UNIVERSIDAD DE SEVILLA
FACULTAD DE CIENCIAS

" ALGUNAS CUESTIONES SOBRE TEORIA DE LA INFORMACION "

ANTONIO PASCUAL ACOSTA

Memoria para optar al grado de Doctor en Ciencias Matematicas realizada bajo la direccion del Prof.Dr.D.RAFAEL INFANTE MACIAS.

Sevilla, 1975

UNIVERSIDAD DE SEVILLA

FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADISTICA
E INVESTIGACION OPERATIVA

"ALGUNAS CUESTIONES SOBRE TEORIA DE LA INFORMACION"

Visado en Sevilla a
de Febrero de 1976
EL DIRECTOR DE LA MEMORIA

Firmado: Prof.Dr. D.
Rafael Infante Macias

MEMORIA que, para optar al
grado de Doctor, presenta e
Licenciado en Ciencias Mate
Maticas D.Antonio Pascual A
ceeta.

Sevilla, Febrero 1976

Fdo: Antonio Pascual Acosta

ALGUNAS CUESTIONES SOBRE
TEORIA DE LA INFORMACION

Antonio Pascual Acosta

Quiero expresar mi profundo agradecimiento al Profesor Dr. D. Rafael Infante Macias, director de esta memoria por su constante estimulo y ayuda en la preparacion de la misma.

Asi mismo, quiero hacer llegar mi reconocimiento a todos aquellos que de un modo u otro han contribuido a la realización de este trabajo, al Centro de Calculo de la Facultad de Ciencias y en especial al Departamento de Estadistica e Investigacion Operativa

Sevilla, Febrero de 1976

INTRODUCCION.

INTRODUCCION

El famoso documento de SHANNON "The Mathematical Theory of Commuication" 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 Información, 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 cuestion. El conocimiento completo acerca de una cuestion es la posibilidad de asignar una probabilidad cero (p = 0) a todas las respuestas posibles excepto a una. A una persona que correctamente asigne una probabilidad unidad (p = 1) 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 la incertidumbre de Shannon, a la que él 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 informacion consistió en demostrar la existencia de una medida de informacion que es independiente de los medios empleados para generar la informacion. El contenido informativo de un mensaje es, por tanto, invariante respecto a la forma y no depende de si el mensaje se envia mediante codigo Morse, mediante la impresion de una determinada forma sobre una onda portadora o mediante alguna forma de criptografia.

Para Shannon el contenido informativo de un mensaje es una medida del cambio en el conocimiento del receptor (del conocimiento X antes del mensaje al conocimiento X' despues del mensaje). Un mensaje que dice lo que ya se sabe no produce cambio alguno ni en cuanto al conocimiento (X sigue siendo igual) ni en cuanto a la asignacion de probabilidades, por lo que no transmite ninguna informacion.

La medida de Shannon fue concebida para resolver un problema concreto: Cómo obtener una medida util de lo que se transmite

por un canal de comunicaciones. Ha demostrado tambien ser la única funcion capaz de satisfacer ciertas necesidades Basicas de la teoria de la informacion. Apenas ha pasado un cuarto de siglo de la publicacion de Shannon y ya se han escrito miles de trabajos al respecto, sin que ninguno de ellos haya encontrado una funcion sustitutiva, ni siquiera la necesidad de ella. Por el contrario, se han descubierto muchos otros caminos para llegar a dicha medida.

La medida de entropia de Shannon es fundamental en la teoria de la informacion y a ella y a su aplicacion a diferentes problemas estadisticos dedicamos el primer capitulo de esta Memoria.

La mas clasica de las caracterizaciones axiomaticas de la entropia de Shannon fué dada por FADDEEV en 1.958, como sigue:

Si notamos por P^n el conjunto de todas las distribuciones de probabilidad discretas $P = (p_1 \ldots p_n)$ con $p_i \geq 0$ $(i = 1,\ldots n)$ y $\sum_{i=1}^{n} p_i = 1$ y consideramos una funcion $H_m(P)$ definida para todo $P \in P^m$ que satisface las siguientes condiciones:

A) $h(p) = H_2(p, 1-p)$ es continua para $0 \leq p \leq 1$

B) $H_2(1/2, 1/2) = 1$

C) $H_n(p_1, p_2, \ldots, p_n)$ es una funcion simetrica de sus argumentos

D) Para cualquier λ $/0 \leq \lambda \leq 1$:

$$H_{m+1}[p_1, \ldots p_{m-1}, \lambda p_m, (1-\lambda)p_n] = H_m(p_1 p_2 \ldots p_n) + p_m H_2(\lambda, 1-\lambda)$$

entonces

$$H_m(p_1, p_2, \ldots p_n) = \sum_{i=1}^{m} p_i \log_2 \frac{1}{p_i}$$

Esta axiomatica simplifica los conjuntos de axiomas dados originalmente por Shannon y posteriormente por Khinchin.

En 1.964, KENDALL, usando metodos completamente diferentes, demuestra que estas cuatro propiedades:

A') $h(p) = H_2(p, 1-p)$ es no decreciente para $p \leq 1/2$.

B') $H_2(1/2, 1/2) = 1$

C') $H_2(p_1, p_2)$ y $H_3(p_1, p_2, p_3)$ son funciones reales simetricas con $p_1, p_2, p_3 > 0$.

D') Para cualquier λ / $0 \leq \lambda \leq 1$

$$H_3(\lambda p_1, (1-\lambda)p_1, p_2) = H_2(p_1, p_2) + p_1 H_2(\lambda, 1-\lambda) \quad (\text{con } p_1, p_2 > 0$$

definen univocamente H_2 y H_3 como sigue:

$$H_2 = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

y

$$H_3 = -\sum_{i=1}^{3} p_i \log_2 p_i$$

Continuando las investigaciones de Kendall, LEE llega, en 1.964, a la siguiente caracterizacion de la entropia de Shannon:

Dada una funcion $H_n(P)$ con las siguientes propiedades:

A'') $h(p) = H_2(p, 1-p)$ es una funcion real, finita medible Lebesgue para $p \in (0,1)$.

B'') $H_2(1/2, 1/2) = 1$

C'') Analogo a C'.

D'') Analogo a D'.

Entonces H (P) esta univocamente determinado para todo n y viene dada por la expresion

$$H_m(p_1 p_2 \ldots p_m) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$$

TVEHERG demuestra la unicidad de la funcion de entropia bajo la hipotesis de la integrabilidad segun Lebesgue de la funcion $h(p) = H_2(p, 1-p)$.

En 1.962 INGARDEM y URBANIK dan una definicion axiomatica para la funcion de informacion sin usar la nocion de probabilidad. Definen la informacion como una funcion real sobre un conjunto de un campo de Borel. Llegan a demostrar que la informacion puede ser expresada por la formula de Shannon por medio de una medida de probabilidad condicionada, definida univocamente, la existencia de la cual se sigue de la existencia de la informacion.

Esta caracterizacion axiomatica que en principio parece desligar las nociones de informacion y probabilidad, demuestran que el orden logico usual de definir la informacion por medio de probabilidades, puede ser invertido, introduciendo primero la nocion de informacion sin probabilidades, las cuales aparecen inmediatamente como consecuencia inevitable. Las nociones de informacion y probabilidad por tanto, no pueden ser separadas.

RAJSKI, investiga las propiedades metricas de la funcion de entropia H (X) y de la informacion mutua I(X,Y) para esquemas probabilisticos discretos. Observa que la funcion

$$d(X,Y) = 1 - \frac{I(X,Y)}{H(X,Y)}$$

define una distancia en $\{X\}$ para todas las posibles distribuciones discretas; mientras que

$$\frac{I(X,Y)}{H(X,Y)} = \sqrt{1 - d^2(X,Y)}$$

puede ser usada como una medida de dependencia entre las variables aleatorias definidas por los esquemas X y Y.

RENYI, en 1.960 propone otra forma de funcion de entropia pa-

ra distribuciones de probabilidad generalizadas que son definidas como una sucesion $P = (p_1 \ldots p_n)$ de numeros no negativos tales que $0 < \sum p_i \leq 1$.

La entropia de Renyi de orden α viene dada por

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

Años mas tarde DAROCZY dá una caracterizacion axiomatica comun para las entropias de Renyi y de Shannon.

Una de las primeras generalizaciones de la funcion de entropia al caso continuo y en general para variables aleatorias arbitrarias es introducida por KOLMOGOROV:

Dadas dos variables aleatorias X e Y se define

$$H(X) = \inf_{P_{X,Y}} I(X,Y) = \inf \int_x \int_y P_{X,Y}(x,y) \log \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)} dx dy$$

donde $I(X,Y)$ es llamada informacion mutua entre las variables X e Y y el inferior es tomado para todas las posibles distribuciones de probabilidad conjuntas del par (X,Y).

Junto con Gelfand y Yaglom define los conceptos de entropia diferencial y ε-entropia.

Mas tarde ROSENBLATH y ROTH dan una relacion entre la entropia diferencial y la correspondiente entropia discreta.

A partir de aqui son muchos los autores que se dedican a profundizar en este nuevo concepto de ε- entropia.

En el año 1.960 aparece la obra de PINSKER: "Information and Stability of Ramdom Variables and Processes" que reune una serie de trabajos anteriores del autor sentando las bases de las aplicaciones y relaciones entre la teoria de la informacion y la teoria de los Procesos Estocasticos.

En cuanto al estudio de la Informacion y sus relacion con la Estadistica, la obra de KULLBACK es, sin duda, el punto de arran

que de todas las investigaciones actuales sobre el tema.

Digamos para terminar este breve bosquejo historico que el desarrollo de la Teoria de la Informacion ha sido tan enorme y su aplicacion tan variada que en poco mas de veinticinco años los tratados sobre el tema son numerosos, citemos, por ejemplo, las obras de Fano, Brouillin, Ash, Feinstein, Roubine, Goldman, Young, Jelinek, etc. y el numero de publicaciones y trabajos sobre el tema se cuentan por millares.

El objeto de esta Memoria se centra en el estudio de las clasicas medidas de informacion matematica en relacion con diversos problemas de la estadistica. A lo largo de ella emplearemos simultaneamente los terminos incertidumbre e informacion pues para nosotros es lo mismo hablar de la incertidumbre relativa a los resultados de un experimento que hablar de la informacion producida por la realizacion de este experimento.

En el primer capitulo s_e estudia la informacion de Shannon y su aplicacion a problemas de estadistica como son la regresion y el muestreo. Se deducen nuevas relaciones entre esta cantidad y las informaciones de Fisher y Kullback.

El segundo capitulo está dedicado a la incertidumbre de Renyi. Se consigue generalizar la entropia de orden α de Renyi, obteniendose una funcion que tiene propiedades bastante interesantes.

Partiendo del concepto de incertidumbre de Shannon en el tercer capitulo llegamos a encontrar una medida de la informacion proporcionada por un experimento estadistico, estudiandose la relacion entre la medida encontrada y las cantidades de informacion de Kullback, Shannon y Renyi. Se emplea esta medida como método para comparar experimentos y se estudia la analogia en-

tre el concepto utilitarista de valor de la informacion asociado con un experimento y la medida encontrada por nosotros.

Al final de la Memoria añadimos un apendice donde se incluye un programa realizado mediante ordenador para el calculo de la entropia generalizada de la de Renyi de orden α, aplicandose dicho programa para la obtencion de resultados en diferentes casos particulares.

CAPITULO PRIMERO

LA INFORMACION DE SHANNON Y SU APLICACION A PROBLEMAS ESTADISTICOS

CONTENIDO.

1.1 - Intepprtacion de la incertidumbre de Shannon como la medida de un conjunto.

- Generalizacion.

1.2 - Regresion y transmision de informacion.

- Transmision de informacion.

- Transmision de informacion bivariante.

- Regresion y transmision de informacion.

- Teoremas.

1.3 - Relaciones entre las distintas cantidades de informacion.

- Caracterizacion.

- Teoremas.

- Algunas consideraciones.

1.4 - La informacion de Shannon en los problemas de muestreo.

CAPITULO PRIMERO.

LA INFORMACION DE SHANNON Y SU APLICACION A PROBLEMAS ESTADISTICOS.

Se estudia la analogia entre la incertidumbre de Shannon y una funcion de medida.

Se aplican las expresiones obtenidas en la transmision de informacion bivariante para conocer la regresion entre dos variables.

Se encuentra una expresion que relaciona las cantidades de informacion de Fisher, Kullback y Shannon y se dan relaciones particulares entre dichas cantidades.

Al final se aplica la informacion de Shannon para obtener diseños de muestreo.

1.1. INTERPRETACION DE LA INCERTIDUMBRE DE SHANNON COMO LA MEDIDA DE UN CONJUNTO.

ABRAMSON (1.963) deja entrever la posibilidad de analogia entre la incertidumbre de Shannon y una funcion de medida. Veamos que para un tipo especial de espacio probabilistico la incertidumbre de Shannon puede interpretarse como la medida de un conjunto (1)

---

(1) REZA (1.961) interpreta las desigualdades fundamentales de la incertidumbre de Shannon mediante teoria de conjuntos

Sea $(S, \mathcal{A}, P)$ un espacio probabilistico y sean α y β dos esquemas de probabilidad finitos

$$\alpha = \begin{pmatrix} A_1 & A_2 & \dots & A_n \\ P(A_1) & P(A_2) & \dots & P(A_n) \end{pmatrix} \; ; \; \beta = \begin{pmatrix} B_1 & B_2 & \dots & B_m \\ P(B_1) & P(B_2) & \dots & P(B_m) \end{pmatrix}$$

con $\sum_{i=1}^{n} P(A_i) = 1 \quad ; \quad \sum_{j=1}^{m} P(B_j) = 1$

y sea $\alpha\beta$ el esquema probabilistico producto de α y β, con sucesos A_iB_j y probabilidades conocidas asociadas $P(A_iB_j)$. Se define tradicionalmente

$$H(\alpha) = - \sum_i P(A_i) \log P(A_i)$$

$$H(\beta) = - \sum_i \sum_j P(A_iB_j) \log P(A_iB_j)$$

$$H(\alpha/\beta) = - \sum_i \sum_j P(A_iB_j) \log P(A_i/B_j)$$

y la informacion mutua $I(\alpha,\beta)$ viene dada por

$$I(\alpha,\beta) = H(\alpha) - H(\alpha/\beta)$$

Estas cantidades satisfacen las siguientes relaciones -- (KHINCHIN, 1.957):

$$H(\alpha,\beta) = H(\alpha) + H(\beta) - I(\alpha,\beta)$$

$$H(\alpha,\beta) = H(\alpha) + H(\alpha/\beta).$$

$$0 \leq H(\alpha) \leq H(\alpha,\beta) \leq H(\alpha) + H(\beta)$$

$$0 \leq H(\alpha/\beta) \leq H(\alpha)$$

$$H(\alpha,\beta) = H(\alpha) + H(\beta) \text{ si } \alpha \text{ y } \beta \text{ son independientes}$$

Consideremos ahora dos conjuntos medibles A y B de un espacio medible, con medidas $\mu(A)$ y $\mu(B)$. Se verifican las siguientes relaciones:

$$\mu(A \cup B) = \mu(A) + \mu(B) - \mu(A \cap B)$$

$$\mu(A \cup B) = \mu(B) + \mu(A - B)$$

$$0 \leq \mu(A) \leq \mu(A \cup B) \leq \mu(A) + \mu(B)$$

$$0 \leq \mu(A - B) \leq \mu(A)$$

$\mu(A \cup B) = \mu(A) + \mu(B)$ si A y B son disjuntos.

A la vista de la analogia entre ambos grupos de relaciones pedemos interpretar:

H(α) como la medida de un conjunto A, $\mu(A)$

H(β) como la medida de un conjunto B, $\mu(B)$

H(α,β) como la medida de su union A$\cup$B, $\mu(A \cup B)$

H(α/β) como la medida de su diferencia A - B, $\mu(A - B)$

I(α,β) como la medida de su interseccion A$\cap$B, $\mu(A \cap B)$

1.1.1. Ejemplo. Supongamos que el espacio S consta de estos 16 elementos que suponemos equiprobnbles:

0000	0100	1000	1100
0001	0101	1001	1101
0010	0110	1010	1110
0011	0111	1011	1111

Sea α un esquema de probabilidad que consta de estos 8 sucesos:

$A_1 = \{0000, 0001\}$ $A_2 = \{0010, 0011\}$

$A_3 = \{0100, 0101\}$ $A_4 = \{0110, 0111\}$

$A_5 = \{1000, 1001\}$ $A_6 = \{1010, 1011\}$

$A_7 = \{1100, 1101\}$ $A_8 = \{1110, 1111\}$

con probabilidades $P(A_i) = 1/8$ para todo i tal que $1 \leq i \leq 8$

Para determinar a qué subconjunto A_i pertenece un elemento de S basta observar solo las tres primeras cifras de dicho elemento.

Sea β un esquema probabilistico de S que consta de los siguientes sucesos:

$B_1 = \{ 0000, 1000 \}$ $\quad$ $B_2 = \{ 0001, 1001 \}$

$B_3 = \{ 0100, 1100 \}$ $\quad$ $B_4 = \{ 0101, 1101 \}$

$B_5 = \{ 0010, 1010 \}$ $\quad$ $B_6 = \{ 0011, 1011 \}$

$B_7 = \{ 0110, 1110 \}$ $\quad$ $B_8 = \{ 0111, 1111 \}$

con probabilidades $P(B_i) = 1/8$ para todo i tal que $1 \leq i \leq 8$.

Para determinar a que subconjunto B_i pertenece un elemento de S bastara con observar las tres ultimas cifras de dicho elemento.

Evidentemente

$$H(\alpha) = H(\beta) = \log_2 8 = 3$$

Sea γ un esquema probabilistico que consta de estos 4 subconjuntos:

$$G_1 = (0000, 0001, 1000, 1001)$$

$$G_2 = (0010, 0011, 1010, 1011)$$

$$G_3 = (0100, 0101, 1100, 1101)$$

$$G_4 = (0110, 0111, 1110, 1111)$$

con probabilidades $P(G_i) = 1/4$ para todo i tal que $1 \leq i \leq 4$.

Los subconjuntos G_i estan caracterizados por el hecho de tener sus elementos iguales el segundo y tercer dígitos.

Evidentemente

$$H(\gamma) = \log_2 4 = 2$$

Si consideramos dos observadores independientes, uno que puede realizar α es decir observar los tres primeros dígitos y el otro realizar β, es decir observar los tres ultimos dígitos.

la informacion mutua de α y β sobre un elemento son precisamente los dígitos segundo y tercero. Por tanto podemos escribir

$$H(\gamma) = I(\alpha, \beta)$$

Sea δ un esquema probabilistico que consta de estos dos sucesos:

$D_1 = \{ 0000, 0001, 0010, 0011, 0100, 0101, 0110, 0111 \}$

$D_2 = \{ 1000, 1001, 1010, 1011, 1100, 1101 \ 1110, 1111 \}$

con probabilidades $P(D_i) = 1/2$ para $i = 1, 2$

Los subconjuntos D_i estan caracterizados por tener todos sus elementos igual el primer dígito.

Evidentemente

$$H(\delta) = \log_2 2 = 1$$

De la propia definicion de α y β se deduce

$$H(\alpha/\beta) = H(\delta) = 1$$

Sea λ un esquema probabilistico que consta de 16 subconjuntos que contienen cada uno un elemento distinto, siendo $P(L_i) = 1/16$ para todo i tal que $1 \leq i \leq 16$.

Evidentemente

$$H(\lambda) = \log_2 16 = 4$$

De la propia definicion de α y β se deduce

$$H(\alpha, \beta) = H(\lambda) = 4$$

Sea L el espacio formado por $\{ 1, 2, 3, 4 \}$ y sea $\mu(L)$ = numero de elementos de L. Consideremos los subconjuntos de L siguientes:

$A = \{ 1, 2, 3 \}$ $B = \{ 2, 3, 4 \}$ $G = \{ 2, 3 \}$ $D = \{ 1 \}$

Existe una correspondencia natural entre los subconjuntos de $L: (A, B, G, D, L)$ y los esquemas probabilisticos de $S: (\alpha, \beta, \gamma, \delta, \lambda)$

"Los elementos de los diferentes subconjuntos de L son los dígitos que permanecen iguales para todos los elementos de las diferentes particiones de λ ".

Evidentemente

$$L = A \cup B \qquad G = A \cap B \qquad D = A - B$$

y por la definicion de μ

$$\mu(A) = 3; \ \mu(B) = 3; \ \mu(L) = 4; \ \mu(G) = 2; \ \mu(D) = 1$$

Hemos comprobado, pues,

$$H(\alpha) = \mu(A)$$

$$H(\beta) = \mu(B)$$

$$H(\gamma) = H(\alpha, \beta) = \mu(A \cup B) = \mu(L)$$

$$I(\alpha, \beta) = \mu(A \cap B) = \mu(G)$$

$$H(\delta) = H(\alpha/\beta) = \mu(A - B) = \mu(D)$$

Este ejemplo puede generalizarse inmediatamente:

Supongamos que S consta de 2^n n-uplas $(X_1, X_2, \ldots, X_n)$ siendo $X_i = 0$ ó 1 y siendo todas las n-uplas equiprobables. Sea L el conjunto $(1, 2, \ldots, n)$ y sean A y B los subconjuntos $(i_1, i_2, \ldots i_r)$ y $(j_1, j_2, \ldots, j_r)$ de L. A determina un esquema probabilistico de Ω con 2^r conjuntos con probabilidades $1/2^r$. Los conjuntos de α estan formados por n-uplas que tienen los dígitos $X_{i_1}, X_{i_2} \ldots X_{i_r}$ iguales. Analogamente B, $A \cup B$, $A \cap B$ y $A - B$ determinan los esquemas, β, γ, δ y λ

Como α consta de 2^r conjuntos igualmente probables, $H(\alpha) = \log_2 2^r = r = \mu(A)$, $H(\beta) = \log_2 2^r = \mu(B)$, ...

1.1.2 Generalizacion. Consideremos tres conjuntos medibles A, B y D y tres esquemas de probabilidad α, β y δ . Ahora la analogia puede perderse debido a la diferencia entre los conceptos:

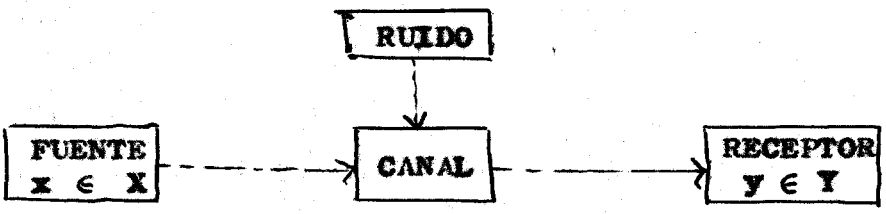
"Ser independientes" y "Ser disjuntos", que aparecen en la ultima relacion de cada uno de los grupos de relaciones indicados anteriormente.

Consideremos, por ejemplo, un conjunto D disjunto de A y de B. Entonces $D \cap (A \cup B) = \emptyset$ y podemos escribir $\mu(A \cup B \cup D) = \mu(A \cup B) + \mu(D)$. Por otro lado, si δ es un esquema de probabilidad independiente de α y β , ello no implica que δ sea independiente del esquema producto $\alpha\beta$ (2) y por tanto no puede escribirse $H(\alpha, \beta, \delta) = H(\alpha, \beta) + H(\delta)$.

Sin embargo en el caso particular de reemplazar la funcion de medida por una funcion de conjuntos aditiva, esta analogia ha sido extendida por HU GUO DING (1.962). Incluso llega a definir unas cantidades de informacion generalizadas que representa por $H[\varphi(A^1, A^2, \ldots A^n)]$ en donde Q representa cualquier combinacion finita de las operaciones union, interseccion y diferencia de conjuntos, basandose en la analogia existente entre estas cantidades de informacion y los valores de una funcion de conjuntos $\varphi[\varphi(A^1, A^2, \ldots A^n)]$.

1.2 - REGRESION Y TRANSMISION DE INFORMACION.

1.2.1. <u>Transmision de informacion.</u> Consideremos un canal de comunicacion con sus fuente y receptor.



---

(2) Basta con que existan tres sucesos que sean independientes dos a dos pero que entre los tres sean dependientes.

La informacion transmitida mide la cantidad de asociacion entre la fuente y el receptor del canal. Si lo emitido por la fuente y lo recibido por el receptor son independientes, no se transmite ninguna informacion. Por otro lado, si ambos estan perfectamente correlacionados, toda la informacion de la fuente es transmitida a lo largo del canal. En la mayoria de los casos, naturalmente, la informacion transmitida se encuentra entre ambos extremos.

Estamos interesados en la cantidad de informacion transmitida. Supongamos que tenemos una distribucion de probabilidad bivariante con funcion de densidad p(x, y). Esto significa que la variable fuente emite un valor o señal x, entonces el ruido del canal lo altera recibiendose en el receptor un valor entre y e y + dy con probabilidad p(y/x) dy siendo

$$P(y/x) = \frac{p(x,y)}{\int p(x,y)\,dy}$$

Significa tambien que la fuente emite las señales de manera que estas toman valores entre x y x + dx con probabilidad

$$p(x)\,dx = dx \int p(x,y)\,dy$$

Bajo estas condiciones y si las sucesivas señales son independientes, la cantidad de informacion transmitida por señal es definida por SHANNON (1.948) como

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

siendo

$$H(X) = \int p(x) \log p(x)\,dx$$

$$H(XY) = \int p(x,y) \log p(x,y)\,dx\,dy$$

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

$$R(x,y) = H(X) - H_Y(X) = H(Y) - H_X(Y) \qquad (ii)$$

siendo

$$H_x(Y) = \int\int p(x,y) \log p(y/x)\, dy\, dx \qquad (3)$$

1.2.2. Transmision de informacion bivariante. Consideremos ahora el caso donde son dos las fuentes que transmiten a un receptor. Si tomamos la variable fuente como bidemensional tendremos

$$R[(x,y),z] = H(x,y) + H(z) - H(x,y,z) =$$

$$= H(x,y) + H(z) - [H(z) + H_z(x,y)] =$$

$$= H(x,y) - H_z(x,y) = H(z) - H_{x,y}(z) \qquad (iii)$$

siendo

$$H_z(x,y) = \int\int\int p(x,y,z) \log p(x,y/z)\, dx\, dy\, dz$$

$$H_{xy}(z) = \int\int\int p(x,y,z) \log p(z/x,y)\, dx\, dy\, dz$$

Vamos a expresar $R[(x, y), z]$ como una combinacion de las transmisiones de informacion 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 informacion transmitida entre Y y Z podemos escribir

$$R_x(y,z) = H_x(y) + H_x(z) - H_x(y,z) =$$

$$= H_x(y) - H_{x,z}(y) = H_x(z) - H_{x,y}(z) \qquad (iv)$$

De (i), (ii), (iii) y (iv) deducimos

$$R[(x,y),z] = R(x,z) + R_x(y,z) = R(y,z) + R_y(x,z)$$

La expresion anteriormente escrita nos expresa que el teorema de adicion (4) es valido tambien para la razon de transmision

---

(3) FEINSTEIN (1.958).

(4) El teorema de adicion para entropias nos dice que la incertidumbre de la realizacion simultanea de dos esquemas de probabilidad es igual a la incertidumbre de un esquema mas la incertidumbre del otro condicionada por el primero (KHINCHIN, 1.957)

de informacion.

De (ii) y (iv) deducimos:

$$R(y,z) - R_x(y,z) = [H(y) - H_y(z)] - [H_x(z) - H_{x,y}(z)] =$$

$$= [H(y) - H_z(y)] - [H_x(y) - H_{x,z}(y)]$$

que es una cantidad igual mayor o menor que cero y nos indica la ganancia o perdida en la informacion transmitida entre Y y Z debido al conocimiento de la variable X.

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

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

cantidades que Mc GILL (1.954) ha denominado con el nombre de informacion mutua entre las tres variables, pues nos indica la interacion entre ellas.

Una vez extendido el concepto de transmision de informacion de Shannon al caso de tres variables pasamos a relacionar este concepto con el de regresion.

1.2.3 <u>Regresion y transmision de informacion.</u> 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 funcion de densidad condicionada de Y a X.

Sea $\varphi(x)$ la linea de regresion de Y sobre X, es decir,

$$\varphi(x) = \int_Y y\, p_2(y/x)\, dy$$

Introducimos una nueva variable Z definida en la forma

$$z/x = y/x - \varphi(x)$$

cuya funcion de densidad será

$$p_3(z) = \int q(x,z)\,dx = \int p_3(z/x)\,p_1(x)\,dx = \int p_2(z+\varphi(x)/x)\,p_1(x)\,dx$$

por ser

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

de la expresion (iv) deducimos

$$R_z(x,y) = H_z(x) = H_z(y) \qquad \text{(vii)}$$

$$R_y(x,z) = H_y(x) = H_y(z) \qquad \text{(viii)}$$

Por otro lado de (vi) se obtiene

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

y teniendo en cuenta (vii) y (viii)

$$R(x,y) = R(x,z) + H_z(y) - H_y(z) =$$

$$= R(x,z) + H(y) - H(z)$$

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

Siguiendo a FERON Y FOURGEAUD (1.951) llamaremos a R(x,z) parte elastica y a H(Y) - H(Z) parte dura de la informacion transmitida R(x,y).

1.2.3.1. Teorema. $R(x,z) = 0$ si y solo si $p_2(y/x) = f[y-\varphi(x)]$ donde f(z) es alguna funcion de densidad con media cero.

En efedto:

Decir que $R(x,z) = 0$ es equivalente a decir que X y Z son independientes, es decir, $p_3(z/x) = p_3(z)$ o sea que la distribucion de Z condicionada a X no depende de X. Razonamos:

A) Si X y Z son independientes entonces

$$P_3(z/x) = P_2(z + \varphi(x)/x) = P_3(z)$$

y haciendo el cambio $z/x = y/x - \varphi(x)$ tenemos

$$P_2[y - \varphi(x)] = P_2(y)$$

siendo $p_2[y - \varphi(x)]$ una funcion de densidad de media cero.

Al reves:

B) Si $$P_2(y/x) = f[y - \varphi(x)]$$

y hacemos el cambio $y/x = z/x + \varphi(x)$, podremos escribir

$$P_3(z + \varphi(x)/x) = f(z + \varphi(x) - \varphi(x) = f(z) = P_3(z/x)$$

de donde se deduce que $p_3(z/x)$ es independiente de X y por tanto X y Z son independientes.

Analogamente sea $\psi(y)$ la linea de regresion de X sobre Y, es decir

$$\psi(y) = \int x\, P_1(x/y)\, dx$$

introducimos una nueva variable W definida en la forma

$$W/y = X/y - \psi(y)$$

cuya funcion de densidad marginal será de la forma

$$P_4(w) = \int P_2(w/y) P_2(y)\, dy = \int P_1(w + \varphi(y)/y) P_2(y)\, dy$$

Haciendo razonamientos analogos a los anteriores llegariamos a demostrar

$$R(x, y) = R(y, w) + H(X) - H(W)$$

en donde $R(y,w) \geq 0$ y $H(X) - H(W) \geq 0$.

A R(x,w) se le llama parte elastica y a H(X) - H(W) parte du-

ra de la informacion transmitida R(x,y). Podemos enunciar un teorema similar al anteriormente demostrado.

1.2.3.2. Teorema. R(y,w) = 0 si y solo si la funcion de densidad condicionada de X a Y puede escribirse en la forma

$$P_1(x/y) = g[x - \psi(y)]$$

siendo g(w) una funcion de densidad de media a cero.

Basandonos en los teoremas 1.2.3.1 y 1.2.3.2 podemos enunciar: "Una distribucion de probabilidad bivariante p(x,y) tiene correlacion dura cuando R(x,z) = R(y,w) = 0, es decir, cuando la razon de transmision de informacion R(x,y) coincide con las partes duras H(y) - H(z) y H(X) - H(W) que en este caso son iguales".

1.2.3.3. Teorema. La distribucion normal bivariante es la unica que posee correlacion lineal dura entre sus variables.

A) Supongamos que p(x,y) es la funcion de densidad de una ley normal bivariante y queremos demostrar

$$P_2(y/x) = f(y - \beta_1 x)$$

$$P_1(x/y) = g(x - \beta_2 y)$$

donde f y g son dos funciones de densidad con media nula.

En efecto:

Si p(x,y) es la funcion de densidad de una variable aleatoria bidimensional que sigue una ley normal bivariante de media

$$\mu = (0,0)$$

y matriz de covarianzas

$$M = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

entonces las distribuciones condicionadas vendrán dadas por:(5)

---

(5) p.e. WILKS (1.962)

$$p_2(y/x) = \frac{1}{\sigma_2\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\tfrac{1}{2}\left(\frac{y-\rho\frac{\sigma_2}{\sigma_1}x}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right\}$$

y

$$p_1(x/y) = \frac{1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} \exp\left\{-\tfrac{1}{2}\left(\frac{x-\rho\frac{\sigma_1}{\sigma_2}y}{\sigma_1\sqrt{1-\rho^2}}\right)^2\right\}$$

que podemos escribir en la forma

$$p_2(y/x) = f(y-\beta_1 x)$$

y

$$p_1(x/y) = g(x-\beta_1 y)$$

siendo $\beta_1 = \rho\frac{\sigma_2}{\sigma_1}$ y $\beta_2 = \rho\frac{\sigma_1}{\sigma_2}$ los coeficientes de las rectas de regresion de Y sobre X y de X sobre Y, respectivamente y f y g dos funciones de densidad con momentos de primer orden nulos.

B) Suponemos que existe correlacion lineal dura, es decir

$$p_2(y/x) = f(y-\beta_1 x)$$

$$p_1(x/y) = g(x-\beta_1 y)$$

Vamos a probar que (X,Y) sigue una ley normal bivariante. (6)

En efecto:

Sean $q_1(x)$ y $q_2(y)$ las funciones de densidad marginales de las variables X e Y, respectivamente. Podemos escribir

$$p(x,y) = q_1(x)\,p_2(y/x) = q_2(y)\,p_1(x/y)$$

o bien

$$q_1(x)\,f(y-\beta_1 x) = q_2(y)\,g(x-\beta_2 y)$$

y tomando logaritmos neperianos

---

(6) Por la desigualdad de Schwartz sabémos que $|\beta_1\beta_2| = |\rho| \leq 1$ aqui suponemos el caso no trivial en que $|\rho| < 1$

$$\ln q_1(x) + \ln f(y-\beta_1 x) = \ln q_2(y) + \ln g(x-\beta_2 y) \quad \text{(ix)}$$

Si notamos

$$A(y-\beta_1 x) = \ln f(y-\beta_1 x)$$

$$B(x-\beta_2 y) = \ln g(x-\beta_1 y)$$

y derivamos ambos miembros de la igualdad (ix) con respecto a X y posteriormente con respecto a Y, obtenemos

$$\frac{q_1'(x)}{q_1(x)} + \frac{\partial A(y-\beta_1 x)}{\partial x} = \frac{\partial B(x-\beta_2 y)}{\partial x}$$

$$\frac{\partial^2 A(y-\beta_1 x)}{\partial y \partial x} = \frac{\partial^2 B(x-\beta_2 y)}{\partial y \partial x} \quad \text{(x)}$$

Haciendo $A(z) = A(y-\beta_1 x)$; $B(w) = B(x-\beta_2 y)$ **encontramos que**

$$\frac{\partial A(y-\beta_1 x)}{\partial x} = -\beta_1 A'(z)$$

$$\frac{\partial^2 A(y-\beta_1 x)}{\partial y \partial x} = -\beta_1 A''(z)$$

y

$$\frac{\partial B(x-\beta_2 y)}{\partial y} = B'(w)$$

$$\frac{\partial^2 B(x-\beta_2 y)}{\partial y \partial x} = -\beta_2 B''(w)$$

y teniendo en cuenta (x)

$$-\beta_1 A''(z) = -\beta_2 B''(w)$$

o bien

$$\beta_1 A''(y-\beta_1 x) = \beta_2 B''(x-\beta_2 y)$$

Para x = 0

$$A''(y) = \frac{\beta_2}{\beta_1} B''(-\beta_2 y)$$

Para $y = 0$

$$B''(x) = \frac{\beta_1}{\beta_2} A''(-\beta_1 x)$$

Podemos, por tanto, expresar

$$A''(z) = \frac{\beta_2}{\beta_1} B''(-\beta_2 z) = \frac{\beta_2}{\beta_1} \cdot \frac{\beta_1}{\beta_2} A''(\beta_2 \beta_1 z) =$$

$$= A''(\beta_2^2 \beta_1^2 z) = \cdots = A''(\beta_1^n \beta_2^n z) = A''(0)$$

Analogamente obtendriamos

$$B''(w) = B''(0)$$

Por ser A(z) una funcion logaritmica y por tanto convexa pode-
mos escribir

$$A''(z) = -1/a \text{ con } a > 0 \qquad (7)$$

Integrando obtenemos

$$A(z) = \frac{-z^2}{2a} + \ln k$$

y como A(z) = ln. f(z)

$$f(z) = k\, e^{-z^2/2a}$$

siendo $k = 1/\sqrt{2\pi a}$ para que f(z) sea una funcion de densidad.

En definitiva

$$f(y - \beta_1 x) = \frac{1}{\sqrt{2\pi a}} e^{-\frac{(y-\beta_1 x)^2}{2a}}$$

Si suponemos

$$B''(w) = -1/b \quad \text{con } b > 0$$

integrando y operando de forma analoga tendremos

$$g(x - \beta_2 y) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-\beta_2 y)^2}{2b}}$$

---

(7) Baste señalar que si $A''(z) = 0 \Rightarrow A'(z) = c \Rightarrow f'(z)/f(z) = c$
de donde $f'(z) = c\, f(z)$ e integrando, c = 0 y esto seria absurdo.

y por las hipotesis podemos escribir

$$p_2(y/x) = \frac{1}{\sqrt{2\pi a}}\, e^{-\frac{(y-\beta_1 x)^2}{2a}}$$

$$p_1(x/y) = \frac{1}{\sqrt{2\pi b}}\, e^{-\frac{(x-\beta_2 y)^2}{2b}}$$

Para calcular la funcion de densidad conjunta, calculemos en primer lugar las funciones de densidad marginales. Sabemos

$$\frac{p_2(y/x)}{p_1(x/y)} = \frac{q_2(y)}{q_1(x)}$$

de donde

$$q_1(x) = \frac{p_1(x/y)}{p_2(y/x)}\, q_2(y)$$

podemos, por tanto, escribir

$$q_1(x) = \left(\frac{a}{b}\right)^{1/2} \exp\left\{-\frac{(x-\beta_2 y)^2}{2b} + \frac{(y-\beta_1 x)^2}{2a}\right\} q_2(y)$$

y para $y = 0$

$$q_1(x) = q_2(0)\cdot\left(\frac{a}{b}\right)^{1/2} \exp\left\{-\frac{x^2}{2b} + \frac{\beta_1^2 x^2}{2a}\right\}$$

es decir

$$q_1(x) = q_2(0)\left(\frac{a}{b}\right)^{1/2} \exp\left\{-\frac{x^2}{2\left(\frac{ab}{a-b\beta_1^2}\right)}\right\}$$

siendo

$$q_2(0) = \frac{1}{\sqrt{2\pi}}\,\frac{\sqrt{a-b\beta_1^2}}{a}$$

para que $q_1(x)$ sea una funcion de densidad. En definitiva

$$q_1(x) = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{a-b\beta_1^2}{ab}} \exp\left\{-\frac{x^2}{2\left(\frac{ab}{a-b\beta_1^2}\right)}\right\} \quad (xi)$$

De forma analoga obtendriamos

$$q_2(y) = q_1(0)\cdot\left(\frac{b}{a}\right)^{1/2} \exp\left\{-\frac{y^2}{2\left(\frac{ab}{b-a\beta_2^2}\right)}\right\}$$

siendo

$$q_1(0) = \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{b - a\beta_2^2}}{b} \qquad (xii)$$

para que $q_2(y)$ sea funcion de densidad.

Haciendo x = 0 en (xi) e igualando con lo obtenido en (xii)

$$\frac{a}{b} = \beta_1^2 \; ; \; \frac{b}{a} = \beta_2^2$$

y por ser $\beta_1 = \rho \frac{\sigma_1}{\sigma_2}$ y $\beta_2 = \rho \frac{\sigma_2}{\sigma_1}$

se deduce (8)

$$\frac{a}{b} = \frac{\sigma_1^2}{\sigma_2^2} \qquad (xiii)$$

Por otro lado de la expresion de $q_1(x)$ se obtiene

$$\frac{ab}{a - b\beta_1^2} = \sigma_1^2$$

y teniendo en cuenta (xiii)

$$b = \sigma_1^2 (1 - \rho^2)$$

Del mismo modo llegariamos a obtener

$$a = \sigma_2^2 (1 - \rho^2)$$

Por tanto la distribucion conjunta del par (X,Y) será

$$p(x,y) = p_2(y/x)\, q_1(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma_1^2} - 2\rho \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right) \right\}$$

que es la funcion de densidad de una distribucion normal bivariante de media

$$\mu = (0,0)$$

y matriz de varianzas

---

(8) Tambien se podria haber deducido de la expresion

$$A''(z) = \beta_2/\beta_1 \, B''(-\beta_2 z)$$

pues $A''(z) = -\frac{1}{a}$ y $B''(z) = -\frac{1}{b}$ cualquiera que sea z.

$$\mathcal{M} = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

con lo que el teorema está demostrado.

1.3. RELACION ENTRE LAS DISTINTAS CANTIDADES DE INFORMACION.

Las tres definiciones clasicas de informacion estadistica, de Fisher, de Shannon y de Kullback estan fuertemente vinculadas a propiedades asintoticas y a un principio general por el cual la cantidad de informacion debería ser aditiva.

FISHER (1.925) define la cantidad de informacion para un parametro θ y una funcion de densidad $p(x;\theta)$ que satisfaga las condiciones de regularidad de Cramer-Rao, en la forma:

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

SHANNON (1.948) propone una definicion de informacion (incertidumbre) para la teoria de la comunicacion. En su primitiva forma nos indica la variacion en una distribucion; con un cambio de signo, la concentracion. Es esta segunda forma la que usaremos. Notaremos

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

KULLBACK (1.959) considera una definicion de informacion para discriminar en favor de una hipotesis $H_1(\theta_1)$ contra otra $H_2(\theta_2)$:

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

1.3.1. Caracterizacion. Consideremos un modelo probabilistico cuya funcion de densidad $p(x;\theta)$ depende del valor de un parametro θ que toma sus valores en un espacio parametrico Ω. Como medida de informacion sobre el parametro θ dado un valor x de la variable aleatoria X asociada al modelo, definimos

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

Si notamos por θ_0 el verdadero valor del parametro, calculando la media para todos los posibles valores de X obtenemos para la informacion sobre el parametro θ , cuando el verdadero valor es θ_0 , la expresion

$$I(\theta, \theta_0) = \int p(x;\theta_0)\ \log\ p(x;\theta)\, dx$$

Consideremos algunos aspectos de esta funcion de informacion.

A) La informacion sobre el verdadero valor del parametro es:

$$I(\theta_0, \theta_0) = \int p(x;\theta_0)\ \log\ p(x;\theta_0)\, dx = I_S(\theta)$$

y es ademas el maximo valor que puede alcanzar esta funcion.

B) La informacion sobre el verdadero valor θ_0 excede de la informacion sobre otro cualquier valor del parametro θ en:

$$I(\theta_0, \theta_0) - I(\theta, \theta_0) = \int p(x;\theta_0) \log p(x;\theta_0)\, dx - \int p(x;\theta_0) \log p(x;\theta)$$

$$= \int p(x;\theta_0)\ \log\ \frac{p(x;\theta_0)}{p(x;\theta)}\, dx = I_K(\theta_0, \theta)$$

C) Estudiemos ahora la culvatura de la funcion de informacion en $\theta = \theta_0$. Suponiendo que se verifican las condiciones de regularidad de Cramer podemos escribir

$$\frac{\partial}{\partial\theta} I(\theta, \theta_0) = \int \frac{\partial}{\partial\theta} \log\ p(x;\theta)\ \ p(x;\theta_0)\, dx$$

de donde

$$\left[\frac{\partial}{\partial\theta} I(\theta, \theta_0)\right]_{\theta=\theta_0} = 0$$

y la curvatura de la funcion de informacion en el punto $\theta = \theta_0$ será

$$-\left[\frac{\partial^2}{\partial\theta^2} I(\theta, \theta_0)\right]_{\theta=\theta_0} = \left[\int -\frac{\partial^2}{\partial\theta^2} \log p(x;\theta)\ p(x;\theta_0)\, dx\right]_{\theta=\theta_0} = I_F(\theta_0)$$

Resumiendo, diremos que la informacion $\mathbb{I}(\vartheta, \vartheta_0)$ como funcion de ϑ tiene su maximo valor $I_S(\vartheta_0)$ en el verdadero valor ϑ_0, tiene curvatura $I_F(\vartheta_0)$ en dicho valor y una discrepancia, con respecto a cualquier otro valor de Ω, de $I_K(\vartheta_0, \vartheta)$

De la relacion

$$I_S(\vartheta_0) - I_K(\vartheta_0, \vartheta) = I(\vartheta, \vartheta_0)$$

deducimos la siguiente expresion que relaciona las tres cantidades de informacion mas conocidas de la literatura estadistica:

$$-\left[\frac{\partial^2}{\partial\vartheta^2}\left[I_S(\vartheta_0) - I_K(\vartheta_0, \vartheta)\right]\right]_{\vartheta=\vartheta_0} = I_F(\vartheta_0)$$

A continuacion expresamos mediante teoremas relaciones particulares entre estas cantidades de informacion.

1.3.2 <u>Relacion entre las informaciones de Kullback y de Fisher. Teorema.</u> Las cantidades de informacion de Fisher y Kullback definidas anteriormente verifican la siguiente relacion:

$$I_K(\vartheta, \vartheta+\Delta\vartheta) = 1/2\, I_F(\vartheta)\cdot\Delta\vartheta^2 + o(\Delta\vartheta^2)$$

En efecto:

$$I_K(\vartheta, \vartheta+\Delta\vartheta) = \int_X p(x/\vartheta) \log \frac{p(x/\vartheta)}{p(x/\vartheta+\Delta\vartheta)}\, dx =$$

$$= -\int_X p(x/\vartheta)\left[\log p(x/\vartheta+\Delta\vartheta) - \log p(x/\vartheta)\right] dx$$

y como suponemos se verifican las condiciones de regularidad, obtenemos mediante un desarrollo de Taylor

$$\log p(x/\vartheta+\Delta\vartheta) = \log p(x/\vartheta) + \frac{\partial \log p(x/\vartheta)}{\partial\vartheta}\Delta\vartheta + \frac{1}{2}\frac{\partial^2 \log p(x/\vartheta)}{\partial\vartheta^2}\cdot\Delta\vartheta^2$$

de donde

$$I_K(\vartheta, \vartheta+\Delta\vartheta) = -\int_X p(x/\vartheta)\left[\frac{\partial \log p(x/\vartheta)}{\partial\vartheta}\cdot\Delta\vartheta + \frac{1}{2}\frac{\partial^2 \log p(x/\vartheta)}{\partial\vartheta^2}\Delta\vartheta^2 + o(\Delta\vartheta^2)\right]$$

y como por las condiciones de regularidad el campo de variacion de X no depende de θ, resulta:

$$\int p(x/\theta) \frac{\partial \log p(x/\theta)}{\partial \theta} \Delta\theta \, dx = 0$$

$$-\int p(x/\theta) \frac{\partial^2 \log p(x/\theta)}{\partial \theta^2} dx = I_F(\theta)$$

de donde, finalmente, (9)

$$I_K(\theta, \theta + \Delta\theta) = \tfrac{1}{2} I_F(\theta) \Delta\theta^2 + o(\Delta\theta^2)$$

1.3.3. <u>Relacion entre la transmision de informacion de Shannon y la informacion de Kullback. Teorema.</u> 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 parametro para el que

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

la razon de transmision de informacion de Shannon $R(\theta, x)$ y la informacion de Kullback $I_K(\theta_1, \theta_2)$ verifican la expresion:

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

En efecto:

$$R(\theta, x) = -\sum_{i=1}^{2} p(\theta_i) \log p(\theta_i) + \int_{\mathcal{X}} \left[\sum_{i=1}^{2} \frac{p(\theta_i)\, p(x/\theta_i)}{p(x, \theta^\circ)} \log \frac{p(\theta_i)\, p(x/\theta_i)}{p(x\theta^\circ} \right] \cdot p(x;\theta^\circ)\, dx =$$

$$= -\sum_{i=1}^{2} p(\theta_i) \log p(\theta_i) + \int_{\mathcal{X}} \sum_{i} p(\theta_i)\, p(x/\theta_i) \log p(\theta_i)\, dx +$$

$$+ \sum_{i} p(\theta_i) \int_{\mathcal{X}} p(x/\theta_i) \log \frac{p(x/\theta_i)}{p(x;\theta^\circ)} dx$$

y por ser $p(x/\theta_1)$ y $p(x/\theta_2)$ funciones de densidad, resulta

---

(9) FUCHS-LETTA (1.970) llegan a la misma expresion para la "ganancia de informacion" de Shannon.

$$R(\theta, x) = -\sum_{i=1}^{\tilde{2}} p(\theta_i) \log p(\theta_i) + \sum_{i=1}^{\tilde{2}} p(\theta_i) \log p(\theta_i) +$$

$$+ \sum p(\theta_i) \int_{\mathcal{X}} p(x/\theta_i) \log \frac{p(x/\theta_i)}{p(x/\theta_0^\circ)} dx =$$

$$= p(\theta_1) \int_{\mathcal{X}} p(x/\theta_1) \log \frac{p(x/\theta_1)}{p(x;\theta^\circ)} dx + p(\theta_2) \int_{\mathcal{X}} p(x/\theta_2) \log \frac{p(x/\theta_2)}{p(x;\theta^\circ)}$$

$$= p(\theta_1) I_u(\theta_1, \theta^\circ) + p(\theta_2) I_u(\theta_2, \theta^\circ)$$

con lo que el teorema queda demostrado.

1.3.4. <u>Relacion entre la cantidad de informacion de Fisher y la razon de transmision de informacion de Shannon.</u> Se define

$$R(\theta, x_1 \cdots x_n) =$$

$$= \int_{x_1} \cdots \int_{x_n} \int_{\Omega} p(\theta) \prod_{i=1}^{n} p(x_i/\theta) \log \frac{\prod_{i=1}^{n} p(x_i/\theta)}{\int p(u) \prod_{i=1}^{n} p(x_i/u) du} d\theta dx_1 \cdots dx_n$$

haciendo el cambio

$$u = \theta + \frac{v}{\varphi(n)}$$

resulta

$$R(\theta, x_1 \cdots x_n) =$$

$$= \int_{x_1} \cdots \int_{x_n} \int_{\Omega} p(\theta) \prod_{i=1}^{n} p(x_i/\theta) \log \frac{\varphi(n)}{\int p(\theta + v/\varphi(n)) z_n(v) dv} d\theta dx_1 \cdots dx_n$$

siendo

$$z_n = \frac{\prod_{i=1}^{n} p(x_i/\theta + v/\varphi_n)}{\prod_{i=1}^{n} p(x_i/\theta)}$$

$$R(\theta, x_1, x_2 \cdots x_n) =$$

$$= \log \varphi(n) - \int_{x_1} \cdots \int_{x_n} \int_{\theta} p(\theta) \prod_{i=1}^{n} p(x_i/\theta) \log \int p\left(\theta + \frac{v}{\varphi(n)}\right) z_n(v)\, dv$$

$$d\theta\, dx_1 \cdots dx_n =$$

$$= \log \varphi(n) - \int_{x_1} \cdots \int_{x_n} \int_{\theta} p(\theta) \prod_{i=1}^{n} p(x_i/\theta) \log \left[p(\theta) \int \frac{p(\theta + v/\varphi(n))}{p(\theta)} z_n(v)\, dv \right]$$

$$d\theta\, dx_1 \cdots dx_n =$$

$$= \log \varphi(n) - \int_{\Omega} p(\theta) \log p(\theta) - E\left[\log \int \frac{p(\theta + v/\varphi(n))}{p(\theta)} z_n(v)\, dv \right]$$

$$= \log \varphi(n) - I_S(\theta) - E\left[\log \int \frac{p(\theta + v/\varphi(n))}{p(\theta)} z_n(v)\, dv \right]$$

Nos basamos ahora en un teorema de IBRAGIMOV (1.962) que afirma que bajo determinadas condiciones para la funcion de densidad $p(x/\sigma)$, las razones de verosimilitud $Z_n(v)$ anteriormente definidas convergen cuando $n \to \infty$, a una funcion $Z(v)$ dada por la expresion

$$Z(v) = \exp \left\{ v\, I_F^{1/2}(\theta)\, y - \frac{v^2}{2} I_F(\theta) \right\}$$

tomando $\varphi(n) = \sqrt{n}$ y siendo Y una variable aleatoria independiente de θ que sigue una ley normal de media 0 y varianza 1.

Tomando limite cuando n tiende a infinito en la ultima expresion obtenida anteriormente para $R(\sigma, x_1, x_2, \ldots, x_n)$ tendremos:

$$R(\theta, x_1, \cdots x_n) = \log \sqrt{n} - I_S(\theta) - E\left[\log \int Z(v)\, dv \right] + o(1)$$

Por otro lado

$$\int z(v)dv \int \exp\left\{ v\, I_F^{1/2}(\theta)\, y - \frac{v^2}{2} I_F(\theta) \right\} dv =$$

$$= \left\{ e^{y^2/2} \int \exp\left[-\frac{I_F(\theta)}{2} \left(v - \frac{y}{I_F^{1/2}(\theta)} \right)^2 \right] dv \right. =$$

$$= e^{y^2/2} \sqrt{\frac{2\pi}{I_F(\theta)}}$$

y tomando logaritmos

$$\log \int z(v)\, dv = \frac{y^2}{2} \log e + \log \sqrt{2\pi} - \log I_F^{1/2}(\theta)$$

y calculando ahora la esperanza matematica en la expresion anterior se deduce

$$E\left[\log \int z(v)\, dv \right] = \tfrac{1}{2} \log e + \log \sqrt{2\pi} - E\left[\log I_F^{1/2}(\theta) \right] =$$

$$= \log \sqrt{2\pi e} - E\left[\log I_F^{1/2}(\theta) \right]$$

En definitiva (cuando $n \to \infty$)

$$R(\theta, x_1, x_2 \cdots x_n) =$$

$$= \log \sqrt{n} - \log \sqrt{2\pi e} - I_S(\theta) + E\left[\log I_F^{1/2}(\theta) \right] + o(1)$$

$$= \log \frac{\sqrt{n}}{\sqrt{2\pi e}} + \int p(\theta) \log \frac{I_F^{1/2}(\theta)}{p(\theta)} d\theta + o(1)$$

En el limite y bajo determinadas condiciones, hemos encontrado una relacion entre la informacion de Fisher y la transmision de informacion de Shannon.

1.3.5. <u>Algunas consideraciones.</u> Exponemos a continuacion algunas diferenciasyy analogias entre las cantidades de informacion de Shannon y de Fisher.

Son dos las diferencias principales:

A) El concepto de Shannon lleva consigo implicita la introduccion de una distribucion a priori para el parametro desconocido. Es, pues, bayesiano mientras que la definicion de Fisher no lo es

B) La cantidad de informacion 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 total de X para un θ fijado.

Si la primera diferencia dá cierta ventaja a la nocion de - Fisher, una vez que la probabilidad a priori para θ ha sido admitida, el segundo apartado hace que la informacion de Shannon sea mas facil de aplicar.

Tienen, sin embargo, las informaciones de Shannon y de Fisher un curioso rasgo en comun. Ambos, consideran el futuro (antes de que ocurra el suceso o experimento) o bien el pasado (cuando el experimento ha sido realizado). La definicion de Fisher nos dá el rigor con que un parametro desconocido puede ser definido mediante experimentos, bien antes de ellos o despues de realizados. En la informacion de Shannon uno puede calcular antes o una vez que el mensaje ha sido enviado, la razon a la que la informacion será transmitida en un codigo dado.

Es conveniente hacer mencion del ingenioso, aunque no demasiado convincente, razonamiento que utiliza BARNARD (1.951) para relacionar las informaciones de Shannon y de Fisher.

1.4. LA INFORMACION DE SHANNON EN LOS PROBLEMAS DE MUESTREO.

Un problema que se presenta frecuentemente en la Estadistica es el de determinar el valor de un parametro; sin embargo este valor no puede ser determinado directamente sino a traves de la

informacion proporcionada por una serie de observaciones.

La primera pregunta que el estadistico se plantea es saber cuándo una serie de observaciones contiene toda la informacion necesaria para encontrar el verdadero valor del parametro.

La medida de informacion que se utiliza normalmente es la de Shannon por ser la que mejor se adapta al punto de vista bayesiano, pues aún en el caso de que no tuvieramos un conocimiento a-prioristico sobre θ, si solo conocieramos el conjunto de los posibles valores de θ, parece logico atribuir a θ la distribucion 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 logico 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 (1.964) 1.967, 1.968) y de VAJDA (1.967, 1.968) dan condiciones necesarias para que esta pregunta anteriormente planteada sea respondida satisfactoriamente, es decir, para que la cantidad de informacion residual, diferencia entre la incertidumbre del espacio parametrico y la informacion que la muestra nos aporta sobre éste, converga a cero. Todos estos teoremas vienen a corroborar el hecho de que la razon de transmision de informacion de Shannon $R(\theta,x)$ es una funcion 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

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

y en general

$$R(\sigma, x_1 x_2 \ldots x_n) \leq R(\sigma, x_1 x_2 \ldots x_{n-1})$$

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

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

Si se verifica

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

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

Si se verifica

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

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

Si se verifica

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

se continua el proceso.

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

$$H(\theta/X_1X_2 \ldots X_n) \leq \alpha$$

es decir, cuando

$$\sum p(\theta/x_1 x_2 \cdots x_m) \log p(\theta/x_1 x_2 \cdots x_m) \leq \alpha$$

En el caso de que el espacio parametrico Ω conste de solo dos elementos, esta regla de muestreo secuencial coincide con el test de razon de verosimilitud de Wald (10)

---

(10) GIRSHICK (1.946).

Otro procedimiento para obtener un diseño optimo de muestreo podria consistir en considerar una funcion de probabilidad de - error asociada con las decisiones aconsejadas por la informacion recibida (RENYI, 1.967). Limitando esta probabilidad de error llegariamos a obtener un criterio de detension de un muestreo secuencial.

REFERENCIAS DEL CAPITULO

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

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

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

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

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

FUCHs-LETTA (1.970)."L'inegalité de Kullback. Aplication a la Theorie de l'estimation". Ed. Springer Verlag.

GIRSHICK (1.946)."Contributions to the theory of sequential analysis I. Ann. Math. Stat. 17 (123-143).

HU GUO DING (1.962)."On Information Quantity" Teor Verojatnost. i Primenen, 7 (447-455)

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

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

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

MCGILL (1.954)."Multivariate information transmission" Trans IRE (93-111)

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

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

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

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

SHANNON (1.948)."A mathematical Theory of Communication" Bell. System Technical Journal 27 (379-423, 623-656). Reeditado por University of Illinois Press en 1.969 (Shannon-Weaver, "A Mathematical Theory of Communication").

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)

WILKS (1.962). "Mathematical Statistics" Ed. J. Wiley.

CAPITULO SEGUNDO

ENTROPIA DE RENYI

DE ORDEN ALFA

CONTENIDO.

2.1. Caracterizacion axiomatica de la entropia de Shannon para distribuciones de probabilidad generalizadas.

- Teorema

2.2. Entropia de Renyi de orden α.

-Teorema

2.3. Generalizacion

- Definicion

- Caracterizacion

- Teoremas

2.4. Cantidad de Informacion de orden α.

- Definicion

- Caracterizacion

- Teorema

2.5. Generalizacion

- Definicion

- Caracterizacion

CAPITULO SEGUNDO

ENTROPIA DE RENYI DE ORDEN ALFA

Se generaliza la entropia de Renyi de orden α . Se caracteriza axiomaticamente siguiendo los trabajos de Renyi y al final se deducen algunas propiedades y teoremas interesantes, para la entropia generalizada.

2.1./CARACTERIZACION DE LA ENTROPIA DE SHANNON PARA DISTRIBUCIONES DE PROBABILIDAD GENERALIZADAS.

El postulado D de la axiomatica de FADEEV (1.956) incluido para obtener la propiedad de aditividad para la incertidumbre de Shannon parecia demasiado fuerte.

Si notamos PQ el producto de dos distribuiciones de probabilidad

$$P = (p_1, \ldots p_n) \text{ y } Q = (q_1, \ldots, q_n)$$

de la definicion de entropia de Shannon, obtenemos

$$H_n(PQ) = H_n(P) + H_n(Q) \qquad (1)$$

que expresa una de las mas importantes propiedades de la entro-

pia, llamada, aditividad: La entropia de un experimento compuesto que consiste en la realizacion de dos experimentos independientes es igual a la suma de las entropias de estos dos experimentos (1). Esta propiedad es evidentemente mucho mas debil que el postulado D de la axiomatica de Fadeev:

$$H_{m+1}(p_1, \dots, p_{m-1}, \lambda p_m, (1-\lambda)p_m) = H_m(p_1 \dots p_m) + p_m H_2(\lambda, 1-\lambda)$$

para todo λ /$0 \leq \lambda \leq 1$.

Cantidades del tipo

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

con $0 < \alpha < 1$, verifican los postulados A, B y C. de Fadeev y la propiedad de adicion (i). Eléo induce a RENYI (1.960) a definir $H_\alpha(P)$ como la entropia de orden α de una distribucion P. Sin embargo de A, B, C y (i) no se deduce (ii), es necesario imponer algun postulado mas; para ello Renyi introduce el concepto de distribucion de probabilidad generalizada (2).

Consideremos un espacio probabilistico $(S, \mathcal{A}, P)$. Notemos por $\xi = \xi(x)$ una funcion medible con respecto a $\mathcal{A}$ definida para $x \in S_1$, donde $S_1 \in \mathcal{A}$. Si $P(S_1) > 0$, se dice que ξ es una variable aleatoria generalizada. Si $P(S_1) = 1$, ξ es una variable aleatoria completa, mientras que si $0 < P(S_1) < 1$, diremos que es una variable aleatoria incompleta.

Siguiendo la idea para distribuciones completas podemos escribir

$$w(P) = \sum_{i=1}^{n} p_i \qquad 0 < w(P) \leq 1$$

---

(1) Ver p.e. YAGLOM-YAGLOM (1.969) o KHINCHIN (1.957).

(2) Llamadas tambien distribuciones de probabilidad incompletas.

$w(P)$ puede ser considerado como el peso de la distribucion. Este peso sera igual a la unidad para distribuciones completas y menor que la unidad para distribuciones incompletas.

Renyi propone los siguientes postulados, en lugar de los anteriormente enunciados de Fadeev:

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

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

Postulado 3: $H[\{1/2\}] = 1$

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

$$w(P) + w(Q) \leq 1$$

Postulado 5:

$$H_m(P \cup Q) = \frac{w(P) H_n(P) + w(Q) H_n(Q)}{w(P) + w(Q)}$$

donde $P = (p_1, \ldots, p_n)$ y $Q = (q_1, \ldots, q_n)$ son dos distribuciones arbitrarias del espacio de todas las distribuciones de probabilidad generalizadas discretas;

$$P \cup Q = (p_1, \ldots, p_n, q_1, \ldots, q_n)$$

siendo $w(P) + w(Q) \leq 1$.

Los tres primeros postulados son los mismos que los de Faddev, y el postulado D es reemplazado por los postulados 4 y 5. El postulado 5 nos dice que la entropia de la union de dos distribuciones incompletas es la media ponderada de las entropias de las dos distribuciones, siendo los pesos los coeficientes de ponderacion.

Renyi demuestra el siguiente teorema:

2.1.1. Teorema. Si H(P) es definida para todas las distribuciones de probabilidad generalizada discretas y satisface los postulados 1, 2, 3, 4 y 5, anteriormente enunciados, entonces

$$H(P) = \frac{\sum p_i \log_2 \frac{1}{p_i}}{\sum p_i}$$

2.2. ENTROPIA DE RENYI DE ORDEN ALFA.

Si se quiere generalizar el postulado 5 para otro tipo de media que no sea la aritmetica se introduce la funcion de Kolmogorov-Naguno y se reemplaza el postulado 5 por este otro.

1. Postulado 5': Existe una funcion estrictamente monotoma y continua y = g(x) tal que si w(P) + w(Q) ≠ 1 tenemos

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

Las funciones g(x) del tipo $g_x = 2^{(\alpha-1)x}$ con $\alpha \neq 1$ son compatibles con el postulado 4. Renyi demuestra el siguiente teorema:

2.2.1. Teorema. Si H(P) es definida para todas las distribuciones de probabilidad generalizadas discretas y satisface los postulados 1, 2, 3, 4 y 5' con $g(x) = 2^{(\alpha-1)x}$, entonces

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

La cantidad $H_\alpha(P)$ es llamada por Renyi entropia de orden α de la distribucion generalizada P que en el caso particular en que α tiende a 1 coincide con la entropia de Shannon.

2.3. GENERALIZACION.

Si en lugar de restringir el intervalo de variacion 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 variacion (0,N) con N ≥ 1, obtenemos una generalizacion de la entropia de Renyi de orden α .

2.3.1. Definicion. Sea $P = (p_1,\dots,p_n)$ una distribucion de probabilidad generalizada con peso $w(P) \leq 1$. Definimos entropia de orden (α, N) de la distribucion generalizada P, y la anotaremos $H_\alpha^N(P)$ por la expresion

$$H_\alpha^N(P) = \frac{N}{N-\alpha} \log_2 \frac{\sum_{i=1}^{n} p_i^{\alpha/N}}{\sum_{i=1}^{n} p_i}$$

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

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

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

$$\lim_{\alpha \to N} H_\alpha^N(P) = \lim_{\alpha \to N} \frac{N \log \frac{\sum p_i^{\alpha/N}}{\sum p_i}}{N-\alpha} = \frac{0}{0}$$

y aplicando la regla de L'Hoppital

$$\lim_{\alpha \to N} H_\alpha^N(P) = -\sum p_i \log p_i = H(P)$$

2.3.2. Caracterizacion. La entropia de orden (α, N) puede ser caracterizada univocamente por los postulados 1, 2, 3, 4 y 5' anteriormente enunciados, eligiendo como funcion de Kolmogorov-Nagumo una funcion exponencial del tipo

$$g(x) = 2^{(\alpha/N - 1)x} \qquad \text{con} \quad \alpha/N > 1$$

En el apendice de esta Memoria incluimos un programa para el calculo mediante ordenador de la funcion entropia $H_\alpha^N(P)$ para distintos valores de α y de N. En particular y para cuatro distribuciones de probabilidad se obtiene el valor de $H_\alpha^N(P)$ en unas tablas de doble entrada para los diferentes valores de α y N.

Exponemos a continuacion, en forma de teoremas, propiedades encontradas para la funcion de entropia de orden (α, N) algunas de las cuales son casi inmediatas.

2.3.3. Teorema. La entropia de orden (α, N), $H_\alpha^N(P)$ es una funcion monotona creciente de N.

En efecto:

$$\frac{d}{dN} H_\alpha^N = \frac{-\alpha}{(N-\alpha)^2} \log_2 \frac{\sum p_i^{\alpha/N}}{\sum p_i} - \frac{\alpha}{N(N-\alpha)} \frac{\sum p_i^{\alpha/N} \log_2 p_i}{\sum p_i^{\alpha/N}} =$$

$$= \frac{-\alpha}{(N-\alpha)^2} \left[\log_2 \frac{\sum p_i^{\alpha/N}}{\sum p_i} + \frac{\sum p_i^{\alpha/N} \log_2 p_i^{1-\alpha/N}}{\sum p_i^{\alpha/N}} \right]$$

pero en virtud de la desigualdad de Jensen (3)

$$\frac{\sum p_i^{\alpha/N} \log p_i^{1-\alpha/N}}{\sum p_i^{\alpha/N}} \leq \log_2 \frac{\sum p_i^{\alpha/N} p_i^{1-\alpha/N}}{\sum p_i^{\alpha/N}} = \log_2 \frac{\sum p_i}{\sum p_i^{\alpha/N}}$$

por tanto

$$\frac{d}{dN} H_\alpha^N \geq \frac{-\alpha}{(N-\alpha)^2} \left[\log \frac{\sum p_i^{\alpha/N}}{\sum p_i} + \log \frac{\sum p_i}{\sum p_i^{\alpha/N}} \right] = 0$$

de lo cual se deduce el teorema (4).

Corolario. Para cualquier distribucion de probabilidad generalizada se verifica

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

---

(3) HARDY-LITTLEWOD-POLYA (1.952).

(4) En las tablas 1, 2, 3 y 4 del apendice puede comprobarse la validez del teorema, si para cada α determinado avanzamos hacia la derecha.

dandose la igualdad si y solo si N = 1.

La demostracion es inmediata en virtud del teorema anterior por ser para todo N, $\alpha/N \leq \alpha$

2.3.4. Teorema. La entropia de orden (α, N). $H_\alpha^N(P)$ es una funcion monotona decreciente de α .

En efecto:

$$\frac{d}{d\alpha} H_\alpha^N(P) = \frac{N}{(N-\alpha)^2} \log_2 \frac{\sum p_i^{\alpha/N}}{\sum p_i} + \frac{1}{N-\alpha} \frac{\sum p_i^{\alpha/N} \log p_i}{\sum p_i^{\alpha/N}} =$$

$$= \frac{N}{(N-\alpha)^2} \left[\log_2 \frac{\sum p_i^{\alpha/N}}{\sum p_i} + \frac{\sum p_i^{\alpha/N} \log p_i^{1-\alpha/N}}{\sum p_i^{\alpha/N}} \right]$$

y aplicando al segundo sumando de la derecha la desigualdad de Jensen como antes lo hicimos, obtenemos

$$\frac{d}{d\alpha} H_\alpha^N(P) \leq \frac{N}{(N-\alpha)^2} \left[\log \frac{\sum p_i^{\alpha/N}}{\sum p_i} + \log \frac{\sum p_i}{\sum p_i^{\alpha/N}} \right] = 0$$

de lo cual se deduce el teorema (5).

2.3.5. Teorema. Sea $P = (p_1, \ldots, p_n)$ una distribucion de probabilidad generalizada con peso $w(P) = \lambda \neq 1$. La entropia $H_\alpha^N(P)$ verifica la siguiente desigualdad

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

En efecto:

Para $\alpha = 0$, se tiene

$$H_0^N(P) = \log_2 n/\lambda$$

y por el teorema anterior

$$\alpha \geq 0 \Rightarrow H_0^N(P) \geq H_\alpha^N(P)$$

deducimos, por tanto,

---

(5) La validez de este teorema puede comprobarse experimentalmente si para un N fijo nos movemos en las tablas 1, 2, 3 y 4 del apendice en sentido descendente.

$$H_\alpha^N(P) \leq \log_2 m/\lambda$$

Por otro lado, por ser $\alpha/N < 1$, $p_i \leq 1$ $\forall i$

$$p_i^{\alpha/N} \geq p_i$$

y sumando en i

$$\sum p_i^{\alpha/N} \geq \sum p_i$$

de donde deducimos

$$\log \frac{\sum p_i^{\alpha/N}}{\sum p_i} \geq 0$$

y por tanto $H_\alpha^N(P) \geq 0$.

En definitiva hemos demostrado

$$0 \leq H_\alpha^N(P) \leq \log_2 m/\lambda$$

en particular si P es completa se verifica

$$0 \leq H_\alpha^N(P) \leq \log_2 m$$

2.3.6. <u>Teorema.</u> Si existe una distribucion $P = (p_1,\ldots,p_n)$ siendo $p_1 = p_2 = \ldots p_n = p$, la entropia $H_\alpha^N(P)$ alcanza el maximo $\log_2 n/\lambda$ independientemente de los valores de α y N, siendo $w(P) = \lambda$.

En efecto:

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

$$= \frac{N}{N-\alpha} \log_2 p^{\alpha/N-1} = -\log_2 p = \log_2 n/\lambda$$

En el caso particular en que P sea una distribucion completa $w(P) = 1$ y $H_\alpha^N(P) = \log_2 n$ (6).

Es interesante notar que en las condiciones del teorema, las entropias de Shannon, Renyi de orden α y la de orden (α, N),

---

(6) Ver tabla 1 del apendice.

coinciden, es decir

$$H_s(P) = H_\alpha(P) = H_\alpha^N(P) = -\log_2 P$$

2.4. CANTIDAD DE INFORMACION DE ORDEN α.

2.4.1. Definicion. Dadas dos distribuciones de probabilidad generalizadas $P = (p_1 \dots p_n)$ y $Q = (q_1 \dots q_n)$ (7) define Renyi, informacion de orden α obtenida cuando la distribucion P es reemplazada por la distribucion Q como la cantidad

$$I_\alpha(Q/P) = \frac{1}{1-\alpha} \log_2 \frac{\sum \frac{q_i^\alpha}{p_i^{\alpha-1}}}{\sum q_i}$$

Cuando $\alpha \to 1$

$$I_1(Q/P) = \frac{\sum q_i \log_2 \frac{q_i}{p_i}}{\sum q_i}$$

que coincide con las clasicas definiciones (KULLBACK, 1.959) para distribuciones de probabilidad generalizadas.

2.4.2. Caracterizacion. Si una cantidad I(Q/P) verifica los siguientes postulados:

Postulado 6: I(Q/P) es invariante ante las reordenaciones de los elementos de Q y P siempre que la correspondencia uno a uno entre ellas no varie.

Postulado 7: Si $P = (p_1, \dots p_n)$ y $Q = (q_1, \dots q_n)$ y $p_i \leq q_i$ entonces $I(Q/P) \geq 0$, mientras que si $p_i \geq q_i$ entonces $I(Q/P) \leq 0$.

Postulado 8: $I\left[\frac{1}{2}, \frac{1}{2} / \frac{1}{4}, \frac{1}{4}\right] = 1$

Postulado 9: Si $I(Q_1/P_1)$ y $I(Q_2/P_2)$ estan definidas y si $P = P_1P_2$ y $Q = Q_1Q_2$ y la correspondencia entre los elementos de P y Q esta inducida por los de P_1 y Q_1, P_2 y Q_2, entonces

$$I(Q/P) = I(Q_1/P_1) + I(Q_2/P_2)$$

Postulado 10: Si existe una funcion monotona estrictamente creciente $y = g(x)$ y si $I(Q_1/P_1)$ y $I(Q_2/P_2)$ estan definidas siendo $0 < w(P_1) + w(P_2) \leq 1$ y $0 < w(Q_1) + w(Q_2) \leq 1$ y la correspondencia entre $P_1 \cup P_2$ y $Q_1 \cup Q_2$ esta determinada por las correspondientes entre los elementos de P_1 y Q_1 y los de P_2 y Q_2 entonces

$$I[Q_1 \cup Q_2 / P_1 \cup P_2] = g^{-1}\left[\frac{w(Q_1)\, g[I(Q_1/P_1)] + w(Q_2)\, g[I(Q_2/P_2)]}{w(Q_1) + w(Q_2)}\right]$$

Renyi demuestra el siguiente teorema:

2.4.3. Teorema. Si la cantidad $I(Q/P)$ satisface los postulados 6,7,8,9 y 10 y tomamos como funcion $g(x) = 2^{(\alpha-1)x}$ entonces

$$I_\alpha(Q/P) = \frac{1}{1-\alpha} \log_2 \frac{\sum q_i^\alpha / p_i^{\alpha-1}}{\sum p_i}$$

2.5. GENERALIZACION.

2.5.1. Definicion. Definimos informacion de orden (α, N) cuando la distribucion P es reemplazada por Q a la cantidad

$$I_\alpha^N(Q/P) = \frac{N}{N-\alpha} \log_2 \frac{\sum \frac{q_i^{\alpha/N}}{p_i^{\alpha/N-1}}}{\sum p_i}$$

con $0 < \alpha < N$, $N \geq 1$, $0 < w(P) + w(Q) \leq 1$

Cuando $\alpha \to N$ se tiene, evidentemente

$$I_\alpha^N(Q/P) = \frac{\sum q_i \log_2 q_i/p_i}{\sum q_i}$$

---

(7) Si ocurre que P y Q no tienen el mismo numero de componentes añadimos los ceros necesarios para que el numero de componentes de ambos coincida y asi poder establecer una correspondencia univoca entre los elementos de P y Q.

2.5.2. Caracterizacion. Los postulados 6, 7, 8, 9 y 10 y la funcion de Kolmogorov-Naguno $g(x) = 2^{(\alpha/N-1)x}$ con $\alpha/N > 1$ caracterizan la cantidad de informacion aqui definida.

REFERENCIAS DEL CAPITULO

FADEEV (1.956)."On the concept of the entropy for a finite probability model" Uspehi Mat. Nauk. 11 (227-231)

HARDY-LITTLEWOOD-POLYA (1.934)"Inequalities" Ed. Cambridge University Press.

KHINCHIN (1.957)."Mathematical Foundations of Information Theory" ed. Dover.

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

RENYI (1.960)."On measures of entropy and information". Proceedings 4th. Berkeley Symposium on Probability and Statistics Vol I (547-561)

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. Ma Hung. 2 (249-256)

SHANNON (1.948)." A mathematical Theory of Communication" Bell System Technical Journal 27 (379-423, 623-656). Reeditado por University of Illinois Press en 1.969 Shannon-Weaver: A Mathematical Theory of Communication).

YAGLOM-YAGLOM (1.969). "Probabilite et Information". Ed. Dunod.

CAPITULO TERCERO

I N F O R M A C I O N Y

E X P E R I M E N T A C I O N

CONTENIDO.

3.1 - Definicion de experimentos.

3.2 - Medida de la informacion propercionado por un experimento.

- Relaciones: Informacion de Shannon.

Informacion de Kullback.

Informacion de Renyi.

- Propiedades: No negatividad.

Generalmente finita.

Invarianza.

Suficiencia.

3.3 - Union de experimentos.

- Teorema de adicion.

- Otros teoremas.

3.4 - Media ponderada de dos experimentos.

- Teoremas.

3.5 - Mixtura de dos experimentos.

- Teoremas.

3.6 - Comparacion de experimentos.

- Teoremas.

3.7 - Suficiencia de experimentos.

- Teoremas.

3.8 - Valor de la informacion de un experimento.

- Analogia.

3. 9 - Capacidad de un experimento.

CAPITULO TERCERO.

INFORMACION Y EXPERIMENTACION.

Partiendo del concepto de incertidumbre de Shannon llegamos a encontrar una medida de la información proporcionada por un experimento estadistico. Estudiamos la relacion entre la medida encontrada y las cantidades de informacion de Kullback, Shannon y Renyi.

Demostramos a continuación algunas propiedadesy teoremas sobre la medida obtenida, empleandola despues como metodo para comparar experimentos. Mas adelante se estudia la analogia entre el concepto utilitarista de valor de la informacion asociada con un experimento y la medida aqui obtenida.

Al final del capitulo, maximizando esta informacion, definimos capacidad de un experimento.

3.1 DEFINICION DE EXPERIMENTO.

Consideremos una variable aleatoria X. A la observacion de esta variable se le llama realizacion de un experimento con dicha variable aleatoria. Sea S el espacio formado por todos los resultados de las posibles observaciones de X, que llamaremos espacios de resultados. Sea $\mathcal{A}$ el σ-algebra definido sobre S. Sea Ω un espacio parametrico 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 continuas con respecto a una medida dominante μ sobre $\mathcal{A}$. Podemos describir cada medida de probabilidad por una funcion 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 Lebesge, escribiremos

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

La cupla

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

caracteriza un experimento.

3.2 MEDIDA DE LA INFORMACION PROPORCIONADA POR UN EXPERIMENTO.

Supongamos que existe una distribucion a priori sobre Ω, y supondremos de nuevo que puede ser descrita por una funcion de densidad $p(\theta)$ (1), con respecto a una medida dominante y que por sencillez notamos $d\theta$. Esta $p(\theta)$ resumira las opiniones iniciales del experimentador sobre el valor del parametro, antes de realizar el experimento.

Existe, por tanto, una incertidumbre inicial sobre el valor de θ que podemos expresar por

$$I(\theta) = -\int_\Omega p(\theta) \log p(\theta)\, d\theta \qquad (2)$$

Una vez realizado el experimento E y obtenido un resultado x, el experimentador tendra una incertidumbre sobre el valor de

que vendra dada por la expresion

$$I(\theta/x) = -\int_\Omega p(\theta/x) \log p(\theta/x)\, dx \qquad (3)$$

---

(1) Por comodidad notaremos $p(\cdot)$ a todas las funciones de probabilidad (funciones de densidad en el caso continuo).

(2) LINDLEY (1.956) no introduce el signo menos al definir la incertidumbre, pues segun él, la maxima informacion en sentido estadistico sera obtenida cuando toda la distribucion de probabilidad este concentrada sobre un unico valor θ y la informacion sera menor cuando la distribucion de θ se disperse; situacion inversa a la que ocurre en teoria de la communicacion, donde la concentracion sobre un unico valor implicaria la no existencia de eleccion en el mensaje transmitido y por tanto la incertidumbre seria minima.

donde $p(\theta/x)$ en virtud del teorema de Bayes viene dada por

$$p(\theta/x) = \frac{P(x/\theta)\, p(\theta)}{p(x)} \quad \text{, siendo} \quad p(x) = \int_\Omega p(x/\theta)\, p(\theta)\, d\theta$$

Es logico definir la informacion 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 informacion que este - puede proporcionar vendra dada por el valor medio extendido a todos los posibles resultados de E, de la informacion proporcionada por uno de ellos. Por tanto podemos definir:

3.2.1. Definicion. La incertidumbre del experimentador sobre el - valor del parametro antes de realizar un experimento E, o lo - que es equivalente, la informacion obtenida una vez realizado E, suponiendo que existe una distribucion a priori $p(\theta)$ sobre el parametro viene dada por

$$I(\theta \parallel X) = E_x \left[I(\theta) - I(\theta/x) \right]$$

Desarrollando tendremos

$$I(\theta \| X) = \int_S \left[-\int_\Omega p(\theta) \log p(\theta)\, d\theta + \int_\Omega p(\theta/x) \log p(\theta/x)\, d\theta \right] p(x)\, dx =$$

$$= -\int_\Omega p(\theta) \log p(\theta)\, d\theta + \int_S \int_\Omega p(\theta, x) \log p(\theta/x)\, d\theta\, dx =$$

$$= -\int_S \int_\Omega p(\theta, x) \log p(\theta)\, d\theta\, dx + \int_S \int_\Omega p(\theta, x) \log p(\theta/x)\, d\theta\, dx =$$

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

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

---

(3) Notese que esta informacion o incertidumbre no tiene por que ser positiva. Puede presentarse un resultado porpresa que nos haga dudar aun mas sobre θ una vez realizado el experimento, que antes de su realizacion.

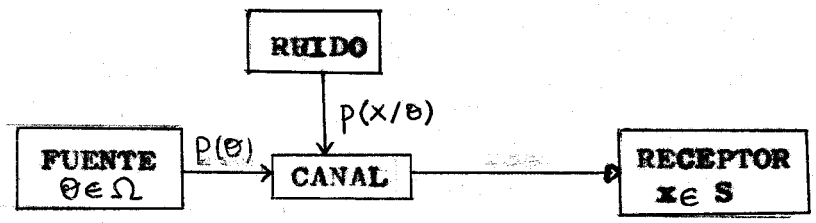
$$I(\theta \| X) = \int_S \int_\Omega p(\theta,x) \log \frac{p(\theta,x)}{p(\theta)\,p(x)}\, d\theta\, dx =$$

$$= \underset{x,\theta}{E}\left[\log \frac{p(\theta,x)}{p(\theta)\,p(x)}\right] = \underset{x,\theta}{E}\left[i(\theta,x)\right]$$

siendo $i(\theta,x)$ la densidad de informacion de las variables θ y X (DOBRUSHIN, 1.959).

3.2.2 Relaciones.

3.2.2.1 - Informacion de Shannon. Podemos establecer una correspondencia o analogia entre la nocion de experimento dada anteriormente y la de un sistema de comunicacion con ruido. En efecto, consideremos como fuente el conjunto Ω de simbolos θ. Estos simbolos son emitidos por la fuente de manera que la eleccion de es hecha con una probabilidad $p(\theta)$, siendo independientes las sucesivas elecciones. Se supone que el ruido del canal es tal que los simbolos son perturbados segun $p(x/\theta)$ obteniendose 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 transicion $p(x/\theta)$ que nos indican la probabilidad de obtener un resultado x cuando sea emitido el simbolo θ de Ω.



La razon de transmision de un canal con ruido vendra dada (SHANNON 1.948) por la diferencia entre la razon de produccion (es decir la incertidumbre de la fuente) y la razon media de incertidumbre condicional (llamada por conveniencia equivocacion), es decir

$$R(\theta,X) = \int_\Omega p(\theta) \log p(\theta)\, d\theta - \int_S \int_\Omega p(x,\theta) \log p(\theta/x)\, dx =$$

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

La razon de transmision de informacion de Shannon coincide - pues con la informacion que sobre θ nos aporta el experimento E.

3.2.2.2. Informacion de Kullback. Si tomamos como hipotesis H_2 que las variables θ y x son independientes con distribuciones de probabilidad marginales $p(\theta)$ y $p(x)$ respectivamente y como - hipotesis H_1 que las variables θ y x son dependientes con distribucion de probabilidad conjunta $p(\theta,x)$, entonces la informacion de Kullback para discriminar en favor de H_1 contra H_2 (KULLBACK, 1.959) vendra dada por

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

expresion que coincide con la medida de informacion encontrada por nosotros.

3.2.2.3. Informacion de Renyi. Si notamos por P la distribucion de una variable aleatoria X y por Q la distribucion de X condicionada por el hecho de que un suceso A ha ocurrido, RENYI (1960) define la cantidad de informacion sobre X contenida en la observacion del suceso A por

$$I(Q\|P) = \int \log h \, dQ = \int h \log h \, dP$$

siendo $Q \gg P$ y h la derivada de Radon-Nikodym de Q con respecto a P.

Esta cantidad de informacion es analoga a la definida por nosotros. La unica diferencia consiste en considerar un unico suceso mientras que aqui nosotros consideramos un experimento del cual se pueden obtener un numero finito o infinito de resultados y por tanto tomamos como medida de informacion la media extendida a todos los posibles resultados del experimento, es decir $E[I(Q\|P)]$

3.2.3. Ejemplos.

A) - Supongamos un canal de transmision con ruidos. Supongamos que los simbolos son emitidos por la fuente segun una dis-

tribucion Beta de parametros α y β . Si a lo largo del canal estos simbolos son perturbados segun una distribucion Bernouilli de parametro θ , vamos a calcular la razon de transmision de informacion del canal.

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{Be(\alpha,\beta)} \quad ; \quad p(x/\theta) = \theta^{x}(1-\theta)^{1-x}$$

$$p(x) = \int_0^1 p(\theta)\, p(x/\theta)\, d\theta = \frac{1}{Be(\alpha,\beta)} \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta-x} d\theta$$

$$= \frac{Be(\alpha+x\,;\,\beta-x+1)}{Be(\alpha,\beta)}$$

y por tanto,

$$p(x) = \frac{Be(\alpha,\beta+1)}{Be(\alpha,\beta)} = \frac{\beta}{\alpha+\beta} \qquad \text{para } x = 0$$

$$p(x) = \frac{Be(\alpha+1,\beta)}{Be(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta} \qquad \text{para } x = 1$$

$$p(x/\theta) = \frac{1-\theta}{\beta/\alpha+\beta} \qquad \text{para } x = 0$$

$$p(x/\theta) = \frac{\theta}{\alpha/\alpha+\beta} \qquad \text{para } x = 1$$

podemos escribir

$$\sum_x p(x/\theta) \log \frac{P(x/\theta)}{p(x)} = (1-\theta)\log\frac{1-\theta}{\beta/\alpha+\beta} + \theta \log \frac{\theta}{\alpha/\alpha+\beta}$$

Lema. Si X es una variable aleatoria distribuida segun una Be(α,β) entonces

$$E(\log x) = \Psi(\alpha) - \Psi(\alpha+\beta)$$

$$E[\log(1-x)] = \Psi(\beta) - \Psi(\alpha+\beta)$$

viniendo la funcion Ψ definida por

$$\Psi(z) = \frac{d}{dz}\,\Gamma(z)$$

funcion que es conodida con el nombre de "funcion digamma" (TRIBUS, 1.972).

Teniendo en cuenta el lema anterior, podremos escribir:

$$I(\theta\|X) = \int_0^1 \frac{\theta^{\alpha-1}(1-\theta)^{\beta}}{Be(\alpha,\beta)} \log(1-\theta)\, d\theta +$$

$$+ \int_0^1 \frac{\theta^{\alpha}(1-\theta)^{\beta-1}}{Be(\alpha,\beta)} \log\theta\, d\theta +$$

$$+ \log\frac{\alpha+\beta}{\beta} \int_0^1 \frac{\theta^{\alpha-1}(1-\theta)^{\beta}}{Be(\alpha,\beta)} d\theta +$$

$$+ \log\frac{\alpha+\beta}{\alpha} \int_0^1 \frac{\theta^{\alpha}(1-\theta)^{\beta-1}}{Be(\alpha,\beta)} d\theta =$$

$$= \frac{Be(\alpha,\beta+1)}{Be(\alpha,\beta)} \left[\Psi(\beta+1) - \Psi(\alpha+\beta+1)\right] +$$

$$+ \frac{Be(\alpha+1,\beta)}{Be(\alpha,\beta)} \left[\Psi(\alpha+1) - \Psi(\alpha+\beta+1)\right] +$$

$$+ \frac{Be(\alpha,\beta+1)}{Be(\alpha,\beta)} \log\frac{\alpha+\beta}{\beta} + \frac{Be(\alpha+1,\beta)}{Be(\alpha,\beta)} \log\frac{\alpha+\beta}{\alpha}$$

en definitiva,

$$I(\theta\|X) = \frac{\beta}{\alpha+\beta}\Psi(\beta+1) + \frac{\alpha}{\alpha+\beta}\Psi(\alpha+1) - \Psi(\alpha+\beta+1) +$$

$$+ \frac{\beta}{\alpha+\beta}\log\frac{\alpha+\beta}{\beta} + \frac{\alpha}{\alpha+\beta}\log\frac{\alpha+\beta}{\alpha}$$

B) Consideremos un experimento E, caracterizado por la cupla

$$E = \left[(\mathbb{R},\mathcal{B}),\ \left\{P(x/\theta) = \frac{1}{b} \quad / x - b/2 < \theta < x + b/2\ ;\ \theta \in \mathbb{R}^+\right\}\right]$$

y supongamos que sobre el espacio parametrico hay definida una distribucion a priori dada por una ley exponencial negativa de parametro a, es decir,

$$p(\theta) = a\, e^{-a\theta} \qquad 0 \leq \theta < \infty$$

Vamos a calcular la informacion que sobre el valor del parametro θ nos proporciona el experimento E.

$$p(x) = \int_0^{\infty} p(\theta)\, p(x/\theta)\, d\theta =$$

$$P(x)=\begin{cases} 0 & \text{para } x < -b/2 \\ \int_0^{x+b/2} \frac{a}{b} e^{-a\theta} d\theta = \frac{1}{b}\left(1-e^{-a(x+b/2)}\right) & \text{para } -b/2 < x < b/2 \\ \int_{x-b/2}^{x+b/2} \frac{a}{b} e^{-a\theta} d\theta = \frac{e^{-ax}}{b}\left(e^{ab/2}-e^{-ab/2}\right) & \text{para } x > b/2 \end{cases}$$

por tanto

$$\log \frac{P(x/\theta)}{P(x)} \begin{cases} -\log\left(1-e^{-a(x+b/2)}\right) & \text{para } -b/2 < x < b/2 \\ -\log\left[e^{-ax}\left(e^{ab/2}-e^{-ab/2}\right)\right] & \text{para } x > b/2 \end{cases}$$

de lo que se sigue

$$I(\theta\|X) = -\int_{-b/2}^{b/2}\int_{0}^{x+b/2} a e^{-a\theta} \frac{1}{b} \log\left(1-e^{-a(x+b/2)}\right) d\theta\, dx + -\int_{-b/2}^{\infty}\int_{x-b/2}^{x+b/2} a e^{-a\theta} \frac{1}{b} \log\left[e^{-ax}\left(e^{ab/2}-e^{-ab/2}\right)\right] d\theta\, dx = I_1 + I_2$$

Resolvamos en primer lugar I_1. Haciendo el cambio

$$\left.\begin{array}{rl} 1-e^{-a[x+b/2]} &= u \\ \theta &= \theta \end{array}\right\} \quad J = \frac{1}{a} e^{a[x+b/2]} = \frac{1}{a(1-u)}$$

y los nuevos limites de integracion para u, seran

$$x = b/2 \Longrightarrow u = 1-e^{-ab}$$

$$x = -b/2 \Longrightarrow u = 0$$

por tanto

$$I_1 = -\int_0^{1-e^{-ab}} \int_0^{-\frac{\ln(1-u)}{a}} a e^{-a\theta} \frac{1}{b} \log u \frac{1}{a(1-u)} d\theta\, du =$$

$$= -\int_0^{1-e^{-ab}} \frac{1}{ab} \frac{u \log u}{1-u} du$$

Resolvamos ahora I_2. Haciendo el cambio

$$\left.\begin{array}{rl} e^{-ax}\left[e^{ab/2}-e^{-ab/2}\right] &= u \\ \theta &= \theta \end{array}\right\} \quad J = \frac{1}{au}$$

y los nuevos limites de integracion para u seran

$$x = b/2 \implies u = 1 - e^{-ab}$$

$$x = \infty \implies u = 0$$

por tanto

$$I_2 = -\int_{1-e^{-ab}}^{0} \int_{-\frac{1}{a}\ln\frac{u\,e^{ab}}{e^{ab}-1}}^{-\frac{1}{a}\ln\frac{u\,e^{-ab}}{1-e^{-ab}}} \frac{a}{b} e^{-a\theta} \log u \, \frac{1}{au} \, d\theta \, du =$$

$$= \frac{-1}{ab} \int_0^{1-e^{-ab}} \frac{\log u}{u} \, u \, du = \frac{-1}{ab} \int_0^{1-e^{-ab}} \log u \, du$$

y sumando ambas integrales I_1 y I_2

$$I(\theta \| X) = \frac{-1}{ab} \int_0^{1-e^{-ab}} \left(\log u + \frac{u \log u}{1-u} \right) du$$

$$= \frac{-1}{ab} \int_0^{1-e^{-ab}} \frac{\log u}{1-u} \, du$$

haciendo el cambio $(1 - u) = t$

los nuevos limites de integracion seran

$$u = 0 \implies t = 1$$

$$u = 1 - e^{-ab} \implies t = e^{-ab}$$

por tanto

$$I(\theta \| X) = \frac{1}{ab} \int_1^{e^{-ab}} \frac{\log(1-t)}{t} \, dt$$

Desarrollando en serie de MacLaurin $\log(1-t)$

$$\int_1^{e^{-ab}} \frac{\log(1-t)}{t} = \sum_{k=1}^{\infty} \frac{e^{-kab} - 1}{k^2}$$

y esto siempre que la serie sea absolutamente convergente, para lo cual bastara con que impogamos que b es un infinitesimo, con lo que $ab \longrightarrow 0$.

En definitiva tendremos

$$I(\theta \| X) = \frac{1}{ab} \sum_{k=1}^{\infty} \frac{1 - e^{-kab}}{k^2} \qquad (\text{con } b \longrightarrow 0)$$

6) - Consideremos el siguiente experimento:

$$E = \left[(\mathbb{R}, \mathcal{B}); \left\{ p(x/\theta) \longrightarrow N(\theta, \sigma_2^2) \;/\; \theta \in \mathbb{R} \right\} \right]$$

y supongamos que sobre θ hay definida una ley normal de parametros μ y σ_1^2, a priori. Calculemos la informacion que sobre el parametro nos proporciona el experimento.

$$p(\theta) = \frac{1}{\sigma_1 \sqrt{2\pi}} \; e^{-\frac{(\theta-\mu)^2}{2\sigma_1^2}}$$

$$p(x/\theta) = \frac{1}{\sigma_2 \sqrt{2\pi}} \; e^{-\frac{(x-\mu)^2}{2\sigma_2^2}}$$

$$p(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_1 \sigma_2 \sqrt{2\pi}} \exp\left\{ -1/2 \left[\frac{(\theta-\mu)^2 \sigma_1^2 + (x-\theta)^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \right] \right\} d\theta =$$

$$= \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sigma_1 \sigma_2 \sqrt{2\pi}} \; \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left\{ -1/2 \left(\theta - \frac{\mu\sigma_2^2 + x\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \right)^2 \right\} =$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left\{ -\frac{1}{2} \left[\frac{(x-\mu)^2}{\sigma_1^2 + \sigma_2^2} \right] \right\}$$

$$\frac{p(x/\theta)}{p(x)} = \sqrt{1 + \frac{\sigma_1^2}{\sigma_2^2}} \exp\left\{ 1/2 \left[\frac{(x-\theta)^2}{\sigma_2^2} + \frac{(x-\mu)^2}{\sigma_1^2 + \sigma_2^2} \right] \right\}$$

$$\log \frac{p(x/\theta)}{p(x)} = \frac{1}{2} \left\{ \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right) \right\} + \frac{1}{2} \left\{ \frac{-(x-\theta)^2}{\sigma_2^2} + \frac{x^2}{\sigma_1^2 + \sigma_2^2} + \frac{\mu^2}{\sigma_1^2 + \sigma_2^2} - \frac{2x\mu}{\sigma_1^2 + \sigma_2^2} \right\} \log$$

$$\int_X p(x/\theta) \log \frac{p(x/\theta)}{p(x)} dx = \frac{1}{2} \left\{ \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right) + \frac{(\theta-\mu)^2}{\sigma_1^2 + \sigma_2^2} \log e - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \log e \right\}$$

En definitiva

$$I(\theta \| X) = \frac{1}{2} \int_{\Omega} p(\theta) \left\{ \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right) + \left[\frac{(\theta-\mu)^2}{\sigma_1^2+\sigma_2^2} - \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \right] \log e \right\} d\theta =$$

$$= \frac{1}{2} \left\{ \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right) + \left[\frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} - \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} \right] \log e \right\}$$

$$= \frac{1}{2} \log\left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right)$$

3.2.4. Propiedades.

3.2.4.1. No negatividad. Veamos que la cantidad de informacion anteriormente definida es no negativa. Para ello necesitamos probar el siguiente lema:

Lema 3.2.4.1. Sean $p(x,y)$ y $q(x,y)$ dos funciones tales que

$$\iint p(x,y)\,dx\,dy = 1 \quad y \quad \iint q(x,y)\,dx\,dy = 1$$; se verifica que

$$\iint p(x,y) \log \frac{q(x,y)}{p(x,y)}\,dx\,dy \leq 0$$

dandose la igualdad si y solo si $p(x,y) = q(x,y)$ salvo conjuntos de medida nula.

En efecto:

Evidentemente si $f(x,y)$ es una funcion real se verifica que $\ln f(x,y) = f(x,y) - 1$, con igualdad si solo si $f(x,y) = 1$. Tenemos

$$\ln \frac{q(x,y)}{p(x,y)} \leq \frac{q(x,y)}{p(x,y)} - 1 = \frac{q(x,y) - p(x,y)}{p(x,y)}$$

$$p(x,y) \ln \frac{q(x,y)}{p(x,y)} \leq q(x,y) - p(x,y)$$

dandose la igualdad si y solo si $p(x,y) = q(x,y)$ salvo conjuntos de medida nula. Integrando en x y en y, tendremos

$$\iint p(x,y) \ln \frac{q(x,y)}{p(x,y)}\,dx\,dy \leq \iint q(x,y)\,dx\,dy - \iint p(x,y)\,dx\,dy = 1 - 1 = 0$$

y ya que los $f(x,y) = \log e \ \ln f(x,y)$, podemos generalizar la desigualdad anterior a un logaritmo de cualquier base. Es decir,

$$\iint p(x,y) \log \frac{q(x,y)}{p(x,y)}\,dx\,dy \leq 0$$

dandose la igualdad si y solo si p(x,y) = q(x,y) salvo conjuntos de medida nula.

<u>Teorema 3.2.4.1.</u> La cantidad de informacion $I(\theta \| X)$ es una cantidad no negativa dandose la igualdad a cero si y solo si $p(x/\theta)$ es independiente de θ , salvo conjuntos de medida nula.

En efecto:

En virtud del lema 3.2.4.1. escribiremos:

$$\iint p(x,\theta) \log \frac{P(x)\, p(\theta)}{p(x,\theta)}\, dx\, d\theta \leq 0$$

de donde deducimos inmediatamente

$$I(\theta \| X) = \iint p(x,\theta) \log \frac{P(x,\theta)}{P(x)\, p(\theta)} \geq 0$$

dandose la igualdad si y solo si $p(x,\theta) = p(x)\, p(\theta)$ lo que equivale a decir que $p(x/\theta) = p(x)$, salvo conjuntos de medida nula.

Podemos tambien demostrar el teorema sin necesidad de recurrir al lema anterior en la forma siguiente:

$$I(\theta \| X) = \int p(x) \int p(\theta/x) \log \frac{P(\theta/x)}{p(\theta)}\, d\theta\, dx$$

consideremos ahora la funcion $f(x,\theta)$ definida como sigue:

$$f(x,\theta) = p(\theta/x) \log_e p(\theta/x) - p(\theta/x) \log_e p(\theta) - p(\theta/x) + p(\theta)$$

por ser $\int_\Omega p(\theta/x)\, d\theta = \int_\Omega p(\theta)\, d\theta = 1$, de lo anterior se sigue

$$\int_\Omega f(x,\theta)\, d\theta = \int_\Omega p(\theta/x) \log_e \frac{P(\theta/x)}{p(\theta)}\, d\theta$$

que tenemos que probar, es no negativo.

Derivando en la expresion de $f(x,\theta)$, con respecto a $p(\theta/x)$ para calcular sus maximos o minimos, tenemos

$$\frac{\partial f(x,\theta)}{\partial p(\theta/x)} = \log_e p(\theta/x) - \log_e p(\theta)$$

$$\frac{\partial^2 f(x,\theta)}{\partial\, p(\theta/x)^2} = \frac{1}{P(\theta/x)}$$

Vemos que la primera derivada se anula para $p(\theta/x) = p(\theta)$ y de la expresion de la segunda derivada se sigue que en este caso

tenemos un minimo para $f(x,\theta)$. Pero cuando $p(\theta/x) = p(\theta)$ entonces $f(x,\theta) = 0$. Deducimos, por tanto, que nunca puede ocurrir que $f(x,\theta) < 0$. De donde concluimos $I(\theta \| X) \geq 0$.

3.2.4.2. Generalmente finita. Veamos que $I(\theta \| X)$ es generalmente finita. Consideremos la expresion

$$\log \frac{P[\theta < \Theta < \theta+\Delta\theta,\ x < X < x+\Delta x]}{P[\theta < \Theta < \theta+\Delta\theta]\,P[x < X < x+\Delta x]}$$

que es una extension directa del caso en que Ω y S sean discretos. Cuando $\Delta\theta$ y Δx se hacen tender a cero, cada uno de estos terminos probabilisticos tienden a cero (4). Mientras que cada uno de estos terminos individuales tienden a cero, la razón anterior permanece finita para todos los casos interesantes (KOLMOGOROV, 1.956). De hecho, la expresion anterior se convierte en el limite en

$$\log \frac{P(\theta, x)}{p(\theta)\,p(x)}$$

Es razonable excluir casos degenerados correspondientes a densidades que no son absolutamente continuas.

3.2.4.3. Invarianza. Veamos que $I(\theta \| X)$ permanece invariante bajo todas las transformaciones lineales-escalares en θ y x. Sea

$$\omega = a_1\theta + a_2$$
$$y = b_1 x + b_2$$

y sean $p_1(\omega)$ y $p_1(y)$ las funciones de densidad asociadas con las variables ω e y, respectivamente y sea $p_1(\omega,y)$ su funcion de densidad conjunta. Entonces teniendo en cuenta las anterieres transformaciones seran

$$P_1(\omega) = \frac{P(\theta)}{|a_1|} \quad ; \quad P_1(y) = \frac{P(x)}{|b_1|}$$

---

(4) Esta es la razón por la que la incertidumbre $I(\theta)$ puede hacerse infinita en el caso continuo, cosa que no puede ocurrir en el modelo discreto.

$$P_1(\omega,y) = \frac{P(\theta,x)}{\begin{vmatrix} a_1 & 0 \\ 0 & b_1 \end{vmatrix}} = \frac{P(\theta,x)}{|a_1 b_1|}$$

y por tanto escribiremos

$$I(\omega) = \int_\Omega P_1(\omega) \log P_1(\omega)\, d\omega = I(\theta) + \log|a_1|$$

$$I(y) = \int_S P_1(y) \log P_1(y)\, dy = I(x) + \log|b_1|$$

$$I(\omega,y) = \int_S \int_\Omega P_1(\omega,y) \log P_1(\omega,y)\, d\omega dy = I(\theta,x) + \log|a_1 b_1|$$

de donde deducimos finalmente

$$I(\omega \| Y) = I(\theta \| X)$$

3.2.4.4.- <u>Suficiencia.</u> Consideremos una funcion medible g definida sobre el σ-algebra $\mathcal{A}$ de elementos de S. La variable aleatoria g(X) sera pues un estadistico. Veamos que si despues de observar X solo consideramos el valor del estadistico g(X) y tenemos en cuenta unicamente la informacion contenida en g(X) y nos olvidamos de la contenida en X, obtenemos una perdida de informacion que sera nula si y solo si g(X) es un estadistico suficiente. Para probar esto, necesitamos del siguiente lema que es una consecuencia inmediata de la definicion de probabilidad condicional.

<u>Lema 3.2.4.2.</u> Si f es una funcion medible y A cualquier elemento del σ-algebra de subconjuntos definida sobre Ω y P la medida de probabilidad definida sobre este σ-algebra, se verifica

$$E\left\{f[g(X)]P[A/g(X)]\right\} = E\left\{f[g(X)]P[A/X]\right\}$$

<u>Teorema 3.2.4.2.</u> Si g es una funcion medible definida sobre el σ-algebra $\mathcal{A}$ de subconjuntos de S, se verifica que

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

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

En efecto:

En virtud del lema 3.2.4.1 podemos escribir

$$\int_S \int_\Omega p(x)\ p(\theta/x) \log \frac{p(x)\ p[\theta/g(x)]}{p(x)\ p(\theta/x)}\ d\theta\, dx \leq 0$$

dandose la igualdad si y solo si p /g(x) = p(/x), salvo conjuntos de medida nula.

De lo anterior se deduce

$$\int_S \int_\Omega p(x)\, p(\theta/x) \log p[\theta/g(x)]\, d\theta\, dx \leq \int_S \int_\Omega p(x)\, p(\theta/x) \log p(\theta/x)\, d\theta\, dx$$

y en virtud del lema 3.2.4.2.

$$E_x\left\{-I[\theta/g(x)]\right\} \leq E_x\left\{-I(\theta/x)\right\}$$

a lo que es equivalente

$$E_x\left\{I(\theta/x)\right\} \leq E_x\left\{I[\theta/g(x)]\right\}$$

de donde se deduce trivialmente

$$I[\theta \| g(X)] \leq I(\theta \| X)$$

dandose la igualdad si y solo si p $[\theta/g(x)]$ = p(θ/x) que es precisamente la caracterizacion bayesiana de estadistico suficiente, que es equivalente a las clasicas definiciones de suficiencia de Fisher y Neyman (WILKS, 1.962; RAIFFA-SCHLAIFER, 1.961 y ZACKS, 1.971)

3.3. UNION DE DOS EXPERIMENTOS.

3.311. <u>Definicion.</u> Consideremos dos experimentos E_1 y E_2 caracterizado por las cuplas

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

ambos con el mismo espacio parametrico Ω.

Llamamos union de estos dos experimentos al experimento compuesto (E_1, E_2) caracterizado por la cupla

$$(E_1, E_2) = \left[(S_1 \times S_2\ ,\ \mathcal{A}_1 \times \mathcal{A}_2); \left\{p(x_1, x_2/\theta) / \theta \in \Omega\right\}\right]$$

donde $p(x_1 x_2/\theta)$ tiene como distribuciones de probabilidad margi-

nales

$$\int_{S_i} p(x_1 x_2/\theta)\, dx_i = p(x_{3-i}/\theta) \quad (i = 1,2)$$

3.3.2. Informacion proporcionada por la union de dos experimentos.

La informacion que sobre θ nos aporta el experimento compuesto (E_1, E_2) y que notaremos por $I(\theta \| X_1, X_2)$ vendra dada por:

$$I(\theta \| X_1, X_2) = \iiint p(\theta, x_1, x_2) \log \frac{p(\theta, x_1, x_2)}{p(\theta)\, p(x_1)\, p(x_2)}\, d\theta\, dx_1\, dx_2 =$$

$$= \iiint p(\theta)\, p(x_1 x_2/\theta) \log \frac{p(x_1 x_2/\theta)}{p(x_1 x_2)}\, d\theta\, dx_1 dx_2 =$$

$$= \underset{x_1 x_2}{E} \left\{ I(\theta) - I(\theta/x_1 x_2) \right\}$$

3.3.3. Teorema de adicion. La cantidad de informacion que sobre θ me aporta el experimento compuesto (E_1, E_2) puede descomponerse como suma de dos cantidades de informacion: Una la que E_1 aporta sobre θ y otra la media extendida a todos los valores del experimento E_1 de la informacion que el experimento E_2 condicionado por el valor x_1 obtenido al realizar E_1, me aporta sobre θ y que no-tenemos $I_{x_1}(\theta \| X_2)$. Es decir,

$$I(\theta \| X_1 X_2) = I(\theta \| X_1) + I_{x_1}(\theta \| X_2)$$

En efecto:

$$I(\theta \| X_1 X_2) = \iiint p(\theta)\, p(x_1 x_2/\theta) \log \frac{p(x_1 x_2/\theta)}{p(x_1 x_2)}\, d\theta\, dx_1\, dx_2 =$$

$$= \iiint p(\theta)\, p(x_1 x_2/\theta) \log \frac{p(x_1/\theta)\, p(x_2/\theta x_1)}{p(x_1)\, p(x_2/x_1)}\, d\theta\, dx_1 dx_2 =$$

$$= \iiint p(\theta)\, p(x_1 x_2/\theta) \log \frac{p(x_1/\theta)}{p(x_1)}\, d\theta\, dx_1 dx_2 +$$

$$+ \iiint p(\theta)\, p(x_1 x_2/\theta) \log \frac{p(x_2/\theta x_1)}{p(x_2/x_1)}\, d\theta\, dx_1 dx_2 =$$

$$= \iint p(\theta)\, p(x_1/\theta) \log \frac{p(x_1/\theta)}{p(x_1)}\, d\theta\, dx_1 +$$

$$+ \iiint p(\theta)\, p(\theta/x_1)\, p(x_2/\theta x_1) \log \frac{p(x_2/\theta, x_1)}{p(x_2/x_1)}\, d\theta\, dx_1 dx_2 =$$

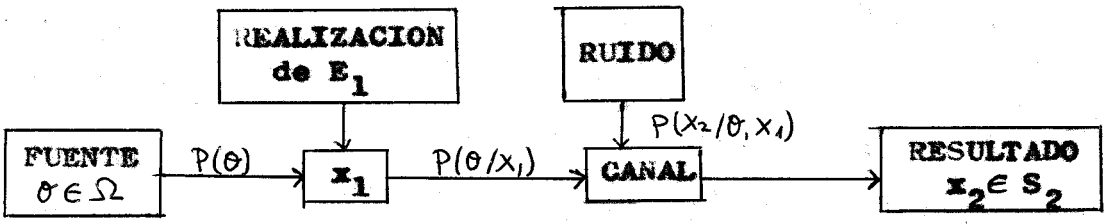
$$I(\theta \| X_1 X_2) = I(\theta \| X_1) + \underset{x_1 x_2 \theta}{E} \left\{ \log \frac{p(x_2/\theta x_1)}{p(x_2/x_1)} \right\} =$$

$$= I(\theta \| X_1) + \underset{x_1}{E} \left\{ I(\theta \| X_2/x_1) \right\}$$

siendo $p(\ /x_1)$ la distribucion a priori para el experimento E_2 y segun la notacion anteriormente indicada

$$I(\theta \| X_1, X_2) = I(\theta \| X_1) + I_{x_1}(\theta \| X_2)$$

Nótese que $I_{x_1}(\theta \| X_2)$ representa la media extendida a todos los resultados x_1 del experimento E_1 de la razon de transmision de informacion del canal representado en el diagrama adjunto.



<u>Corolario.</u> Por el teorema 3.2.4.1. $I_{x_1}(\theta \| X_2) \geq 0$ y por tanto del teorema anterior deducimos

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

3.3.4. <u>Definicion.</u> Diremos que dos experimentos E_1 y E_2 son independientes cuando $p(x_1 x_2/\theta) = p(x_1/\theta)\ p(x_2/\theta)$.

Esto no quiere decir que si E_1 y E_2 son independientes las variables aleatorias X_1 y X_2 son estadisticamente independientes.

Si E_1 y E_2 son independientes las expresiones E_2/x_1 y E_2 son equivalentes y entonces podemos escribir $I_{x_1}(\theta \| X_2)$ en la forma $E_{x_1}[I(\theta \| X_2)]$ siendo $p(\theta/x_1)$ la distribucion a priori sobre θ para E_2.

3.3.5. <u>Teorema.</u> Si dos experimentos E_1 y E_2 son independientes entonces

$$I(\theta \| X_1 X_2) \leq I(\theta \| X_1) + I(\theta \| X_2)$$

mo espacio parametrico Ω y con espacios de resultados S_1 y S_2 disjuntos, llamaremos media ponderada de E_1 y E_2 (con peso λ sobre E_1) al experimento

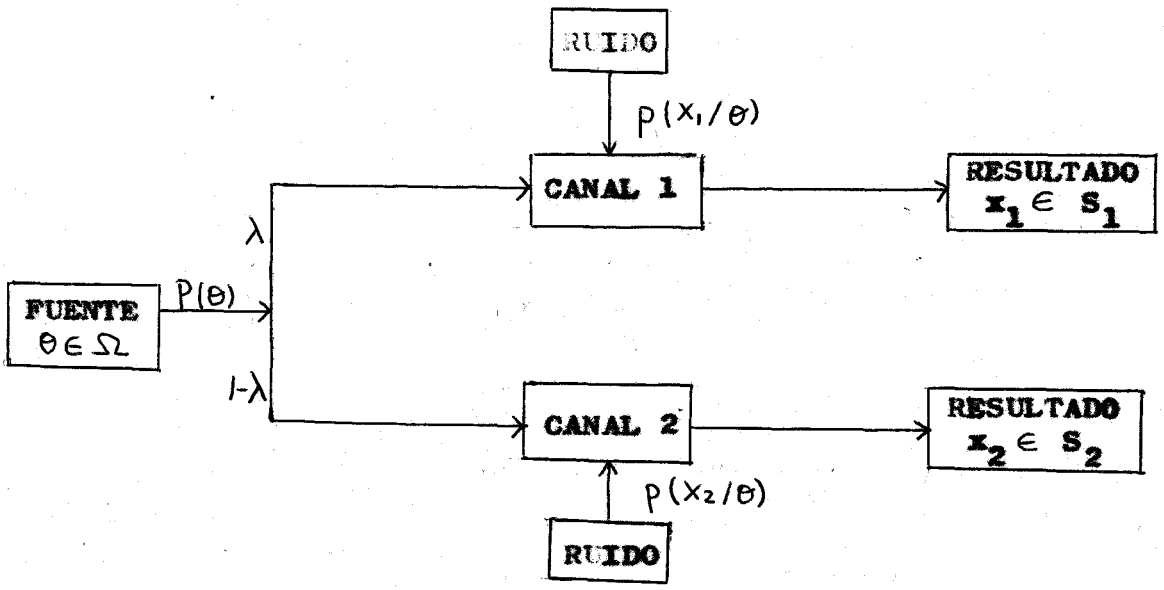
$$\lambda E_1 + (1-\lambda) E_2 = [(S_1 \cup S_2, \mathcal{A}), \{P(x/\theta) / \theta \in \Omega\}]$$

donde

$$P(x/\theta) = \begin{cases} \lambda\, p(x_1/\theta) & \text{para } x = x_1 \quad S_1 \\ (1-\lambda)\, p(x_2/\theta) & \text{para } x = x_2 \quad S_2 \end{cases}$$

siendo $\lambda / 0 \leq \lambda \leq 1$

Podemos representar este experimento mediante el siguiente diagrama:



Con probabilidad λ , un resultado x_1 es obtenido, siendo $p(x_1) = \int_\Omega p(\theta)\, p(x_1/\theta)\, d\theta$; con probabilidad $1 - \lambda$, un resultado x_2 es sobtenido, siendo $p(x_2) = \int_\Omega p(\theta)\, p(x_2/\theta)\, d\theta$

El experimentador conoce no solo un resultado x_1 ó x_2, sino tambien qué suceso el de probabilidad λ ó $(1 - \lambda)$ ha ocurrido.

De la propia definicion se deduce inmediatamente:

$$I[\theta \| \lambda X_1 + (1-\lambda) X_2] = \lambda I[\theta \| X_1] + (1-\lambda) I[\theta \| X_2]$$

dandose la igualdad si y solo si X_1 y X_2 son estadisticamente independientes.

En efecto : Por ser E_1 y E_2 independientes

$$I(\theta \| X_1 X_2) = \iiint p(\theta)\, p(x_1 x_2 / \theta) \log \frac{P(x_1/\theta)\, P(x_2/\theta)\, P(x_2)}{P(x_1)\, p(x_2/x_1)\, P(x_2)} \, d\theta \, dx_1 dx_2$$

$$= \iiint p(\theta)\, p(x_1 x_2/\theta) \left[\log \frac{P(x_1/\theta)}{p(x_1)} + \log \frac{P(x_2/\theta)}{p(x_2)} + \log \frac{P(x_2)}{p(x_2/x_1)} \right] d\theta \, dx_1 dx_2$$

e integrando en Ω, quedara

$$I(\theta \| X_1 X_2) = I(\theta \| X_1) + I(\theta \| X_2) + \iint p(x_1 x_2) \log \frac{P(x_2)}{p(x_2/x_1)} \, dx_1 dx_2 =$$

$$= I(\theta \| X_1) + I(\theta \| X_2) + \iint p(x_1 x_2) \log \frac{P(x_1) P(x_2)}{p(x_1 x_2)} \, dx_1 dx_2$$

y como consecuencia del lema 3.2.4.1 el tercer sumando del segundo miembro es una cantidad no positiva, que será nula si y solo si $p(x_1)\, p(x_2) = p(x_1, x_2)$ o sea si X_1 y X_2 son independientes.

3.3.8. Teorema: Si E_1 y E_2 son dos experimentos independientes

$$I_{X_1}(\theta \| X_2) \leq I(\theta \| X_2)$$

con igualdad si y solo si X_1 y X_2 son independientes.

En efecto :

$I(\theta \| X_1 X_2) = I(\theta \| X_1) + I_{X_1}(\theta \| X_2)$, por el teorema de adición.

$I(\theta \| X_1 X_2) = I(\theta \| X_1) + I(\theta \| X_2) - R(x_1, x_2)$, por el teorema 3.5.3

siendo $$R(x_1 x_2) = \iint p(x_1 x_2) \log \frac{P(x_1 x_2)}{p(x_1)\, p(x_2)} \, dx_1 dx_2$$

De las dos expresiones anteriores deducimos

$$I_{X_1}(\theta \| X_2) = I(\theta \| X) - R(x_1, x_2)$$

por el lema 3.2.4.1, $R(x_1, x_2) \geq 0$, con igualdad si y solo si X_1 y X_2 son estadisticamente independientes. Esto demuestra el teorema.

3.4 MEDIA PONDERADA DE DOS EXPERIMENTOS

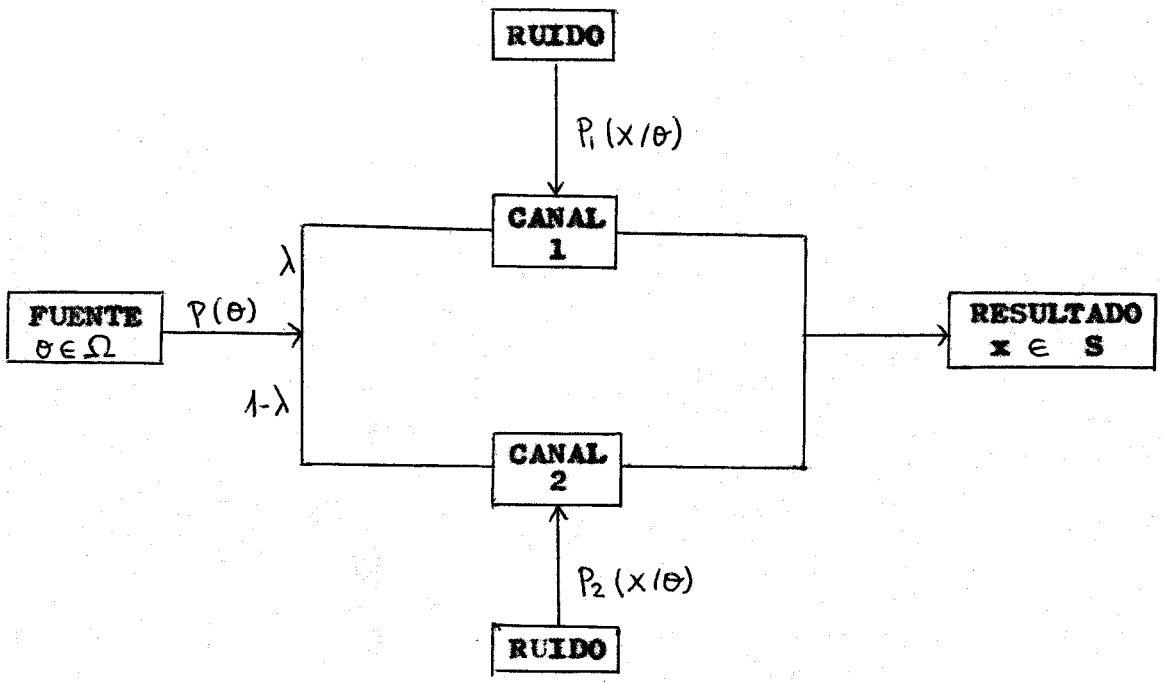
Siguiendo a Lindley, dados dos experimentos E_1 y E_2 con el mi

3.5 MIXTURA DE DOS EXPERIMENTOS.

3.5.1. Definicion. Sean E_1 y E_2 dos experimentos con el mismo espacio parametrico Ω y el mismo espacio de resultados S. Definimos mixtura de E_1 y E_2 (con peso λ sobre E_1) al experimento

$$\lambda E_1 * (1-\lambda) E_2 = \left[(S, \mathcal{A}) ; \{\lambda p_1(x/\theta) + (1-\lambda) p_2(x/\theta); \theta \in \Omega\}\right]$$

Podemos representar este experimento mediante el siguiente diagrama



Un valor x es obtenido segun las probabilidades $p_1(x/\theta)$ y $p_2(x/\theta)$ con probabilidades λ y $1-\lambda$. Ahora el experimentador conoce solo el resultado x y no qué suceso el de probabilidad λ ó $1-\lambda$, ha ocurrido al realizar el experimento.

3.5.2. Teorema. La informacion proporcionada por un experimento es convexa en la mixtura de experimentos. Es decir,

$$I(\theta \| \lambda X_1 * (1-\lambda) X_2) \leq \lambda I(\theta \| X_1) + (1-\lambda) I(\theta \| X_2)$$

En efecto:

Si representamos por E un experimento mediante el que conocemos cual de los dos sucesos, el de probabilidad λ ó el de $1-\lambda$ ha ocurrido, entonces

$$[\lambda E_1 * (1-\lambda) E_2, E] = \lambda E_1 + (1-\lambda) E_2$$

y por tanto podremos escribir

$$I[\theta \| \lambda X_1 * (1-\lambda) X_2] \leq I[\theta \| \lambda X_1 * (1-\lambda) X_2, X] =$$

$$= I[\theta \| \lambda X_1 + (1-\lambda) X_2] = \lambda I[\theta \| X_1] + (1-\lambda) I[\theta \| X_2]$$

3.5.3. <u>Teorema.</u> La cantidad de informacion es una funcion concava en la distribucion a priori $p(\theta)$, es decir, si $p_1(\theta)$ y $p_2(\theta)$ son dos distribuciones de probabilidad a priori y λ siendo $0 \leq \lambda \leq 1$, entonces se verifica la desigualdad

$$\iint p(x/\theta)[\lambda p_1(\theta) + (1-\lambda) p_2(\theta)] \log \frac{p(x/\theta)}{p(x)} \, d\theta \, dx \geq$$

$$\lambda \iint p(x/\theta) p_1(\theta) \log \frac{p(x/\theta)}{p_1(x)} \, d\theta dx +$$

$$+ (1-\lambda) \iint p(x/\theta) p_2(\theta) \log \frac{p(x/\theta)}{p_2(x)} \, d\theta dx$$

con

$$p_i(x) = \int_\Omega p(x/\theta) p_i(\theta) \, d\theta \qquad (i = 1,2)$$

y

$$p(x) = \lambda p_1(x) + (1-\lambda) p_2(x)$$

En efecto:

Si llamamos

$$I = \iint p(x/\theta)[\lambda p_1(\theta) + (1-\lambda) p_2(\theta)] \log \frac{p(x/\theta)}{p(x)} \, d\theta \, dx -$$

$$-\lambda \iint p(x/\theta)\, p_1(\theta) \log \frac{p(x/\theta)}{p_1(x)}\, d\theta\, dx -$$

$$-(1-\lambda) \iint p(x/\theta)\, p_2(\theta) \log \frac{p(x/\theta)}{p_2(x)}\, d\theta\, dx$$

veamos que $I \geq 0$

Simplificando en la expresion anterior, tendremos

$$I = \lambda \iint p(x/\theta)\, p_1(\theta) \log \frac{p_1(x)}{p(x)}\, d\theta\, dx$$

$$+(1-\lambda) \iint p(x/\theta)\, p_2(\theta) \log \frac{p_2(x)}{p(x)}\, d\theta\, dx$$

e integrando en Ω

$$I = \lambda \int p_1(x) \log \frac{p_1(x)}{p(x)}\, dx + (1-\lambda) \int p_2(x) \log \frac{p_2(x)}{p(x)}\, dx$$

y en virtud del lema 3.2.4.1 ambas integrales son no negativas y por tanto $I \geq 0$, con lo que queda demostrado el teorema.

3.6. COMPARACION DE EXPERIMENTOS.

3.6.1. <u>Definicion.</u> Dados dos experimentos E_1 y E_2 con el mismo espacio parametrico Ω caracterizados por las cuplas

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

diremos que E_1 es <u>no menos informativo</u> que E_2 y lo escribiremos $E_1 \succcurlyeq E_2$ cuando

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

para todas las posibles distribuciones a priori $p(\theta)$ definidas sobre Ω (5).

3.6.2. <u>Definicion.</u> Dados dos experimentos E_1 y E_2 con el mismo espacio parametrico Ω diremos que E_1 es <u>tan informativo como</u> E_2, y lo notaremos $E_1 \sim E_2$ cuando $E_1 \succcurlyeq E_2$ y $E_2 \succcurlyeq E_1$.

---

(5) Al decir para todas las distribuciones $p(\theta)$ no excluimos las que no estan dominadas por una medida prefijada.

3.6.3. Propiedad. La relacion $\succsim$ anteriormente definida es una relacion de orden parcial. En efecto, verifica las propiedades

Reflexiva $\quad E_1 \succsim E_1$

Transitiva $\quad \left.\begin{matrix} E_1 \succsim E_2 \\ E_2 \succsim E_3 \end{matrix}\right\} \Rightarrow E_1 \succsim E_3$

Antisimetrica $\quad \left.\begin{matrix} E_1 \succsim E_2 \\ E_2 \succsim E_1 \end{matrix}\right\} \Rightarrow E_1 \sim E_2$

pero no verifica la completitud, pues dados dos experimentos E_1 y E_2 no tiene porque ocurrir $E_1 \succsim E_2$ ó $E_2 \succsim E_1$, ya que estos dos experimentos no tienen por qué ser comparables.

3.6.4. Definicion. Dados dos experimentos E_1 y E_2 con el mismo espacio parametrico Ω, diremos que E_1 es mas informative que E_2 y lo notaremos $E_1 \succ E_2$ cuando ocurre $E_1 \succsim E_2$ pero no ocurre que $E_1 \sim E_2$. (6).

---

(6). Aparte del metodo que aqui vamos a seguir, existen otros metodos para comparar experimentos.

El mas antiguo debido a Bohnenblust, Shapley y Sherman (1.949) dice que E_1 es masinformative que E_2 si cada funcion de perdida alcanzable con un resultado de E_2 lo es tambien con E_1 (SAVAGE 1.954 y BLACKWELL 1.951).

Blackwell dados dos experimentos discretes E_1 y E_2 dice que E_1 es mas informative que E_2 y lo nota $E_1 \supset E_2$ si para cada conjunto cerrado, acotado y convexo A de acciones se verifica

$$B(E_1, A) \supset B(E_2, A)$$

siendo $B(E_i, A)$ el conjunto de todos los posibles vectores riesgo $b(E,d)$ donde d varia para todas las posibles funciones de decision (BLACKWELL-GIRSHIK 1.954).

Lehmann supone que experimentos diferentes son validos para contrastar una hipotesis simple H contra una alternativa K. Un experimento E_1 resultaba al observar una variable aleatoria X_1 que tiene funciones de densidad p_1 y q_1 bajo H y K, respectivamente; otro experimento E_2 conduce a la observacion de X_2 con funciones de densidad p_2 y q_2 bajo H y K, respectivamente. Si se nota $\beta_i(\alpha)$ la potencia del test de nivel alfa mas potente basado en X_i, entonces E_1 es mas informative que E_2 cuando $\beta_2(\alpha) \leq \beta_1(\alpha)$ (LEHMANN 1.959).

3.6.5. Ejemplo. Consideremos una poblacion de individuos que poseen o no cada una de dos caracteristicas H y G. Supongamos que conocemos las proporciones h y g de individuos con las caracteristicas H y G., pero que no conocemos la proporcion w de individuos que tienen ambas caracteristicas. Tomamos $\theta = w/h$ como un parametro desconocido. Supongamos sin perdida de generalidad que $0 \leq h \leq g \leq 1 - g \leq 1 - h \leq 1$. Se pueden considerar cuatro tipos diferentes de experimentos que notamos E(H), $E(\bar{H})$, E(G) y $E(\bar{G})$. Una realizacion del experimento E(H) puede consistir, por ejemplo, en observar N individuos con la caracteristica H y de ellos contar el numero de individuos que no poseen la caracteristica G, (7).

Resumiendo en una tabla:

Caracteristicas	G	$\bar{G}$	Totales
H	w		h
$\bar{H}$			1 - h
totales	g	1 - g	1

Veamos que el experimento E(H) es no menos informativo que cualquiera de los otros tres experimentos.

En efecto: Los cuatro experimentos son binomiales con espacio parametrico $\Omega = [0, 1]$ y con probabilidades de la forma $\lambda\mu + (1-\lambda)\theta$ donde μ es una proporcion independiente de que varia para cada experimento ($0 \leq \mu \leq 1$).

El experimento E definido por

$$E = [(S, \mathcal{A}); \{p(x/\theta) / \theta \in [0,1]\}]$$

donde las $p(x/\theta)$ siguen una ley Bernouilli de parametro $\lambda\mu + (1-\lambda)\theta$, caracteriza cualquiera de los cuatro experimentos an-

(7) BLACKWELL (1.951) o bien BLACKWELL-GIRSHICK (1.954).

teriores.

Consideremos dos experimentos E_1 y E_2 definidos por

$$E_i = \left[(S, \mathcal{A}) ; \{ P_i(x/\theta) \quad / \theta \in [0,1] \} \right] \quad (i = 1,2)$$

siendo

$$P_1(x/\theta) = \mu^x (1-\mu)^{1-x}$$

$$P_2(x/\theta) = \theta^x (1-\theta)^{1-x}$$

podemos escribir

$$E \sim \lambda E_1 * (1-\lambda) E_2$$

y en virtud del teorema 3.5.2

$$I[\theta \| \lambda X_1 * (1-\lambda) X_2] \leq \lambda I[\theta \| X_1] + (1-\lambda) I[\theta \| X_2]$$

pero $I(\theta \| X_1) = 0$, pues $p_1(x/\theta)$ no depende de θ ,

por tanto

$$I(\theta \| X) \leq (1-\lambda) I(\theta \| X_2) \leq I(\theta \| X_2)$$

pero para $\lambda = 0$, E(H) coincide con el experimento E_2. Hemos, pues, demostrado que E(H) es no menos informative que cualquier de los otros tres experimentos y como habiamos impuesto $0 \leq h \leq g \leq 1 - g \leq 1 - h \leq 1$, deducimos finalmente, que el experimento asociado con la caracteristica mas rara es el que másinformacion nos proporciona.

3.6.6. <u>Teorema.</u> Sean E_1, E_2 y E_3 tres experimentos con el mismo espacio parametrico Ω . Si E_3 es independiente de E_1 y E_2, entonces

$$E_1 \succcurlyeq E_2 \Rightarrow (E_1, E_3) \succcurlyeq (E_2, E_3)$$

En efecto:

Para cualquier $p(\theta)$, por el teorema 3.3.3.

$$I(\theta \| X_1 X_3) = I(\theta \| X_3) + E_{x_3}[I(\theta \| X_1/x_3)]$$

con distribucion a priori $p(\theta/x_3)$ para E_1.

como E_1 y E_3 son independientes

$$I[\theta \| X_1 X_3] = I[\theta \| X_3] + \underset{X_3}{E}\left[I(\theta \| X_1)\right]$$

con distribucion a priori $p(\theta/x_3)$ para E_1. Pero $E_1 \succcurlyeq E_2$ implica que

$$I[\theta \| X_1] \geq I[\theta \| X_2]$$

y en particular esto ocurre para la distribucion a priori $p(\theta/x_3)$ para cualquier $x_3 \in S_3$. Se deduce inmediatamente

$$I[\theta \| X_1 X_3] \geq I[\theta \| X_2 X_3]$$

lo que prueba el teorema.

3.6.7. <u>Definicion.</u> Dados dos experimentos E_1 y E_2 con el mismo espacio parametrico Ω diremos que E_1 es <u>fuertemente no menos informativo</u> que E_2 y lo notaremos $E_1 \overset{(f)}{\succcurlyeq} E_2$, cuando

$$\int_S p_1(x/\theta_0) \log \frac{p_1(x/\theta_0)}{\int p(\theta) p_1(x/\theta) d\theta} dx \geq \int_S p_2(x/\theta_0) \log \frac{p_2(x/\theta_0)}{\int p(\theta) p_2(x/\theta) d\theta} dx$$

para todo $\theta_0 \in \Omega$ y $p(\theta)$.

Notemos que si tomamos esperanza matematica en θ en la desigualdad anterior obtenemos $I(\theta \| X_1) \geq I(\theta \| X_2)$, es decir, que E_1 es no menos informativo que E_2, de ahi que hayamos utilizado en esta nueva definicion 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.

3.7. SUFICIENCIA DE EXPERIMENTOS.

Otro metodo 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 parametrico Ω , se dice que el experimento E_1 es suficiente para E_2 si existe una funcion no negativa f sobre el espacio producto $S_1 \times S_2$ para la cual se satisfacen las siguientes relaciones:

$$P_2(x_2/\theta) = \int_{S_1} f(x_2, x_1) P_1(x_1/\theta)\, dx_1 \qquad \text{para} \quad \theta \in \Omega,\ x_2 \in S_2 \qquad (i)$$

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

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

Una funcion no negativa f que satisface (ii) es llamada transformacion estocastica de X_1 a X_2.

jPara cada resultado $x_1 \in S_1$ fijado, la funcion $f(\cdot, x_1)$ es una funcion de densidad sobre S_2. Ya que esta funcion no depende del parametro θ, un resultado $x_2 \in S_2$ puede ser generado de acuerdo con esta funcion de densidad por medio de una aleatorizacion auxiliar. Por tanto la ecuacion (i) nos dice que E_1 es suficiente para E_2 si haciendo caso omiso del valor del parametro θ, una observacion de E_1 y una auxiliar aleatorizacion hace posible generar una variable aleatoria que tiene la misma distribucion que X_2. La ecuacion (iii) no es otra cosa que una condicion de integrabilidad de f sobre S_1.

Intuitivamente esta claro, que si E_1 es suficiente para E_2, el estadistico nunca debera 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 transformacion aleatoria que solo puede oscurecer cualquier informacion sobre el valor del parametro que podria estar contenida en este resultado de E_1. El proximo teorema formaliza la afirmacion anterior de que E_1 es no menos informativo que E_2.

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

En efecto:

Si E_1 es suficiente para E_2, existe una transformacion estoca

tica de X_1, que llamaremos X'_2 tal que $X'_2 \in S_2$ y X'_2 y X_2 estan identicamente distribuidas para cada $\theta \in \Omega$.

Representemos por E'_2 el experimento en el cual observamos la variable aleatoria X'_2 . Evidentemente

$$I(\theta \| X_2) = I(\theta \| X'_2)$$

Consideremos el experimento compuesto $E = (E_1, E'_2)$. Por ser X_1 suficiente para (X_1, X'_2), en el sentido de Neyman, y por el teorema 3.2.4.2. tendremos

$$I(\theta \| X_1 X'_2) = I(\theta \| X_1)$$

Por otro lado en virtud del teorema 3.3.3.

$$I(\theta \| X_1 X'_2) \geq I(\theta \| X'_2)$$

de donde se deduce finalmente

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

para todo $p(\theta)$. Es decir $E_1 \succsim E_2$, con lo que queda demostrado el teorema.

3.7.2. <u>Teorema.</u> Sean E_1 y E_2 dos experimentos con el mismo espacio parametrico Ω . Si E_1 es suficiente para E_2, entonces E_1 es fuertemente no menos informativo que E_2.

En efecto:

Continuando con la notacion utilizada a lo largo del capitulo, si E_1 es suficiente para E_2 entonces:

Existe $p(x_2/x_1)$ y es independiente de , siendo

$$\int_{S_2} p(x_2/x_1)\, dx_2 = 1$$

y ademas

$$p_2(x_2/\theta) = \int_{S_1} p(x_2/x_1)\, p_1(x_1/\theta)\, dx_1$$

para todo θ y x_2.

Para cualquier $\theta_0 \in \Omega$ y $p(\theta)$, tenemos

$$\int_{S_2} P_2(x_2/\theta_0) \log \frac{P_2(x_2/\theta_0)}{\int_\Omega p(\theta) P_2(x_2/\theta)\, d\theta}\, dx_2 =$$

$$= \int_{S_2} \left\{ \left[\int_{S_1} p(x_2/x_1) P_2(x_1/\theta_0) dx_1 \right] \log \frac{\int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1}{\int_\Omega p(\theta) \int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1 d\theta} \right\} dx_2$$

$$= \int_{S_2} \left\{ \left[\int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1 \right] \log \frac{\int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1}{\int_{S_1} p(x_2/x_1) \int_\Omega p(\theta) P_1(x_1/\theta) d\theta\, dx_1} \right\} dx_2$$

$$= \int_{S_2} \left\{ \left[\int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1 \right] \log \frac{\int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) dx_1}{\int_{S_1} p(x_2/x_1) P_1(x_1)\, dx_1} \right\} dx_2 \leq$$

$$\leq \int_{S_2} \int_{S_1} p(x_2/x_1) P_1(x_1/\theta_0) \log \frac{P_1(x_1/\theta_0)}{P_1(x_1)}\, dx_1 dx_2 =$$

$$= \int_{S_1} P_1(x_1/\theta_0) \log \frac{P_1(x_1/\theta_0)}{P_1(x_1)} \left[\int_{S_2} p(x_2/x_1) dx_2 \right] dx_1 =$$

$$= \int_{S_1} P_1(x_1/\theta_0) \log \frac{P_1(x_1/\theta_0)}{P_1(x_1)}\, dx_1 =$$

$$= \int_{S_1} P_1(x_1/\theta_0) \log \frac{P_1(x_1/\theta_0)}{\int_\Omega p(\theta) P_1(x_1/\theta)\, d\theta}\, dx_1$$

con lo que el teorema ha sido demostrado.

3.7.3. Ejemplo. Consideremos dos experimentos gaussianos caracterizados por las cuplas :

---

(8) HARDY-LITTLEWOD-POLYA (1952), teorema 205.

$$E_i = \left[(\mathbb{R}, B) \; ; \; P_i(x/\theta) = \frac{1}{\sqrt{2\pi}\, v_i} e^{-\frac{1}{2}(x_i - \theta)^2} \; , \; \theta \in \mathbb{R} \right] \quad i = 1,2$$

suponiendo las varianzas v_{ii}^2 conocidas y tales que $v_1^2 < v_2^2$.

Evidentemente, por ser su varianza menor, una observacion de X_1, o sea una realizacion del experimento E_1, aporta mas informacion sobre el valor del parametro que una realizacion del experimento E_2. No obstante vamos a demostrar formalmente que E_1 es suficiente para E_2 y por tanto en virtud del teorema 3.7.2 habremos demostrado que E_1 es fuertemente no menos informative que E_2.

En efecto:

Sea Y una variable aleatoria independiente de X_1 y de θ, distribuida segun una ley normal de media 0 y varianza $v_2^2 - v_1^2$. Para cualquier valor de θ, la variable aleatoria $X_1 + Y$ tiene la misma distribucion que la variable aleatoria X_2. Por tanto E_1 es suficiente para E_2 (9).

3.8 VALOR DE LA INFORMACION DE UN EXPERIMENTO.

Si comparamos experimentos por las utilidades asociadas con las decisiones finales adoptadas, llegamos a la nocion de valor de informacion de un experimento. Ordinariamente es el incremento en utilidad al pasar de un experimento menos informative a otro mas informative.

Hemos construido una medida de informacion que, como hemos comprobado, manifiesta un comportamiento monotomo con respecto a la union de experimentos (10). Veamos que el valor de la informacion asociados con un experimento concepto utilitarista, como

---

(9) Si se supone que $P(\theta)$ sigue una distribucion normal, los resultados obtenidos en el apartado 3.2.3.C. corroboran que $E_1 \succsim E_2$ es decir, E_1 es fuertemente no menos informativo que E_2. Es mas en este caso particular $E_1 \succ E_2$.

(10) Ver corolario del teorema 3.3.3.

dijimos antes, tiene una expresion analitica completamente análoga a la medida de informacion aqui encontrada, y como ella tiene el mismo comportamiento monotomo con respecto a la union de experimentos.

Para llegar a ello, formularemos el problema de decision en terminos del valor de la informacion (RAIFFA-SCHALAIFER, 1.961) en lugar de hacerlo en termino de la funcion de utilidad.

Los componentes de un problema de decision estadistica segun la formulacion clasica de WALD (1.950) son:

-Un espacio parametrico o espacio de estados de la naturaleza Ω con elementos $\theta \in \Omega$

-Un espacio de experimentos $\mathcal{E}$ con elementos $E \in \mathcal{E}$

-Un espacio de resultados del experimento E : $S = \{x\}$

-Un espacio de decisiones finales $\mathcal{D}^f$ con elementos $d^f \in \mathcal{D}^f$

-Una probabilidad a priori $p(\theta)$ sobre Ω .

-Una probabilidad $p(x/\theta)$ de que se presente el resultado x dado que θ es el estado de la naturaleza cuando se ha realizado el experimento E.

-Una funcion de utilidad $u(E,x,d^f,\theta)$ definida sobre $\mathcal{E} \times S \times \mathcal{D}^f \times \Omega$

La utilidad $u(E,x,d^f,\theta)$ puede ser descompuesta en otras dos utilidades:

-Utilidad muestral $u_m(E,x)$.

-Utilidad final $u_f(d^f,\theta)$.

Vamos a introducir el concepto de valor de la informacion (RAIFFA SCHALAIFER, 1.961) para lo cual impondremos que se verifica

$$u(E,x,d^f,\theta) = u_m(E,x) + u_f(d^f,\theta) \qquad (i)$$

y ademas supondremos que la utilidad muestral es igual y de signo contrario al coste de muestreo:

$$u_m(E,x) = - c(E,x) \qquad (ii)$$

Dado $\theta \in \Omega$ notamos d^f_θ a aquella decision final para la cual se alcanza el maximo de las utilidades finales. Es decir

$$u_f(d^f_\theta, \theta) = \max_{d^f \in \mathbb{D}^f} u_f(d^f, \theta)$$

Supongamos que un experimento ideal E_∞ con resultado asociado x_∞, y coste muestral $c(E_\infty)$ nos produce la informacion perfecta sobre el espacio parametrico Ω.

Teniendo en cuenta (i) y (ii) la utilidad de tomar la decision final d^f_θ con informacion perfecta vendra dada por

$$u(E_\infty, x_\infty, d^f_\theta, \theta) = u_f(d^f_\theta, \theta) - c(E_\infty) \qquad \text{(iii)}$$

Sea E_o el experimento nulo (11) que tiene asociados el resultado nulo x_o. La utilidad de tomar la decision d^f_o con informacion nula, teniendo en cuenta (i) y (ii) vendra dada por

$$u(E_o, x_o, d^f_o, \theta) = u_f(d^f_o, \theta) \qquad \text{(iv)}$$

viniendo d^f_o definida por

$$E\left[u_f(d^f_o, \theta)\right] = \max_{d^f \in \mathbb{D}^f} E\left[u_f(d^f, \theta)\right]$$

$u(E_\infty, x_\infty d^f_\theta, \theta) - u(E_o, x_o, d^f_o, \theta)$ sera positiva siempre que $c(E_\infty)$ sea menor que el regret (12) final de la decision optima d^f_θ:

$$u_f(d^f_\theta, \theta) - u_f(d_o, \theta)$$

Por consiguiente se puede definir

$$v(\theta) = r_f(d^f_o, \theta) = u_f(d^f_\theta, \theta) - u_f(d^f_o, \theta)$$

como el valor de la informacion perfecta dado un $\theta \in \Omega$. Y teniendo en cuenta la distribucion a priori $p(\theta)$ sobre Ω, el valor

---

(11) Queremos indicar con E_o el hecho de tomar una decision final sin experimentar. Merecera la pena realizar un experimento E cuando su utilidad o su informacion sea estrictamente mayor que la de E_o. Representa, pues, el cero en la escala de utilidad. Evidentemente suponemos nulo el coste asociado a E_o.

(12) WALD(1.950) y RAIFFA-SCHLAIFFER(1.961) utilizan el termino "opportunity loss" pero hemos preferido emplear el termino "regret" introducido por SAVAGE (1.954). El regret viene dado por la diferencia entre la ganancia obtenida cuando el verdadero estado

...

esperado de la informacion perfecta V(θ) sera

$$V(\theta) = \int_{\Omega} v(\theta) p(\theta) d\theta$$

Supongamos ahora que realizamos un experimento E y obtenemos un resultado $x \in S$. Las probabilidades a priori $p(\theta)$ sobre Ω son modificadas por la realizacion de E y el conocimiento de x a unas probabilidades a posteriori $p(\theta/x)$. Si definimos d_x^f por la expresion

$$\int u_f (d_x^f, \theta)\ p(\theta/x)\, d\theta = \max_{d^f \in \mathbb{D}^f} \int u_f (d^f, \theta) p(\theta/x) d\theta$$

el regret obtenido al tomar la decision d_x^f vendra dado por

$$v(\theta/x) = u_f(d_\theta^f, \theta) - u_f(d_x^f, \theta)$$

y calculando la media para todos los $\theta \in \Omega$

$$V(\theta/x) = \int_{\Omega} v(\theta/x)\, p(\theta/x)\, d\theta$$

Por tanto el valor de la informacion que el resultado x bbtenido al realizar el experimento E, me aporta sobre Ω vendra dado por

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

y calculando la media para todos los posibles resultados x del experimento E podemos definir finalmente:

3.8.1. Definicion. El valor de la informacion asociado con el experimento E viene dado por la expresion:

$$V(\theta \| X) = E_x \left[V(\theta) - V(\theta/x) \right] =$$

$$= \int_x \int_\theta \left[u_f (d_\theta^f, \theta) - u_f (d_\theta^f, \theta) \right] p(\theta)\, p(x)\, d\theta dx -$$

(12 - continuacion) de la naturaleza es conocido y la ganancia actual conseguida. Sera siempre positivo o nulo, dandose este ultimo caso cuando el estado de la naturaleza verdadero es conocido.

$$- \int_x \int_\Omega \left[u_f(d^f_{o,}, \theta) - u_f(d^f_{x,}, \theta) \right] p(\theta/x)\, p(x)\, d\theta\, dx$$

$$= \int_x \int_\Omega \left[u_f(d^f_{x,}, \theta) - u_f(d^f_{o,}, \theta) \right] p(\theta, x)\, d\theta\, dx =$$

$$= \underset{x,\theta}{E} \left[u_f(d^f_{x,}, \theta) - u_f(d^f_{o,}, \theta) \right]$$

De la propia definicion se deduce inmediatamente que el valor de la informacion asociada a un experimento es una cantidad no negativa, es decir, $V(\theta \| X) \geq 0$.

3.8.2. <u>Valor de la informacion asociado con la union de experimentos.</u> Sean E_1 y E_2 dos experimentos con el mismo espacio parametrico y sea (E_1, E_2) el experimento union de E_1 y E_2 segun lo definimos en el apartado 3.3.

El valor de la informacion asociado con el experimento $(E_1\ E_2)$ sera

$$V(\theta \| X_1, X_2) = \underset{x_2 x_2, \theta}{E} \left[V(\theta) - V(\theta/x_1 x_2) \right]$$

siendo

$$V(\theta/x_1, x_2) = \int_\Omega \left[u_f(d^f_{\theta,}, \theta) - u_f(d^f_{(x_1,x_2)}, \theta) \right] p(\theta/x_1 x_2)\, d\theta$$

viniendo $d^f_{(x_1,x_2)}$ definida por la expresion

$$\int_\Omega u_f(d^f_{(x_1,x_2)}\ \theta)\, p(\theta/x_1, x_2)\, d\theta = \max_{d^f \in \mathcal{D}^f} \int_\Omega u_f(d^f, \theta)\, p(\theta/x_1 x_2)\, d\theta$$

Supongamos que hemos realizado el experimento E_1 y obtenido un resultado x_1 con una utilidad final asociada $u_f(d^f_{x_1}, \theta)$. El valor de la informacion del experimento E_2/x_1 sera

$$V(\theta \| X_2/x_1) = \int_{X_2} \int_\theta \left[u_f(d^f_{x_1,x_2,}, \theta) - u_f(d^f_{x_1,}, \theta) \right] p(\theta/x_1)\, p(x_2/\theta, x_1)\, d\theta\, dx_2$$

y calculando la media para todos los posibles resultados x_1 del experimento E_1, tendremos:

$$V_{X_1}(\theta \| X_2) = E_{x_1}\left[V(\theta \| X_2/x_1)\right] =$$

$$= \int_{x_1}\int_{y_2}\int_{\theta}\left[u_f(d^f_{x_1x_2},\theta) - u_f(d^f_{x_1},\theta)\right] p(x_2/\theta x_1)\, p(\theta/x_1) p(\theta)\, d\theta\, dx_1\, dx_2$$

Por ser el integrando no negativo deducimos $V_{X_1}(\theta \| X_2) \geq 0$

De las expresiones de $V_{X_1}(\theta \| X_2)$ y $V(\theta \| X_1X_2)$ se deduce inmediatamente el siguiente teorema.

3.8.3. <u>Teorema.</u> El valor de la informacion asociado con el experimento (E_1,E_2) puede descomponerse en suma de dos componentes: Una el valor de la informacion asociado con E_1 y otra la media extendida a todos los resultados x_1 del experimento E_1 del valor de la informacion del experimento condicionado E_2/x_1. Es decir,

$$V(\theta \| X_1X_2) = V(\theta \| X_1) + V_{X_1}(\theta \| X_2)$$

<u>Corolario.</u> Por ser $V_{X_1}(\theta \| X_2) \geq 0$, se deduce trivialmente del teorema anterior que

$$V(\theta \| X_1X_2) \geq V(\theta \| X_1)$$

Podemos resumir finalmente la analogia entre la medida de informacion de un experimento y el valor de la informacion asociado con un experimento en la siguiente tabla:

MEDIDA DE LA INFORMACION	VALOR DE LA INFORMACION
Autoinformacion $i(\theta) = -\log p(\theta)$	**Regret final** $V(\theta) = u_f(d^f_\theta, \theta) - u_f(d^f_o, \theta)$
Medida de la incertidumbre total $I(\theta) = -\int_\Omega i(\theta)\, p(\theta)\, d\theta$	**Valor esperado de la informacion perfecta** $V(\theta) = \int_\Omega V(\theta)\, p(\theta)\, d\theta$
$I(\theta/x) = -\int_\Omega p(\theta/x) \log p(\theta/x)\, d\theta$	$V(\theta/x) = \int_\Omega V(\theta)\, p(\theta/x)\, d\theta$ **mismo** $V(\theta/x) = u_f(d^f_\theta, \theta) - u_f(d^f_x, \theta)$
Medida de la informacion asociada al resultado x $I(\theta) - I(\theta/x)$	**Valor de la informacion asociado al resultado x** $V(\theta) - V(\theta/x)$
Medida de la informacion del experimento E $I(\theta \mid\mid X) = E_x[I(\theta) - I(\theta/x)] = E_{x,\theta}[i(\theta, x)]$	**Valor de la informacion asociado con el experimento E** $V(\theta \mid\mid X) = E_x[V(\theta) - V(\theta/x)] = E_{x,\theta}[u_f(d^f_x, \theta) - u_f(d^f_o, \theta)]$
Teorema de adicion $I(\theta \mid\mid X_1 X_2) = I(\theta \mid\mid X_1) + I_{X_1}(\theta \mid\mid X_2)$	**Teorema de adicion** $V(\theta \mid\mid X_1 X_2) = V(\theta \mid\mid X_1) + V_{X_1}(\theta \mid\mid X_2)$
Monotonia $I(\theta \mid\mid X_1) \leq I(\theta \mid\mid X_1 X_2)$	**Monotonia** $V(\theta \mid\mid X_1) \leq V(\theta \mid\mid X_1 X_2)$

3.9. CAPACIDAD DE EXPERIMENTOS.

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

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

3.9.2. Ejemplo. Consideremos un experimento gaussiano E definido en la forma

$$E\left[(\mathbb{R}, \mathcal{B}) \; ; \; \{p(x/\theta) \,/\, -\infty < \theta < \infty\}\right]$$

donde $p(x/\sigma) = \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2\sigma_2^2}}$

Si tomamos

$$p(\theta) = \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\sigma_1^2}}$$

entonces segun deducimos en el ejemplo C del apartado 3.2.3

$$I(\theta \| \mathcal{X}) = \frac{1}{2} \log \left(1 + \frac{\sigma_1^2}{\sigma_2^2}\right)$$

Sea $P = \{$funciones de densidad $p(\theta)/ \; p(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \cdot e^{-\frac{(\theta-\mu)^2}{\sigma_1^2}}$ siendo $-\infty < \mu < \infty \; ; \; 0 < \sigma_1 \leq V \}$

Si limitamos nuestra atencion unicamente a las distribuciones $p(\theta)$ de P, tenemos

$$\max_{p(\theta)\in P} I(\theta \| \mathcal{X}) = \frac{1}{2} \log \left(1 + \frac{V^2}{\sigma_2^2}\right)$$

Este resultado es obtenido cuando se calcula la capacidad de un canal gaussiano con ruido aditivo, potencia de ruido σ_2^2 conocida y potencia de señal limitada a V^2 (FANO, 1.949).

Se deduce de la expresion anterior que la capacidad de un canal gaussiano es estrictamente decreciente en la potencia de rui-

do y creciente en la potencia de señal.

3.9.3. Teorema. Sean E_1 y E_2 dos experimentos con el mismo espacio parametrico Ω . Si E_1 es suficiente para E_2 entonces

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

En efecto:

Si E_1 es suficiente para E_2 entonces en virtud del teorema 3.7.1 se verifica que

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

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

En particular esto tambien se verificara para aquellas distribuciones que me alcancenel maximo. En definitiva

$$\max_{p(\theta)\in P} I(\theta \| X_1) \geq \max_{p(\theta)\in P} I(\theta \| X_2)$$

es decir,

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

3.9.4. Capacidad de un experimento con espacio parametrico finito

El calculo de la capacidad de un experimento en el caso de espacio parametrico continuo es un problema bastante dificil y - creemos que ningun metodo general que cubra todas las circunstancias puede ser dado.

Consideremos ahora un experimento E con espacio parametrico Ω finito, $\Omega = \{\theta_1, \theta_2, \dots, \theta_k\}$, caracterizado por la cupla

$$E = \left[(S, A) ; \{p(x/\theta_1), p(x/\theta_2), \dots, p(x/\theta_k) / \theta_i \in \Omega\} \right]$$

Si notamos $p(x/\theta_i) = q_i(x)$ para todo $i/1 \leq i \leq k$, la cantidad de informacion proporcionada por E será

$$I(\theta \| X)_p = \sum_{i=1}^{k_2} p_i \int_S q_i(x) \log \frac{q_i(x)}{\sum_{i=1}^{k} p_i q_i(x)} dx$$

siendo $(p_1, p_2, \dots p_k) = [p(\theta_1), p(\theta_2), \dots, p(\theta_k)] = p \in P^k$,

un k-vector de probabilidades a priori que nos representa el conocimiento a prioristico sobre Ω. La capacidad del experimento E vendra dada por

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

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

$$\int_S q_i(x) \log \frac{q_i(x)}{\sum_i p'_i q_i(x)} dx$$

es independiente de i, entonces p maximiza la informacion $I(\theta \| \mathcal{X})$ y ademas

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

En efecto:

En virtud del teorema 3.5.3, $I(\theta \| X)$ es estrictamente concava en la distribucion a priori $p \in P^k$, de manera que si las ecuaciones de Lagrange nos dan una solucion que no este en la frontera de P^k, esta solucion nos proporciona el maximo.

Sea

$$\Phi(p_i, \lambda) = \sum_{i=1}^{k} p_i \int_S q_i(x) \log \frac{q_i(x)}{\sum p_i q_i(x)} dx - \lambda \left(\sum p_i - 1 \right)$$

derivando con respecto a p_j, tenemos

$$\frac{\partial \Phi(p_i, \lambda)}{\partial p_j} = \int_S q_j(x) \log q_j(x) dx -$$

$$- \int_S q_j(x) \log \left[\sum_i p_i q_i(x) \right] dx -$$

$$- \sum p_i \int_S q_i(x) \frac{q_j(x)}{\sum_{i=1}^{k} p_i q_i(x)} dx - \lambda$$

y si notamos p_j^* los valores para los que $\frac{\partial \Phi(p_i, \lambda)}{\partial p_j} = 0$ podemos escribir

$$\lambda = \int q_j(x) \log \frac{q_j(x)}{\sum p'_i q_i(x)} dx - \sum p'_i \int \frac{q_i(x) q_j(x)}{\sum p'_i q_i(x)} dx =$$

$$= \int q_j(x) \log \frac{q_j(x)}{\sum p'_i q_i(x)} dx - \int_S q_j(x) dx$$

es decir,

$$\lambda = \int q_j(x) \log \frac{q_j(x)}{\sum p'_i q_i(x)} - 1 \qquad (i)$$

multiplicando por p' y sumando en j en la expresion anterior tendremos

$$\lambda \sum_{j=1}^{k} p'_j = I(\theta \| X)_{p'} - \sum_{j=1}^{k} p'_j$$

$$\lambda = I(\theta \| X)_{p'} - 1 \qquad (ii)$$

de (i) y (ii) deducimos

$$I(\theta \| X)_{p'} = \int q_j(x) \log \frac{q_j(x)}{\sum p'_i q_i(x)} dx$$

siendo

$$p' \in P^k \ / \ C(X) = \max_{p \in P^k} I(\theta \| X)_{p'}$$

lo que demuestra el teorema

---

(13) Introducimos el sub-indice p en la notacion de la medida de informacion $I(\theta \| X)_p$, para indicar cúal es el vector de probabilidades a priori respecto del cual hemos calculado la informacion producida por el experimento E.

REFERENCIAS DEL CAPITULO

BLACKWEL (1.953). "Comparisons 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)

BLACKWEL-GIRSHIK (1.954) "Theory of Games and Statistical Decisions". Ed. J. Wiley.

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

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.

LEHMANN (1.959). "Testing Statistical Hipothesses" Ed. J. Wiley

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

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

RENYI (1.960). "On measure of entropy and informations". Proceedings of the Fourth Berkeley Symposium on Math. Stat. and Prob. (547-561)

SAVAGE (1.954). "The Foundations of Statistics". Ed. J. Wiley

SHANNON (1.948)."A mathematical theory of communication". Bell. Syst. Tech. Journal 27 (379-423, 623-656). Reeditado por University of Illinois Press en 1969 (Shannon-Weaver," A Mathematical Theory of Information)

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

WALD (1.950). "Statistical Decision Functions". Ed. J. Wiley

WILKS (1.962). "Mathematical Statistics". Ed. J. Wiley

ZACKS (1.971). The Theory of Statistical Inference". Ed. J. Wiley

APENDICE

CALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN ALFA/

El programa adjunto calcula la entropia generalizada de la de Renyi de orden α , según la fórmula

$$H_{\alpha}^{N}(P) = \frac{N}{N-\alpha} \log \frac{\sum_{i=1}^{M} P_i^{\alpha/N}}{\sum_{i=1}^{M} P_i}$$

con $0 < \alpha < N$; $N \geq 1$ y $\sum_{i=1}^{M} p_i \leq 1$, para cuatro esquemas de probabilidad distintos.(Ver tablas 1, 2, 3 y 4)

En cada una de estas tablas de doble entrada, las filas representan la entropia para un α fijo y las columnas la entropía para cada n. La primera columna nos da, para los diferentes valores de α ($\alpha \leq 1$), el valor de la entropia de Renyi de orden α , $H_{\alpha}(p)$.

La primeratabla corresponde al caso de igualdad de las probabilidades. Se observa que en este caso el valor de $H_{\alpha}^{N}(p)$ es independiente de los valores de α y N.

Las tablas 2 y 4 corresponden a esquemas probabilisticos con $\sum_{i=1}^{M} p_i = 1$, y la tabla 3 a un esquema de probabilidad generalizado ($\sum_{i=1}^{M} p_i < 1$). En todas ellas se observa la monotonia de $H_{\alpha}^{N}(p)$ tanto en α como en n. A medida que α aumenta $H_{\alpha}^{N}(p)$ decrece y a medida que n aumenta, $H_{\alpha}^{N}(p)$ crece.

El programa completo consta de un programa principal y una subrutina FUNCTION.

El programa principal se ocupa fundamentalmente de la preparación y lectura de los datos, asi como de la salida de los resultados, y la subrutina calcula, a instancias del programa -- principal, la entropia para cada par (α, n).

SUBRUTINA FUNCTION

```
1*              FUNCTION H(P,R,T)
2*              DIMENSION P(5)
3*              S=0.
4*              S1=0.
5*              DO 1 I=1,5
6*              S=S+P(I)**(RNT)
7*            1 S1=S1+P(I)
8*              Z=ALOG(S/S1)
9*              H=(T-R)*ZNALOG(2.)
10*             RETURN
11*             END
```

END OF COMPILATION: NO DIAGNOSTICS.

```
      C     ESTE PROGRAMA CALCULA LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN
      C     ALFA.
      C
            DIMENSION FORM(24)
            DIMENSION ENE(7),ALFA(9),A(9,7),P(5)
      C
      C     INTRODUCCION DE DISTINTOS VALORES DE ALFA
      C
            DATA (ALFA(I),I=1,9)/0.3333,0.5,0.6666,0.75,1.,1.5,2.,3.,5./
      C
      C     INTRODUCCION DE DISTINTOS VSLORES DE N
      C
            DATA(ENE(I),I=1,7)/1.,1.5,2.,3.,5.,10.,30./
      C
      C     PARTE FIJA DEL FORMATO VARIABLE
      C
            DATA(FORM(I),I=1,5)/6H(5X,7(,6H1H+,8X,6H),1H+/6H1X,F5.,6H3,
            DATA (FORM(I),I=20,24)/3H1H+,8X,7(1H+,8X),1H+/70(1H+)/
            NNN=0
          1 DO 50 II=6,18,2
            FORM(II)=6H1H+,F7
         50 FORM(II+1)=*,4,1X,*
            BLANC=*       *
      C
      C     ENCABEZAMIENTO Y LECTURA DE DATOS PARA CADA TABLA
      C
            READ(5,10) P
            WRITE(6,20) P
            WRITE(6,30)ENE
      C
      C     CALCULO DE LA ENTROPIA A TRAVES DE LA FUNCION H PARA CADA PAR
      C     (ALFA,N), CON ALFA<N
      C
            DO 91 I=1,9
            IF(I-4)4,4,5
          4 L=1
            GO TO 80
          5 L=I-3
            M=I+L-1
      C
      C     REDEFINICION EN CADA CASO DEL FORMATO VARIABLE
      C
            FORM(M)=6H*,*,A8
            FORM(M+1)=*,      *
            M=I-4
            DO 70 II=1,M
         70 A(I,II)=BLANC
         80 DO 92 J=L,7
         92 A(I,J)=H(P,ALFA(I),ENE(J))
         91 WRITE(6,FORM) ALFA(I),(A(I,J),J=1,7)
            NNN=NNN+1
            WRITE(6,100) NNN
            IF(NNN-4)1,2,2
          2 STOP
         10 FORMAT(5F10.0)
         20 FORMAT('1CALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORD
           *EN ALFA'/65('*')/' PARA LAS PROBABILIDADES:',4(F7.4,','),F7.4/65
           *('*'))
         30 FORMAT(6X,7('*',F6.1,2X),'*'/6X,7('*',8X),'*'/70('*'))
         40 FORMAT(6X,7('*',8X),'*'/6X,F5.3,7('*',F7.4,1X),'*'/6X,7('*',8X),'*
           *'/70('*'))
        100 FORMAT(32X,'TABLA',I2,31X)
            END

END OF COMPILATION:       NO  DIAGNOSTICS.
```

CALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN ALFA

PARA LAS PROBABILIDADES: .2000, .2000, .2000, .2000, .2000

α \ N	1.0	1.5	2.0	3.0	5.0	10.0	30.0
.333	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219
.500	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219
.667	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219
.750	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219	2.3219
1.000		2.3219	2.3219	2.3219	2.3219	2.3219	2.3219
1.500			2.3219	2.3219	2.3219	2.3219	2.3219
2.000				2.3219	2.3219	2.3219	2.3219
3.000					2.3219	2.3219	2.3219
5.000						2.3219	2.3219

TABLA 1

ALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN ALFA

ARA LAS PROBABILIDADES: .3333, .1666, .0833, .2500, .1666

α \ N	1.0	1.5	2.0	3.0	5.0	10.0	30.0
.333	2.2749	2.2903	2.2982	2.3061	2.3125	2.3174	2.3206
.500	2.2522	2.2749	2.2864	2.2982	2.3077	2.3149	2.3198
.667	2.2304	2.2597	2.2749	2.2903	2.3029	2.3125	2.3190
.750	2.2197	2.2522	2.2691	2.2864	2.3006	2.3113	2.3186
.000		2.2304	2.2522	2.2749	2.2935	2.3077	2.3174
.500		2.2197	2.2522	2.2795	2.3006	2.3149	
.000			2.2304	2.2657	2.2935	2.3125	
.000				2.2390	2.2795	2.3077	
.000					2.2522	2.2982	

TABLA 2

CALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN ALFA

PARA LAS PROBABILIDADES: .3750, .3333, .1666, .0833, .0416

α \ N	1.0	1.5	2.0	3.0	5.0	10.0	30.0
.333	2.1880	2.2306	2.2528	2.2754	2.2939	2.3080	2.3175
.500	2.1282	2.1880	2.2197	2.2528	2.2800	2.3009	2.3151
.667	2.0736	2.1476	2.1880	2.2306	2.2663	2.2939	2.3127
.750	2.0482	2.1282	2.1725	2.2197	2.2595	2.2904	2.3115
1.000		2.0736	2.1282	2.1880	2.2394	2.2800	2.3080
1.500			2.0482	2.1282	2.2005	2.2595	2.3009
2.000				2.0736	2.1634	2.2394	2.2939
3.000					2.0948	2.2005	2.2800
5.000						2.1282	2.2528

TABLA 3

CALCULO DE LA ENTROPIA GENERALIZADA DE LA DE RENYI DE ORDEN ALFA

PARA LAS PROBABILIDADES: .0500, .2000, .2000, .1500, .4000

α/N	1.0	1.5	2.0	3.0	5.0	10.0	30.0
.333	2.2329	2.2614	2.2761	2.2911	2.3033	2.3125	2.3188
.500	2.1922	2.2329	2.2542	2.2761	2.2941	2.3079	2.3172
.667	2.1540	2.2055	2.2329	2.2614	2.2850	2.3033	2.3157
.750	2.1358	2.1922	2.2225	2.2542	2.2806	2.3010	2.3149
1.000		2.1540	2.1922	2.2329	2.2672	2.2941	2.3125
1.500			2.1358	2.1922	2.2413	2.2806	2.3079
2.000				2.1540	2.2163	2.2672	2.3033
3.000					2.1690	2.2413	2.2941
5.000						2.1922	2.2761

TABLA 4

@FIN

FACULTAD DE CIENCIAS

nido el Tribunal integrado por los abajo firman
l día de la fecha, para juzgar la Tesis Docto
D. Antonio Pascual Acosta
ulada "Algunos Resultados sobre Teoría
de la Información"

oró otorgarle la calificación de SOBRESALIENTE
CUM LAUDE

Sevilla, 15 de MARZO 1.976

El Vocal, El Vocal, El Vocal,

El Presidente, El Secretario, El Doctorando

UNIVERSIDAD DE SEVILLA · FACULTAD DE CIENCIAS ·