptein, etc.). Aún más interesante, teórica y prácticamente hablando, sería abordar sistemáticamente la cuestión sobre clases de distribuciones multivariantes continuas o discretas, para las que los métodos de generación son hoy relativamente bien conocidos, tanto a nivel de los métodos clásicos (extensiones multivariantes de los sistemas continuos de Pearson, vía ecuaciones diferenciales de las clases) como vía Polinomios Zonales aplicables en el caso discreto (ver por ejemplo Gutiérrez-Hermoso, 1989).

A este respecto es de resaltar la reciente consideración de funciones de información en las clases Elíptica y Esférica de distribuciones multivariantes continuas, que, como es sabido, desempeñan un papel básico en todos los esfuerzos puestos en marcha para la extensión del Análisis Multivariante Normal a otras distribuciones. En este aspecto Quan-Fang-Teng (1990) consideran dos poblaciones elípticas π_1 y π_2 con densidades multivariantes f y g y la «función de información»

$$I(f, g) = \int f(x) \log \frac{f(x)}{g(x)} dx$$

Con esta divergencia y bajo determinadas condiciones, la consideración del vector aleatorio $z \subset ECD_p(\mu, \Sigma; f)$, permite establecer interesantes consecuencias entre la información citada y la distancia de Mahalanobis. Técnicamente el problema lo resuelven por una reducción al caso esférico, acotando la información, obteniendo cotas del tipo $I(f, g) < \varepsilon$ y resolviendo además, en dicha clase, el problema general de Discriminación.

Creo que podríamos sugerir algunos «problemas abiertos» de gran interés en mi opinión, en este campo, en los cuales podrían desempeñar un interesante papel, a su vez, las aportaciones originales de L. Pardo y su grupo de trabajo. Esencialmente dos tipos de problemas:

(i) La caracterización de resultados ya obtenidos para la clase elíptica, como los antes mencionados, en términos del Parámetro de Curtosis de dicha clase (parámetro básico en el enfoque tipo Muirhead del Análisis Multivariante sobre esta clase); así como conocer «distancias» en información entre distribuciones concretas de la clase en términos de dicho parámetro. Todo ello si fuera posible dentro de los planteamientos de las (h, ϕ)-divergencias.

(ii) La consideración de otras divergencias de entre las que han mostrado utilidad, manejadas a través de las (h, ϕ)-divergencias que nos expone L. Pardo para el caso univariante, en situación multivariante y en concreto para la clase Elíptica. De manera ideal cabría plantearse hasta dónde es posible extender lo realizado para Constrastes de Hipótesis Multivariantes vía (h, ϕ)-divergencias, a la clase elíptica Multivariante, lo que en mi modesta opinión constituiría un avance notable en este campo.

5. INFORMACION Y DIVERGENCIA EN PROCESOS ESTOCASTICOS

En mi opinión el culmen del papel de la Información en Estadística es su utilización en la Inferencia Estadística sobre procesos Estocásticos, así como en la propia fundamentación y aplicación de los mismos.

Los conceptos básicos de Información y Divergencia desempeñan un importante papel en la teoría de Procesos Estacásticos; y ello a dos niveles: A nivel teórico y a nivel de la inferencia estadística sobre Procesos. Aunque respecto de los límites del trabajo de L. Pardo, sólo el segundo nivel está directamente relacionado, es muy ilustrativo el considerar, siquiera sea brevemente, algunos aspectos conceptuales de la Información en lo que a la fundamentación teórica de Procesos se refiere, sobre todo porque, al igual que en Inferencia Estadística en procesos sucede, la extensión de la Información sobre variables aleatorias a procesos no es ni mucho menos trivial.

Dado un proceso estacionario continuo o discreto, en tiempo real $\{\xi(t): t \ll T\}$ la información desempeña en esta clase un importante papel a la hora de estudiar su regularidad. De forma más precisa se denomina «regular en información» si el llamado «coeficiente de regularidad en información», que está basado en la divergencia de Kullback, tiende a cero.

En otro tipo de Procesos, como los Gaussianos de fundamental importancia cuando los Procesos se utilizan en la modelización de Sistemas de Comunicación, la Información en general ha de aplicarse con adaptaciones de los conceptos generales «ad hoc» para el tipo considerado. Es interesante pues observar como en los distintos tipos de procesos se precisa ciertas partes concretas de la Información y de la Información Estadística en general. En los mencionados Gaussianos juega un papel fundamental la extensión a Procesos de la ε-entropía, por ejemplo, que en los de tipo markoviano no tiene papel relevante. La ε–entropía de un proceso, en general, $\{\xi(t): t \in T\}$, se define como: $H_\varepsilon(\xi(t)) = \inf I(\xi; \eta)$; ($I(\xi, \eta)$ información mutua) en donde el inferior es tomado para todos los pares de procesos: $(\xi; \eta)$; $(\xi(t); \eta(t); 0 \leq t \leq T)$ que satisfacen:

$$\int_0^T E(\xi(t) - \eta(t))^2 \, dt \leq \varepsilon^2$$

Esta definición adaptada a Procesos fue tomada por Pinsker (1964), sobre la base del concepto de ε-entropía de Shannon (1959) para variables aleatorias reales, $H_\varepsilon(\xi)$, respecto del criterio del error cuadrático medio. Ya Shannon probó que si la variable aleatoria tiene densidad $p_\xi(x)$, entonces:

$$H(\xi) + \frac{1}{2} \ln \frac{1}{2\pi e \xi^2} \leq H_\varepsilon(\xi) \leq \frac{1}{2} \ln \frac{\sigma^2}{\xi^2}$$

en donde σ^2 es la varianza de ξ y $H(\xi) = -\int P_\xi(x) \ln p_\xi(x) \, dx$. Cuando $\varepsilon \to 0$ se sabe que

$$H_\varepsilon(\xi) = H(\xi) + \frac{1}{2} \ln \frac{1}{2\pi e \xi^2} + o(1)$$

Estos resultados se extendieron a variables aleatorias multidimensionales con densidad (Gerrish y Schultheis, 1964; Linkov, 1965 y otros) y más tarde a procesos estocásticos gaussianos y no gaussianos con enorme dificultad técnica (Binia, 1974, Binia-Zakai-Ziv, 1974) constituyendo un clásico ejemplo de la dificultad inherente a la extensión de la Información e Información Estadística a determinados tipos de Procesos Estocásticos.

No conocemos que se haya utilizado en relación con la ε-entropía sobre Procesos Estocásticos, en particular gaussianos, otras entropías distintas a la inicialmente propuesta de Shannon; ni tampoco conocemos investigaciones sobre la utilización de medidas de divergencia en el contexto de estos procesos, distintas a la clásica de Kullback, estando por verse la interpretación y aplicación en Teoría de la Comunicación de divergencias generalizadas tipo (h, ϕ).

Finalmente en lo que al papel de la Información en Procesos se refiere y en concreto sobre Información Estadística en Procesos, creo que es hoy imprescindible hacer mención del papel de la Información Fisher en la inferencia asintótica sobre Procesos de Difusión, tanto a nivel de muestreo discreto como continuo por trayectorias muestrales. Es bien sabido que ambos tipos de esquemas de muestreo sobre las Difusiones, tienen que sèr tratados técnicamente de manera muy diferente, a saber: A través de Inferencia sobre verosimilitudes asociadas a las observaciones de las densidades de transición, en el caso discreto; y mediante recursos de ecuaciones estocásticas diferenciales tipo Ito, en el continuo. En el caso de Difusiones Multidimensionales (de Orstein-Uhlenbeck; Lognormales, etc.) y por muestreo discreto, es bien conocido el comportamiento asintótico en la Inferencia correspondiente (ver por ejemplo Gutiérrez y otros, 1991). Y en el caso continuo, la formulación de la estimación y contrastes de hipótesis también es conocida en términos de la información de Fisher apropiada (ver el texto básico de Basawa y Rao 1980, Cap. 9), para el caso del parámetro Drift. ¿Sería posible introducir otro tipo de información para estudiar contrastes de hipótesis sobre los parámetros tendencia de Difusiones, al menos, relacionadas con el Wiener, es decir en definitiva sobre procesos de variables normales multivariantes o transformadas? Es otro posible problema a abordar en el futuro, a través de resultados como los que L. Pardo ha obtenido, junto con otros autores, para contrastes en situación normal multivariante y con divergencias generalizadas.

Finalmente quiero terminar agradeciendo a L. Pardo su trabajo, que desde ahora, junto a importantes trabajos publicados por el autor, creo que debe constituir una referencia obligada para conocer una panorámica muy completa y sugerente de lo hecho y por hacer en muchos campos estadísticos y probabilísticos en los cuales la información desempeña un importante papel. Y animarle a ampliarlo tanto desde el punto de vista de su exposición sistemática como desde el de continuar investigando en la línea que se ha trazado en los últimos años y en la que tan interesantes resultados ha obtenido.

BIBLIOGRAFIA

AMARI, S. I. (1987). «Differential Geometrical Theory of Statistics». Cap. 2 del libro *Differential Geometry in Statistical Inference,* Barnddorff-Nielsen.

BASAWA, I y RAO, L. S. P. (1980). *Statistical Inference for Stochastic Processes,* Academic Press.

BINIA, J. (1974). «On the ε-entropy of certain gaussian Processes», *IEEE Trans. Inform. Theory,* vol. IT-20, 2, 190-196.

BINIA, J., ZAKAI, M. y ZIV, J. (1974). «On the ε-entropy and the rate distortion Function of certain Non-gaussian Processes». *IEEE Trans. Inform. Theory,* vol. IT-20, **4,** 517-524.

CORMACK, R. M. (1971). «A review of Classification». *J. Roy. Statist. Soc.,* Serie A, vol. 134, Parte 2.

GUTIÉRREZ, R.; ANGULO, J. M.; PÉREZ, R.; GONZÁLEZ, A. (1991). «Inference in Log-normal Difussion Processes with exogenous factors. Application to Modelling in Economics». *Applied Stoch. Models and Data Analysis,* vol. 7, 4, 295-316.

PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes,* Holden-Day.

QUANG, H.; FANG, K. y TENG, Ch. (1990). «The Application of Information Function for Spherical Distributions», *Stastistical Inference in Elliptically contored and Related Distribucion.* Editado por K. Fang y T. W. Anderson. Allerton Press, New York.

GUTIERREZ, R. y HERMOSO, J. (1989): «An Application of Zonal Polinomial to generation of Probability Distributions» Linear Alg. and Applic. Vol. 119, 74-80.

Estimación por máxima entropía

HENRYK GZYL

Universidad Central de Venezuela

DANIEL PEÑA

Universidad Carlos III de Madrid

En primer lugar, queremos felicitar a Leandro Pardo por su excelente trabajo de revisión de un campo de gran importancia en Estadística y que no recibe habitualmente la atención adecuada. Nuestro comentario va a centrarse en las

aplicaciones estadísticas del principio de maximización de la entropía, y, en concreto, en su utilización para la estimación paramétrica de modelos estadísticos.

Supongamos un modelo $F(x, \theta)$, donde θ es un parámetro (posiblemente vectorial) desconocido, y una muestra aleatoria simple $x_1,...,x_n$ de F. Algunos autores (Kapur y Kesavin (1992), Jiménez y Palacios (1993)) han propuesto aplicar el principio de máxima entropía para estimar θ de la forma siguiente: (1) Sea

$$x_{(1)} \leq x_{(2)} \ldots \leq x_{(n)}$$

la muestra ordenada; (2) definamos

$$P_1(\theta) = F(x_{(1)};\theta)$$

$$P_i(\theta) = F(x_{(i)};0) - F(x_{(i-1)};\theta) \qquad i=2,...,n$$

$$P_{n+1}(\theta) = 1-F(x_{(n)};\theta);$$

(3) construyamos la entropía

$$S(\theta) = -\sum_{i=1}^{n+1} P_i(\theta) \ln P_i(\theta) \tag{1}$$

y maximicemos esta función con la restricción $\sum P_i(\theta)=1$. La solución de este problema proporcionaría un estimador de θ que llamaremos de máxima entropía.

Una interpretación estadística de este procedimiento (que se fundamenta habitualmente por teoría de la información) es la siguiente. Recordemos que los estadísticos ordenados dividen el área bajo la distribución en n + 1 partes de manera que cada parte tiene la misma área esperada. La distribución discreta con máxima entropía es la uniforme. En consecuencia, maximizar (1) implica elegir θ de manera que las variables aleatorias $P_i(\theta)$ estén lo más próximo posible a su valor esperado 1/(1+n). Esta aplicación del método de máxima entropía ofrece una solución razonable del problema que: (1) utiliza toda la información en la muestra, ya que el conjunto de estadísticos ordenados es siempre suficiente; (2) se basa en una propiedad central de la distribución de las observaciones muestrales.

Como ejemplo de este método consideremos que x sigue una distribución uniforme en $(0,\theta)$. Entonces, $F(x;\theta) = x/\theta$, y:

$$S(\theta) = -\sum_{i=1}^{n} (x_{(i)} - x_{(i-1)})\, \theta^{-1} \ln (x_{(i)} - x_{(i-1)})\, \theta^{-1} - \left(1 - \frac{x_{(n)}}{\theta}\right) \ln \left(1 - \frac{x_{(n)}}{\theta}\right) \tag{2}$$

con $x_{(0)} = 0$. Como $\sum P_i(0) = 1$ para cualquier valor de θ. La estimación por máxima entropía se reduce en la maximización sin restricciones de (2). Derivando $S(\theta)$ respecto a θ e igualando a cero y utilizando que $\sum_{i=1}^{n} (x_{(i)} - x_{(i-1)}) = x_{(n)}$, se obtiene:

$$\hat{\theta}_{ME} = x_n (1 + \Delta) \tag{3}$$

donde

$$\Delta = \exp \left\{ \frac{1}{x_{(n)}} \sum_{i=2}^{n} (x_{(i)} - x_{(i-1)}) \ln (x_{(i)} - x_{(i-1)}) \right\}$$

Jiménez y Palacios (1993) han estudiado este estimador comparándolo con el máximo verosímil ($x_{(n)}$) y el obtenido por Cheng y Amin (1983). Los tres son consistentes, pero el estimador (3) parece tener un error cuadrático medio menor.

Para variables absolutamente contínuas el principio de maximización de la entropía podría plantearse como maximizar

$$S(\theta) = -\sum f(x_i; \theta) \ln f(x_i; \theta) \tag{4}$$

que puede escribirse

$$S(\theta) = -\sum \omega_i(x_i; \theta) \ln f(x_i; \theta) \tag{5}$$

donde $\omega_i \geq 0$ y $\sum \omega_i = 1$. Entonces el método de máxima verosimilitud sería un caso particular donde $\omega_i = n^{-1}$, es decir, todos los términos $\ln f(x_i;\theta)$ tendrían el mismo peso, mientras que en el método de máxima entropía las observaciones recibirían un peso proporcional a su probabilidad de aparición. Es interesante resaltar que el método de máxima entropía sería, en consecuencia, más robusto que el método de máxima verosimilitud ya que las observaciones extremas serían descontadas, al entrar con un peso menor que las observaciones centrales. En este sentido recuerda los M estimadores, aunque desde un enfoque diferente.

Por ejemplo, la estimación de μ para variables normales $N(\mu, 1)$, aplicando (5) conduce a la ecuación

$$\sum (x_i - \bar{\mu})\omega_i = \sum (x_i - \bar{\mu})^3 \omega_i/2 \tag{6}$$

donde

$$\omega_i = \exp\left\{ -\frac{1}{2} (x_i - \hat{\mu})^2 \right\} \tag{7}$$

La ecuación (6) muestra que $\hat{\mu}$ es estimado tratando de hacer compatibles las propiedades de centralización de $\bar{x}$ (que supone $\sum (x_i - \bar{x}) = 0$) y simetría en la distribución (que supone $\sum (x_i - \hat{\mu})^3 = 0$), dando además una ponderación a las

observaciones de manera que las más atípicas tengan menor peso, lo que muestra las propiedades de robustez antes indicadas.

En resumen, el método de máxima entropía aparece como una alternativa interesante a explorar en problemas de estimación estadística. Sería interesante investigar las propiedades estadísticas de los estimadores así obtenidos y desarrollar procedimientos computacionales para calcularlos en problemas complejos.

REFERENCIAS

CHEN, R. C. H. y AMIN, N. A. K. (1983), «Estimating parameters in continuous univariate distributions with a shifted origin». *J. R. Estatist. Soc. B*, **45**, 394-403.

JIMÉNEZ, R. y PALACIOS, J. L. (1993), «Shanon's measure of ordered samples: some asymptotic results». Documento de Trabajo. Universidad Simón Bolívar.

KAPUR, J. N. and KESAVAN, H. K. (1992), *Entropy optimization principles with applications*. Academic Press.

JULIAN DE LA HORRA

Universidad Autónoma de Madrid

Es para mí un placer el poder añadir algunos comentarios a este interesantísimo trabajo donde se recogen las principales ideas y aplicaciones de la Teoría de la Información Estadística. Quizás mi mejor contribución pueda ser la de intentar ampliar un poco la parte de la Teoría de la Información que mejor conozco: su aplicación al problema de encontrar distribuciones de referencia (mínimo informativas) en los métodos bayesianos de inferencia. De modo que trataré de resumir a continuación algunas de las principales ideas.

a) Distribuciones de referencia usando la entropía de Shannon

Como el autor indica en la Sección 4, la obtención de distribuciones mínimo informativas es una de las más importantes aplicaciones del principio de maximización de la entropía.

Siguiendo a Bernardo (1979), la distribución a priori de referencia para el parámetro desconocido θ se define como «aquella que maximiza la información asintótica que se espera obtener de la muestra, entendiendo por información (en este caso) la diferencia entre la entropía (de Shannon) de la distribución a priori y la entropía (de Shannon) esperada de la distribución a posteriori».

En el caso de que el espacio paramétrico sea finito, la distribución de referencia es, como indica el autor, la distribución uniforme.

Si el espacio paramétrico es continuo entonces, bajo condiciones de regularidad, la distribución a priori de referencia es la obtenida mediante la conocida regla de Jeffreys. Para la justificación rigurosa de esto último véase Ghosh y Mukerjee (1992) y las referencias allí contenidas.

La situación se complica enormemente cuando existen paramétros perturbadores. Este caso ha sido abordado en el trabajo de Berger y Bernardo (1992).

b) Distribuciones de referencia usando otras medidas de entropía

Aunque la entropía de Shannon es, con diferencia, la más popular, hay otras medidas de entropía que son también perfectamente respetables. Las más importantes son citadas por el autor en la Sección 2. Por tanto, parece muy razonable considerar distribuciones de referencia que estén basadas en una medida de entropía cualquiera, en vez de estar basadas necesariamente en la entropía de Shannon. Esta extensión de la idea de distribución de referencia se encuentra en García-Carrasco (1986).

En el caso de espacio paramétrico finito, es fácil obtener que la distribución a priori de referencia es la distribución uniforme, siempre que el concepto de entropía utilizado cumpla unos requisitos mínimos; concretamente, siempre que sea una función cóncava, no negativa e invariante por permutaciones (cosa que ocurre en la mayor parte de las medidas de entropía razonables, por no decir en todas).

c) Distribuciones de referencia bajo restricciones

Un problema muy natural es el de buscar la distribución a priori de referencia dentro de las que cumplen una serie de condiciones previas. Pero entonces, puede darse el caso de que la distribución de referencia dependa de la medida de entropía que estemos utilizando; esto es lo que ocurre cuando imponemos restricciones sobre los momentos.

El problema de obtener la distribución a priori de referencia sujeta a restricciones sobre los cuantiles ha sido estudiado en el trabajo de García-Carrasco y De la Horra (1988), para espacios paramétricos finitos. Este estudio es interesante, en primer lugar porque las restricciones sobre los cuantiles son, posiblemente, las más naturales y las más sencillas de especificar en casos prácticos, y en segundo lugar porque se probaba en dicho trabajo que la distribución de referencia es independiente de la medida de entropía empleada, siempre que dicha medida cumpla los requisitos mínimos indicados en el apartado anterior (cóncava, no negativa e invariante por permutaciones).

Espero que estos breves comentarios contribuyan a completar la excelente panorámica ofrecida por el Profesor Pardo sobre la Teoría de la Información Estadística.

REFERENCIAS EN LOS COMENTARIOS

BERGER, J. O.; BERNARDO, J. M. (1992), «On the development of reference priors», Bayesian Statistics 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith, eds.). Oxford University Press, 35-60 (with discussion).

BERNARDO, J. M. (1979), «Reference posterior distributions for Bayesian inference», *J. R. Statist. Soc.* **B 41,** 113-147 (with discusion).

GARCÍA CARRASCO, M. P. (1986), «Distribuciones mínimo informativas; caso de espacio paramétrico finito», *Qüestió* **10,** 7-12.

GARCÍA-CARRASCO, M. P.; DE LA HORRA, J. (1988), «Maximizing uncertainty fucntions under constraints on quantiles», *Statistics & Decisions* **6,** 275-282.

GHOSH, J. K.; MUKERJEE, R. (1992), «Non-informative priors», *Bayesian Statistics* 4 (J. M. Bernardo, J. O. Berger, A. P. Dawid y A. F. M. Smith, eds.). Oxford University Press, 195-210 (with discusion).

Contestación

Es mi deseo expresar mi más sincero agradecimiento al Profesor Ricardo Vélez, miembro del Consejo de Redacción de Estadística Española, por su propuesta inicial para que elaborara una perspectiva del estado actual de la Teoría de la Información Estadística. Agradecimiento que hago extensivo a todo el equipo editorial de Estadística Española por la buena acogida de la misma y en especial a su director, el Profesor Daniel Peña, por haber tenido además la amabilidad de brindarme unos interesantes comentarios al trabajo.

Vaya también mi felicitación, no exenta de agradecimiento, a los Profesores Ramón Ardanuy, Carlos Cuadras, Pedro Gil, Ramón Gutiérrez, Henryk Gzyl, Julián de la Horra y al ya citado anteriormente Profesor Daniel Peña ya que sus comentarios han enriquecido y ampliado la perspectiva dada en mi trabajo sobre la Teoría de la Información Estadística. Al centrarse sus comentarios en aspectos diferentes del trabajo contestaré, o mejor dicho, expresaré mi punto de vista por separado a cada uno de ellos. Si bien, en un primer comentario de carácter general, quiero señalar que algunos de los tópicos mencionados por ellos, tuvieron que ser excluidos en aras de no aumentar más un trabajo que ya de por sí ha resultado bastante extenso. Otros temas han constituido una novedad y en ellos incluso se sugieren algunas ideas originales que podrían constituir motivos de interesantes investigaciones.

Me congratula el hecho de que importantes investigadores en Estadística resalten, unánimemente, la importancia e interés de la Teoría de la Información en Estadística e incluso los Profesores Henryk Gzyl y Daniel Peña señalen, a mi modo de ver muy acertadamente, que no recibe la atención adecuada. Este sentir por parte de nosotros, los Estadísticos, hacia la Teoría de la Información se está empezando a extender, no sólo en la comunidad científica estadística española, sino en la internacional. Hoy en día, ya a nadie le parece extraño ver trabajos publicados de Teoría de la Información Estadística en las revistas estadísticas de mayor índice de impacto o incluso de otras áreas. Sin ir más lejos, mientras estoy escribiendo esta contestación ha aparecido una interesante recopilación sobre la aplicación de diversas técnicas de Teoría de la Información en Economía y Econometría (E. Maasoumi, 1993). Esto hace unos años, muy pocos años, era

impensable. ¿Qué ha ocurrido para que se produzca este cambio? Hasta hace muy pocos años la Teoría de la Información era una materia que tenía prácticamente como única aplicación la transmisión de mensajes lo que llevaba consigo que inicialmente casi, únicamente, fuese objeto de estudio por parte de Ingenieros. Posteriormente fue objeto de estudio por parte de Matemáticos que se dedicaron, casi exclusivamente, a dar una gran cantidad de diversas y variadas axiomáticas y a definir nuevas medidas de entropía y divergencia para posteriormente presentar una caracterización de las mismas. Durante todo este desarrollo, no intervinieron de forma activa Estadísticos, por ello la casi total ausencia de aplicaciones estadísticas en la literatura científica de Teoría de la Información durante este período de tiempo. A mi modo de ver el interés por parte de los Estadísticos nace con los trabajos de Kullback, De Groot, Lindley, Rao, etc., en los que se abordaron diversos problemas estadísticos a través de la Teoría de la Información. No obstante, yo estoy firmemente convencido que en los próximos años tendrá un desarrollo espectacular en lo que se refiere a las aportaciones de la misma en Estadística y la Teoría de la Información Estadística tendrá una atención adecuada por parte de los Estadísticos.

El Profesor R. Ardanuy hace un interesante y completo estudio de las distribuciones de máxima entropía en el intervalo (a, b) con momentos de segundo orden dados y plantea, presentando algunos resultados, la utilización de las medidas de divergencia en problemas de estimación paramétrica. Este segundo tópico no ha sido suficientemente estudiado hasta la fecha y por el interés que tiene voy a dedicar unas líneas al problema de estimación puntual desde esta perspectiva: Sea $(\aleph, \beta_{\aleph}, P_\theta)$, $\theta \in \Theta$, un espacio estadístico, $X_1, \ldots, X_n$ una muestra aleatoria simple de P_{θ_0} con $\theta_0 \in \Theta$ y desconocido. Para cada θ, $\theta_0 \in \Theta$ se considera la función de pérdida no negativa $L(\theta_0, \theta)$ y la familia de (h, ϕ)-divergencias no negativas $D_\phi^h (P_{\theta_0}, P_\theta)$. Una medida de (h, ϕ)-divergencia y una función de pérdida se dice que son equivalentes si para cada θ_0, $\theta \in \Theta$, se verifica

$$D_\phi^h (P_{\theta_0}, P_\theta) = L (\theta_0, \theta).$$

Vajda (1989) estableció que si se considera la distancia

$$D (F_{\theta_0}, F_\theta) = \int_0^1 (F_{\theta_0}^{-1} - F_\theta^{-1})^2 \, dW$$

con W distribución de probabilidad sobre (0, 1), verificando

$$\int_0^1 (F^1)^2 \, dW < \infty,$$

si el parámetro es de localización ($F_\theta(x) = F(x - \theta)$, $\theta \in \Theta \subset \mathbb{R}$), entonces se tiene que la distancia D y la pérdida cuadrática son equivalentes. Eligiendo convenientemente las funciones h y ϕ es posible encontrar funciones de pérdida usuales equivalentes a las (h, ϕ)-divergencias correspondientes.

Desde un punto de vista clásico el estimador $\hat{\theta}_n$ minimiza en algún sentido la función de riesgo

$$R(\theta_0, \hat{\theta}_n) = E_{\theta_0}(L(\theta_0, \hat{\theta}_n))$$

o lo que es lo mismo

$$E_{\theta_0}(L(\theta_0, \hat{\theta}_n)) = E_{\theta_0}(D(P_{\theta_0}, P_{\hat{\theta}_n})).$$

en el caso de que D sea una distancia equivalente con L. Si reemplazamos la distribución desconocida P_{θ_0} por un buen estimador no paramétrico P_n basado en $X_1, \ldots, X_n$ se obtendrán estimadores que minimizan

$$E_{\theta_0}(D(P_n, P_{\hat{\theta}_n}))$$

Es claro que esta esperanza se minimiza si

$$D(P_n, P_{\hat{\theta}_n}) = \inf_{\theta \in \Theta} D(P_n, P_\theta) \qquad \text{c.s.}$$

Este estimador representa la contrapartida al estimador minimax dentro del método de la mínima distancia,

$$\max_{\theta_0 \in \Theta} R(\theta_0, \hat{\theta}_n) = \inf_{\theta \in \Theta} \max_{\theta_0 \in \Theta} R(\theta_0, \theta).$$

Si,

$$\sup_{\theta \in \Theta} |E_{\theta_0}(D(P_{\theta_0}, P_\theta)) - E_{\theta_0}(D(P_n, P_\theta))|$$

tiende a cero suficientemente rápido, los dos estimadores son equivalentes.

Un razonamiento análogo se puede hacer para estimadores Bayes o centrados uniformemente de mínima varianza.

Otro punto de interés es la importancia de la divergencia de Kullback como función de pérdida en la estimación de densidades. Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de una densidad f y sea $\hat{f}$ un estimador de la forma

$$\hat{f}(x, h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}$$

con ventana h y núcleo K. Habema, Herman y Van der Broek (1978) proponen como parámetro de suavizado el valor h que minimiza

$$CV(h) = n^{-1} \sum_{i=1}^{n} \log \hat{f}_i (X_i, h)$$

donde

$$\hat{f}_i (x, h) = \{(n-1)\, h\}^{-1} \sum_{j \neq i} K \{(x - X_j)/h\}.$$

La justificación de este procedimiento como señala Hall (1987) se encuentra en que CV(h) es un estimador insesgado de

$$E(D(f, \hat{f})) \qquad \text{(D divergencia de Kullback)}$$

para muestra de tamaño n – 1 y en consecuencia el valor de h que maximiza CV(h) minimizará asintóticamente, bajo determinadas condiciones (Hall 1987), la pérdida esperada de Kullback.

En relación con los interesantes comentarios del Profesor C. Cuadras quisiera comenzar señalando que así como la entropía de Shannon, como muy bien él indica, tiene una interpretación desde el punto de vista de la Física, otras también lo tienen, como por ejemplo la energía informacional en términos de la Energía Cinética de la Mecánica Clásica (Guiasu 1977, p. 61). Por otra parte, como se pone de manifiesto en Pardo, L. y Taneja (1991) y Morales, Pardo y Quesada (1987) la resolución de los denominados «problemas lógicos» se puede hacer con cualquier medida de entropía, no únicamente con la entropía de Shannon. Quizá en este sentido sea conveniente señalar que mientras la energía informacional o la entropía cuadrática admiten un estimador analógico centrado con una expresión sencilla, la entropía de Shannon no. No es mi intención sobrevalorar la energía informacional o la entropía cuadrática sino únicamente señalar algunas características de las mismas que pueden ser de interés para el lector.

En la tabla 1, el Profesor C. Cuadras presenta un cuadro de proporciones de grupos sanguíneos para diferentes razas humanas y algunas medidas de diversidad en el que se aprecia una relación monótona entre la entropía de Shannon, la de Gini-Simpson y la desviación típica. Esta circunstancia le conduce a señalar que como medidas de incertidumbre las entropías del tipo

$$H_\phi (P) = \sum_{i=1}^{n} p_i \, \phi (p_i)$$

vienen a indicar lo mismo. A mi entender la relación es causal ya que si por ejemplo se consideran las distribuciones

$$P = (0.30, 0.24, 0.22, 0.12, 0.09, 0.03)$$
$$Q = (0.36, 0.21, 0.16, 0.12, 0.08, 0.07)$$

se tiene

$$GS(P) = 0.7806 > GS(Q) = 0.7750$$
$$H(P) = 1.61315 < H(Q) = 1.63138.$$

Una cuestión importante en relación con las medidas de diversidad es la inconsistencia entre ellas. En general, como señala Hulbert (1971) diferentes medidas de diversidad pueden dar lugar a órdenes inconsistentes. Una situación donde puede evitarse la inconsistencia es cuando las distribuciones de probabilidad bajo estudio están mayorizadas: Dadas las distribuciones $P = (p_1, ..., p_M)$ y $Q = (q_1, ..., q_M)$ se dice que P está mayorizada por Q ($P <_m Q$) si se verifica

$$\sum_{i=1}^{r} p_{(i)} \leq \sum_{i=1}^{r} q_{(i)} \qquad r = 1, ..., M - 1$$

con $p_{(1)} \geq p_{(2)} \geq ... \geq p_{(M)}$ y $q_{(1)} \geq q_{(2)} \geq ... \geq q_{(M)}$.

En este caso si las medidas de diversidad consideradas son schur-cóncavas se mantiene el orden dado por la relación de mayorización (intuitivamente $P <_m Q$ significa que P está más separada en el sentido de una mayor variabilidad que la Q):

$$\text{Diversidad}(P) > \text{Diversidad}(Q).$$

Buena parte de las medidas de entropía usuales son schur-cóncavas: Shannon, Havrda y Charvat, Rényi, Arimoto, etc. Un estudio interesante acerca de ordenaciones con diversidades puede verse en Patil y Taille (1982).

Una cuestión sí me gustaría señalar en relación con las entropías generalizadas y que hace referencia al sentido que para nosotros tiene el funcional de las (h, ϕ)-entropías (lo mismo se podría señalar en relación con las (h, ϕ)-divergencias o las R_ϕ^h–divergencias). Nuestro objetivo al introducir tales funciones fue poder estudiar de forma unificada y no entropía por entropía (o divergencia por divergencia) características y aplicaciones comunes en todas ellas. Es por ello y para evitar que pudiera pensarse que se trata de construir nuevas medidas de entropía (o divergencia) que en nuestros últimos trabajos hemos querido dejar esto suficientemente claro, por lo que textualmente se dice lo siguiente: «Under the above formula, a wide variety of divergence measures, can be enumerated; however, our aim is not to introduce a new measure. We have tried to give a very general functional, which can be used to do global studies, instead of measure

to measure individualized studies, therefore we would like to consider the (h, ϕ)-entropy measures as a simple way to summarize previously definied entropy measures, in such a way that when we talk about (h, ϕ)-entropies is because we have in mind an already existing and studied entropy. The final purpose is to save time and work».

Finalmente, en relación a la interesante cuestión que el Profesor Cuadras plantea acerca de la elección de la medida de divergencia (lo mismo se puede preguntar acerca de las medidas de entropía) adecuada para una determinada situación, mi opinión personal es que ésta dependerá del problema que se trate de resolver. Yo no puede compartir las palabras que el Profesor Cuadras cita sobre la opinión acerca de este tema de R. R. Sokal. Hasta la fecha, éste es un problema que nos ha preocupado y que hemos abordado en el caso de poblaciones multinomiales cuando se han utilizado las (h, ϕ)-divergencias en la construcción de contrastes de hipótesis para los problemas de bondad de ajuste, homogeneidad e independencia. En este sentido, aunque para cada uno de los contrastes se tiene la función de potencia no resulta muy útil su utilización, en procedimientos asintóticos, para la comparación de diversos contrastes de hipótesis. Por ello, parece adecuado utilizar la eficiencia tipo Pitman o tipo Bahadur para abordar el problema. Zografos (1992) ha establecido que dentro de la clase de medidas de divergencia de Csiszar la eficiencia Pitman es la misma cuando se utilizan en los problemas de contraste de independencia. Por otra parte, se sabe que dentro de esta familia de divergencias, la de Kullback es la que da lugar a un contraste más eficiente en el sentido de Bahadur. Estos resultados se han extendido, dentro de la familia de las (h, ϕ)-divergencias, para los constrastes de bondad de ajuste, homogeneidad e independencia en Menéndez, Morales, Pardo y Salicrú (1993).

En el caso de familias de divergencias más reducidas es posible comparar los miembros de la familia atendiendo a los momentos y a las funciones de distribución. Centremos nuestra atención en la familia de divergencias de orden r y grado s, D_r^s de Sharma y Mittal. Bajo la hipótesis (por ejemplo) de bondad de ajuste, H_0: $P = P_0$, se sabe que

$$T_{r,s,n} = \frac{2n}{r} D_r^s (\hat{P}, P_0)$$

se distribuye asintóticamente según una χ^2_{M-1}. Un método de analizar la velocidad de convergencia a la distribución asintótica consiste en calcular los desarrollos de Taylor de segundo orden de los momentos respecto al origen de los estadísticos $T_{r,s,n}$. Los tamaños de los términos de corrección nos darán alguna información sobre el error de aproximación que se comete al usar la distribución asintótica χ^2_{M-1} en lugar de la distribución de probabilidad exacta del estadístico.

Más concretamente, se tendrá

$$E\,(T_{r,\,s\,n}) = (M - 1) + \frac{1}{n} f_1\,(r, s, M, P_0) + O(n^{-3/2})$$
$$E\,(T^2_{r,\,s\,n}) = (M^2 - 1) + \frac{1}{n} f_2\,(r, s, M, P_0) + O(n^{-3/2})$$

y dados M y P_0 se seleccionarían aquellos valores de r y s que hagan

$$f_1(r, s, M, P_0) = f_2(r, s, M, P_0) = 0$$

Dichos valores harán que la convergencia de los momentos exactos a los asintóticos sea más rápida.

También es interesante utilizar esta metodología en las hipótesis alternativas contiguas, $H_{1,\,n}$: $P = P_0 + c/n^{1/2}$, donde $c = (c_1, \ldots, c_M)^t$ y $\sum_{i=1}^{M} c_i = 0$.

Obsérvese que ahora la distribución asintótica de $T_{r,s,n}$ es $\chi^2_{M-1}(\delta)$, donde el parámetro de no centralidad es $\delta = c^t \operatorname{diag} (P_0)^{-1} c$. Además,

$$E\,(T_{r,\,s\,n}) = (M - 1 + \delta) + n^{-1/2}\, g_1\,(r, s, M, P_0) + O(n^{-1})$$
$$E\,(T^2_{r,\,s\,n}) = (M^2 - 1 + 2\,(M + 1)\,\delta + \delta^2) + n^{-1/2}\, g_1\,(r, s, M, P_0) + O(n^{-1}).$$

Cressie y Read (1984) establecieron en el caso de tomar

$$\phi(x) = (x^{a+1} - 1)\,/a\,(a + 1) \qquad y \qquad h(x) = x$$

con estas técnicas, que el rango de valores recomendado para a era el intervalo [1/3, 2/3] con un especial énfasis en el valor a = 2/3.

Al centrarse el trabajo en aplicaciones estadísticas de medidas cuantitativas de incertidumbre y divergencia no se mencionaron, como muy bien señala el profesor P. Gil las medidas ponderadas de entropía y divergencia. Es claro, como puede evidenciarse en las dos siguientes situaciones, que el enfoque cuantitativo no agota todos los aspectos de la información: La primera se refiere a una leyenda de la mitología griega. Teseo, partiendo para una expedición, ha prometido a su padre Egeo que, si consigue su hazaña, remplazará a su vuelta la vela negra de su barco por una blanca. Podemos ahora hacernos la pregunta: ¿Cuál es la cantidad de información contenida en las velas? La segunda situación se reduce a la cuestión trivial: ¿Cuál es la información que obtenemos cuando lanzamos una moneda? Aunque completamente diferentes, estos dos problemas ponen en juego el mismo esquema. Las dos velas como las dos caras de la moneda representan dos sucesos equiprobables y encierran la misma cantidad de información, $\log_2 2 = 1$. Pero para Egeo, entre la información proporcionada por la vela negra y la proporcionada por la vela blanca había una distinción

importante; así al olvidarse Teseo de sustituir la vela negra por la blanca, su padre de desesperación se lanzó al mar desde lo alto de una roca. Sin embargo, desde un enfoque cuantitativo de la Teoría de la Información esta distinción no es esencial.

En un sistema cibernético (biológico o técnico) la actividad está dirigida hacia la realización de un fin cualquiera. El sistema debe disponer enconces de un criterio para poder diferenciar los sucesos. El criterio cibernético para una diferenciación cualitativa de los sucesos consiste en la importancia, la significación o la utilidad de la información que reportan respecto al fin. La aparición de un suceso elimina una doble «incertidumbre»: Una de orden cuantitativo respecto a la probabilidad de aparición y otra de orden cualitativo relativa a su utilidad en la realización del fin. Esto motivó la necesidad de conjuntar en una expresión estos dos conceptos fundamentales: probabilidad y utilidad. Belis y Guiasu (1968) adaptaron la entropía de Shanon a este nuevo esquema. P. Gil (1975) la redefinió de una forma más razonable y M. A. Gil (1979) la caracterizó axiomáticamente. L. Pardo (1981) hizo lo propio con la entropía cuadrática, siendo J. A. Pardo (1986) quien la caracterizó axiomáticamente. Posteriormente se han realizado estudios adaptando diversas medidas de entropía a este esquema y estudiando algunas interesantes aplicaciones. Entre estos trabajos merecen una especial mención los de Emptoz (1976), M. A. Gil; Pérez y P. Gil (1989), Hooda (1984), Singh (1983), Sharma, Mitter y Mohan (1978), Mohan y Mitter (1978), Singhal, Tuteja y Jain (1988), Khan y Autar (1979).

En lo que hace referencia a las medidas de divergencia, Taneja (1985) adapta la divergencia de Kullback y posteriormente Frank, Menéndez y L. Pardo hacen lo propio con la ϕ-divergencia estudiando además su distribución asintótica y su aplicación al contraste de hipótesis.

En relación con el otro punto que señala el Profesor P. Gil: La Teoría axiomática de la Información, si bien me parece de interés desde un punto de vista matemático le aplicaría una frase sacada textualmente de sus propios comentarios: Las «florituras» matemáticas son, evidentemente, valiosas como aportación al conocimiento general, pero poco rentables para el conocimiento del mundo que nos rodea.

Las reflexiones del profesor R. Gutiérrez son interesantes, importantes y quizá muchas de ellas puedan ser el germen de futuras investigaciones si se tiene en cuenta, como muy bien señala, la etapa de madurez en la que en estos momentos se encuentra la Teoría de la Información Estadística. A la luz de los comentarios del apartado 3 siento no haber expresado en algún momento mi punto de vista sobre algunas cuestiones tratadas y que como consecuencia de ello se pueda deducir un panorama idílico de lo tratado. Así, tras una breve incursión en la Geometría Diferencial, quizá por eso me falte tener un horizonte más amplio, e incluso tras publicar algunos trabajos en este campo, mi opinión sea totalmente coincidente con la dada

por el Profesor R. Gutiérrez: Pocos han sido los aspectos innovadores que la Geometría Diferencial ha aportado a la Estadística. Una gran parte de los trabajos en este área y relacionados con la Teoría de la Información Estadística han estado enfocados a la obtención de nuevas métricas Riemannianas y al cálculo de la distancia en diversas familias de distribuciones de probabilidad, quedando en un segundo lugar la aplicación de las mismas a la obtención de resultados concretos que de alguna manera las justifiquen. Estoy completamente de acuerdo con el profesor R. Gutiérrez, y tomo buena nota de ello, que un tema pendiente de la Teoría de la Información Estadística es construir una visión integradora entre lo que es Información en Estadística Matemática y la utilización de la Información en el Análisis de Datos o Estadística sin Modelo probabilístico.

En relación a las medidas de información y divergencia en el Análisis Multivariante quiero señalar que si bien no se ha mencionado en el trabajo por motivos de extensión sí que existen interesante trabajos en lo que respecta al Análisis Discriminante. Así, en Kailath (1967), Kanal (1974), Chen (1976), Boekee y Van der Lubbe (1979), Taneja (1985), Toussaint (1974), Vicente (1993), Menéndez y otros (1993), etc., se presentan cotas para la probabilidad de clasificación errónea en términos de medidas de información y divergencias si el individuo, w, se asigna a la clase C_i cuando su probabilidad a posteriori, una vez observado $X = (X_1, ..., X_n)$ es la más alta; es decir, se clasifica de acuerdo a la regla Bayes. Considerar alguna clase restringida de divergencias como por ejemplo las de orden r y grado s de Sharma y Mittal o las de Cressie y Read, construir en base a estas medidas de divergencias un clasificador para las familias esféricas y elípticas de distribuciones multivariantes y analizar si existe algún valor de r y s dentro de las medidas de divergencia de orden r y grado s o algún valor de r dentro de las de Cressie y Read que den lugar a una probabilidad de clasificación errónea menor que el clasificador obtenido con la de Kullback por Quan, Fang y Teng (1990) parece interesante.

Por otra parte, como señala el profesor R. Gutiérrez, será necesario estudiar si al hacer contrastes en general con las familias de distribuciones esféricas y elípticas de distribuciones multivariantes, de una forma similar a los procedimientos que hemos venido desarrollando con poblaciones normales, conducen a unos resultados tan satisfactorios como con estas poblaciones.

El profundo conocimiento que el profesor R. Gutiérrez tiene sobre procesos estocásticos le hace plantear de una forma clara y precisa el papel que ha jugado hasta la fecha la Teoría de la Información Estadística en el campo de los procesos Estocásticos, dejando bien claro, como en realidad es, que éste es un campo en el que los estudiosos de la Teoría de la Información Estadística deberíamos volcarnos como culminación del interés e importancia de ésta en Estadística. Si bien éste es un tema que no hemos tratado en profundidad, quizás ésta sea una línea de encuentro para una investigación futura conjunta entre nuestro grupo y

el del profesor R. Gutiérrez, sí que tenemos algunos resultados que mencionaré posteriormente. Antes quiero comentar un resultado interesante debido a Blackwell (1956) y por otra parte curioso ya que como es conocido a Blackwell siempre se le asocia con la Teoría de la Decisión: Sea $\{X_n, -\infty < n < \infty\}$ un proceso estocástico ergódico estacionario con un número finito de estados, $i = 1, ..., I$. Para cualquier secuencia finita $s = (i_1, ..., i_k)$, $1 \leq i_j \leq I$, $k = 1, 2, ...$, sea $q(s) = P(X_1 = i_1, ..., X_k = i_k)$, y sea $Z_k = q(X_1, ..., X_k)$. McMillan (1953) estableció que la sucesión de variables aleatorias $\{Z_n\}$ es asintóticamente consistente en el sentido de que existe una constante $H \geq 0$, llamada entropía del proceso $\{X_n\}$, tal que $E(|n^{-1} \log Z_n + H|) \xrightarrow[n \to \infty]{} 0$. Este resultado implica que para valores grandes de n, cualquier realización del vector aleatorio $(X_1, ..., X_n)$ tiene una probabilidad aproximada de ocurrencia de 2^{-nH}, la constante H es fundamental para el desarrollo de la Teoría de la Información. Además, si $U(i) = P(X_1 = i/X_0, X_{-1}, ...,)$ y $V = U(X_1)$, McMillan probó que $H = -E(\log V)$. Si $\{X_n\}$ es un proceso de Markov con $\lambda_i = P(X_n = i)$, $m(i, j) = P(X_{n+1} = j/X_n = i, X_{n-1}, ...,)$, se tiene $U(i) = m(X_0, i)$, $V = m(X_0, X_1)$, con lo cual

$$H = \sum_{i,j} \lambda_i \, m(i, j) \log m(i, j).$$

Entonces la entropía de un proceso de Markov es fácilmente calculable. Sin embargo, si ϕ es una función definida en $1, ..., I$ con valores $1, ..., A$ no existía una expresión para al entropía de $\{Y_n = \phi(X_n)\}$ tan simple como la dada anteriormente. Blackwell estableció que la entropía del proceso $\{Y_n\}$ venía dada por

$$H = -\int \sum_{\alpha} r_{\alpha}(w) \log r_{\alpha}(w) \, dQ(w)$$

donde Q es una distribución de probabilidad sobre los conjuntos de Borel del conjunto W de vectores $w = (w_1, ..., w_I)$ con

$$w_i \geq 0, \ \sum_i w_i = 1 \text{ y } r_a(w) = \sum_{i=1}^{I} \sum_{j/\phi(j)=a} w_i \, m(i, j).$$

La distribución Q está concentrada sobre los conjuntos $W_1, ..., W_A$, donde W_α está formada por todos los $w \in W$ con $w_i = 0$ para $\phi(i) \neq a$ y satisface la ecuación

$$Q(E) = \sum_a \int_{f_a^{-1}(E)} r_a(w) \, dQ(w)$$

donde f aplica W en W_a con la j-ésima coordenada de $f_a(w)$ dada por $\sum_i w_i \, m(i, j)/r_a(w)$ para $\phi(j) = a$. Además Blackwell establece bajo determinadas condiciones de regularidad que la solución de la ecuación integral que obtiene para Q es única.

Esta idea sirvió posteriormente a McFadden (1965) para definir y estudiar las propiedades de la entropía de un proceso puntual idea desarrollada posteriormente con la energía informacional en Morales, Pardo y Quesada (1985). El cálculo de la divergencia entre dos procesos puntuales y el estudio de sus propiedades se ha realizado en Morales y Pardo (1992).

Un problema poco tratado, o al menos desconocido para mí, es la utilización de medidas de divergencia y entropía con el fin de hacer inferencias sobre procesos estocásticos. En este sentido en Morales, Pardo y Quesada (1991), se adapta la χ^2-divergencia para cuantificar la información que una muestra aleatoria simple $X_1, ..., X_n$ proporciona sobre un proceso de Dirichlet, F(t), con parámetros $\alpha(\cdot)$ en un contexto bayesiano no paramétrico y se utiliza para establecer una regla de parada en el muestreo secuencial en aquellos problemas donde el objetivo del estadístico no sea alcanzar una decisión sino recoger información acerca de la función de distribución aleatoria. El trabajo de Barndorr-Nielsen y Sorensen (1991) resulta de interés en relación a la problemática que se plantea en los procedimientos de inferencia en procesos estocásticos con medidas de información, mientras que el libro de Liese y Vajda (1987) efectúa un tratamiento riguroso de procesos con incrementos independientes a través de divergencias.

El interés del método de estimación paramétrica basado en la maximización de la entropía mencionada en los comentarios de los Profesores H. Gzyl y D. Peña es claro y como ellos señalan sería muy interesante investigar sus propiedades. Resulta también digna de mención la utilización del principio de maximización de la entropía en la estimación no paramétrica de densidades. A este respecto quiero señalar los procedimientos debidos a Theil y Laitinen (1980), Kapur (1983) y Rodríguez y Van Ryzin (1985).

Sea $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ la muestra ordenada y $\xi_1, ..., \xi_{n-1}$ valores intermedios a determinar y que luego señalaremos su forma de obtención. Estos n – 1 puntos determinan n intervalos abiertos de la forma:

$$I_1 = (-\infty, \xi_1);\ I_2 = (\xi_1, \xi_2);\ ...;\ I_{n-1} = (\xi_{n-2}, \xi_{n-1});\ I_n = (\xi_{n-1}, \infty)$$

Kapur (1983) bajo las restricciones: (i) La masa de probabilidad en cada I_i, ha de ser 1/n. (ii) La media de la variable objeto de estudio en cada intervalo I_i es $x_{(i)}$, encuentra que la distribución que maximiza la entropía viene dada por

$$f(x) = \begin{cases} \dfrac{1}{n}\dfrac{1}{\xi_1 - x_{(1)}} \exp\left(\dfrac{x - \xi_1}{\xi_1 - x_{(1)}}\right) & \text{si } -\infty < x < \xi_1 \\ A_i \exp(a_i x) & \text{si } \xi_{i-1} < x < \xi_i \quad i = 2, ..., n-1 \\ \dfrac{1}{n}\dfrac{1}{x_{(n)} - \xi_{n-1}} \exp\left(\dfrac{\xi_{n-1} - x}{x_{(n)} - \xi_{n-1}}\right) & \text{si } \xi_{n-1} < x < \infty \end{cases}$$

donde los a_i se obtienen mediante la resolución de las ecuaciones

$$a_i = \frac{1 - \exp(a_i(\xi_{i-1} - \xi_i))}{(\xi_i - x_{(i)}) - \exp(a_i(\xi_{i-1} - \xi_i))(\xi_{i-1} - \xi_i)}, \qquad i = 2, \ldots, n-1$$

y los A_i mediante

$$A_i = \frac{\exp(-a_i \xi_i)}{n\left((\xi_i - x_{(i)}) - \exp(a_i(\xi_{i-1} - \xi_i))(\xi_{i-1} - x_{(i)})\right)} \qquad i = 2, \ldots, n-1$$

En relación con la determinación de los $\xi_1, \ldots, \xi_{n-1}$, Kapur propuso tres métodos: (a) Si se impone que f(x) sea continua en todos los puntos se tienen n – 1 ecuaciones a partir de las cuales se pueden determinar ξ_i, i = 1, ..., n – 1,

$$A_i \exp(a_i \xi_i) = A_{i+1} \exp(a_{i+1} \xi_i), \qquad i = 1, \ldots, n-1.$$

(b) Si nos centramos no en la continuidad de la densidad, f(x), sino en que tenga la máxima entropía posible, será necesario maximizar

$$H(\xi_1, \ldots, \xi_{n-1}) = \sum_{i=1}^{n} \left(\log \frac{A_i}{n} + a_i \frac{x_i}{n} \right).$$

Conociendo A_i y a_i en función de ξ_i, podemos elegir éstos de tal forma que maximicen $H(\xi_1, \ldots, \xi_{n-1})$ sujetos a

$$-\infty < x_{(1)} < \xi_1 < x_{(2)} < \ldots < \xi_{n-1} < x_{(n)} < \infty.$$

(c) Elegir $\xi_i = \frac{x_{(i)} + x_{(i+1)}}{2}$, i = 1,..., n – 1.

En lo sucesivo estos tres métodos se denotarán por DMEI, DMEII y DMEIII respectivamente. Se comprueba que la varianza de la variable aleatoria cuya densidad es la de máxima entropía obtenida mediante cualquier de los tres procedimientos anteriores excede a la varianza muestral en una cantidad que es el promedio de las varianzas dentro de cada uno de los intervalos.

En el procedimiento establecido por Theil y Laitinen (1980), la primera condición impuesta es análoga a la señalada anteriormente en el procedimiento de Kapur, mientras que en lugar de la restricción (ii) proponen la restricción: (ii') La función de densidad estimada, f(x), de la variable aleatoria objeto de estudio X se debe construir de forma que la media de X en cada intervalo I_i sea una función lineal convexa de los valores $x_{(i)}$ asociados a los extremos del intervalo y tal que la media global o general se conserve (sea igual a la muestral).

Una cuestión importante es que los valores ξ_i son de la forma $\xi_i = \xi(x_{(i)}, x_{(i+1)})$, i = 1, ..., n, donde $\xi(.,.)$ es una función simétrica y diferenciable en sus dos argumentos, cuyo valor está dentro del rango definido por ellos.

Con estas hipótesis Theil y Latinen obtienen como densidad de máxima entropía bajo las restricciones (i) y (ii'), la siguiente

$$f(x) = \begin{cases} \dfrac{1}{n}\dfrac{1}{\frac{1}{4}(x_{(2)} - x_{(1)})} \exp\left(\dfrac{4x - 2x_{(1)} - 2x_{(2)}}{x_{(2)} - x_{(1)}}\right) & \text{si } -\infty < x < \dfrac{x_{(1)} + x_{(2)}}{2} \\ \dfrac{2}{n(x_{(i+1)} - x_{(i-1)})} & \text{si } \dfrac{x_{(i-1)} + x_{(i)}}{2} < x < \dfrac{x_{(i)} + x_{(i+1)}}{2} \\ \dfrac{1}{n}\dfrac{1}{\frac{1}{4}(x_{(n)} - x_{(n-1)})} \exp\left(\dfrac{2x_{(n-1)} + 2x_{(n)} - 4x}{x_{(n)} - x_{(n-1)}}\right) & \text{si } \dfrac{x_{(n-1)} + x_{(n)}}{2} < x < \infty \end{cases}$$

que evidentemente es uniforme en los intervalos acotados I_2, ..., I_{n-1} y exponencial en I_1 e I_n. Este método se designará por DMEIV. En este caso la varianza de la variable aleatoria asociada a la densidad estimada viene dada por

$$\frac{1}{n}\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2 - \frac{1}{4n}\sum_{i=1}^{n-1}(x_{(i+1)} - x_{(i)})^2 - \frac{1}{24n}\sum_{i=2}^{n-1}(x_{(i+1)} - x_{(i-1)})^2$$

que obviamente es menor que la varianza muestral.

Theil y O'Brien (1980) consideraron la mediana y los cuartiles de la DMEIV como estimadores de los correspondientes valores poblacionales para muestras aleatorias de poblaciones normales y establecieron que estos estimadores tienen menor error cuadrático esperado que los muestrales para n pequeño. Theil y Lightburn (1981) establecieron que si el rango es (0, ∞) la DEMEIV sobre (0, ξ_1) es una exponencial truncada.

Finalmente se indicará el procedimiento señalado por Rodríguez y Van Ryzin (1985). Sea k un número entero positivo fijo, $m = \left[\frac{n}{k}\right]$ y q = n – mk. Entonces n = mk + q con q ∈ {0, 1, .., k – 1}. Se define también

$$\xi_i = h(x_{(ki-\theta+1)}, x_{(ki-\theta+2)}, \ldots, x_{(ki+\theta)})$$

donde h es una función, que habrá que determinar, de 2θ argumentos tal que $\xi_i \in (x_{(ki-\theta+1)}, x_{(ki+\theta)})$ y donde

$$\theta = \begin{cases} \dfrac{k}{2} & \text{si k es par} \\ \dfrac{k+1}{2} & \text{si k es impar} \end{cases}$$

A partir de aquí se considera la partición

$$I_1 = (-\infty, \xi_1), \ldots, I_i = [\xi_{i-1}, \xi_i), \ldots, I_n = [\xi_{n-1}, \infty)$$

y para obtener la densidad de máxima entropía imponen las dos condiciones siguientes:

(a) Restricciones de conservación de la masa,

$$\int_{I_i} f(x)\,dx = P_f(X \in I_i) = \frac{k}{n} \qquad i = 1,\ldots, m-1$$

$$\int_{I_m} f(x)\,dx = P_f(X \in I_m) = \frac{k+q}{n}$$

(b) Restricciones de conservación de la media. Por un lado se requiere para cada I_i que $E(X/X \in I_i)$ sea una combinación lineal convexa de todos los estadísticos ordenados implicados en la definición de I_i, i = 1, ..., m. Por otro que se conserve la media general

$$\int_{\mathbf{R}} xf(x)\,dx = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}.$$

Con estas restricciones la densidad de máxima entropía viene dada por

$$f_n(x) = \frac{k/n}{\xi_i - \xi_{i-1}} \qquad \text{si } x \in [\xi_{i-1}, \xi_i), \quad i = 2,\ldots, m-1$$

siendo exponencial en los intervalos extremos I_1 e I_m. Además,

$$\xi_i = h(z_1,\ldots, z_{2\theta}) = \begin{cases} \dfrac{1}{2k} z_1 + \dfrac{1}{k}\displaystyle\sum_{j=2}^{k} z_j + \dfrac{1}{2k} z_{k+1} & \text{si k es par} \\ \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} z_j & \text{si k es impar} \end{cases}$$

Este procedimiento se denomina Histograma de Máxima Entropía (HME). Es claro que la DMEIV es un caso particular del HME con k = 1 y por tanto m = n, q = 0 y θ = 1.

Si se elige $k = k(n)$ de forma que $k(n) \xrightarrow[n \to +\infty]{} \infty$ y $n^{-1}k(n) \xrightarrow[n \to +\infty]{} 0$, entonces

$f_n(x) \xrightarrow{c.s} f(x)$ para todo x punto de continuidad de f. Es decir, se tiene garantizada la consistencia fuerte del HME. La normalidad asintótica del HME se tiene si se permite que $k = k(n)$ diverja de forma que $n^{-1}k(n)^{3/2} \xrightarrow[n \to \infty]{} 0$, ya que en este caso

$$k(n)^{1/2} (f_n(x) - f(x)) \xrightarrow{L} N(0, \frac{2}{3} f(x))$$

para todo punto de continuidad de f.

En relación con la entropía de Shannon y los estadísticos ordenados, si bien no con el principio de máxima entropía, es interesante resaltar los tests de normalidad y uniformidad establecidos por Vasicek (1976) y Dudewicz y Van der Meulen (1981). La utilización para estos problemas de otras medidas de información proporciona unos resultados tan aceptables como con la entropía de Shannon como puede verse en Pardo, M. C. (1993).

El Profesor J. de la Horra señala un importante resultado en relación con la obtención de distribuciones de referencia bajo restricciones sobre los cuantiles debido a que la distribución de referencia resultante es independiente de la medida de entropía empleada. Este hecho no es habitual ya que como también señala el Profesor De la Horra en general depende de la medida de entropía utilizada. Es claro que la obtención de distribuciones de referencia con medidas de entropía distintas de la de Shannon es una línea abierta de investigación, si bien con unas complicaciones adicionales. El método de los multiplicadores de Lagrange no admite más que restricciones de igualdad por lo que no cubre las restricciones de no negatividad que debe verificar la distribución de referencia. En el caso de la entropía de Shannon se tiene asegurada la no negatividad de la distribución de referencia por ser ésta de tipo exponencial. El no poder utilizar en general el método de los multiplicadores de Lagrange conduce a que con determinadas restricciones y algunas medidas de entropía haya que recurrir a métodos de optimización más complejos, pero no por ello menos exentos de interés.

En el caso particular de la utilización de la entropía cuadrática para la obtención de distribuciones de referencia el problema se soluciona de forma conveniente mediante la programación cuadrática.

REFERENCIAS

BARNDORFF-NIELSEN, O. E. y SORENSEN, M. (1991). «Information quantities in non-classical setting». *Computational Statistics and Data Analysis*, **12**, 143-158.

BELIS, M. y GUIASU, S. (1968). «A Quantitative-Qualitative Measure of Information in Cybernetics Systems». *IEEE Trans. Inf. Th.* **IT-4,** 593-594.

BLACWELL, D. (1956). «The entropy of functions of finite-State Markov Chains». Trans. 1th Prague Conf. on Inform. Theory, Statist. Dec. Functions, Random Process. Academia-Kluwer, Praha-Dodrecht, 13-20.

BOEKEE, D. E. y VAN DER LUBBE (1979). «Some aspects or error bounds in feature selection». *Pattern Recogn.* **11,** 353-360.

CHEN, C. H. (1976). «On information and distance measures, error bounds, and feature selection». *Inf. Sci.* **10,** 159-171.

CRESSIE, N. y READ, T. R. C. (1984). «Multinomial goodness-of fit tests». *J. Roy. Statist. Soc. Ser. B,* **46,** 440-464.

DUDEWICZ, E. J. y DER MEULEN, E. C. V. (1981). «Entropy-Based Tests of Uniformity». *Journal of the American Statistical Association,* **76,** 967-974.

EMPTOZ, H. (1976). «Information of type β integrant un concept d'utilite». *C. R. Acad. Sci. Paris,* **282A,** 911-914.

FRANK, O., MENÉNDEZ, M. L. y PARDO, L. (1991). «Asymptotic Properties of Weighted Divergence between Distributions». *Tech. Rep. Stockholm Unversity. Department of Statistics.*

GIL, M. A. (1979). «Incertidumbre y utilidad». Tesis Doctoral, Oviedo.

GIL, P. (1975). «Medidas de incertidumbre e información en problemas de decisión estadística». *Rev. Acad. CC. Ex. Fis. Nat. de Madrid,* **LXIX,** 549-610.

GIL, M. A., PÉREZ, R. y GIL, P. (1989). «A family of measures of uncertainty involving utilities: Definitions, Properties and Statistical Inferences». *Metrika,* **36,** 129-147.

GUIASU, S. (1977). *Information Theory with Applications.* McGraw Hill, New York.

HABBEMA, J. D. F., HERMANS, J. y REMME, J. (1978). «Variable kernel estimation in discriminant analysis». in *Compstat* 1978 (L. C. A. Corsten and J. Hermans, eds.) 178-185. Physica, Vienna.

HALL, P. (1987). «On Kullback-Leibler loss and density estimation». *The Annals of Statistics.* vol. **15,** 4, 1491-1519.

HOODA, D. S. (1984). «A non-Additive Generalized Measure of Relative Useful Information». *Pure App. Math. Sci.,* vol. **22,** pp. 141-151.

HULBERT, STUART, H. (1971). «The nonconcept of Species Diversity: A Critique and alternative parameters». *Ecology,* **52,** 577-586.

KAILATH, T. (1967). «The divergence and Battacharyya measures in signal selection». *IEEE Trans. Commum. Tech.*, **1,** 52-60.

KANAL, L. (1974). «Patterns in pattern recognition». *IEEE Trans. Inf. Theory,* **20,** 697-722.

KAPUR, J. N. (1989). «On the concept of Useful information». *Adv. Manag. Studies,* vol. 2, pp. 147-162.

KHAN, A. B. y AUTAR, R. (1979). «On useful information of order α and type β». *Soochow J. Math.*, **5,** 93-99.

LIESE, F. y VAJDA, I. (1987). *Convex Statistical Distances.* Teuber, Leipzig.

MAASONUMI, E. (1993). «A compendium to Information Theory in Economic and Econometrics». *Econometric Reviews,* **12(2),** 137-182.

MCFADDEN, J. A. (1965). «The entropy of a point process». *J. Soc. Indust. Appl. Math.* **12,** 988-994.

MENÉDEZ, M. L., MORALES, D., PARDO, L. y SALICRÚ, M. «Asymptotic behaviour and Statistical Applications of Divergence measures in Multinomial Populations». Enviado.

MENÉNDEZ, M. L., TANEJA, I. J., y PARDO, L. (1993). «Generalized Divergence Measures and the probability of error». *Journal of the Franklin Institute,* **330,** 345-368.

MOHAN, M. y MITTER, J. (1978). «On bounds of useful information measures». *Indian J. Pure and Appl. Math.* **9,** 960-964.

MORALES, D., PARDO, L. y QUESADA, V. (1991). «The Chi-Square Divergence Measure in Random Sampling with Dirichlet Process priors». *Information Sciences,* **55,** 239-249.

MORALES, D., PARDO, L. y QUESADA, V. (1985). «La Energía Informacional como medida de información en un proceso puntual». *Estadística Española,* **107,** 5-13.

MORALES, D., PARDO, L. y QUESADA, V. (1987). «La Energía Informacional en la resolución de problemas lógicos». Tech. Rep. 6/1. Departamento de Estadística e I. O. Universidad Complutense de Madrid.

MORALES, D. y PARDO, L. (1992). «La divergencia entre dos procesos puntuales». Tech. Rep. Departamento de Estadística e I. O.

PARDO, L. (1986). «Order α Useful Information Energy». *Information Science,* **40,** 155-164.

PARDO, J. A. (1986). «Caracterización axiomática de la Energía Informacional Util». *Estadística Española,* **108,** 107-116.

PARDO, M. C. (1993). «Un contraste de normalidad basado en la Energía Informacional». *Questiio,* **17.**

PARDO, L., MORALES, D. y QUESADA, V. (1987). «La Energía Informacional en la resolución de los problemas lógicos». Tech. Rep. 6/1987. Departamento de Estadística e I. O. Facultad de Matemáticas.

PARDO, L. y TANEJA, I. J., (1991). «Information Energy and its Applications». *Advances in Electronic and Electron Physic,* **90,** 165-241.

PATIL, G. P. y TAILLE, C. (1982). «Diversity as a concept and its measurement». *J. Amer. Stat. Assoc.* **77,** 548-567.

RODRÍGUEZ, C. C. y RYZIN, J. V. (1985). «Maximum entropy histograms». *Statistics and Probability Letters,* **3,** 117-120.

SHARMA, B. D., MITTER, J. y MOHAN, M. (1978). «On Measures of "Useful" Information». *Information and Control,* **39,** 323-336.

SINGH, R. P. (1983). «On information measure of type (α, β) with preference». *Caribb. J. Math.,* **2,** 25-37.

SINGHAL, L. C., TUTEJA, R. K. y JAIN, P. (1988). «On Measures of relative information with preference». *Commun. Statist. (Theory and Meth),* **17(5),** 1449-1464.

TANEJA, H. C. (1985). «On the Mean and the Variance of estimates of Kullback information and relative "useful" information measures». *Aplikace Matematiky,* **30,** 166-175.

TANEJA, I. J. (1985). «Generalized error bounds in Pattern Recognition». *Pattern Recog. Lett.,* **13,** 261-386.

THEIL, H. y O'BRIEN, P. C. (1980). «The median of the maximum entropy distribution». *Econometrics Letter,* **5,** 345-347.

THEIL, H. y LAITINEN, K. (1980). «Singular moment matrix in applied Econometrics». In *Multivariate Analysis V.* P. R. Krishnaiah ed., 629-649. North Holland. Amsterdam.

THEIL y LIGHTBURN (1981). «The positive maximum entropy distribution distribution». *Econometrics Letter,* **6,** 231-239.

TOUSSANINT, G. T. (1974). «On the divergence between two distributions and the probability of misclassfication of several decision rules». *Proc. 2nd Int. Joint Conf. on Pattern Recognition, 1-8 Aug.* Copenhagen.

VAJDA, I. (1989). *Theory of Statistical Inference and Information.* Kluvwer Academic Press, Dordrecht, The Netherlands.

VASICEK, O. (1976). «A Test for Normality Based on Sample Entropy». *Journal of the Royal Statistical Society,* **69,** 730-737.

VICENTE, M. L. (1993). «Information of degree β and probability of error». *Journal of the Franklin Institute,* **330,** 229-242.

ZOGRAFOS, K. (1992). «Asymptotic properties of φ-divergence statistic and its applications in contingency tables». Technical Report 185. University of Ioannina.