La Serie Universitaria de la Fundación Juan March presenta resúmenes, realizados por el propio autor, de algunos estudios e investigaciones llevados a cabo por los becarios de la Fundación y aprobados por los Asesores Secretarios de los distintos Departamentos.

El texto íntegro de las Memorias correspondientes se encuentra en la Biblioteca de la Fundación (Castello, 77. Madrid-6).

La lista completa de los trabajos aprobados se presenta, en forma de fichas, en los Cuadernos Bibliográficos que publica la Fundación Juan March.

Estos trabajos abarcan las siguientes especialidades: Arquitectura y Urbanismo; Artes Plásticas; Biología; Ciencias Agrarias; Ciencias Sociales; Comunicación Social; Derecho; Economía; Filosofía; Física; Geología; Historia; Ingeniería; Literatura y Filología; Matemáticas; Medicina, Farmacia y Veterinaria; Música; Química; Teología. A ellas corresponden los colores de la cubierta.

Edición no venal de 300 ejemplares, que se reparte gratuitamente a investigadores, Bibliotecas y Centros especializados de toda España.

Este trabajo fue realizado con una Beca de España, 1975. Departamento de Matemáticas.

Fundación Juan March

AD·MAIVS·HISPANIAE·LVMEN

Algunos tópicos sobre teoría de la información/Antonio Pascual Acosta

22

SERIE UNIVERSITARIA

Fundación Juan March

Algunos tópicos sobre teoría de la información

Antonio Pascual Acosta

22

Fundación Juan March

Serie Universitaria

22

Algunos tópicos sobre teoría de la información

Antonio Pascual Acosta

Fundación Juan March
Castelló, 77. Teléf. 225 44 55
Madrid - 6

La Fundación Juan March no se solidariza necesariamente con las opiniones de los autores cuyas obras publica.

Depósito Legal: M - 6714 – 1977
I. S. B. N.: 84 - 7075 - 045 - 3
Ibérica, Tarragona, 34.–Madrid-7

I N D I C E

Página

Introduccion

El famoso documento de Shannon "The Mathematical Theory of Communication" publicado en 1.948 marca el punto de partida de una nueva rama de la Ciencia Matematica conocida con el nombre de Teoria de la Informacion.

Hoy dia el concepto de probabilidad desempeña un papel fundamental en cualquier teoria del conocimiento. En la moderna Teoria de la Informacion, las probabilidades se consideran como una codificacion numerica de un estado de conocimiento. El conocimiento que uno tiene sobre una cuestion en particular puede representarse mediante la asignacion de una cierta probabilidad p a las varias respuestas concebidas para la cuestión. El conocimiento completo acerca de una cuestion es la posibilidad de asignar una probabilidad cero a todas las respuestas posibles excepto a una. A una persona que correctamente asigne una probabilidad unidad a una respuesta en particular, no le queda, evidentemente, nada que aprender sobre la cuestion. Observando que el conocimiento puede, por tanto, codificarse en una distribucion de probabilidad, la informacion puede definirse como algo que produce un reajuste en una asignacion de probabilidades. Numerosos trabajos han demostrado que la medida de incertidumbre de Shanon, a la que el llamaba entropia, mide cuanto cabe esperar aprender sobre una cuestion cuando todo lo que se conoce es un conjunto de probabilidades.

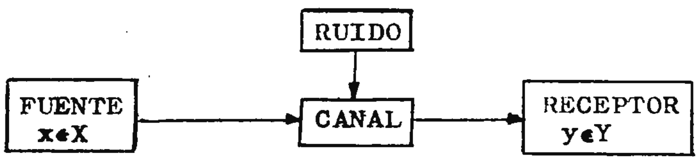
La contribucion de Shannon a la Teoria de la Información consistió en demostrar la existencia de una medida de información que es independiente de los medios empleados para generar la información. El contenido informativo de un mensaje es por tanto, invariante respecto a la forma y no depende de cómo se envía el mensaje.

La medida de Shannon fue concebida para resolver un problema concreto: Cómo obtener una medida útil de lo que se transmite por un canal de comunicaciones. Ha demostrado tambien ser la unica función capaz de satisfacer ciertas necesidades básicas de la Teoría de la Información. Apenas ha pasado un cuarto de siglo de la publicación de Shannon y ya se han escrito miles de trabajos al respecto, sin que ninguno de ellos haya encontrado una función sustitutiva, ni siquiera la necesidad de ella. Por el contrario, se han descubierto muchos otros caminos para llegar a dicha medida.

La medida de entropía de Shannon es fundamental en la Teoría de la Información y a ella y a su aplicación a diferentes problemas estadísticos dedicamos la primera parte del trabajo.

Regresión y transmisión de información

Consideremos un canal de comunicación con su fuente y receptor.



La información transmitida mide la cantidad de asociación entre la fuente y el receptor del canal. Shannon define la cantidad de información transmitida por señal como

$$R(x,y) = H(X) + H(Y) - H(X,Y) \qquad (1)$$

donde $H(\cdot)$ es la funcion de entropia.

$R(x,y)$ es una cantidad no negativa e igual a cero si y solo si X e Y son independientes y puede expresarse en la forma

$$R(x,y) = H(X) - H_y(X) = H(Y) - H_x(Y) \qquad (2)$$

Consideremos ahora el caso en que son dos las fuentes que transmite a un receptor.Si tomamos la variable fuente como bidimensional,tendremos

$$\begin{aligned} R\left[(x,y),z\right] &= H(X,Y) + H(Z) - H(X,Y,Z) = \\ &= H(X,Y) - H_z(X,Y) = H(Z) - \\ &- H_{x,y}(Z) \qquad (3) \end{aligned}$$

siendo

$$H_z(X,Y) = -\iiint p(x,y,z)\log p(x,y,z)dxdydz$$

$$H_{x,y}(Z) = -\iiint p(x,y,z)\log p(z/x,y)dxdydz$$

donde $p(z/x,y)$ es la probabilidad de recibirse en el receptor una señal z habiéndose emitido por la fuente una señal (x,y).

Vamos a expresar $R\left[(x,y),z\right]$ como una combinación de las transmisiones de información entre X y Z e Y y Z.Si definimos $R_x(y,z)$ como la media extendida a todos los valores de X,de la informa ción transmitida entre Y y Z podemos escribir

$$R_x(y,z) = H_x(Y) - H_{x,z}(Y) = H_x(Z) - H_{x,y}(Z)$$

De donde

$$R\left[(x,y),z\right] = R(x,z) + R_x(y,z) =$$
$$= R(y,z) + R_y(x,z) \qquad (4)$$

La expresión anteriormente escrita nos expresa que el teorema de adición de la entropia es válido también para la razón de transmisión de información.Por otro lado

$$R(y,z) - R_x(y,z) = \left[H(Z) - H_y(Z)\right] - \left[H_x(Z) - H_{x,y}(Z)\right] = \left[H(Y) - H_z(Y)\right] - \left[H_x(Y) - H_{x,z}(Y)\right]$$

cantidad que nos indica la ganancia o pérdida en la informacion transmitida entre Y y Z debido al conocimiento de X.

Las identidades anteriores demuestran la simetria del primer miembro en los argumentos X e Y y X y Z.Pero de (1) y (3) deducimos la simetría entre Y y Z,y por tanto tenemos

$$R(y,z)-R_x(y,z)=R(x,z)-R_y(x,z)=R(x,y)-R_z(x,y)$$

Una vez extendido el concepto de transmisión de informacion al caso de tres variables pasamos a estudiar su relacion con el concepto de regresion.

Sean X e Y dos variables aleatorias con funciones de densidad marginales $p_1(x)$ y $p_2(y)$ y funcion de densidad conjunta $p(x,y)$. Notaremos $p_1(x/y)$ la funcion de densidad condicionada de X a Y, y por $p_2(y/x)$ la función de densidad condicionada de Y a X.

Sea $m(x)$ la línea de regresión de Y sobre X es decir

$$m(x) = \int y p_2(y/x) dy$$

Introducimos una nueva variable Z definida como

$$Z/x = Y/x - m(x)$$

cuya función de densidad será

$$p_3(z) = \int p_3(z/x) p_1(x) dx = \int p_2(z - m(x)/x) p_1(x) dx$$

por ser

$$H_{x,y}(Z) = H_{y,z}(X) = H_{z,x}(Y) = 0$$

podemos escribir

$$R_z(x,y) = H_z(X) = H_z(Y) \qquad (5)$$

$$R_y(x,z) = H_y(X) = H_y(Z) \qquad (6)$$

por otro lado

$$R(x,y) = R(x,z) + R_z(x,y) - R_y(x,z)$$

y teniendo en cuenta (5) y (6)

$$R(x,y) = R(x,z) - H(Y) - H(Z)$$

siendo $R(x,z) \geq 0$ y $H(Y)-H(Z) \gtreqless 0$

Siguiendo a Feron y Fourgeaud, llamaremos a $R(x,z)$ parte elástica y a $[H(Y) - H(Z)]$ parte dura de la información transmitida $R(x,y)$.

TEOREMA. $R(x,z)=0$ si y solo si $p_2(y/x)=f(y-m(x))$ donde $f(z)$ es alguna función de densidad con media cero.

TEOREMA. La distribución Normal bivariante es la única que posee correlacion dura entre sus variables, es decir, es la única para la cual la razón de transmisión de información $R(x,y)$ coincide con las partes duras.

Relacion entre distintas cantidades de información

Las tres definiciones clásicas de información de Fisher, de Shannon y de Kullback estan fuertemente vinculadas a propiedades asintóticas y a un principio general por el cual la cantidad de información debería ser aditiva.

Fisher define la cantidad de información para un parámetro θ y una función de densidad $p(x;\theta)$ que satisfaga las condiciones de regularidad de Cramer-Rao en la forma

$$I_F(\theta)=\int\left[\frac{\partial}{\partial\theta}\log\ p(x;\theta)\right]^2 p(x;\theta)dx$$

Shannon propone una definición de informacion (incertidumbre) para la teoría de la comunicación. En su primitiva forma nos indica la variacion en una distribución;con un cambio de signo,la concentracion.Es esta segunda forma la que usaremos notándola

$$I_S(\theta)=\int\log\ p(x;\theta).p(x;\theta)\ dx$$

Kullback considera una definicion de informacion para discriminar en favor de una hipotesis $H_1(\theta_1)$ contra otra $H_2(\theta_2)$:

$$I_K(\theta_1;\theta_2)\ =\int\log\ \frac{p(x;\theta_1)}{p(x;\theta_2)}p(x;\theta_1)\ dx$$

CARACTERIZACION. Consideremos un modelo probabilístico cuya funcion de densidad $p(x;\theta)$ depende del valor de un parámetro θ que toma sus valores en un espacio parametrico Ω .Como medida de in - formación sobre el parámetro θ,dado un valor x de la variable aleatoria X asociada al modelo, definimos

$$I(\theta,x) = \log p(x;\theta)$$

Si notamos θ_o el verdadero valor del parámetro calculando la media para todos los posibles valores de X obtenemos para la información sobre el parámetro θ, cuando el verdadero valor es θ_o, la expresión

$$I(\theta,\theta_o)= \int \log p(x;\theta) . p(x;\theta_o)dx$$

Consideremos algunos aspectos de esta función de información.

A) La información sobre el verdadero valor θ_o del parámetro es:

$$I(\theta_o,\theta_o)= \int \log\ p(x;\theta_o) . p(x;\theta_o)dx= I_S(\theta_o)$$

y es ademas el máximo valor que puede alcanzar esta función.

B) La información sobre el verdadero valor θ_o excede de la información sobre otro cualquier valor del parametro θ en:

$$I(\theta_o,\theta_o)-I(\theta,\theta_o) = I_K(\theta_o,\theta)$$

C) Estudiemos ahora la curvatura de la fun -

ción de información en $\theta = \theta_o$ .Suponiendo que se verifican las condiciones de regularidad de Cramer podemos escribir

$$\frac{\partial}{\partial\theta} I(\theta,\theta_o) = \int \frac{\partial}{\partial\theta}\log\ p(x;\theta).p(x;\theta_o)dx$$

de donde

$$\left[\frac{\partial}{\partial\theta} I(\theta,\theta_o)\right]_{\theta=\theta_o} = 0$$

y la curvatura de la función de información en el punto $\theta=\theta_o$, será

$$-\frac{\partial^2}{\partial\theta^2} I(\theta,\theta_o)\Big|_{\theta=\theta_o} = \int -\frac{\partial^2}{\partial\theta^2}\log\ p(x;\theta).p(x;\theta_o)dx\ \Big|_{\theta=\theta_o} =$$

$$= I_F(\theta_o)$$

Resumiendo,diremos que la informacion $I(\theta,\theta_o)$ como funcion de θ tiene su maximo valor $I_S(\theta_o)$ en el verdadero valor θ_o,tiene curvatura $I_F(\theta_o)$ en dicho valor y una discrepancia,con respecto a cualquier otro valor de θ,de $I_K(\theta_o,\theta)$.

De la relacion

$$I_S(\theta_o) - I_K(\theta_o,\theta) = I(\theta,\theta_o)$$

deducimos la siguiente expresión que relaciona las tres cantidades de información más conocidas de la literatura estadistica:

$$-\frac{\partial^2}{\partial\theta^2}\left[I_S(\theta_o) - I_K(\theta_o,\theta)\right]\Big|_{\theta=\theta_o} = I_F(\theta_o)$$

A continuacion expresamos mediante teoremas relaciones particulares entre estas cantidades.

TEOREMA. Las cantidades de informacion de Fisher y Kullabck definidas anteriormente verifican la siguiente relación:

$$I_K(\theta,\theta+\Delta\theta) = 1/2.I_F(\theta) + 0(\Delta\theta^2)$$

TEOREMA. Si el parametro θ puede tomar solo dos valores θ_1 y θ_2 con probabilidades $p(\theta_1)$ y $p(\theta_2)$ y notamos por θ° un valor ficticio del parame - tro para el que

$$p(x)=p(x/\theta^\circ)=p(\theta_1)p(x/\theta_1) + p(\theta_2)p(x/\theta_2)$$

la razón de transmisión de información de Shannon $R(\theta,x)$ y la información de Kullback $I_K(\theta_1,\theta_2)$, verifican la expresión:

$$R(\theta,x)=p(\theta_1)I_K(\theta_1,\theta^\circ) + p(\theta_2)I_K'(\theta_2,\theta^\circ)$$

TEOREMA. En el límite y bajo determinadas condiciones para la función de densidad $p(x/\theta)$, la información de Fisher y la razón de transmisión de Shannon verifican la siguiente relación:

$$R(\theta,x_1,\dots,x_n)= \log \frac{\sqrt{n}}{\sqrt{2\pi e}} - I_S(\theta) + E\left[\log I_F(\theta)\right] + O(1) = \log \frac{\sqrt{n}}{\sqrt{2\pi e}} +$$

$$+ \int p(\theta)\log \frac{I_F^{1/2}(\theta)}{p(\theta)}\, d\theta + O(1)$$

Son dos las diferencias principales entre las informaciones de Shannon y de Fisher:

A) El concepto de Shannon lleva consigo implicito la introducción de una distribución a priori para el parámetro desconocido.Es pues bayesiano, mientras que la definición de Fisher no lo es.

B) La cantidad de información de Shannon cuando se aplica a la construccion de esquemas de muestreo usa la verosimilitud de la muestra obtenida mediante el teorema de Bayes,mientras la de Fisher emplea la distribucion de probabilidad de X para un θ fijado.

Si la primera diferencia da cierta ventaja a la noción de Fisher,una vez que la probabilidad a priori para θ ha sido admitida,el segundo apartado nos indica que la información de Shannon es mas facil de aplicar.

Tienen,sin embargo,las informaciones de Shannon y de Fisher,un curioso rasgo en comun.Ambas consideran el futuro (antes de que ocurra el suceso o experimento) o bien el pasado (cuando el experimento ha sido realizado).La definición de Fisher nos da el rigor con que un parámetro desconocido puede ser definido mediante experimentos.En la información de Shannon uno puede calcular antes o una vez que el mensaje ha sido enviado,la razón a la que la información será transmitida en un código dado.

La información de Shannon en los problemas de muestreo

Un problema que se presenta frecuentemente en la Estadística es el de determinar el valor de un parámetro;sin embargo este valor no puede ser determinado sino a través de lainformacion proporcionada por una serie de observaciones.

La primera pregunta que el estadístico se plantea es saber cuando una serie de observaciones contiene toda la información necesaria para encontrar el verdadero valor del parámetro.

La medida de información que se utiliza normalmente es la de Shannon por ser la que mejor se adapta al punto de vista bayesiano,pues aun en el caso de que no tuvieramos un conocimiento a priori sobre θ,si solo conociéramos el conjunto de de los posibles valores de θ, parece lógico atribuir a θ la distribución a priori, sobre el conjunto de todos los valores admisibles de θ, que tenga mayor entropia, o sea que corresponda a una incertidumbre maximal. Incluso si el valor del parametro es desconocido para nosotros, pensamos que es lógico atribuir a θ una distribucion a priori si ésta es necesaria para comparar diferentes experimentos o diferentes estadisticos y escoger cual es el mejor para nuestros propositos.

Los trabajos de Renyi (1964, 1967, 1968) y de Vajda (1967, 1968) resuelven satisfactoriamente la pregunta anteriormente planteada dando condiciones necesarias para que la cantidad de informacion residual, diferencia entre la incertidumbre del espacio paramétrico y la información que la muestra nos aporta sobre éste converja a cero. Estos trabajos vienen a corroborar el hecho de que la razón de transmisión de Shannon $R(\theta,x)$ es una función creciente con n, que evidentemente respeta la acotacion de esta cantidad de informacion por la entropia del espacio parametrico, pues

$$H(\theta) \geq R(\theta, x_i) \quad \text{para cualquier } i$$

y ademas

$$R(\theta,x_1x_2) \geq R(\theta,x_1)$$

y en general

$$R(\theta,x_1,\ldots,x_n) \geq R(\theta,x_1,\ldots,x_{n-1})$$

El segundo problema que se plantea y que es el interesante en la práctica,consiste en encontrar un diseño óptimo de muestreo,es decir, un criterio de detención de un muestreo secuencial. Ideamos el siguiente:

Dado un nivel "a" fijado de antemano hacemos una observacion X_1 y calculamos $H(\theta/X_1)$.

Si se verifica

$$H(\theta/X_1) > a$$

hacemos una nueva observacion X_2 y calculamos $H(\theta/X_1X_2)$.Si se verifica

$$H(\theta/X_1X_2) > a$$

hacemos una nueva observacion X_3 y calculamos $H(\theta/X_1X_2X_3)$.

Si se verifica

$$H(\theta/X_1X_2X_3) > a$$

se continua el proceso.

El proceso se detiene cuando para algun valor de n se verifica

$$H(\theta/X_1X_2X_3\ldots X_n) \leq a$$

es decir, cuando

$$\sum p(\theta/x_1,\ldots,x_n)\log p(\theta/x_1,\ldots,x_n) \leq a$$

En el caso de que el espacio paramétrico conste de solo dos elementos, esta regla de muestreo secuencial coincide con el test de razón de verosimilitud de Wald.

Otro procedimiento para obtener un diseño óptimo de muestreo podría consistir en considerar una función de probabilidad de error asociada con las decisiones aconsejadas por la información recibida. Limitando esta probabilidad de error llegariamos a obtener un criterio de detencion de un muestreo secuencial.

Generalización de la entropía de Renyi.

En 1958 Fadeev da una de las más clásicas caracterizaciones axiomáticas de la entropía de Shannon. Estudiando dicha axiomática Renyi comprueba que cantidades del tipo

$$H_\alpha(P) = \frac{1}{1-\alpha}\log_2\sum p_i^\alpha$$

con $0<\alpha<1$, verifican los postulados A, B y C de Fadeev y el teorema de adición de la entropía . Ello le induce a definir $H_\alpha(P)$ como la entropía de orden α de una distribución P, sin embargo, no se verifica el axioma D de Fadeev, siendo necesario imponer algun postulado más. Por ello Renyi introduce el concépto de distribución de probabilidad generalizada y define la entropía de orden α de una distribucion generalizada P, $H_\alpha(P)$ en la forma

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \frac{\sum p_i^\alpha}{\sum p_i}$$

que en el caso particular en que α tiende a 1 coincide con la entropía de Shannon.

Si en lugar de restringir el intervalo de variación de α al $(0,1)$ con lo cual limitamos los valores de α a valores fraccionarios y menores que 1, suponemos para α un intervalo de variación $(0,N)$ con $N\geq 1$, obtenemos una generalizacion de la entropía de Renyi de orden α.

DEFINICION. Sea $P=(p_1,\ldots,p_n)$ una distribución de probabilidades generalizada con peso $w(P)= 1$. Definimos entropía de orden α (α,N) de la dis - tribución generalizada P, y la notaremos $H_\alpha^N(P)$

por la expresion

$$H_{\alpha}^{N}(P) = \frac{N}{N-\alpha}\log_2\frac{\sum p_i^{\alpha}}{\sum p_i}$$

siendo $0<\alpha<N$, $N\geq 1$ y $w(P) = \sum p_i \leq 1$

Notemos que segun nuestra definición la entropía de Renyi de orden α no es otra cosa que la entropia de orden $(\alpha,1)$.

Cuando $\alpha \rightarrow N$ la entropia de orden (α,N) coincide con la de Shannon.

CARACTERIZACION. La entropia de orden (α,N) puede ser caracterizada univocamente por los postulados siguientes:

Postulado 1: $H(P)$ es una funcion simétrica de los elementos de P.

Postulado 2: Si notamos por $\{p\}$ la distribución de probabilidad generalizada que consta de una única probabilidad p, entonces $H[\{p\}]$ es una función contínua de p en el intervalo $0<p\leq 1$.

Postulado 3: $H(1/2,1/2) = 1$

Postulado 4: $H(PQ) = H(P) + H(Q)$

donde P y Q son dos distribuciones de probabilidad generalizadas con pesos $w(P)$ y $w(Q)$ tales que

$w(P) + w(Q) = 1$

Postulado 5: Existe una función estrictamente monótona y contínua $y=g(x)$ tal que si $w(P) + w(Q) =1$, tenemos

$$H\{P\cup Q\}=g^{-1}\left\{\frac{w(P)\,gH(P)+w(Q)\,g[H(Q)]}{w(P)+w(Q)}\right\}$$

eligiendo como función de Kolmogorov-Naguno $g(x)$ una función exponencial del tipo

$$g(x)= 2^{(\alpha/N-1)x} \quad \text{con } 0<\frac{\alpha}{N}<1$$

Exponemos a continuación, en forma de teoremas propiedades encontradas para la función de entropía de orden (α,N).

TEOREMA. La entropía de orden (α,N) $H_\alpha^N(P)$ es una función monótona creciente de N, para cada α fijo

Corolario. Para cualquier distribucion de probabilidad generalizada se verifica

$$H_\alpha^N(P) = H_\alpha(P)$$

dándose la igualdad si y solo si $N=1$.

TEOREMA. La entropía de orden (α,N) $H_\alpha^N(P)$ es una función monótona decreciente de α, para cada N fijo.

TEOREMA. Sea $P = (p_1,\dots,p_n)$ una distribución de probabilidad generalizada con peso $w(P) = u \leq 1$. La entropía $H_\alpha^N(P)$ verifica la siguiente desigualdad

$$0 \leq H_\alpha^N(P) \leq \log_2(n/u)$$

TEOREMA. Si existe una distribucion $P=(p_1,\dots p_n)$ siendo $p_1=p_2=\dots=p_n$, la entropia $H_\alpha^N(P)$ alcanza el máximo $\log_2(n/u)$ independientemente de los valores de α y N, siendo $w(P) = u$.

Es interesante notar que en las condiciones de este último teorema, las entropías de Shannon, Renyi de orden α y la de orden (α,N) coinciden.

Razón de transmisión de información generalizada

Supongamos un canal de comunicación con una fuente discreta que emite simbolos x_i con probabilidades $p(x_i)$ y sea $p(y_j/x_i)$ la probabilidad de transición de recibir y_j cuando se ha emitido x_i.

DEFINICION. Si en lugar de emplear la entropía de Shannon utilizamos la entropia de orden (α,N)

podemos generalizar la razón de transmision de Shannon $R(x,y)$ y definir razón de transmisión generalizada para un canal con fuente discreta mediante las expresiones

$$R_{\alpha}^{N}(x,y) = \frac{N}{\alpha - N} \log_2 \sum_i \sum_j \frac{P^{\alpha/N}(x_i,y_j)}{P^{\alpha/N-1}(x_i)\, P^{\alpha/N-1}(y_j)}$$

$$\bar{R}_{\alpha}^{N}(x,y) = H_{\alpha}^{N}(X) + H_{\alpha}^{N}(Y) - H_{\alpha}^{N}(X,Y)$$

$$\bar{\bar{R}}_{\alpha}^{N}(x,y) = H_{\alpha}^{N}(X) - H_{\alpha}^{N}(X/Y)$$

$$H_{\alpha}^{N}(X) = \frac{N}{N-\alpha} \log_2 \sum_i P^{\alpha/N}(x_i)$$

$$H_{\alpha}^{N}(X,Y) = \frac{N}{N-\alpha} \log_2 \sum_i \sum_j P^{\alpha/N}(x_i,y_j)$$

$$H_{\alpha}^{N}(X/Y) = \frac{N}{N-\alpha} \log_2 \left\{ \sum_j P(y_j) \sum_i P^{\alpha/N}(x_i/y_j) \right\}$$

y por tanto podemos escribir

$$\bar{R}_{\alpha}^{N}(\bar{x},\bar{y}) = \frac{N}{N-\alpha} \log_2 \frac{\sum_i \sum_j P^{\alpha/N}(x_i) \cdot P^{\alpha/N}(y_j)}{\sum_i \sum_j P^{\alpha/N}(x_i,y_j)}$$

$$\bar{\bar{R}}_{\alpha}^{N}(x,y) = \frac{N}{N-\alpha} \log_2 \frac{\sum_i P^{\alpha/N}(x_i)}{\sum_i \sum_j P^{1-\alpha/N}(y_j)\, P^{\alpha/N}(x_i,y_j)}$$

Enunciamos a continuación, en forma de teoremas, propiedades encontradas para estas funciones

TEOREMA. Para cada N fijo, $R_{\alpha}^{N}(x,y)$ es una función monótona creciente de α.

TEOREMA. Para cada α fijo, $R_{\alpha}^{N}(x,y)$ es una fun -

ción monótona decreciente de N.

TEOREMA. La razón de transmisión generalizada es siempre no negativa.

TEOREMA. La razón de transmisión de información generalizada $\bar{\bar{R}}_{\alpha}^{N}(x,y)$ es siempre no negativa.

TEOREMA. Si las distribuciones de entrada $p(x_i)$ y de salida $p(y_j)$ son independientes, entonces

$$R_{\alpha}^{N}(x,y) = \bar{R}_{\alpha}^{N}(x,y) = \bar{\bar{R}}_{\alpha}^{N}(x,y) = 0$$

Razón de distorsión de una fuente discreta.

Consideremos un canal de comunicación con una fuente discreta que emite símbolos x_i con probabilidad $p(x_i)$, y sean $p(y_j/x_i)$ las probabilidades de transición de recibir un símbolo y_j cuando se ha emitido x_i. Si notamos por d_{ij} un valor numérico asignado a la distorsión cuando se ha emitido x_i y recibido y_j, siendo $d(x_i,x_i) = 0$ para todo i y $d_{ij} > 0$ para $i \neq j$, definimos "medida de distorsión de orden t" por la expresión

$$D_t(x,y) = \frac{1}{t} \log_2 \sum_i \sum_j p(x_i) p(y_j/x_i) 2^{t d_{ij}}$$

cuando t tiende a cero coincide con la medida de distorsion de Shannon.

TEOREMA. Existen valores t_1 y t_2 de t tales que D_t es convexa para $t < t_1$ y cóncava para $t > t_2$.

DEFINICION. Supongamos una fuente discreta con medida de distorsión D_t y sea $\bar{\bar{R}}_\alpha^N(x,y)$ la razón de transmisión de información del canal.Se define <u>función de razón de distorsión relativa a la medida de distorsión D_t</u>, y la notaremos $\bar{\bar{R}}_\alpha^N(D_t^*)$,mediante la expresión

$$\bar{\bar{R}}_\alpha^N(D^*) = \min_{p(y_j/x_i)} \bar{\bar{R}}^N(x,y)$$

sujeto a la restricción

$$\frac{1}{t}\log_2 \sum_i \sum_j p^{\alpha/N}(x_i)\ p^{\alpha/N}(y_j/x_i).2^{td_{ij}} \leq D^*$$

Propiedades:

A). Por la continuidad de $\bar{\bar{R}}_\alpha^N(x,y)$ con respecto a $p(y_j/x_i)$ en la región de variación de las $p(y_j/x_i)$ que es cerrada,$R_\alpha^N(x,y)$ alcanza su mínimo, $R_\alpha^N(D^*)$.

B) De la propia definición se deduce que $R_\alpha^N(D)$ es una función monótona decreciente de D.

Sería interesante demostrar la convexidad de esta función de razon de distorsion para le medida D_t y extender las acotaciones dadas por Berger(1968),Pinkston(1969) y Wyner-Ziv(1971) para la función de razón de distorsión de Shannon a la función aquí definida.

Medida de la información proporcionada por un experimento.

Consideremos una variable aleatoria X.A la observacion de esta variable se le llama realización de un experimento con dicha variable aleatoria. Sea S el espacio formado por todos los resultado de las posibles observaciones de X, que llamaremos espacio de resultados. Sea $\mathcal{A}$ el σ-álgebra definido sobre S. Sea Ω un espacio paramétrico con elementos $\theta \in \Omega$. Para cada $\theta \in \Omega$ definimos una medida de probabilidad sobre el espacio medible $(S, \mathcal{A})$. Supongamos que estas medidas de probabilidad son absolutamente contínuas con respecto a una medida dominante μ sobre $\mathcal{A}$. Podemos describir cada medida de probabilidad por una función de densidad $p(x/\theta)$, de manera que para un subconjunto $A \in \mathcal{A}$

$$P(A) = \int_A p(x/\theta)\, d\mu(x)$$

y por sencillez de notacion, tomando la medida de Lebesgue, escribiremos

$$P(A) = \int_A p(x/\theta)dx$$

La dupla

$$E=\left\{(S, \mathcal{A}) \;,\; [p(x/\theta);\; \theta \in \Omega]\right\}$$

caracteriza un experimento.

Supongamos que existe una distribución a priori sobre Ω, y supondremos de nuevo que puede ser descrita por una función de densidad $p(\theta)$, con respecto a una medida dominante y que por sencillez notamos $d\theta$. Esta $p(\theta)$ resumirá las opiniones iniciales del experimentador sobre el valor del parámetro, antes de realizar el experimento.

Existe, por tanto, una incertidumbre inicial sobre el valor de θ, basándonos en el concepto de incertidumbre de Shannon, que podemos expresar como

$$I(\theta) = -\int_{\Omega} p(\theta)\log\, p(\theta)d\theta$$

Una vez realizado el experimento E y obtenido un resultado x, el experimentador tendrá una incertidumbre sobre el valor de θ que vendrá dada por la expresión

$$I(\theta/x) = -\int_{\Omega} p(\theta/x)\log p(\theta/x)d\theta$$

Es lógico definir la información proporcionada sobre θ por el resultado x obtenido al realizar el experimento por

$$I(\theta) - I(\theta/x)$$

que es precisamente la incertidumbre que sobre θ ha permitido el experimento E eliminar al experimentador.

Evidentemente, antes de realizar E, la información que éste puede proporcionar vendrá dada por el valor medio extendido a todos los posibles resultados de E, de la información proporcionada por uno de ellos. Por tanto podemos definir:

DEFINICION. La incertidumbre del experimentador sobre el valor del parámetro θ antes de realizar un experimento E, o lo que es equivalente, la información obtenida una vez realizado E, suponiendo que existe una distribución a priori $p(\theta)$ sobre el parámetro, viene dada por

$$I(\theta \| X) = E_x\left\{I(\theta) - I(\theta/_x)\right\}$$

o bien

$$I(\theta \| X) = \int_S \int_\Omega p(\theta,x) \log \frac{p(\theta,x)}{p(\theta)\,p(x)}\, d\theta dx$$

RELACIONES:

A) <u>Información de Shannon</u>. Podemos establecer una correspondencia o analogía entre la noción de experimento dada anteriormente y la de un sistema de comunicación con ruido. En efecto, consideremos como fuente el conjunto Ω de símbolos θ. Estos símbolos son emitidos por la fuente de manera que la elección de θ es hecha con una probabilidad $p(\theta)$, siendo independientes las sucesivas elecciones. Se supone que el ruido del canal es tal que los símbolos son perturbados segun $p(x/\theta)$ obteniéndose en el receptor el resultado $x \in S$, dependiente precisamente de dichas probabilidades. Podemos pues decir que el canal viene descrito por unas probabilidades de transición $p(x/\theta)$ que nos indican la probabilidad de obtener un resultado x cuando sea emitido el símbolo θ de Ω.

La razón de transmisión de un canal con ruido vendrá dada por

$$R(\theta,x)=\int_S\int_\Omega p(\theta,x)\ \log \frac{p(\theta,x)}{p(\theta)\ p(x)}d\theta dx$$

La razón de transmisión de informacion de Shannon coincide pues con la información que sobre θ nos aporta el experimento E.

B) <u>Información de Kullback.</u> Si tomamos como hipótesis H_1 que las variables θ y x son dependientes con distribución de probabilidad conjunta $p(\theta,x)$ y como hipótesis H_2 que las variables θ y x son independientes con distribuciones de probabilidad marginales $p(\theta)$ y $p(x)$ respectiva - mente, entonces la información de Kullback para discriminar en favor de H_1 contra H_2 vendrá dada por

$$I(1:2) = \int_S \int_\Omega p(\theta,x) \log \frac{p(\theta,x)}{p(\theta)p(x)} d\theta dx$$

expresión que coincide con la medida de información anterior.

PROPIEDADES.

A) La cantidad de información $I(\theta \| X)$ es no negativa

B) La cantidad $I(\theta \| X)$ es "generalmente finita".

C) $I(\theta \| X)$ permanece invariante bajo todas las transformaciones lineales escalares en θ y x.

D) Si g es una función medible definida sobre el σ-álgebra $\mathcal{A}$ de subconjuntos de S, se verifica que

$$I(\theta \| X) \geq I\left[\theta \| g(X)\right]$$

dandose la igualdad si y solo si $g(X)$ es un estadístico suficiente.

Comparación de experimentos.

DEFINICION. Dados dos experimentos E_1 y E_2 con el mismo espacio paramétrico caracterizados por las duplas

$$E_i = \left\{ (S_i, \mathcal{A}_i);\ [p_i(x_i/\theta);\theta\in\Omega] \right\} \qquad i=1,2$$

diremos que E_1 es no menos informativo que E_2 y lo escribiremos $E_1 \succsim E_2$ cuando

$$I(\theta \| X_1) \geq I(\theta \| X_2)$$

para todas las posibles distribuciones a priori $p(\theta)$ definidas sobre Ω.

DEFINICION. Dados dos experimentos E_1 y E_2 con el mismo espacio paramétrico Ω diremos que E_1 es tan informativo como E_2, y lo notaremos $E_1 \sim E_2$ cuando $E_1 \succsim E_2$ y $E_2 \succsim E_1$.

PROPIEDAD. La relación anteriormente definida es una relación de orden parcial.

DEFINICION. Dados dos experimentos E_1 y E_2 con el mismo espacio paramétrico Ω, diremos que E_1 es más informativo que E_2 y lo notaremos $E_1 \succ E_2$ cuando ocurre $E_1 \succsim E_2$ pero no ocurre que $E_1 \sim E_2$.

TEOREMA. Sean E_1, E_2 y E_3 tres experimentos con el mismo espacio paramétrico Ω. Si E_3 es independiente de E_1 y E_2, entonces

$$E_1 \succsim E_2 \Longrightarrow (E_1, E_3) \succsim (E_2, E_3)$$

DEFINICION. Dados dos experimentos E_1 y E_2 con el mismo espacio paramétrico Ω diremos que E_1 es <u>fuertemente no menos informativo</u> que E_2 y lo notaremos $E_1 \overset{f}{\succsim} E_2$, cuando

$$\int_S p_1(x/\theta_o) \log \frac{p_1(x/\theta_o)}{p(\theta) p_1(x/\theta) d\theta} dx \geq$$

$$\geq \int_S p_2(x/\theta_o) \log \frac{p_2(x/\theta_o)}{p(\theta) p_2(x/\theta) d\theta} dx$$

para todo $\theta_o \in \Omega$ y $p(\theta)$.

Notemos que si tomamos esperanza matemática en θ en la desigualdad anterior obtenemos que $I(\theta \| X_1) \geq I(\theta \| X_2)$, es decir, que E_1 es no menos informativo que E_2, de ahí que hayamos utilizado en esta nueva definición el calificativo "fuertemente", pues ahora exigimos que la desigualdad se verifique para cada uno de los $\theta \in \Omega$ y no solo para la media.

Otro método para comparar experimentos se basa en el concepto de suficiencia entre experimentos debido a Blackwell (1.951,1.953).Sean E_1 y E_2 dos experimentos con el mismo espacio paramétrico Ω ,se dice que el experimento E_1 es suficiente para E_2,si existe una función no negativa f sobre el espacio producto S_1xS_2 para la cual se satisfacen las siguientes relaciones:

$$p_2(x_2/\theta)=\int_{S_1} f(x_2,x_1)p_1(x_1/\theta)dx_1 \quad \text{para } \theta\in\Omega \quad x_2\in S_2 \qquad (1)$$

$$\int_{S_2} f(x_2,x_1)dx_2=1 \qquad \text{para} \quad x_1\in S_1 \qquad (2)$$

$$0<\int_{S_1} f(x_2,x_1)dx_1<\infty \quad \text{para } x_2\in S_2 \qquad (3)$$

Una función no negativa f que satisface (2) es llamada transformacion estocástica de X_1 a X_2

Para cada resultado $x_1\in S_1$ fijado,la función $f(\cdot,x_1)$ es una función de densidad sobre S_2. Ya que esta función no depende del parámetro θ, un resultado $x_2\in S_2$ puede ser generado de acuerdo con esta función de densidad por medio de una aleatorizacion auxiliar.Por tanto la ecuación (1) nos dice que E_1 es suficiente para E_2 si haciendo caso omiso del valor del parámetro θ,una observación de E_1 y una auxiliar aleatorizacion ha ce posible generar una variable aleatoria que tiene la misma distribución que X_2.La ecuacion (3) no es otra cosa que una condición de integrabilidad de f sobre S_1.

Intuitivamente esta claro,que si E_1 es suficiente para E_2,el estadístico nunca deberá realizar el experimento E_2 cuando sea posible realizar E_1,pues realizar E_2 es equivalente a realizar E_1 y someter el resultado a una transformación aleatoria que solo puede oscurecer cualquier información sobre el valor del parámetro que podría estar contenida en este resultado de E_1. El próximo teorema formaliza la afirmación anterior de que E_1 es no menos informativo que E_2.

TEOREMA. Sean E_1 y E_2 dos experimentos con el mismo espacio paramétrico Ω ;si E_1 es suficiente para E_2,entonces E_1 es no menos informativo que E_2.

TEOREMA. Sean E_1 y E_2 dos experimentos con el mismo espacio paramétrico Ω .Si E_1 es suficiente para E_2,entonces E_1 es fuertemente no menos informativo que E_2.

Capacidad de experimentos.

DEFINICION. Dado un experimento E se define capacidad de dicho experimento y lo notaremos $C(X)$, a la máxima información que el experimento nos proporciona sobre Ω para todas las posibles distribuciones a priori $p(\theta)$ sobre Ω. Es decir,

$$C(X) = \max_{p(\theta)\in P} I(\theta \| X)$$

TEOREMA. Sean E_1 y E_2 dos experimentos con el mismo espacio paramétrico Ω. Si E_1 es suficiente para E_2, entonces

$$C(X_1) \geq C(X_2)$$

TEOREMA. Consideremos un experimento E con espacio paramétrico finito $\Omega=\{\theta_1,\ldots,\theta_n\}$, caracterizado por la cupla

$$E=\left\{ (S,\mathcal{A}); \left[p(x/\theta_1),\ldots,p(x/\theta_k)\ /\ \theta_i\in\Omega\right]\right\}$$

y notemos $p(x/\theta_i) = q_i(x)$ para todo i tal que $1 \leq i \leq k$.

TEOREMA. Si existe un vector $p'\in P^k$ con todas sus componentes estrictamente positivas y tal que

$$\int_S q_i(x) \log \frac{q_i(x)}{\sum p_i' q_i(x)} dx$$

es independiente de i, entonces p' maximiza la información y además

$$C(X) = I(\theta \| X)_{p'} = \int_S q_i(x) \log \frac{q_i(x)}{p_i' q_i(x)} \, dx$$

TEOREMA. Consideremos un experimento E con espacio paramétrico $\Omega = \{\theta_1, \theta_2\}$, caracterizado por la cupla

$$E = \{ (S, \mathcal{A}) \; ; [q_1(x), q_2(x) \; / \; \theta_1, \theta_2 \in \Omega] \}$$

La capacidad de E está acotada por $1/4 J_k(\theta_1, \theta_2)$, siendo $J_k(\theta_1,\theta_2) = I_k(\theta_1,\theta_2) - I_k(\theta_2,\theta_1)$ la divergencia de Kullback.

Bibliografía

ABRAMSON (1.963)."Information Theory and Coding" Ed. Mc Graw Hill (Editado en castellano por Ed. Paraninfo en 1.966)

ASH (1.967)."Information Theory" Ed. Wiley Interscience.

BARNARD (1.951)."The Theory of Information" J. Roy.Statist.So.Serie B,13 (46-64)

BERGER (1.968)."Rate distortion theory for sources with abstract alphabets and memory". Infor. and Control 13 (254-273).

BLACKWEL (1.951)."Comparisos of experiments".Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability,ed. Jerzy Neyman,Berkeley University of California Press.

BLACKWEL (1.953)."Equivalent comparisons of experiments".Ann. Math.Stat. 24 (265-272).

DOBRUSHIN (1.959)."General formulation of Shannon's Basic Theorems of the Theory of Information". USP. MATH.NAUK vol.14 nº 6 (3-104)

FANO (1.961). "Transmission of Information".Ed. MIT Press.

FEINSTEIN (1.958)."Foundations of Information Theory"Ed. Mc Graw Hill.

FERON-FOURGEAUD (1.951)."Information et regresion".C.R. Acad. Sci. Paris 232,(1636 - 1638).

FISHER (1.925)."The Theory of Statistical Estimation". Proc. Cambridge Philo.Soc.22 (700 -725).

IBRAGIMOV (1.962)."Some limiting theorems for stationary processes".Teor.Verojatnosr. i Primenen. 7 (361-392)

IBRAGIMOV-HASMISNKII. "Asymptotic behavior of generalized Bayes estimates".Dok. Akad.Nak. (1.970) nº 2.

KHINCHIN (1.957)."Mathematical foundation of Information Theory" Ed. Dover

KOLMOGOROV (1.956)."On the Shannon Theory of Information in the Case of Continuous signals". IEEE Trans.Information Theory (102-108).

KULLBACK (1.959)."Information Theory and Statistics".Ed J. Wiley.

LINDLEY (1.956). "A measure of the information provided by an experiment".Ann. Math.Sta. 27 (986-1005)

PINKSTON (1.969)."Rate distortion theory to a converse to the coding theorems" IEEE Trans Infor, Theory 15 (66-71)

RAIFFA-SCHLAIFER (1.961)."Applied Statistical Decision Theory".Division of research Harward business school.

RENYI (1.964)."On the amount of information concerning an uknown parameter in a sequence of observations".Publ.Math.Inst.Hung Acad. Scie. 9 (617-624)

RENYI (1.965)."On the foundations of Information Theory" Rev. of the Inst. Sta. 33(1-14)

RENYI (1.967)."Statistics and Information Theory" Studia Sci. Math.Hung. 2 (249-256).

RENYI (1.967)."On some basic problems of Statistics from the point of view of Information Theory".Proceedings of the 5th Ber keley Symposium.Vol I (531-543)

RENYI (1.968)."On some problems of Statistics from the point of view of Information Theory".Colloquium of Information Theory Debrecen (343-357)

REZA (1.961)."An Introduction to Information Theory" Ed. Mac Graw Hill.

SHANNON (1.948)."A mathematical Theory of Communication" Bell. System Technical Journal 27 (379-656).Reeditado por University of Illinois Press en 1.969 (Shannon-Weaver, "A Mathematical Theory of Communication").

SHANNON (1.959)."Codings theorems for a discrete source with a fidelity criterion" IRE Nat Conv. Rec. 4 (142-163),presentado al Symposium Purdue University on Information and Decission Processes" Ed. Machol (93-126), 1.960.

TRIBUS (1.972)."Decisions rationelles dans l'incertain".Ed. Masson et Cie.

VAJDA (1.967)."Rate of convergence of the Information in a sample concerning a parameter"

Czechoslovak Mathematical Journal 17 (225-230).

VAJDA (1.968)."On the convergence of information in a sequence of observations".Colloquium of Information Theory.Debrecen. (489-501)

WYNER-ZIV (1.971)."Bounds on the Rate-Distortion function for stationary sources with memory".IEEE Trans. Inf. Theory 17 (508-513)

FUNDACION JUAN MARCH

SERIE UNIVERSITARIA

Titulos Publicados:

15.– *Evaluación de Proyectos de Inversión en una Empresa de producción y distribución de Energía Eléctrica. Felipe Ruíz López (Ingeniería. Extranjero, 1974)*

16.– *El significado teórico de los términos descriptivos./Carlos Solís Santos (Filosofía. España, 1973)*

17.– *Encaje de los modelos econométricos en el enfoque objetivos-instrumentos relativos de política económica./Gumersindo Ruíz Bravo (Sociología. España, 1971)*

18.– *La imaginación natural (estudio sobre la literatura fantástica norteamericana)./Pedro García Montalvo (Literatura y Filología. Extranjero, 1974)*

19.– *Estudio sobre la hormona Natriurética./Andrés Purroy Unanua (Medicina, Farmacia y Veterinaria. Extranjero, 1973)*

20.– *Análisis farmacológico de las acciones miocárdicas de bloqueantes Beta–Adrenérgicos./José Salvador Serrano Molina (Medicina, Farmacia y Veterinaria. España, 1970)*

21.– *El hombre y el diseño industrial./Miguel Durán–Lóriga (Artes Plásticas. España, 1974)*