DISTRIBUCIONES DE REFERENCIA

1980

Estadística e Investigación de Operaciones

TESIS QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS, PRESENTA

RAUL RUEDA DIAZ DEL CAMPO

A RENATA

INDICE

INTRODUCCION

El problema de asignar una distribución apriori o inicial que "deje hablar a los datos", ha sido tratado desde hace tiempo (Bayes, 1763; Laplace, 1820). Numerosos trabajos han sido escritos en ese sentido sin llegar a unificar criterios en como definir a tal distribución inicial; y aún, en muchos casos, este tipo de enfoques no son conocidos, debido a que aparecen en revistas científicas de diversa índole. Es por esto y por la relevancia que este tema tiene, el que una revisión bibliográfica se hace necesaria.

Este trabajo tiene como objetivo principal, el presentar los métodos más comunes que se han desarrollado, para encontrar una distribución inicial que "deje hablar a los datos" a lo que comúnmente se le llama no informativa o de referencia.

La revisión comprende desde el trabajo de Bayes (1763), pasando por importantes proposiciones como son las de Jeffreys (1937, 1946), Perks (1947), Barnard (1952), Hartigan (1964; 1965), Stone (1965, 1970), Jaynes (1968), etc., concluyendo con un nuevo enfoque propuesto por Bernardo (1979), basado en la Teoría de la Información. Se comentará cada una de estas técnicas marcando los aciertos y desaciertos que contenga y finalmente, la técnica de Bernardo (1979) será ilustrada con algunos ejemplos.

Con esto, se pretende dar un panorama más o menos amplio sobre el tema, que de ninguna manera trata de ser exhaustivo por la complejidad de rastrear toda la bibliografía referente al mismo.

La necesidad de contar con una distribución no informativa o de referencia, surge a partir de que una forma coherente de hacer inferencias sobre los parámetros de una distribución, es asignando una distribución inicial y usando el teorema de Bayes, encontrar la distribución final (Lindley, 1971; Bernardo,1977). Sin embargo, como se menciona en la literatura muchas veces es útil tener un punto de referencia contra el cual comparar las inferencias que pueden ser obtenidas a partir de los datos, sin un conocimiento prejuiciado de los parámetros. Otras veces la falta de recursos o tiempo puede volver prohibitiva la designación de una distribución inicial informativa. Cuando un problema es atacado por un grupo de personas, en donde la distribución inicial informativa no es necesariamente la misma para todos, el usar una distribución de referencia puede ayudar a resolver el problema de conflicto.

Los términos "no información", "conocimiento vago", "dejar hablar a los datos", son usados frecuentemente, como sinónimos de "referencia", sin embargo son términos ambiguos pues la "no información" no está definida en forma única, al igual que los demás; pues no información y conocimiento vago, son conceptos que dependen de la persona que los utiliza, ya que es obvio que un conocimiento vago que pueda tener un físico sobre un

problema de su área, será muy distinto al conocimiento vago o a la no información que, sobre el mismo problema, pueda tener un médico o una ama de casa , o aún más, otro físico. El término "dejar hablar a los datos", presenta también problemas sobre su significado, pues, ¿qué debe entenderse por dejar hablar a los datos? ¿No introducir una apriori, a menos que sea constante, y hacer siempre la distribución final proporcional a la verosimilitud? o ¿simplemente es asignar una apriori que no contradiga ni apoye a la información?.

El término de distribución de referencia es tal vez, el más adecuado y el uso de las distribuciones iniciales de referencia en los reportes técnicos, parecen indicarlo, pues son usadas en el sentido expuesto anteriormente, y mencionado por la literatura. Así, el término que se usará en este trabajo para las distribuciones iniciales que no sean informativas, será el de distribuciones de referencia y el mismo será utilizado para las distribuciones finales obtenidas a partir de una distribución inicial de referencia, esto es, se les llamará distribuciones finales de referencia.[1]

1.- Los términos inicial y final son equivalentes a los términos apriori y posteriori, por lo cual se usarán indistintamente.

CAPITULO I

En este capítulo se presentarán los principios básicos en los que se basa la metodología bayesiana. Se planteará el problema de decisión en forma general. Se establecerán los principios de coherencia, que dictarán una forma de comportamiento. A partir de estos principios, se definirá la probabilidad como un grado de creencia y se dará una definición de función de utilidad. Se verá que la forma coherente de actuar será seleccionar la decisión que maximice la utilidad esperada.

Se dará una descripción somera de dos tipos de análisis que pueden ser utilizados para resolver un problema de decisión: Análisis en forma normal y análisis en forma extensiva. Se introducirá el teorema de Bayes, especificando a la distribución inicial, la distribución final y la verosimilitud, explicando la importancia de cada una de ellas. Se planteará el principio de verosimilitud, mostrando como una violación al principio de verosimilitud, se traduce a una violación de los principios de coherencia.

Finalmente, se dará una breve descripción de los problemas que la metodología bayesiana resuelve, concluyendo con una exposición sobre las diferencias entre inferencia y decisión.

I.1 ANTECEDENTES HISTORICOS

Aunque los objetivos de este trabajo no comprenden una reseña histórica del desarrollo de la probabilidad y de la estadística, se dará un pequeño bosquejo histórico con el fín de centrar la discusión y mostrar la importancia que las distribuciones de referencia han tenido dentro del desarrollo de la metodología bayesiana.

Para la mayoría de la gente es claro que la interpretación subjetivista de la probabilidad, concebida como grados de creencia, no es una interpretación nueva, aunque aparentemente, no tan antigua como lo es la interpretación objetivista, ya sea clásica o frecuentista. Esta última empezó a desarrollarse a partir del trabajo de J. Bernoulli publicado después de su muerte en 1713, titulado Ars Conjectandi; mientras que la primera empezó a trabajarse en los tiempos de Fermat y Pascal (1654).

Uno de los trabajos más importantes enfocado a interpretar subjetivamente la probabilidad fué escrito por Thomas Bayes (1763), en el cual establece su famoso teorema, conocido como teorema de Bayes, y que dá nombre a la estadística bayesiana por razones que se mencionarán más adelante. Sin embargo, este artículo no fué presentado por él, sino por Richard Price quien lo mandó a John Canton, F.R.S.* después de la muerte de Bayes.

*(Fellow of the Royal Society: Asesor de la Real Sociedad de Estadística de Inglaterra).

En este trabajo, se muestra como calcular probabilidades de sucesos en - base a otros sucesos observados (probabilidades condicionales), y se mencionа el famoso postulado de Bayes, del cual se hablará con más detalle un poco adelante; el postulado va relacionado con el teorema, pues en este último se menciona que para conocer la probabilidad de un suceso a - partir de los resultados de un experimento, es necesaria una proposición apriori sobre la probabilidad de dicho suceso; si ésta no se tiene, no - puede deducirse una proposición probabilística del suceso a menos que el postulado sea utilizado. Dicho postulado establece que en caso de ignorancia del fenómeno, una distribución uniforme debe ser asociada al fenómeno, apriori.

Dos cosas son importantes de hacer notar: la primera es la concepción de Bayes de una distribución apriori no informativa o de referencia, en los casos de ignorancia. De aquí que el problema de encontrar una distribución de referencia es tan antiguo como importante, siendo éste, un problema que muchos han querido resolver, por su relevancia en la metodología bayesiana, como se verá más adelante.

La segunda observación importante, es el hecho de que, aparentemente, Bayes no estaba completamente convencido de que su postulado fuese la solución para el problema de no información, y fué precisamente por este problema, la distribución de referencia, que la idea de Bayes fué perdiendo fuerza.

Laplace es, después de Bayes, el segundo "subjetivista" importante. A pesar de que, en la que fué su mejor obra Teoría Análitica de la Probabilidad (1820), define la probabilidad en forma clásica: casos posibles entre casos totales; discute ampliamente el carácter de la probabilidad como un grado de creencia, como muestran varios ejemplos que Laplace menciona en su libro Ensayo Filosófico de la Probabilidad que puede considerarse como una introducción al libro mencionado anteriormente. Al igual que Bayes, Laplace se enfrenta al problema de no información, el cual resuelve aparentemente en forma similar a la de Bayes*.

A pesar de estos grandes trabajos, el desarrollo subjetivista no tuvo gran impulso, aunque fué bien acogida esta interpretación a principios del siglo XIX.

Una de las grandes causas, fueron los trabajos desarrollados por Sir Ronald A Fisher, gran estadístico inglés, a principios del presente siglo, y basados, muchos de ellos, en una interpretación frecuentista. Fisher criticó duramente el postulado, aunque reconoció ampliamente las ideas de Bayes, y como consecuencia introdujo el argumento fiducial en 1930 (J. Fisher, 1978). Lo que más criticaba Fisher, era que si se "desconocía todo" sobre Θ , entonces era claro que lo mismo sucedía con Θ^2 , pero que sin embargo una distribución uniforme sobre Θ no era equivalente

* A pesar de que no se pudo conseguir el libro mencionado, esta afirmación es mencionada por varios autores, por lo que se incluye.

a una sobre θ^2, lo cual llevaba a que la "no información" no podía ser - medida en esa forma. Argumentos estos, que junto con el carácter subjetivo que se le daba a la probabilidad, impidieron un desarrollo de esta corriente.

Sin embargo, hubo gente que empezó a preocuparse en cómo formalizar axiomáticamente el hecho de trabajar la probabilidad como un grado de creencia. Trabajos fundamentales en este sentido son los de Ramsey (1926) y - De Finetti (1937). A pesar de esto, la metodología bayesiana (que se basa en la interpretación subjetivista de la probabilidad) fué opacada por el gran desarrollo que la estadística ortodoxa (interpretación objetiva), tuvo, debido a, como se mencionó anteriormente, los trabajos de Fisher y en general al trabajo de la escuela inglesa.

No fué sino hasta 1954, con la aparición del libro de Leonard J. Savage "The Foundations of Statistic" que los estadísticos volvieron a pensar - en la metodología bayesiana, no como una parte de la estadística como - muchos creen, sino como una alternativa global a la estadística ortodoxa. A partir de la aparición de este libro, la metodología bayesina ha cobrado nuevo impulso, gracias a los trabajos de personas como D. Lindley, M. De Groot, B. De Finetti, J.M. Bernardo, A. Zellner, I.J. Savage, L.J. Savage, etc. Sin embargo, el proolema de las distribuciones de referencia, sigue siendo actual.

I.2 EL PROBLEMA GENERAL DE DECISION

En general, en una situación dada, el problema de tomar una decisión que defina un curso de acción, no es sencillo. Por ejemplo, cuando se desea comprar un auto, habrá que considerar factores como: precio, garantía, rendimiento en gasolina, tamaño, color, etc. y no simplemente comprar cualquier auto. En estos casos se dice que se tiene un problema de decisión y lo que se desea es seleccionar de entre varias, la mejor decisión posible. Esto es, se tienen varias alternativas y se busca elegir la mejor, en cierto sentido. Cuando se tiene un sólo curso de acción, es obvio que el problema de decisión no existe, por lo que se hace necesaria al menos una alternativa para que exista. En general no es bueno considerar una cierta situación y su negación como alternativas, pues en caso de decidirse por la negación, ésta puede no plantear un curso de acción. Como ejemplo, supóngase que una alternativa es comprarse un determinado auto con ciertas especificaciones y considérese la otra alternativa, como la negación; esto es, no comprarse ese auto. Si se decide que la segunda alternativa es la mejor, no se tiene una idea de que acción seguir después. Lo recomendable es construir un conjunto que contenga todas las posibles alternativas u opciones, al que se le llamará espacio de decisiones y se denotará con D. Este conjunto deberá ser exclusivo y exhaustivo por razones obvias; aunque lo segundo no sea fácil de asegurar en la práctica, deberá tratarse de que se cumpla. Ya definido este conjunto es necesario considerar dos casos: el primero a considerar, es

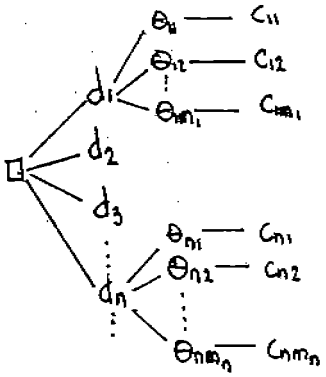
cuando se tiene una situación en la que existe información perfecta. - Ejemplo de éstas, puede ser el juego de ajedrez. Desde el punto de vista teórico, este tipo de problemas son de fácil solución, aunque el número de alternativas es muy grande, para que a nivel práctico lo sea. Sin embargo, se considerará que en estos casos el problema está resuelto.

El segundo caso a considerar y que es el que se tratará en este trabajo, es el problema de la toma de decisiones en ambiente de incertidumbre. Esto es, las consecuencias de tomar cierta decisión, dependerá de ésta y - de la ocurrencia de cierto suceso incierto. Para ejemplificar, supóngase que se quiere tomar una decisión sobre llevar o no consigo un impermea - ble. Si se sabe con certeza que va a llover, la decisión es trivial. El problema reside en no saber con exactitud el estado del tiempo para ese día.

Para poder tomar una decisión, es necesario especificar todos los suce - sos inciertos relevantes al problema, así como las posibles consecuen - cias que se deriven de su ocurrencia. Al igual que en el espacio de decisiones, se debe tener definido un conjunto que contenga a los sucesos inciertos en forma exhaustiva y exclusiva, sea éste Θ, el conjunto de todas las consecuencias posibles asociadas a los elementos de Θ

En forma gráfica la relación existente entre D, Θ y C, puede represen-

tarse como



en donde $d_i \in D$, $\theta_{ij} \in \Theta$ y $c_{ij} \in C \quad \forall j \in J_{m_i}, i \in J_n$.

En resumen, para que un problema de decisión exista, deben estar definidos los conjuntos Θ y D. En algunos casos es necesario realizar algún experimento para tomar una decisión, que estará basada en la información provista por el experimento. En estos casos, además de especificar a Θ y a D, se debe especificar al conjunto de todos los experimentos relevantes, denotado por E, y al conjunto de resultados, Z. Sin embargo estos casos no se considerarán por el momento, sino hasta I.4.

I.3 PRINCIPIOS DE COHERENCIA

Ya teniendo definidos a Ⓗ y D, el siguiente paso es seleccionar la decisión que lleve al "mejor" curso de acción. Es obvio que el adjetivo "mejor" dependerá de muchas cosas, desde la persona que va a tomar la decisión hasta las consecuencias que ésta acarrearía. Una forma de resolver esto, es plantear unos principios de comportamiento coherente y encontrar la mejor decisión a partir de ellos. Esto lo que significa es que quien crea estos principios debe basar su mejor decisión en la que se derive de ellos, lo cual no implica que sea la única forma de definir "mejor". Sin embargo, en este trabajo, la "mejor" decisión será la que se desprenda de los principios y cualquier otra forma de tomar decisiones o es equivalente o se considerará incoherente y por tanto insostenible.

Para poder establecer los principios de coherencia, se considerará primero lo que es una opción. Sea L el conjunto de opciones, en donde una opción se define como

$$\ell = \left\{ C_1|A_1 \, , \, C_2|A_2 , \cdots , C_j|A_j , \cdots \right\}$$

y que denota una situación en la que se obtiene la consecuencia C_i si el suceso incierto A_i sucede, y esto para toda i ; donde además se pide que el conjunto $\{A_i\}_{i \in I}$ forme una partición del conjunto de sucesos inciertos.

Ahora bien, como las decisiones dependen de las consecuencias que aca --

rrearían; para poder tomar una decisión será necesario que estas conse - cuencias sean comparables, por lo que es natural que también las opcio - nes puedan ser comparadas.

Para esto se introduce la siguiente relación de orden

i) Si $l_1 \prec l_2$, se dice que la opción l_2 es preferible a la opción l_1.

y ii) Si $l_1 \sim l_2$, se dice que las opciones son igualmente deseables.

De la misma forma, puede introducirse una relación de orden entre las - consecuencias:

i) Si $C_1 \prec C_2$, la consecuencia C_2 es preferible a C_1

y ii) Si $C_1 \sim C_2$, lasconsecuencias son igualmente deseables.

Ya establecidos estos conceptos, los principios de comportamiento cohe - rente son los siguientes

POSTULADO 1 (Comparabilidad).- $\forall\, l_1, l_2 \in L$ se tiene que $l_1 \prec l_2$ o $l_2 \prec l_1$ o $l_1 \sim l_2$. $\forall\, c \in C\ \exists\ C_*, C^*$ t. $C_* \preccurlyeq C \preccurlyeq C^*$.

POSTULADO 2 (Transitividad).- $\forall\, l_1, l_2, l_3 \in L$ si $l_1 \preccurlyeq l_2$ y $l_2 \preccurlyeq l_3$ entonces $l_1 \preccurlyeq l_3$

POSTULADO 3 (Sustituibilidad).- Si $l_1 \prec l_2$ en presencia de A y $l_1 \prec l_2$ en presencia de A^c, entonces $l_1 \prec l_2$ siempre. Esto es equivalente a de- cir que si en una opción, una conse cuencia igualmente deseable se subs tituye, la nueva opción así genera- da es equivalente a la primera.

POSTULADO 4 (Experimento Auxiliar).- Existe un procedimiento $z \in [0,1]^2$ tal que si $l_1 = \{z \in R_1 : R_1 \subset [0,1] \times [0,1]\}$ y $l_2 = \{z \in R_2 : R_2 \subset [0,1] \times [0,1]\}$ entonces

$$l_1 \prec l_2 \iff A(R_1) < A(R_2).$$

El postulado 1 especifica que cualquier par de opciones pueden ser comparadas, y que además existe la peor consecuencia, C_* y la mejor C^*. Por otro lado, dado que en el problema de decisión es necesario cuantificar el peso de cada una de las consecuencias que se pueden tener al tomar una decisión y que cierto suceso incierto suceda, es forzoso tener procedimientos para asignar esa cuantificación; uno de estos queda establecido en el cuarto postulado, sin embargo, esto no significa que la persona que vaya a tomar una decisión deba diseñar especificamente el experimento, basta que esté consciente de que puede realizarse. Por otro lado, estos principios pueden ser defendidos formalmente como lo muestra Pratt, Raiffa & Schlaifer (1964) en el caso finito, Savage (1954,1961) y Bernardo (1977).

El siguiente paso lógico es definir una medida de probabilidad que mida la credibilidad de los sucesos inciertos, pues es obvio que el grado de posibilidad que tenga cierto suceso incierto, influirá en la toma de decisiones. Esta medida de probabilidad debe ser consistente con los principios de coherencia y además la probabilidad de un suceso A debe estar condicionada al medio ambiente que rodea al suceso, y el cual se denotará con H.

Considérense las siguientes opciones

$$\ell_1 = \{C_* | \bar{A}, c^* | A\} \quad y \quad \ell_2 = \{C_* | \bar{R}, c^* | R\}^2 \text{ con } R \subseteq [0,1] \times [0,1];$$

si se denota con $p(A|H)$ la probabilidad de A en presencia de las condiciones H, una definición consistente con los principios de coherencia es $p(A|H) = A(R)$, en donde R debe ser tal que $\ell_1 \sim \ell_2$. Para encontrar R tal que $\ell_1 \sim \ell_2$, basta definir $R(x)$ en $[0,1] \times [0,1]$ un rectángulo de lado x tal que

$$\ell_x = \{c^* | R(x), C_* | \overline{R(x)}\}$$

Si $x = 0 \Rightarrow \ell_{(0)} = \{c^* | \phi, C_* | [0,1]^2\}$ y si $x = 1$ se tiene que $\ell_{(1)} = \{c^* | [0,1]^2, C_* | \phi\}$,

de donde, por el principio de comparabilidad

$$C_* = \ell_{(0)} \preceq \ell_1 \preceq \ell_{(1)} = c^* .$$

Como ℓ_x es una función contínua en $[0,1] \times [0,1] \Rightarrow \exists\ x \in [0,1] \ni \ell_x = \ell_1$, por lo que siempre existe $R(x) \subset [0,1] \times [0,1]$ para cada suceso A, tal que $\{c^* | A, C_* | \bar{A}\} \sim \{c^* | R(x), C_* | \overline{R(x)}\}$, pudiendo así definirse a $p(A|H)$ como $p(A|H) = A(R(x))$. Además puede demostrarse, usando el principio de transitividad, el que $p(\cdot|H)$ es una función, y que la definición es independiente de las consecuencias C_* y C^* consideradas.

Esta construcción de la medida de probabilidad, muestra que

i) La probabilidad es un grado de creencia y por tanto una apreciación personal; y

2.- Donde $\bar{A}$ denota el complemento de

ii) La probabilidad no puede ser absoluta, siempre está condicionada al medio que la rodea.

A pesar de que esta interpretación de la probabilidad difiere a la ortoxa; la función, considerada como medida, debe cumplir los axiomas de Renyi-Kolmogoroff, que no son más que los axiomas de Kolmogoroff, para el caso de probabilidades condicionales y que fueron propuestas por Renyi (1961).

Estos axiomas son

i) $p(A|H) \in [0,1]$ y $p(H|H) = 1$

ii) Si $ABH = \emptyset \Rightarrow p(A \cup B|H) = p(A|H) + p(B|H)$

iii) $p(AB|H) = p(A|H)\, p(B|AH) = p(B|H)\, p(A|BH)$

Es obvio que el primer axioma es satisfecho por la medida de probabilidad que se ha construído, pues toda área es no negativa y lo más que puede valer es uno, que es el área de $[0,1] \times [0,1]$.

Para verificar el segundo, considérese $R_1 \subset [0,1]\times[0,1]$ $\ni$ $p(A|H) = A(R_1)$ y considérese también a $R_2 \subset [0,1]\times[0,1]$ $\ni$ $p(A \cup B|H) = A(R_2)$, entonces $\{C^*|R_1, C_*|\bar{R}_1\} \sim \{C^*|A, C_*|\bar{A}\}$ y $\{C^*|R_2, C_*|\bar{R}_2\} \sim \{C^*|A \cup B, C_*|\overline{A \cup B}\}$ y como en presencia de H, A y B son incompatibles se sigue que

$$\{C^*|R_2 \setminus R_1, C_*|\overline{R_2 \setminus R_1}\} \sim \{C^*|B, C_*|\bar{B}\}$$

ya que si supone que

$$\{C^* | R_2 \setminus R_1, C_* | \widetilde{R_2 \setminus R_1}\} > \{C^* | B, C_* | \bar{B}\}$$

y sabiendo que

$$\{C^* | R_1, C_* | \bar{R}_1\} \sim \{C^* | A, C_* | \bar{A}\}$$

se sigue que

$$\{C^* | (R_2 \setminus R_1) \cup R_1, C_* | \widetilde{(R_2 \setminus R_1) \cup R_1}\} > \{C^* | B, C_* | \bar{B}\}$$

por los principios de sustituibilidad y transitividad; lo que lleva a

$$\{C^* | R_2, C_* | \bar{R}_2\} > \{C^* | A \cup B, C_* | \widetilde{A \cup B}\}$$

lo cual es una contradicción.

De igual forma, el suponer que $\{C^* | R_2 \setminus R_1, C_* | \widetilde{R_2 \setminus R_1}\} < \{C^* | B, C_* | \bar{B}\}$ lleva a una contradicción, por tanto es cierto $\{C^* | R_2 \setminus R_1, C_* | \widetilde{R_2 \setminus R_1}\} \sim \{C^* | B, C_* | \bar{B}\}$ lo que

$$\Rightarrow p(B|H) = A(R_2 \setminus R_1) = A(R_2) - A(R_1)$$

$$\Rightarrow p(B|H) = p(A \cup B|H) - p(A|H)$$

$$\therefore p(A \cup B|H) = p(A|H) + p(B|H) \quad \text{si} \quad ABH = \phi.$$

Para demostrar que la medida $p(\cdot|H)$ cumple el tercer axioma se procede en forma similar. Sean $R_1, R_2 \subset [0,1] \times [0,1]$ t. $\{C^* | R_1, C_* | \bar{R}_1\} \sim \{C^* | A, C_* | \bar{A}\}$ y $\{C^* | B|A, C_* | \widetilde{B|A}\} \sim \{C^* | R_2, C_* | \bar{R}_2\}$ teniéndose que

$$A(R_1) = p(A|H) \quad \text{y} \quad A(R_2) = p(B|AH)$$

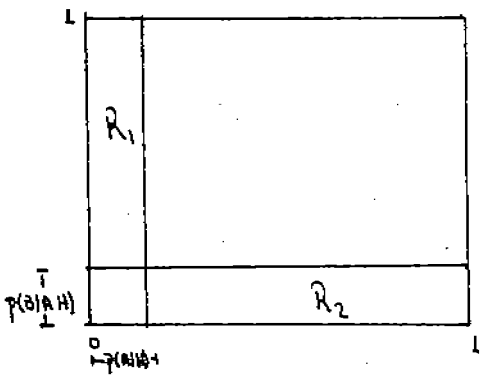
Usando el principio de sustituibilidad se tiene

$$\{C^* | R_1 R_2, C_* | \widetilde{R_1 R_2}\} \sim \{C^* | A(B|A), C_* | \widetilde{A(B|A)}\}$$

lo cual nuevamente por el principio de sustituibilidad, implica que

$$\{C^* | R_1 R_2, C_* | \widetilde{R_1 R_2}\} \sim \{C^* | AB, C_* | \widetilde{AB}\}$$

Ahora bien, como lo muestra el dibujo R_1 y R_2 pueden seleccionarse de forma que $A(R_1R_2) = p(B|AH)\,p(A|H)$ y como



$p(AB|H) = A(R_1R_2)$

$\Rightarrow p(AB|H) = p(A|H)\,p(B|AH)$.

De igual forma, cambiando los papeles de A y B, se demuestra $p(AB|H) = p(B|H)\,p(A|BH)$

$\therefore\ p(AB|H) = p(A|H)\,p(B|AH) = p(B|H)\,p(A|BH)$

que es lo que se quería demostrar. Por lo que $p(\cdot|H)$ definida como un grado de creencia cumple con ser una medida de probabilidad. Además, por el principio de sustituibilidad, las opciones

$$\ell = \{C_1|A_1, \ldots, C_n|A_n, \ldots\} \quad \text{y} \quad \ell' = \{C_1|p(A_1), \ldots, C_n|p(A_n), \ldots\}$$

son equivalentes.

Ya que se ha construído una medida de probabilidad que cuantifica y que permite comparar los sucesos inciertos, se hace necesario definir una medida que cuantifique las consecuencias y de esta forma elegir la decisión u opción que tenga "menos consecuencias" de acuerdo a esta medida. El postulado de comparabilidad transitiva de las opciones, y por tanto de las consecuencias, permite construir una función $u \in [0,1]^C$ que mida la deseabilidad de cada consecuencia C. Es obvio que $u(C^*)=1$ y $u(C_*)=0$. De esta forma, puede definirse a $u(C)$ como la probabilidad que debe asignarse a C^* para que $C \sim \{C^*|R, C_*|\bar{R}\}$, donde $R \subset [0,1]\times[0,1]$ es tal que $A(R) = p(A|H) = u(C)$ con A tal que $C = \{C^*|A, C_*|\bar{A}\}$ y $p(A|H) = A(R)$, esto es, $C \sim \{C^*|R, C_*|\bar{R}\}$ así $C \sim \{C^*|u(C), C_*|1-u(C)\}$.

Debido a los principios de coherencia, y sin necesidad de un nuevo artifi cio, se prueba que $\forall c \in C$ existe $u(c) \in [0,1]$ $\ni$ $c \sim \{c^* | u(c), c_* | 1-u(c)\}$ pues $c_* \preceq c \preceq c^*$. Además la utilidad de c $(u(c))$ queda así bien definida, salvo una trasformación lineal (Bernardo, 1977).

El último paso es asignar un número a cada una de las decisiones posibles, de forma tal que la mejor decisión sea aquella que tenga asociada la mejor consecuencia con la mayor probabilidad. Para esto, se sabe que tomar la decisión d_i es aceptar la opción siguiente

$$\{c_{i1} | \theta_{i1}, \dots, c_{ik} | \theta_{ik}, \dots\}$$

donde $\{\theta_{ij}, j \in J\}$ son los sucesos inciertos que pueden afectar las conse cuencias de la decisión d_i y c_{ij} es la consecuencia que se tiene al tomar d_i y sucede θ_{ij} (nótese que θ_{ij} depende de d_i y c_{ij} de θ_{ij} y d_i). Así el grado de creencia sobre la ocurrencia de θ_{ij} estará dado por - - $p(\theta_{ij} | d_i, H)$.

Ahora bien,

$$c_{ij} \sim \{c^* | u(c_{ij}), c_* | 1-u(c_{ij})\}$$

en donde por construcción

$$u(c_{ij}) = p\{c^* | d_i, \theta_{ij}, H\}$$

Ahora, usando el principio de sustituibilidad

$$d_i \sim \{(c^* | u(c_{i1}), c_* | 1-u(c_{i1})) | p(\theta_{i1} | d_i, H), \dots, (c^* | u(c_{ik}), c_* | 1-u(c_{ik})) | p(\theta_{ik} | d_i, H), \dots\}$$

$$\Rightarrow \quad d_i \sim \{(c^* | u(c_{i1}) p(\theta_{i1} | d_i, H), c_* | (1-u(c_{i1})) p(\theta_{i1} | d_i, H)), \dots, (c^* | u(c_{ik}) p(\theta_{ik} | d_i, H), c_* | (1-u(c_{ik})) p(\theta_{ik} | d_i, H)), \dots\}$$

$$\Rightarrow \quad d_i \sim \{c^* | \sum_{j \in J} u(c_{ij}) p(\theta_{ij} | d_i, H), c_* | 1 - \sum_{j \in J} u(c_{ij}) p(\theta_{ij} | d_i, H)\}$$

Definiendo a $u^*(d_i) = \sum_{j \in J} u(c_{ij})\, p(\theta_{ij} | d_i, H)$, se tiene que $u^*(d_i)$ representa la utilidad esperada de tomar la decisión d_i , pero también

$$u^*(d_i) = \sum_{j \in J} p(c^* | d_i, \theta_{ij}, H)\, p(\theta_{ij} | d_i, H).$$

Entonces, de acuerdo a lo dicho anteriormente con respecto a que la mejor decisión será aquella que de mayor probabilidad a la mejor consecuencia, la decisión óptima (denotada por d^*) será aquella que maximice la utili - dad esperada, esto es

$$d^* = \max_D u^*(d_i) = \max_D \sum_{j \in J} u(c_{ij})\, p(\theta_{ij} | d_i, H)$$

ya que $d_i > d_j$ si y sólo si $u^*(d_i) > u^*(d_j)$

En pocas palabras, toda persona que use los principios establecidos al comienzo de esta sección, como forma de comportamiento, debe tomar como mejor decisión, aquella que maximice la utilidad esperada[3]. Algunos ejemplos de criterios de decisión incoherentes y comentarios sobre este criterio se encuentran en Bernardo (1977) y Lindley (1977).

3.- Aunque la demostración de esto fué realizada para el caso al mas numerable, puede extenderse al caso contínuo, mediante el uso de la integral de Radon-Nikodym definida con respecto a la medida dominante del espacio.

I.4 ANALISIS EN FORMA NORMAL Y EN FORMA EXTENSIVA

En muchas ocasiones no basta el conocimiento que se tiene sobre una situación, para poder tomar una decisión. Por ejemplo, si en un momento dado se desea tomar una decisión sobre si construir una presa en determinado lugar, no basta con el conocimiento que sobre construcción de presas se tenga, sino que es necesario realizar estudios para ver si el terreno es adecuado, calcular en forma apróximada la capacidad de agua que la presa deba contener, etc. En estos casos se hace deseable experimentar e incrementar el conocimiento que se tenga, en base a los resultados que se obtengan del experimento. Esto significa, que debe considerarse además de Θ y Ω, al conjunto E de experimentos potenciales, que servirán para obtener más información acerca de $\theta \in \Theta$, el cual a su vez generará un conjunto Z, que estará formado por los resultados potenciales asociados a cada $e \in E$.

En estos casos no basta asociar una medida de probabilidad sobre Θ, sino que es necesario considerar una distribución de probabilidades sobre $\Theta \times Z$ la cual deberá de depender de elementos de E, puesto que la ocurrencia de un $z \in Z$ dependerá del $e \in E$ elegido.

De la misma forma, la función de utilidad deberá ser extendida a elementos de Z, la cual nuevamente dependerá de E.

Entonces, para cada $e \in E$, $p(\theta, z|e,H)$[4] definirá la probabilidad de obtener el resultado z y que ocurra θ si e se realiza. Análogamente $u(e,z,d,\theta)$[5] denotará la utilidad de realizar e, obtener z, seleccionar d y que suceda θ. De esta forma, el problema general de decisión queda planteada como sigue: estando dados E, Z, D y H, asignar u y $p(\theta,z|e)$, entonces seleccionar $e \in E$ y dado que $z \in Z$ es observado, escoger $d \in D$ de forma que la utilidad esperada se maximice. De esta forma, el problema es seleccionar el experimento que conduzca a la mejor decisión.

Existen dos formas básicas para resolver el problema de decisión planteado: la forma extensiva y la forma normal. En la forma extensiva el mejor curso de acción se construye considerando todos los resultados experimentales posibles a la vez, determinando para cada uno la decisión óptima y entonces seleccionar el experimento óptimo; mientras que en la forma normal se consideran primero todas las decisiones posibles para un $e \in E$ fijo y seleccionar de ahí la óptima para proceder después a encontrar el $e \in E$ óptimo[6].

4.- A fín de abreviar notación, de aquí en adelante se escribirá $p(\theta,z|e)$ en lugar de $p(\theta,z|e,H)$ que es lo correcto.

5.- Formalmente, la función de utilidad debería depender de C, la consecuencia que se dá si θ sucede, pero esa consecuencia depende de que decisión se seleccione.

6.- Los conceptos de optimilidad están definidos en términos de maximizar la utilidad esperada.

En términos un poco más formales, el análisis en forma extensiva consiste en lo siguiente.

Sea $\theta \in \Theta$ para cada terna (e, z, d) considérese

$$E_{\theta|z}\, u(e,z,d,\theta) = \int_{\Theta} u(e,z,d,\theta)\, p(\theta|e,z)\, d\theta^{7} \quad \forall e \in E,\ z \in Z \text{ y } d \in D$$

suponiendo (e, z) fijo, encontrar la decisión óptima, haciendo

$$\max_{D} \int_{\Theta} u(e,z,d,\theta)\, p(\theta|e,z)\, d\theta \quad \forall e \in E \text{ y } \forall z \in Z$$

para seleccionar el experimento óptimo, debe considerarse

$$E_{z|e} E_{\theta|z}\, u(e,z,d,\theta) = \int_{Z} \left(\max_{D} \int_{\Theta} u(e,z,d,\theta)\, p(\theta|e,z)\, d\theta \right) p(z|e)\, dz \quad \forall e \in E$$

y el experimento óptimo será aquel que maximice esta expresión, esto es

$$e^{*} = \max_{e \in E} \left\{ \int_{Z} \left(\max_{D} \int_{\Theta} u(e,z,d,\theta)\, p(\theta|e,z)\, d\theta \right) p(z|e)\, dz \right\}$$

Las primeras dos expresiones suponen que un experimento ha sido realizado y unos resultados han sido obtenidos[8] para encontrar la mejor decisión. A esta parte del análisis se le llama <u>análisis terminal</u>. Las últimas dos expresiones constituyen el llamado <u>análisis preposterior</u> y consiste en encontrar el mejor experimento posible que lleve a esa decisión óptima.

En el análisis en forma normal, el problema es encontrar $e \in E$ y $d \in D$ tal que maximice la expresión siguiente

$$E_{\theta z|e}\, u(e,z,d,\theta) = \int_{\Theta \times Z} u(e,z,d,\theta)\, p(z,\theta|e)\, d\theta \times dz^{9}$$

7.- Las integrales se definen con respecto a la medida dominante del espacio donde se opera.

8.- Uno de los experimentos puede ser "no experimentar"

9.- Aquí la integral se define en el espacio producto $\Theta \times Z$.

Para calcular $E_{\theta,z|e}$ pueden seguirse dos caminos, primero calcular $E_{\theta|z}$ y después $E_z E_{\theta|z}$ o bien, primero $E_{z|\theta}$ y posteriormente $E_\theta E_{z|\theta}$. Debido a que los dos caminos son equivalentes, se ilustrará el segundo, siendo análogo el desarrollo con el otro camino.

Entonces, si $e \in E$ y $d \in D$ están dados,

$$E_{z|e,\theta}\, u(e,z,d,\theta) = \int_{\mathcal{Z}} u(e,z,d,\theta)\, p(z|e,\theta)\, dz$$

define la utilidad condicional esperada de (e,d) para un $\theta \in \Theta$ dado, por lo que

$$E_\theta E_{z|e,\theta}\, u(e,z,d,\theta) = \int_\Theta \left\{ \int_{\mathcal{Z}} u(e,z,d,\theta)\, p(z|e,\theta)\, dz \right\} p(\theta|e)\, d\theta$$

será la utilidad esperada de (e,d) ; entonces el experimento óptimo que lleva a la decisión óptima estará dado por

$$\max_{e \in E} \left\{ \max_{d \in D} \left\{ \int_\Theta \left(\int_{\mathcal{Z}} u(e,z,d,\theta)\, p(z|e,\theta)\, dz \right) p(\theta|e)\, d\theta \right\} \right\}$$

Como puede notarse, el análisis en forma normal requiere de más trabajo pues tiene que evaluar $u(e,z,d,\theta)$ para cada z que es posible que ocurra, en lugar de trabajar con el z que se sabe ya ocurrió como es el caso del análisis en forma extensiva. Puede demostrarse que estos dos tipos de análisis son equivalentes (Raiffa & Schlaifer, 1961; Bernardo, 1977) por lo que el uso de uno u otro dependerá del problema específico que se trate.

1.5 EL TEOREMA DE BAYES

Como se vió en la sección 1.3, la decisión óptima puede ser determinada - con sólo el conocimiento inicial que se tenga del fenómeno, es decir, sin necesidad de tomar información adicional. En otros casos (sección 1.4) - esta información adicional es necesaria para tomar una decisión y un experimento debe ser realizado para recolectar dicha información. Como se ha visto, la forma de expresar el conocimiento que se tiene sobre los ele - mentos de Θ es mediante una distribución de probabilidades. Por otro lado, el conocimiento que sobre un $\theta \in \Theta$ se tenga puede darse a dos niveles distintos: el primero de ellos, es el que se tiene en un momento dado y que puede expresarse sin necesidad de nueva información y que queda representado por $p(\theta)$. El segundo es el que se adquiere en base a los resultados de un nuevo experimento y que se representa por $p(\theta|z)$[10]. Encontrar la forma analítica explícita de $p(\theta|z)$ para una situación específica, es uno de los problemas centrales de la inferencia estadística. Una forma de encontrar esta distribución es mediante el tercer axioma de probabilidad, siempre y cuando se conozca la distribución conjunta de Θ y Z definida sobre $\Theta \times \eta$. En efecto, si $p(\theta,z)$ se conoce, entonces $p(z) = \int_{\Theta} p(\theta,z)\, d\theta$ y usando el tercer axioma

$$p(\theta,z) = p(z)\, p(\theta|z)$$

10.- Nuevamente, con el fín de ahorrar notación se escribirá $p(\theta|z)$ en lugar de $p(\theta|\theta,z)$ que es lo correcto.

y si $p(z) \neq 0$ entonces

$$p(\theta|z) = p(\theta,z) \Big/ p(z)$$

Usando nuevamente el tercer axioma $p(\theta,z) = p(\theta)\, p(z|\theta)$,
de donde $p(\theta|z) = p(\theta)\, p(z/\theta) \,/\, p(z)$
esto significa que si $p(\theta,z)$ no se conoce, basta especificar $p(\theta)$ y $p(z|\theta)$ para conocer $p(\theta|z)$.

La necesidad de conocer $p(\theta|z)$ es natural, pues esta distribución representa lo que se ha aprendido de θ a partir de los resultados z. Dicho de otra forma, $p(\theta)$ representa el conocimiento inicial que se tiene sobre θ en un momento dado y la transformación de este conocimiento mediante los resultados de un experimento queda reflejado por $p(\theta|z)$. Entonces la forma natural y coherente de incorporar esta información es usando la expresión

$$p(\theta|z) = p(\theta)\, p(z|\theta) / p(z)$$

que es equivalente a

$$p(\theta|z) = p(\theta)\, p(z|\theta) \Big/ \int_{\Theta} p(\theta)\, p(z|\theta)\, d\theta$$

A esta expresión se le conoce con el nombre de Teorema de Bayes debido a que fué Bayes quien la introdujo en 1763. Algunas observaciones de este teorema se hacen a continuación.

i) Puede notarse que la expresión $\int_{(\Theta)} p(\theta)\, p(z|\theta)\, d\theta$, no es más que una constante de normalización, esto es, lo único que asegura es que $p(\theta|z) = 1$. Por esta razón, el teorema de Bayes frecuentemente es escrito como:

$$p(\theta|z) \propto p(\theta)\, p(z|\theta)$$

ii) La distribución representada por $p(\theta|z)$ es llamada distribución posterior o final, y expresa el conocimiento que se adquiere sobre θ a partir de los datos z.

iii) Como se mencionó con anterioridad, $p(\theta)$ representa el conocimiento inicial que se tiene sobre cierto fenómeno. Debido a esto, se le llama distribución apriori o inicial. Esta distribución deberá ser fijada por la persona que tiene el problema de tomar una decisión y por ninguna otra. La distribución $p(\theta)$ representa lo que dicha persona cree sobre θ y que no necesariamente coincidirá con lo que otra persona puede pensar sobre el fenómeno. La elección de $p(\theta)$ no es sencillo en la mayoría de los casos, sin embargo existen sugerencias en la literatura para encontrarla (Raiffa & Schlaifer, 1961; De Finetti, 1962; Savage, 1971; Bernardo, 1976). Una de las mayores limitaciones del enfoque bayesiano, radica en el carácter personalista de la toma de decisiones debido a la asignación de la distribución inicial, pues no se plantea una axiomática para resolver problemas de decisión en los que estén involucra

dos dos o más decisores.

La importancia de la distribución apriori y los problemas que se presentan para asignarla, será tratada más extensamente en el siguiente capítulo.

iv) La función $p(z|\theta)$ es llamada la función de verosimilitud, pues considerada como función de Θ, mide que tan verosímiles son los posibles valores de θ para un conjunto de datos z dado. La importancia de esta función radica en que establece el modelo bajo el cual la información es generada, para cada $\theta \in (H)$. Esta función deberá ser determinada por el tomador de decisiones y junto con la distribución inicial describirán directamente la información que se tiene sobre los resultados experimentales.

Supóngase ahora que $z_1, z_2 \in \mathcal{Z}$ son tales que

$$p(z_1|\theta) \propto p(z_2|\theta)$$

por el teorema de Bayes se sigue que

$$p(\theta|z_1) \propto p(\theta)\, p(z_1|\theta) \propto p(\theta)\, p(z_2|\theta) \propto p(\theta|z_2)$$

de donde puede concluirse que las inferencias basadas en z_1 deben ser las mismas que las basadas en z_2; resultado al cual se le llama principio de verosimilitud. Es obvio que una violación al principio de verosimilitud repercute en una violación a los principios de coherencia, pues significará que la

medida de probabilidad construída a partir de ellos no cumple los axiomas de Kolmogoroff-Renyi.

Para finalizar esta sección nótese que después de haber sido obtenido un conjunto de datos z_1, y el conocimiento inicial que sobre θ se tenía modificarse mediante el teorema de Bayes, un nuevo conjunto de datos z_2 independiente de z_1 puede ser tomado para modificar nuevamente el conocimiento actual que sobre θ se tenga. Esto es, aplicaciones sucesivas del teorema de Bayes pueden ser hechas, para conocer más sobre θ. Esto se realiza de la siguiente forma

Sea z_1 y supónganse $p(\theta)$ y $p(\cdot|\theta)$ dadas. Por el teorema de Bayes

$$p(\theta|z_1) \propto p(\theta)\, p(z_1|\theta) \quad \forall \theta \in \Theta$$

Si $z_2 \in \mathcal{Z}_1$ es independiente de z_1, nuevamente aplicando el teorema de Bayes se tiene

$$p(\theta|z_2, z_1) \propto p(\theta)\, p(z_1, z_2|\theta) \quad \forall \theta \in \Theta$$

$$\Rightarrow \quad p(\theta|z_2, z_1) \propto p(\theta)\, p(z_1|\theta)\, p(z_2|\theta)$$

$$\Rightarrow \quad p(\theta|z_2, z_1) \propto p(\theta|z_1)\, p(z_2|\theta) \quad \forall \theta \in \Theta$$

en donde $p(\theta|z_1)$, la distribución final dada la información z_1, juega ahora el papel de distribución inicial para z_2.

En general se tiene que

$$p(\theta|z_n, z_{n-1}, \ldots, z_1) \propto p(\theta|z_{n-1}, \ldots, z_1)\, p(z_n|\theta) \quad \forall \theta \in \Theta$$

partiendo de

$$p(\theta|z_1) \propto p(\theta)\, p(z_1|\theta) \quad \forall \theta \in \Theta$$

Este mecanismo de incorporar la información en el momento que se tenga, llamado proceso de aprendizaje, es una de las grandes ventajas de la metodología bayesiana como se verá en la siguiente sección.

I.6 PANORAMA GENERAL DE LOS METODOS BAYESIANOS

Como se mencionó en la sección I.1, la metodología bayesiana no es una parte más de la estadística ortodoxa, sino una alternativa global de ella. Esto significa que cualquier problema que la estadística ortodoxa resuelve, también lo resuelve la metodología bayesiana. No se contempla dentro de los objetivos de este trabajo el plantear problemas y su solución bayesiana, así como tampoco dar una explicación de ellos. Las obras de Lindley (1965), Winkler (1972), Raiffa & Schlaifer (1961), Box & Tiao (1973), De Groot (1970) y Bernardo (1977) pueden servir al lector interesado como textos introductorios a la metodología bayesiana. Referencias sobre algunos tópicos que se tratan con la metodología bayesiana son:

Problemas en análisis multivariado, han sido trabajados por Geisser (1965), Geisser y Cornfield (1963), Dempster (1963). Dickey (1973) ha trabajado en diseño de experimentos, mientras que Smith (1973a, 1973b) ha desarrollado los modelos lineales, así como aportaciones sobre el análisis de sensitividad (Smith, 1977). Sobre modelos de regresión, los artículos de Lindley (1968, 1969) son fundamentales. Problemas relacionados con estas áreas, como son mínimos cuadrados, ajuste de polinomios, predicción y observaciones discordantes han sido estudiados por Hill (1969), Guttman (1967), Aitchison y Dunsmore (1975) y De Finetti (1961) respectivamente. Dunsmore (1966, 1968 y 1969) ha tratado los problemas de clasificación y calibración. La importancia del muestreo y una presentación bayesiana,

está en Goel y Rambe (1970) y en la misma línea se encuentran los trabajos de Ericson (1965, 1969). En cuanto a aplicaciones a la medicina, sobre todo en el problema de diagnosis. Ascombe (1963), Armitage (1963) y Dawid (1976) han aportado resultados originales. Hartigan (1969) plantea el problema - no paramétrico y Wetherill (1961) el análisis secuencial.

Sobre la relación entre la estadística ortodoxa y la bayesiana, y contra- ejemplos en donde la primera es incoherente, están los artículos de Bar - tholomew (1965) y Stein (1956- 1959) respectivamente. Los conceptos de información y utilidad se discuten en Good (1966, 1971 y 1972). Zellner (1971) ha hecho aportaciones importantes en econometría. El problema de determinar la distribución inicial, ha sido tratado por Winkler (1967, 1968 y 1969).

Un tratamiento sobre estimación, pruebas de hipótesis y regiones creíbles, se encuentra en De Groot (1970) y en Winkler (1972). Una exposición amplia sobre técnicas multivariantes está en Press (1972), mientras que Box y - Tiao (1973) y Zellner (1971) tratan ampliamente el modelo de regresión.

Para finalizar esta sección, es importante mencionar una ventaja que la - metodología bayesiana tiene sobre los métodos clásicos. Esta ventaja, es la claridad con que pueden diferenciarse los problemas de inferencia y decisión. De acuerdo a lo visto a lo largo de este capítulo, el problema de decisión puede ser resuelto con la distribución final o con la inicial si no se posee nueva información, y con una función de utilidad que marque

las preferencias del tomador de decisiones en sus consecuencias. Sin em - bargo, el encontrar la distribución final $p(\theta|z)$, no es un problema de decisión, sino de inferencia, así como también lo son, el calcular momios, factores, etc.[11] La diferencia básica entre estos dos problemas, es entonces la función de utilidad. Si se cuenta con una función de utilidad, puede tomarse una decisión, teniéndose en estos casos un problema de decisión. En caso de que no se tenga una función de utilidad, no puede tomarse una decisión y lo más que puede hacerse es inferencia: encontrar $p(\theta|z)$, calcular momios iniciales o finales, etc. Esta diferencia clara no queda reflejada en los métodos clásicos.

11.- Véase Winkler (1972)

I.7 BIBLIOGRAFIA DEL CAPITULO

Aitchison y Dunsmore (1975)

Anscombe (1963)

Armitage (1963)

Bartholomew (1965)

Bayes (1763)

Bernardo (1976), ; (1977) y

Box y Tiao (1973)

Dawid (1976)

De Finetti (1937) ; (1961) ; (1962)

De Groot (1970)

Dempster (1963)

Dickey (1973)

Dunsmore (1966, 1968 y 1969)

Ericson (1965 y 1969)

Fisher, J (1978)

Geisser (1965)

Geisser y Cornfield (1963)

Godambe (1970)

Good (1966, 1971 y 1972)

Gutlman (1967)

Hartigan (1969)

Hill (1969)

Laplace (1820)

Lindley (1965, 1968 y 1969) ; (1977)

Pratt, Raiffa y Schlaifer (1964)

Press (1972)

Raiffa y Schlaifer (1961)

Ramsey (1926)

Renyi (1961)

Savage, L.J. (1954) ; (1961) ; (1971)

Smith (1973a, 1973b y 1977)

Stein (1956 y 1959)

Wetherill (1961)

Winkler (1967, 1968 y 1969) ; (1972)

Zellner (1971)

CAPITULO II

El papel que la distribución inicial juega en los métodos bayesianos es de vital importancia y por eso la selección de dicha distribución, en general, deber ser realizada cuidadosamente. En este capítulo, además de mostrar la importancia que la distribución inicial tiene, se verán recomendaciones para asignarla. Se verá también como la asignación de la función de verosimilitud depende del tomador de decisiones, al igual que la distribución inicial. Se plantearán situaciones en las que no es posible asignar una distribución inicial informativa, y que el único camino coherente es dar una distribución inicial de referencia. Por otro lado, se verá también como en algunas ocasiones es conveniente dar una distribución de referencia.

II.1 LA DISTRIBUCION INICIAL Y LA FUNCION DE VEROSIMILITUD

Como se vió en el capítulo anterior habiendo definido una función de utilidad, la mejor decisión puede ser determinada usando la distribución inicial sobre Θ o la distribución final sobre H si es que un experimento ha sido llevado a cabo. De igual forma, cierto tipo de inferencias dependerán de $p(\theta)$ o de $p(\theta|z)$ de acuerdo a si se realizó o no un experimento. También se mostró como encontrar $p(\theta|z)$ usando la distribución inicial mediante el teorema de Bayes,

$$p(\theta|z) \propto p(\theta)\, p(z|\theta)$$

en donde $p(z|\theta)$ es la función de verosimilitud.

En ambos casos, el papel que juega la distribución inicial es importante, por lo que una inadecuada selección de esta distribución, puede repercutir en tomar una mala decisión o en hacer inferencias incorrectas.

Debe quedar claro que esto no significa que para un problema específico deba existir una sola distribución inicial, pues recuérdese que la elección de ella depende de los juicios del tomador de decisiones y que estos no necesariamente debe coincidir con los de cualquier otra persona. El carácter subjetivo del enfoque bayesiano permite desarrollar una problema en forma personal, esto es, dependiendo de la persona que va a tomar una decisión o a hacer inferencias, la solución variará. Este subjetivis

mo queda expresado, en parte, en la distribución inicial y dependiendo de ellas, se obtendrá cierta respuesta al problema, ya sea de toma de decisiones o de inferencia. De aquí que una parte fundamental de la metodología bayesiana sea la elección de la distribución inicial.

Además de la distribución inicial, dos funciones más son relevantes en la metodología bayesiana y que expresan también la opinión del tomador de decisiones. La primera es la función de utilidad, que como se recordará mide las consecuencias de tomar una decisión específica. La necesidad de esta función en un problema de toma de decisiones es obvia, pues sin ella no es posible tomar una decisión en forma coherente. Sin embargo, en un problema de inferencia, una función de utilidad no es necesaria como ya se comentó en el capítulo anterior.

La función de verosimilitud, es la otra función importante. Es necesaria si es que un experimento es realizado y como consecuencia de eso, unos resultados son observados. La función de verosimilitud es el modelo que describe el proceso generador de datos, y conjuntamente con la distribución inicial reflejan toda la información que sobre los resultados experimentales se tenga, de aquí su importancia.

La selección de la función de verosimilitud es un problema que también atañe directamente al tomador de decisiones y aunque en algunos casos

pueda ser obvia su elección, debe tenerse cuidado al hacerlo.

Una forma de saber si la distribución inicial y la verosimilitud han sido seleccionadas en forma coherente, es mediante la simulación de datos y obtener, por el teorema de Bayes, la distribución final e ir checando si los resultados que se obtengan de esta forma, resultan creíbles para el decisor. En caso de que así sea, entonces la elección de las funciones mencionadas pueden considerarse coherentes, sin que esto implique que son correctas. En caso contrario, habrá que revisar la distribución inicial o la verosimilitud (o posiblemente ambas) y repetir el proceso, hasta que se tenga una selección coherente.

Existen reportados en la literatura, varias sugerencias para encontrar una distribución inicial que sea coherente con los juicios del decisor. Esto se verá a continuación y en forma breve, en la sección siguiente.

II.2 LA ASIGNACION DE LA DISTRIBUCION INICIAL

El primer paso para asignar una distribución de probabilidades, es definir las variables aleatorias que son de interés para el investigador, ya que será a ellas a quienes se les asociará la distribución, que expresará el conocimiento apriori o inicial que se tenga sobre dichas variables. La forma de expresar este conocimiento estará dada por el experimento auxiliar mencionado en el cuarto principio de coherencia. Como se mencionó en ese momento, no es necesario que el decisor construya el experimento sino que basta que esté consciente que pueda hacerlo. Debido a eso pueden existir, y de hecho existen, varios procedimientos para asignar probabilidades.

Si se cuenta con información anterior al problema, que está dada en términos frecuentistas, es natural esperarse que las probabilidades iniciales se parezcan a dichas fecuencias, pero no es absolutamente necesario, pues como menciona Bernardo (1977) existen situaciones en los casos de diagnosis médica, en los que estas frecuencias deben combinarse con las apreciaciones subjetivas del médico.

Winkler (1967) sugiere una forma de asignar distribuciones de probabilidades mediante el uso de ciertos cuantiles , o calculando probabilidades de algunos intervalos, mediante las cuales se puede construir la distribución. Otros métodos que menciona Winkler (1972) es el uso de momios y lo

terías.

Un método muy utilizado es usando funciones de pago ("scoring rules"), en el que se ofrece un premio si la probabilidad ha sido bien designada y un castigo si pasa lo contrario. Trabajos importantes en este sentido son lo de De Finetti (1962) y Savage (1971).

Lindley (1977a) sugiere el uso de opciones equivalentes en términos de primas de pólizas de seguros, o bien el de calcular varias probabilidades que estén relacionadas entre sí por un ley para verificar la coherencia en la asignación (Lidley, 1977b).

De igual forma, la asignación de la función de verosimilitud puede hacerse mediante el uso de loterías o momios condicionales (Winkler, 1972). En la mayoría de los casos es fácil asociar un modelo probabilístico establecido, sin embargo hay que tener cuidado con las hipótesis y condiciones del modelo y el problema (Winkler, 1972). Algunas veces un modelo conocido no es directamente aplicable y debe modificarse (Lindley, 1965).

En general, el problema de asignar tanto la distribución inicial como la verosimilitud, no es un problema sencillo y debido al papel fundamental que juegan en la metodología bayesiana, debe tenerse mucho cuidado y dedicarle tiempo suficiente para hacerlo, pues de una buena selección dependerá un buen análisis.

II.3 PROBLEMAS EN LA ASIGNACION DE LA DISTRIBUCION INICIAL

A pesar de que existen varios métodos que sirven como guía para asignar probabilidades y en particular probabilidades iniciales, en diversas situaciones no es posible asignar una distribución inicial que exprese las opiniones del decisor ya sea por problemas técnicos o porque podría haber contradicción con los principios de coherencia. En otras ocasiones es conveniente introducir una distribución inicial que no involucre un conocimiento prejuiciado del decisor. En esta sección se darán algunos ejemplos en las que estas situaciones se presentan.

i) Una situación muy común en donde la asignación de una distribución inicial resulta prácticamente imposible es cuando el número de parámetros involucrados en el problema es grande. Como ejemplo, imagínese un problema de regresión en donde la dimensión de la matriz de diseño es $n \times p$; ésto querría decir que habría que asignar una distribución inicial sobre $(n(n+1) + 2np)/2$ parámetros, y si n ó p ó ambos son grandes el problema es practicamente irresoluble.

En un caso no paramétrico, el problema de asignar una distribución inicial es todavía más complicado, pues habría que asignar una distribución de probabilidades a un espacio funcional: el de las distribuciones de probabilidad.

ii) Situaciones en las que un grupo de personas deban de tomar una decisión conjunta no están contempladas axiomáticamente en la metodología bayesiana, esto es, no existe una teoría formal que resuelva situaciones con conflicto. Es obvio que dado un grupo de personas, no todas deben coincidir en sus juicios, no todas deben de pensar lo mismo sobre una situación específica; por lo que la asignación de una distribución inicial no parece sencillo.

iii) En los casos de asesoría estadística, el asesor no es (generalmente) el tomador de decisiones y pocas veces la persona que fungirá como decisor ha determinado su distribución inicial, sino que generalmente tiene una idea del comportamiento del fenómeno debido a la observación de los datos que desea analizar y por lo tanto la elección de una distribución inicial en ese momento no es consistente con la metodología bayesiana. Lo anterior se debe a que hipótesis sugeridas por ciertos datos no deben ser verificadas con los mismos datos.

iv) El tiempo es un factor muy importante en el análisis estadístico. Como se ha mencionado, la elección de una distribución inicial no es sencilla y por lo tanto definir una que sea coherente con los conocimientos iniciales que el decisor tenga,

llevará algún tiempo. Sin embargo, en muchos casos se necesita tomar una decisión importante en un lapso pequeño y no parece conveniente modelar una distribución inicial incoherente o que distorsione la realidad. Otras veces la falta de recursos computacionales o de otro tipo pueden ser el factor de definir una distribución que exprese en forma incoherente las opciones del decisor.

v) En el estudio de ciertas propiedades teóricas de algunas distribuciones de probabilidad, el introducir una distribución inicial que refleje la opinión del decisor[12] puede dar resultados muy particulares y no tendrían por tanto, una aplicación general que es lo que buscaría al estudiar el comportamiento puramente matemático de una distribución.

vi) En general, al hacer un estudio es conveniente dar un punto de referencia contra el cual comparar las inferencias que pueden ser obtenidas a partir de los datos sin un conocimiento prejuiciado de los datos.

Estos ejemplos, que no cubren todos los casos, ilustran la necesidad en

12.- En este caso no se tomarán decisiones, sino sólo se harían inferencias, sin embargo se seguirá llamando a la persona interesada en el problema, decisor.

unas ocasiones y conveniencia en otras, de introducir en el análisis, una distribución inicial que no represente el conocimiento del decisor o que influya en los datos de forma tal que no los apoye ni tampoco los desmienta, es decir que no los sesgue (de aquí el nombre de "distribuciones que 'dejen hablar' a los datos"); o bien una distribución que pueda ser tomada por cada persona como una referencia y que un momento dado pueda comparar sus resultados contra ella.

Una distribución de referencia, puede servir también para tratar de resolver problemas de conflicto, como el mencionado en el punto (ii). Muchas veces se han usado distribuciones de referencia para comparar los enfoques clásico y bayesiano de la estadística.

En resumen puede decirse lo siguiente, siempre es preferible definir una distribución inicial que exprese el conocimiento del decisor, en los casos que esto no sea posible o bien, cuando se deseen hacer comparaciones, lo indicado es trabajar con una distribución de referencia.

En el capítulo siguiente se verán varios métodos para encontrar distribuciones de referencia, asimismo se comentarán las ventajas y desventajas de cada método.

II.4 BIBLIOGRAFIA DEL CAPITULO

Bernardo (1977)

De Finetti (1962)

Lindley (1965, 1977a, 1977b)

Savage (1971)

Winkler (1967)

CAPITULO III

En este capítulo se presentarán diferentes métodos para encontrar distribuciones de referencia. Serán presentados en orden cronológico, para mostrar el desarrollo que se ha tenido. En algunos métodos se darán demostraciones y en otros se omitirán, sobre todo en los casos en que éstas sean muy complejas. Ventajas y desventajas de cada método serán mencionadas, dejándose para el último capítulo, una comparación.

III.1 BAYES, LAPLACE Y JEFFREYS

"Las dificultades y controversias del teorema de probabilidad inversa (teorema de Bayes), aparecen en el caso vital en que las probabilidades iniciales no son conocidas y tiene que hacerse una suposición sobre ellas".

El párrafo anterior, debido a Perks (1947) expone con claridad el problema de las distribuciones de referencia en la estadística bayesiana. Como se recordará, Bayes fué el primero en considerar a la probabilidad desde un punto de vista subjetivo, como grados de creencia; hecho que se ve reflejado en su famoso teorema. Sin embargo, existía el problema de qué hacer en los casos en que se estaba en estado de "completa ignorancia" acerca de los valores de los parámetros. Para resolver esta dificultad, Bayes propone la siguiente regla:[13]

"Si se desconoce todo sobre θ, no existe razón alguna para pensar que algunos valores de θ son más probables que otros" (Bayes, 1763).

Aparentemente Laplace adopta la misma regla, llamada "regla de indiferencia", para subsanar la misma dificultad. Desgraciadamente, no se pudo conseguir la obra de Laplace que permitiría asegurar este hecho, sin embargo, varios autores, entre ellos Perks, le adjudican a Laplace esta forma de actuar (Perks, 1947).

13.- De la cual no estaba muy convencido.

El postulado de Bayes, reconocido posteriormente como la regla de indiferencia de Bayes-Laplace, trata de expresar en términos formales la "ignorancia completa". Según afirman muchos autores, la regla de Bayes-Laplace resiste todo tipo de críticas cuando los parámetros varían en todo $\mathbb{R}$. Esto es, la asignación de una distribución sobre $\mathbb{R}$, no lleva a contradicciones. Pero, si se transforma la variable en forma no lineal, la aplicación de la regla de Bayes-Laplace a la variable transformada lleva a resultados que difieren en mucho de los que se obtendrían al aplicar la regla a la variable original; por ejemplo en el caso de σ^2 en la normal, el usar σ o σ^2 como parámetro de interés, al asignar una distribución uniforme a cada uno de ellos, llevaría a resultados inconsistentes. Este tipo de problemas pueden ser evitados si se pudiese estar seguro de que sólo una variable es de particular interés, esto es, existe una métrica absoluta.[14]

Para evitar este tipo de problemas Jeffreys (1939)[15] propone las siguientes reglas

i) $p(\theta) \propto$ constante, si $\theta \in \mathbb{R}$

y ii) $p(\theta) \propto \frac{1}{\theta}$, si $\theta \in \mathbb{R}^+$

14.- En esta frase, métrica tiene el sentido de unidad de medición y no el significado matemático de distancia.

15.- Nuevamente el problema de no encontrar la primera edición del libro de Jeffreys, hace difícil el dar las razones de porqué propone esas dos reglas, una conjetura es el que trata de aplicar la regla de Bayes-Laplace, pero a variables transformadas. Esto lo menciona Perks (1947) aunque no es muy claro.

La primera no es más que la "regla de indiferencia" de Bayes-Laplace en los casos en que no había problemas. La segunda, es una regla que es invariante bajo transfomraciones del tipo $k = \theta^n$ y que resuelve el problema generado al usar σ en lugar de σ^2. Para ver que en efecto es invariante nótese que

$$dk = n\,\theta^{n-1} d\theta$$

$$\Rightarrow \frac{dk}{k} = n\,\frac{d\theta}{\theta}$$

y por tanto $$\frac{dk}{k} \propto \frac{d\theta}{\theta}$$

Sin embargo, estas dos reglas no cubren todos los casos, Jeffreys (1946) desarrolla una nueva regla, teniendo en mente la idea de invarianza, esto es, asegurar que proposiciones equivalentes tengan la misma probabilidad.

La idea de Jeffreys se presenta a continuación, manteniendo la notación original:

Sea P la distribución que depende de $\{\theta_1, \ldots, \theta_m\}$ y P' la distribución resultante de usar en lugar de $\{\theta_1, \ldots, \theta_m\}$ a $\{\theta_1 + \Delta\theta_1, \ldots, \theta_m + \Delta\theta_m\}$, Considérense las siguientes funciones

$$I_1 = \int (\sqrt{dP} - \sqrt{dP'})^2 \qquad e \qquad I_2 = \int \log_e \frac{dP'}{dP}\, d(P' - P)$$

definidas en el sentido de Stieltjes.

Considérese un partición de $\mathbb{R}$ y toménse incrementos δP y $\delta P'$ para la misma celda δx y haciendo $p_r = \delta P_r$ y $p_r' = \delta P_r'$, se tiene que

$$\sqrt{p_r'} \doteq \sqrt{p_r} + \frac{1}{2\sqrt{p_r}}(p_r' - p_r) - \frac{1}{8\sqrt{p_r^3}}(p_r' - p_r)^2 \ldots \quad (1)$$

desarrollando a $\sqrt{p_r'}$ alrededor de p_r.

Análogamente para $\log_e p_r'$

$$\log_e p_r' \doteq \log_e p_r + \frac{1}{p_r}(p_r' - p_r) - \frac{1}{2p_r}(p_r' - p_r)^2 \ldots \quad (2)$$

De (1):

$$\sqrt{p_r'} - \sqrt{p_r} \doteq \frac{1}{2\sqrt{p_r}}(p_r' - p_r) - \frac{1}{8\sqrt{p_r^3}}(p_r' - p_r)^2 \ldots$$

$$\Rightarrow (\sqrt{p_r'} - \sqrt{p_r})^2 \doteq \frac{1}{4p_r}(p_r' - p_r)^2 + \frac{1}{64p_r^3}(p_r' - p_r)^4 - \frac{1}{8p_r^2}(p_r' - p_r)^3 \ldots$$

Considerando únicamente los términos menores o iguales a dos:

$$(\sqrt{p_r'} - \sqrt{p_r})^2 \doteq \frac{1}{4p_r}(p_r' - p_r)^2$$

$$\Rightarrow \lim_{\|\delta x\| \to 0} \sum_r (\sqrt{p_r'} - \sqrt{p_r})^2 \doteq \lim_{\|\delta x\| \to 0} \sum_r \frac{1}{4p_r}(p_r' - p_r)^2, \text{ pero } \lim_{\|\delta x\| \to 0} \sum_r (\sqrt{p_r'} - \sqrt{p_r})^2 = I_1$$

$$\therefore I_1 = \lim_{\|\delta x\| \to 0} \sum_r \frac{1}{4p_r}(p_r' - p_r)^2$$

En (2):

$$\log_e p_r' - \log_e p_r \doteq \frac{1}{p_r}(p_r' - p_r) - \frac{1}{2p_r^2}(p_r' - p_r)^2 \ldots$$

$$\Rightarrow \log_e \frac{p_r'}{p_r}(p_r' - p_r) \doteq \frac{1}{p_r}(p_r' - p_r)^2 - \frac{1}{2p_r^2}(p_r' - p_r)^3 \ldots$$

omitiendo términos de orden mayor a dos:

$$\log_e \frac{p_r'}{p_r}(p_r' - p_r) \doteq \frac{1}{p_r}(p_r' - p_r)^2$$

$$\therefore I_2 \doteq \lim_{\|\delta x\| \to 0} \sum_r \frac{1}{p_r}(p_r' - p_r)^2$$

Desarrollando ahora a p_r' alrededor de $\theta = (\theta_1, \theta_2, \ldots \theta_m)$, se tiene que

$$p_r' \doteq p_r + \sum_i \frac{\partial p_r'}{\partial \theta_i}\Delta\theta_i + \sum_{i,j} \frac{\partial^2 p_r'}{\partial \theta_i \partial \theta_j}\Delta^2\theta_i \Delta^2\theta_j + \ldots$$

$$\Rightarrow p_r' - p_r \doteq \sum_i \frac{\partial p_r'}{\partial \theta_i}\Delta\theta_i + \sum_{i,j} \frac{\partial^2 p_r'}{\partial \theta_i \partial \theta_j}\Delta^2\theta_i \Delta^2\theta_j + \ldots$$

$$\Rightarrow (p_r' - p_r)^2 \doteq \left(\sum_i \frac{\partial p_r'}{\partial \theta_i}\Delta\theta_i\right)^2 + \left(\sum_{i,j} \frac{\partial^2 p_r'}{\partial \theta_i \partial \theta_j}\Delta^2\theta_i \Delta^2\theta_j\right)^2 + 2\sum_i \frac{\partial p_r'}{\partial \theta_i}\Delta\theta_i \sum_{ij} \frac{\partial p_r}{\partial \theta_i \partial \theta_j}\Delta^2\theta_i \Delta^2\theta_j \ldots$$

nuevamente, considerando una aproximación a orden 2

$$(p_r' - p_r)^2 \doteq \sum_{i,j} \frac{\partial p_r'}{\partial \theta_i}\frac{\partial p_r'}{\partial \theta_j}\Delta\theta_i \Delta\theta_j$$

$$\therefore I_2 \doteq \lim_{\|\delta x\| \to 0} \sum_{r,i,j} \frac{1}{p_r}\frac{\partial p_r'}{\partial \theta_i}\frac{\partial p_r'}{\partial \theta_j}\Delta\theta_i \Delta\theta_j$$

y análogamente

$$I_1 \doteq \frac{1}{4}\lim_{\|\delta x\| \to 0} \sum_{r,i,j} \frac{1}{p_r}\frac{\partial p_r'}{\partial \theta_i}\frac{\partial p_r'}{\partial \theta_j}\Delta\theta_i \Delta\theta_j$$

Si $g_{ij} = \lim_{\|\delta x\| \to 0} \sum_{r,i,j} \frac{1}{p_r}\frac{\partial p_r'}{\partial \theta_i}\frac{\partial p_r'}{\partial \theta_j}$, se tiene: $I_1 \doteq \frac{1}{4} I_2 \doteq g_{ij}\Delta\theta_i \Delta\theta_j$

Si se consideran unos nuevos parámetros $\{\theta_1', \theta_2', \ldots, \theta_m'\}$ se sigue que

$$I_2 \doteq g_{ij}'\Delta\theta_i' \Delta\theta_j', \text{ con } g_{ij}' = g_{ij}\frac{\partial\theta_i}{\partial\theta_k'}\frac{\partial\theta_k}{\partial\theta_s'}$$

$$\Rightarrow \|g_{ij}'\| = \|g_{ij}\| \left\|\frac{\partial\theta_i}{\partial\theta_k'}\right\|^2$$

y como $d\theta_1 d\theta_2 \dots d\theta_m = \left\| \frac{\partial \theta_i}{\partial \theta'_i} \right\| d\theta'_1 \dots d\theta'_m$

$$\Rightarrow \|g_{ij}\|^{1/2} d\theta_1 \dots d\theta_m = \|g'_{ij}\|^{1/2} d\theta'_1 \dots d\theta'_m$$

esto es $\|g_{ij}\|^{1/2} d\theta_1 \dots d\theta_m$ es invariante ante transformaciones de los parámetros, por lo que si $p(\theta_1, \dots, \theta_m) \propto \|g_{ij}\|^{1/2}$ querrá decir que $p(\theta_1, \dots, \theta_m)$ será invariante ante transformaciones de los parámetros.

Se verá un ejemplo, tomando del artículo original de Jeffreys (1947)

Considérese a la distribución normal

$$p_r \doteq \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_r-\alpha)^2}{2\sigma^2}\right\} dx_r$$

entonces $\sqrt{p_r} \doteq \sqrt{\frac{1}{\sqrt{2\pi}}} \sqrt{\frac{1}{\sigma}} \exp\left\{-\frac{(x_r-\alpha)}{4\sigma^2}\right\} dx_r$

$$\Rightarrow I_1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left[\frac{1}{\sqrt{\sigma'}} \exp\left\{-\frac{(x-\alpha')^2}{4\sigma'^2}\right\} - \frac{1}{\sqrt{\sigma}} \exp\left\{-\frac{(x-\alpha)^2}{4\sigma^2}\right\}\right]^2 dx$$

$$\Rightarrow I_1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \left[\frac{1}{\sigma'} \exp\left\{-\frac{(x-\alpha')^2}{2\sigma'^2}\right\} + \frac{1}{\sigma} \exp\left\{-\frac{(x-\alpha)^2}{2\sigma^2}\right\} - \frac{2}{\sqrt{\sigma'\sigma}} \exp\left\{-\frac{(x-\alpha')^2}{4\sigma'^2} - \frac{(x-\alpha)^2}{4\sigma^2}\right\}\right]$$

$$\Rightarrow I_1 = 2\left[1 - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\sigma'\sigma}} \exp\left\{-\frac{(x-\alpha')^2}{4\sigma'^2} - \frac{(x-\alpha)^2}{4\sigma^2}\right\} dx\right]$$

completando cuadrados para tener una densidad normal, se llega a;

$$I_1 = 2\left[1 - \frac{\sqrt{2}}{\sqrt{\frac{\sigma}{\sigma'} + \frac{\sigma'}{\sigma}}} \exp\left\{-\frac{(\alpha-\alpha')^2}{4(\sigma^2+\sigma'^2)}\right\}\right]$$

Haciendo $\sigma = \sigma_0 e^{-\xi}$ y $\sigma' = \sigma_0 e^{\xi}$ con $\xi \in \mathbb{R}$, lo que lleva a suponer que $\sigma = \sigma' e^{-2\xi}$, que siempre es válido por la suprayectividad de la función e, se tiene que

$$I_1 = 2\left[1 - \frac{\sqrt{2}}{\sqrt{e^{-2\xi} + e^{2\xi}}} \exp\left\{-\frac{(\alpha-\alpha')^2}{4(\sigma_0^2 e^{-2\xi} + \sigma_0^2 e^{2\xi})}\right\}\right]$$

$$\Rightarrow I_1 = 2\left[1 - \operatorname{sech}^{1/2} 2\xi \exp\left\{-\frac{(\alpha-\alpha')^2}{8\sigma_0^2 \cosh 2\xi}\right\}\right]$$

Si se desarrolla por series de Taylor alrededor del cero, tanto a la función exponencial como a sech, y se omiten términos mayores a dos, se llega a

$$I_1 \doteq 2\rho^2 + \frac{(\alpha-\alpha')^2}{4\sigma_0^2}$$

y como

$$I_1 \doteq \frac{1}{4} I_2 \quad \Rightarrow \quad I_2 = 8\rho^2 + \frac{(\alpha-\alpha')^2}{\sigma_0^2}$$

$$\therefore \quad I_2 \doteq 2\left(\frac{d\sigma}{\sigma}\right)^2 + \left(\frac{d\alpha}{\sigma}\right)^2 \quad \text{e} \quad I_1 \doteq \frac{1}{2}\left(\frac{d\sigma}{\sigma}\right)^2 + \frac{1}{4}\left(\frac{d\alpha}{\sigma}\right)^2$$

Ahora bien, si α es fijo, se tiene que $g_{\sigma\sigma} \propto \sigma^{-2}$ y por tanto $p(\sigma) \propto \sigma^{-1}$. Si σ es fijo $\Rightarrow p(\alpha) \propto$ constante; y finalmente, si σ y α son desconocidos, el determinante es proporcional σ^{-4} y por tanto $p(\alpha,\sigma) \propto \sigma^{-2}$.

Los comentarios que pueden hacerse a estos resultados y en general a este procedimiento son varios, pero antes de llevarlos a cabo, se dará otra forma de encontrar la distribución de referencia, que es equivalente a ésta y que fué propuesta por el mismo Jeffreys (1961). Un desarrollo similar, puede verse en Box & Tiao (1973).

Dado que $dP \propto L$, donde L es la función de verosimilitud, puede reescribirse I_2 como:

$$I_2 = \lim_{n\to\infty} \frac{1}{n} \sum_i \log_e \frac{L(x;\alpha_i)}{L(x;\alpha_i')} \left\{ L(x;\alpha_i) - L(x;\alpha_i') \right\}$$

y siguiendo un desarrollo similar al anterior, se llega a que

$$g_{ij} \doteq \lim_{n\to\infty} -\frac{1}{n} \frac{\partial^2 \log_e L(x;\underline{\alpha})}{\partial\alpha_i \, \partial\alpha_j}$$

lo que $\Rightarrow g_{ij} \doteq -E\left\{\frac{\partial^2 \log_e L(x,\underline{\theta})}{\partial\theta_i\,\partial\theta_j}\right\}$, que no es más que la matriz de información de Fisher (1922,1925).

De donde, si la distribución inicial es proporcional a $\|g_{ij}\|^{1/2}$ será invariante ante transformaciones no singulares de los parámetros, de acuerdo a lo deducido anteriormente.

Con el fín de comparar con los resultados obtenidos anteriormente, supóngase nuevamente una densidad normal, cuya función de verosimilitud para una muestra de tamaño n, está dada por

$$L(x;\theta) = (2\pi)^{-n/2}\ \sigma^{-n}\ \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_i (x_i-\mu)^2\right]\right\} \quad \text{con } \theta=(\mu,\sigma).$$

Entonces

$$\log_e L(x;\theta) = -\frac{n}{2}\log_e 2\pi - n\log_e\sigma - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$$

de donde

$$\frac{\partial \log_e L(x;\theta)}{\partial\mu} = \frac{1}{\sigma^2}\sum_i (x_i-\mu)\ ,\quad \frac{\partial \log_e L(x;\theta)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2$$

así

$$\frac{\partial^2 \log_e L(x;\theta)}{\partial\mu^2} = -\frac{1}{\sigma^2}\ ;\ \frac{\partial^2 \log_e L(x;\mu)}{\partial\mu\,\partial\sigma} = \frac{\partial^2 \log_e L(x;\theta)}{\partial\sigma\,\partial\mu} = -\frac{2}{\sigma^3}\sum_i (x_i-\mu)\ \text{y}\ \frac{\partial^2 \log_e L(x;\theta)}{\partial\sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_i (x_i-\mu)^2$$

por lo que

$$g_{11} = \frac{1}{\sigma^2} \qquad g_{12} = g_{21} = 0 \qquad \text{y} \qquad g_{22} = \frac{2n}{\sigma^2}$$

Si σ es conocida, la matriz de información es simplemente g_{11}, por lo que la distribución inicial para μ será $p(\mu) \propto$ constante. De la misma forma, si μ es constante, la matriz de información se reduce a g_{22} y por tanto $p(\sigma) \propto \sigma^{-1}$ será la distribución de referencia para σ.

Ahora bien, si tanto μ como σ son desconocidos se tiene que

$$\|g_{ij}\| \propto 2n\sigma^{-4} \quad \text{y entonces} \quad p(\mu,\sigma) \propto \sigma^{-2}.$$

Las distribuciones iniciales de referencia $p(\mu)$ y $p(\sigma)$ encontradas con este procedimiento, coinciden con las establecidas inicialmente por Jeffreys y mencionadas al principio de esta sección.

Sin embargo la distribución de referencia conjunta $p(\mu,\sigma)$ difiere de la obtenida con la regla inicial de Jeffreys[16] que daba $p(\mu,\sigma) \propto \sigma^{-1}$ en lugar de $p(\mu,\sigma) \propto \sigma^{-2}$ que es lo que se obtiene con esta nueva regla.

Por otro lado, si se supone apriori que μ y σ son independientes, puede obtenerse que $p(\mu,\sigma) \propto \sigma^{-1}$ de la siguiente forma $p(\mu,\sigma) = p(\mu)\,p(\sigma)$ y usando la nueva regla separadamente en cada uno de los parámetros, obteniendo $p(\mu,\sigma) \propto \sigma^{-1}$ que conincide con las reglas anteriores. Sin embargo, esto a lo que lleva es que la nueva regla no debe ser utilizada en el caso multivariado a menos que se suponga independencia y se utilice la regla para obtener las distribuciones marginales de referencia, obteniendo la distribución conjunta como el producto de las marginales, pero esto no es aplicar la regla de Jeffreys para el caso multivariado, puesto que no establece diferencias para cuando se tiene independencia y cuando no se tiene. Esto es, la nueva

16.- En este trabajo, se entenderá por regla de Jeffreys, la que dá como distribución de referencia con la matriz de información de Fisher. En esta sección, para diferenciarla de las dos reglas iniciales dadas por Jeffreys, se referirá a ella como la nueva regla.

regla lleva a que $p(\mu,\sigma) \propto \sigma^{-2}$, sean o no independientes.

Para que la nueva regla coincida con las anteriores, Jeffreys (1961) propone que se tome como distribución inicial a

$$p(\mu,\sigma,\alpha_1,\cdots,\alpha_n) \propto \sigma^{-1} \|g_{ik}\|^{1/2} \qquad (*)$$

donde $\|g_{ik}\|$ es encontrada variando únicamente a las α_i's y manteniendo fijas μ y σ. Las α_i's no son más que parámetros numéricos (como los índices en varias leyes de probabilidad del sistema de Pearson o el de la binomial negativa). Esta nueva regla es invariante ante transformaciones del tipo

$$\mu' = \mu + \sigma f(\alpha_i) \;, \quad \sigma' = \sigma g(\alpha_i) \quad y \quad \alpha_j' = \alpha_j'(\alpha_i)$$

Sin embargo esta regla (*); lo único que está estableciendo es que $p(\mu,\sigma) \propto \sigma^{-1}$, pues $\|g_{ik}\|^{1/2}$ quedará como constante, pues no depende de μ y σ.

Este mismo problema lo tratan Box & Tiao (1973) quienes muestran que al suprimir la suposición apriori de independencia entre μ y σ, cualquier transformación que se haga sobre μ, es tal que la información que se tiene sobre la tranformación es proporcional a σ^{-2}, lo que lleva al resultado $p(\mu,\sigma) \propto \sigma^{-2}$. Lo que Box & Tiao (1973) proponen, es suponer independencia siempre y utilizar la regla de Jeffreys para cada parámetro por separado y encontrar la distribución de referencia conjunta como el producto de las marginales. Pero esto no es muy sostenible, porque no siempre tiene que suponerse independencia apriori.

En resumen, puede decirse que la regla de Jeffreys funciona bien en el caso de un sólo parámetro y que cuando son varios, las modificaciones hechas son muy discutibles.

Otro problema que tiene la regla de Jeffreys, es que no es aplicable cuando la verosimilitud no es diferenciable con respecto a los parámetros, debido a la forma como se deduce. Sin embargo Jeffreys (1961) menciona, que el uso de la regla (*) puede llevar a resultados que frecuentemente son satisfactorios en los términos de las reglas anteriores. Esto es debido al hecho de que $\left|a_{ij}\right|^{\frac{1}{2}}$ en (*) no depende de μ y σ. En los casos en que los parámetros pueden tomar un número finito de valores, Jeffreys (1961) menciona que una extensión de la regla es posible en muchos casos, sin embargo no lo hace explícito.

Como último comentario, cabe mencionar que varios de los trabajos que desarrollan una nueva regla para asignar distribuciones de referencias, llevan a la conclusión de que la regla de Jeffreys dá la distribución de referencia apropiada, en el caso de que no existan parámetros de ruido (nuisance parameters), esto es, en el caso univariado. Por otro lado, otros trabajos están dedicados a justificarla por variados argumentos. De aquí la importancia que tiene en el desarrollo de las distribuciones de referencia, tanto por su significado matemático como por el momento histórico en que se sitúa; puede decirse que fué uno de los primeros pasos y muy importante además, en la resolución al problema de las distribuciones de referencia.

III.2 PERKS

Hasta la aparición del artículo de Perks (1947) sólo existían dos formas de encontrar distribuciones de referencia:[17]

i) El principio de razón insuficiente de Bayes-Laplace

y ii) Las dos reglas propuestas por Jeffreys: (a) $p(\theta) \propto$ constante, si $\theta \in \mathbb{R}$ y (b) $p(\theta) \propto \theta^{-1}$ si $\theta \in \mathbb{R}^+$.

Como se comentó en III.1, el principio de razón insuficiente de Bayes, tiene problemas graves, cuando el recorrido de Θ no es todo $\mathbb{R}$. Por otro lado, las reglas (a) y (b) sólo resuelven este problema, para cierto tipo de transformaciones, para los cuales las reglas son invariantes.

El trabajo de Perks (1947) busca una regla única que cubra las reglas pasadas en los casos adecuados, y que además pueda ser aplicada a cualquier clase de parámetro (no sólo cuando $\theta \in \mathbb{R}^+$, sino también cuando $\theta \in (a,b) \subset \mathbb{R}$) y que sea invariante ante cualquier tipo de transformación. Lo único que Perks demuestra es la invarianza (cuestión que se tratará más adelante), pues menciona que su regla debe de tomarse como un postulado.

17.- Aunque el artículo de Jeffreys (1946) apareció antes que el de Perks (1947), este último ya estaba en prensa cuando la aparición del primero. Esto explica el porque se consideran sólo las reglas (i) y (ii) y no la nueva regla de Jeffreys que se trató en la sección III.1.

La regla es la siguiente:

$$p(\theta) \propto \frac{1}{\sigma_\theta}$$

donde θ es un parámetro de un función de distribución, cuyo recorrido puede variar en todo $\mathbb{R}$ o en subconjuntos de $\mathbb{R}$, y para el cual existe una estadística suficiente. σ_θ es la desviación estandar de dicha estadística, la cual depende de θ .

La demostración de que esta regla es invariante ante cualquier transformación, no es ni por mucho satisfactoria. Por otro lado, tanto Solomon (1947) como Kendall (1947) critican el uso de σ_θ dentro de la demostración, pues no queda claro si σ_θ es la desviación estandar de θ o la desviación estandar de un estimador de θ .

Otra objeción a esta relga, es que no especifica que debe de hacerse en el caso de que no se tenga una estadística suficiente para θ . Esto lleva a que la regla no está definida en forma única, pues dependiendo de la estadística suficiente que se está usando, será el valor de σ_θ .

Al final del artículo en un addendum, Perks menciona que ha leído el artículo de Jeffreys (1946) y conviene en que la regla de Jeffreys generada por la invarianza de las funciones

$$I_1 = \int (\sqrt{dP} - \sqrt{dP'})^2 \quad \text{e} \quad I_2 = \int \log_e \frac{dP}{dP'} d(P-P')$$

es una regla más general que la propuesta por él. Aún más, menciona que

si en su regla, en lugar de usar σ_θ utiliza la raíz cuadrada de la cantidad de información de Fisher (1922,1925), su regla es mejor. Pero esto no es más que considerar a

$$I_2 = \lim_{n\to\infty} \frac{1}{n} \sum_i \log_e \frac{L(x; d_i)}{L(x; d_i')} \left(L(x; d_i) - L(x; d_i')\right)$$

cuestión contemplada por Jeffreys (1961).

En resumen, Perks considera que la regla de Jeffreys, es superior a su propia regla. Sin embargo el mérito de Perks, es que revisa muchos aspectos de la metodología bayesiana dentro de su artículo (de los que la regla es un caso), así como una buena crítica a las otras interpretaciones de la probabilidad. Además, comenta acertadamente los problemas de la metodología bayesiana, en lo referente a las distribuciones de referencia.

III.3 LINDLEY

El resultado que presenta Lindley (1961) para encontrar distribuciones de referencia coincide con la regla de Jeffreys. Sin embargo Lindley lo obtiene a partir del concepto de cantidad de información contenida en un experimento, a diferencia de Jeffreys (1946), quien la obtuvo por principios de invarianza.

Lindley (1956) basado en un trabajo de Shannon (1948), considera el espacio parametral (H) en el que $p(\theta)$ existe $\forall \theta \in (H)$ (se supone que existe una medida dominante, con respecto a la cual puede encontrarse $p(\theta)$), define a

$$I^{\circ} = \int_{(H)} p(\theta) \log p(\theta) d\theta$$

como la cantidad de información que se tiene sobre θ (si $p(\theta)=0$, $I^{\circ}=0$).

De la misma manera, si un experimento se realiza y como resultado se obtiene X, puede definirse la cantidad de información que se tiene sobre θ como

$$I(x) = \int_{(H)} p(\theta|x) \log p(\theta|x) d\theta$$

(Nuevamente, si $p(\theta|x)=0$ defínase $I(x)=0$).

De esta forma, es natural definir la cantidad de información que un experimento proporciona, cuando $p(\theta)$ es el conocimiento inicial que se tiene sobre Θ y X es el resultado del experimento, como

$$I(x) - I^{\circ} = \int_{(H)} p(\theta|x) \log p(\theta|x) d\theta - \int_{(H)} p(\theta) \log p(\theta) d\theta$$

y la cantidad de información esperada, será por tanto

$$E_x\left[I(x) - I^\circ\right] = \int_{\mathcal{X}} \int_{\Theta} \left[p(\theta|x) \log p(\theta|x)\right] p(x)\, d\theta\, dx - \int_{\Theta} p(\theta) \log p(\theta)\, d\theta$$

La justificación de usar este tipo de medidas se menciona en Shannon (1948), Lindley (1956) y Fernández (1978); y se basa fundamentalmente en la siguiente idea:

Supóngase que se sabe que $\theta \in H, C(H)$, por lo que se tiene una cantidad de información I. Si después se llega a conocer el valor de θ, se tendrán nuevas cantidades de información I_1 o I_3 dependiendo de si $\theta \in H$, ó $\theta \in H^c$. Si $p = \int_H p(\theta)\, d\theta$, es natural pedir que el total de información que se tiene sobre θ sea

$$I = I_1 + pI_2 + (1-p)I_3$$

Shannon (1948, apéndice II) demuestra que $I = k \int p(z) \log p(z)\, dz$ es la única función que cumple dicha propiedad (una demostración sencilla de este hecho puede verse en Fernández, 1978 p33).

Supóngase ahora que $H \subseteq \mathbb{R}$, esto es, considérese $p(x|\theta)$ una densidad[18] que depende de $\theta \in \mathbb{R}$. El objetivo de la metodología bayesiana es pasar de un estado en el que θ se conoce vagamente, a otro en el que se conozca -

18.- La densidad es en el sentido de Radon-Nykodim con respecto a la medida dominante del espacio.

completamente. Sin embargo esto generalmente es imposible. Lo que sucede en la mayoría de los casos, es llegar a tener una idea aproximada de θ, por ejemplo, que θ esté en un intervalo de longitud $\delta(\theta)$, donde $\delta(\theta)$ es una función que dependerá de θ (Esto quiere decir que $\delta(\theta)$ no es la misma para distintas densidades $p(x|\theta)$). Si se considera una partición del intervalo definido por $\delta(\theta)$, conocer θ significaría qué punto de la partición se obtiene, por lo que una forma de expresar ignorancia acerca de θ, es asignar una distribución inicial proporcional a $\delta(\theta)$, o bien si el rango de θ puede ser substituído por valores discretos en la partición, la distribución inicial sería asignar probabilidades iguales a todos los valores considerados.

Por otro lado, conocer a θ se traduce en un conocimiento sobre la forma de $p(x|\theta)$; entonces una menera natural de saber que tanto difiere θ de $\theta+\delta(\theta)$ es midiendo que tanto difiere $p(x|\theta)$ de $p(x|\theta+\delta(\theta))$. Esta diferencia puede ser calculada, viendo que tanta información proporciona un experimento para distinguir entre θ y $\theta+\delta(\theta)$.

Si esta cantidad de información varía para cada $\theta\in\Theta$, significará que puede distinguirse entre distintos θ, es decir, algunos serán más creíbles que otros, por lo que un estado de ignorancia será aquel en el que la cantidad de información sea constante para todo $\theta\in\Theta$.

Como se vió anteriormente, una medida de la cantidad esperada de informa-

ción proporcionada por un experimento es

$$\int p(x) \int p(\theta|x) \log p(\theta|x)\, d\theta\, dx - \int p(\theta) \log p(\theta)\, d\theta$$

lo que es equivalente a

$$\iint p(x|\theta)\, p(\theta) \log \frac{p(x|\theta)}{p(x)}\, d\theta\, dx$$

Si se considera únicamente a θ y $\theta+\delta(\theta)$ con probabilidades iniciales iguales, y si además puede desarrollarse hasta orden dos, en series de taylor alrededor de $\theta+\delta(\theta)$, esto es que $\delta(\theta)\to 0$, Lindley menciona que la cantidad de información para este caso, es igual a

$$2\delta^2(\theta)\, E\left\{\left[\frac{\partial \log p(x|\theta)}{\partial \theta}\right]^2\right\} = 2\,\delta^2(\theta)\, I(\theta)$$

Entonces por lo discutido anteriormente, se tendrá que $\delta^2(\theta)\, I(\theta) \propto$ constante, de donde $I^{1/2}(\theta) \propto \delta^{-1}(\theta)$, por lo que una distribución de referencia será

$$p(\theta) \propto I^{1/2}(\theta)$$

que es la regla de Jeffreys,

Como comentario, se quiere hacer notar que la exposición de los puntos importantes para el desarrollo del método está dada en una forma vaga y con-

fusa.

Finalmente, en el caso de que $\theta \in \mathbb{R}^n$, Lindley comenta que el argumento anterior no es válido pues no puede darse una partición en forma natural que sugiere una distribución de referencia, además de que la regla de Jeffreys en el caso multivariado tiene bastantes problemas como ya se mencionó.

III.4 HARTIGAN

La búsqueda de reglas para encontrar distribuciones de referencia que resulten ser invariantes ante cierto tipo de transformaciones, como es el caso de la regla de Jeffreys (1946), es el punto central del trabajo de Hartigan (1964), en el que plantea varias proposiciones de invarianza y se encuentran diferentes tipos de distribuciones de referencia que cumplen con algunas de ellas y en ciertos casos, con todas.

Supóngase que X es una variable aleatoria definida en un conjunto abierto S de $\mathbb{R}^n$ y que (H) es un subconjunto abierto de $\mathbb{R}^k$; considérese además que $p(x|\theta)$[19], la función de densidad[20] asociada a X , es contínuamente diferenciable en (H) , para toda $x \in S$ y que F es una familia de dichas densidades, la cual obviamente depende de X .

Si $p(\theta|x)$ es la densidad final de $\theta \in (H)$, defínase a una inversión como cualquier función que asigna una densidad final $p_F(\theta|x)$ a cada F[21]. Análogamente, una inversión kernel es una función que asigna una densidad inicial a cada F y que determina una inversión mediante la relación $p_F(\theta|x) = p_F(\theta)p(x|\theta)$. Esto significa que una inversión kernel no es más que un método para asig

19.- Aquí x denota el valor que toma la variable aleatoria X .

20.- En el sentido de Radon-Nykodin

21.- Se denotará con p a las inversiones y a las densidades indistintamente, ya que la inversión dependerá de la densidad.

nar distribuciones iniciales.

Las propiedades de invarianza que Hartigan (1964) considera sobre las inversiones son:

i) Sean F y F^* dos familias de densidades. Supóngase que existe una transformación biyectiva T de S a S^* diferenciable, tal que

$$p^*(T(x)|\theta)\frac{\partial T}{\partial x} = p(x|\theta) \quad \forall x \in S \text{ y } \theta \in \Theta$$

Si $p_{F^*}(\theta|T(x)) \propto p_F(\theta|x)$ para toda T que cumpla lo anterior, se dirá que la inversión p es S-invariante (S-Labelling invariant).

ii) Sean F y F^* dos familias de densidades y supóngase que existe una transformación biyectiva y diferenciable T de Θ a Θ^*, tal que

$$p^*(x|T(\theta)) = p(x|\theta) \quad \forall x \in S \text{ y } \theta \in \Theta$$

Se dirá que la inversión p es Θ-invariante (Θ-labelling invariant) si

$$p_{F^*}(T(\theta)|x)\frac{\partial T}{\partial \theta} \propto p_F(\theta|x)$$

iii) Sea Θ^* un subconjunto abierto de Θ y F^*, F las familias de densidades correspondientes. Si $p_{F^*}(\theta|x) \propto p_F(\theta|x)$, con $\theta \in \Theta^*$, se dirá que p es una inversión Θ-<u>restringida invariante</u> (Θ-restric

tion invariant)

iv) Una inversión ρ es <u>suficiente-invariante</u> (sufficiency invariant) si $\rho_{F^*}(\theta|T(x)) \propto \rho_F(\theta|x)$, donde T es una transformación suficiente y diferenciable de S en S^*, esto es, $T(x)$ es suficiente para θ , con $x \in S$.

v) Sean X y Y dos variables aleatorias independientes con densidades $\rho_1(x|\theta)$ y $\rho_2(y|\phi)$ respectivamente, tales que $\rho(x,y|\theta,\phi) = \rho_1(x|\theta)\rho_2(y|\phi)$.

Sea F_1 la familia de densidades $\rho_1(x|\theta)$, $\theta \in \Theta_1$ y F_2 la de $\rho_2(y|\phi)$, $\phi \in \Theta_2$ y sea $F = F_1 \times F_2$ la familia de densidades de la forma $\rho(x,y|\theta,\phi)$ con $\theta \in \Theta_1$ y $\phi \in \Theta_2$. Se dirá que una inversión es <u>producto directo invariante</u> (direct product invariant) si

$$\rho_F(\theta,\phi|x,y) \propto \rho_{F_1}(\theta|x)\,\rho_{F_2}(\phi|y)$$

vi) Sea $\theta \in \Theta$ y considérese a la familia F de densidades $\rho(x|\theta)$ y F^* la de densidades $\rho^*(x_1,\ldots,x_m|\theta)$ donde $X_1, \ldots, X_m$ son variables aleatorias independientes con la misma distribución. Una inversión es <u>repetición invariante</u> (repetition invariant) si

$$\rho_{F^*}(\theta|x_1,\ldots,x_m) \propto \rho(x_2,\ldots,x_m|\theta)\,\rho_F(\theta|x_1)$$

Algunos comentarios se hacen necesarios:

La primera propiedad de invarianza lo que asegura es que la información - que sobre Θ proporciona $\mathcal{I}$, es la misma que $T(x)$. Mientras que la segunda dice que las distribuciones finales para F y F^* pueden ser identificadas por $T(\theta)$.

El tercer tipo de invarianza implica que dado $\mathcal{I}$ la distribución final de Θ es independiente de los valores de $\theta \in \mathcal{H}^c$. La cuarta propiedad es similar a la primera, sólo que no pide que T sea una función inyectiva.

A partir de esto, pueden darse las siguientes definiciones:

I.- Supóngase que h es una inversión kernel que determina una inversión - que cumple con i, ii y iii. Sea F una familia de densidades $p(x|\theta)$, con $\theta \in \mathcal{H}$. Sean $\mathcal{H}_1$ y $\mathcal{H}_2$ dos subconjuntos abiertos de $\mathcal{H}$ y considérese una transformación diferenciable y biyectiva T de S en S' y de $\mathcal{H}_1$ en $\mathcal{H}_2$ tal que

$$p(T(x)|T(\theta))\left|\frac{\partial T}{\partial x}\right| = p(x|\theta) \qquad \forall\, x \in S,\ \theta \in \mathcal{H}$$

Entonces, por (ii) y (iii) se sigue que

$$p_F(T(\theta)|T(x))\left|\frac{\partial T}{\partial \theta}\right| \propto p_F(\theta|T(x)) \qquad \forall\, x \in S,\ \theta \in \mathcal{H}_1$$

y por (i)

$$p_F(T(\theta)|T(x))\left|\frac{\partial T}{\partial \theta}\right| \propto p_F(\theta|x)$$

$$\Rightarrow \exists\, c \in \mathbb{R} \ni \quad p_F(T(\theta))\left|\frac{\partial T}{\partial \theta}\right| = c\, p_F(\theta) \qquad \forall\, \theta \in \mathcal{H}_1 \qquad (1)$$

Toda densidad incial p que cumpla (1), para toda T como la considerada, se le llamará densidad inicial relativamente invariante.

II.-Sea F una familia de densidades $p(x|\theta)$ con $x \in S$ y $\theta \in \Theta$. Supóngase que T es el conjunto de operadores biyectivos y diferenciables definidos en S y en Θ tales que

$$p(T(x)|T(\theta))\frac{\partial T}{\partial x} = p(x|\theta) \qquad \forall x \in S \text{ y } \theta \in \Theta$$

Es fácilmente demostrable que T define un grupo bajo la operación producto entre funciones. Supóngase además que η es tal que existe una biyección entre η y Θ que define a un único grupo en Θ insomorfo a η. Como Θ es localmente compacto pues $\mathbb{R}^k$ lo es, puede pensarse en la medida izquierda y la medida derecha de Haar (left y right Haar measure, respectivamente) con respecto al grupo (Choquet, 1976). Dado esto, se dice que p es una densidad inicial invariante izquierda (left invariant prior density) si

$$p_F(\theta'\theta)\frac{\partial(\theta'\theta)}{\partial\theta} = p_F(\theta) \qquad \forall \theta, \theta' \in \Theta$$

donde $\theta'\theta$ denota el producto de θ' y θ en el grupo inducido en Θ por T. Análogamente, p es una densidad inicial invariante derecha si

$$p_F(\theta\theta')\frac{\partial(\theta\theta')}{\partial\theta} = p_F(\theta) \qquad \forall \theta, \theta' \in \Theta$$

III. Sea F una familia de densidades $p(x|\theta)$ con $\theta \in \Theta$, donde Θ es un subconjunto abierto de $\mathbb{R}$, tales que

$$\frac{\partial^r}{\partial\theta^r}\log_e p(x,\theta) \qquad \text{existe } \forall\, r \leq 2,\ \forall x \in S,\ \theta \in \Theta$$

y tienen segundo momento finito, i.e. $E\left\{\left(\frac{\partial^r}{\partial\theta^r}\log_e p(x|\theta)\right)^2\right\} < \infty$, $r \leq 2$.

Supóngase que T_1 y T_2 son dos transformaciones tales que existe T una transformación de S en S que cumplen con

$$p(T(x)|\theta)\frac{\partial T}{\partial x} = p(x|\theta) \quad \forall \theta \in V_{\theta_0}$$

$$T_1\left[\frac{\partial}{\partial\theta}\log_e p(T(x)|\theta)\right] = \frac{\partial}{\partial\theta}\log_e p(x|\theta) \quad \forall \theta \in V_{\theta_0}$$

$$T_2\left[\frac{\partial}{\partial\theta}\log_e p(T(x)|\theta)\right] + T_1^2\left[\frac{\partial^2}{\partial\theta^2}\log_e p(T(x)|\theta)\right] = \frac{\partial^2}{\partial\theta^2}\log_e p(T(x)|\theta) \quad \forall \theta \in V_{\theta_0}$$

donde $\theta_0 \in \Theta$ es el elemento unitario del grupo y V_{θ_0} denota una vecindad de θ_0 .

Se dirá entonces que p es una densidad localmente invariante en V_{θ_0} si

$$\frac{\partial}{\partial\theta}\log_e p(\theta) = \frac{T_2}{T_1(1-T_1)} \quad \text{en} \quad V_{\theta_0}$$

es decir, a la solución de esta ecuación se le llamará densidad inicial localmente invariante. Implícitamente se está pidiendo que p_F la inversión kernel que asigna la distribución inicial p a F , sea tal que genere una inversión Θ-restringida invariante, donde $V_{\theta_0} \subset \Theta$ es el abierto que se pide en la definición. Hasta aquí, la suposición de que $E\left\{\left(\frac{\partial^r}{\partial\theta^r}\log_e p(x|\theta)\right)^2\right\} < \infty$, $r \leq 2$, no se ha utilizado. Esta suposición es necesaria en los casos en que no es posible encontrar T_1 y T_2 tales que exista T que cumpla con las ecuaciones planteadas. En estas situaciones se tiene

IV.- Sea $S_1 = \frac{\partial}{\partial\theta}\log_e p(x|\theta)$ y $S_2 = \frac{\partial^2}{\partial\theta^2}\log_e p(x|\theta)$, $\forall \theta \in V_{\theta_0}$. Si existen T_1 y T_2 que cumplen con

$$T_1(E(S_1)) = E(S_1)$$

$$T_2(E(S_1)) + T_1^2(E(S_2)) = E(S_2)$$

$$T_1^2\left(E(S_1^2)\right) = E(S_1^2)$$

$$T_1\left[T_2\left(E(S_1^2)\right)\right] + T_1^3\left(E(S_1S_2)\right) = E(S_1S_2)$$

$$T_2^2\left[E(S_1^2)\right] + 2T_2\left(T_1^2\left(E(S_1S_2)\right)\right) + T_1^4\left(E(S_2^2)\right) = E(S_2^2)$$

entonces la solución es la ecuación (si existe)

$$\frac{\partial}{\partial\theta}\log_e p(\theta) = -\frac{E(S_1S_2)}{E(S_2)} \quad \text{con} \quad E(S_1)=0$$

es única y se le llamará densidad inicial invariante localmente asintótica. (asymptotically locally invariant, ALI)

Con esto, se tiene la siguiente <u>definición</u>

Sea $\Theta \subset \mathbb{R}^k$ abierto y $\{p(x|\theta), \theta\in\Theta\}$ una familia de densidades. Supóngase que en $\theta=\theta_0$

$$E\left\{\frac{\partial}{\partial\theta_i}\log_e p(x|\theta)\right\}=0 \quad \forall i\in J_k$$

$$E\left\{\frac{\partial}{\partial\theta_i}\log_e p(x|\theta)\,\frac{\partial}{\partial\theta_j}\log_e p(x|\theta)\right\}+E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log_e p(x|\theta)\right\}=0 \quad \forall i,j\in J_k$$

Sea $\{g_{ij}\}_{i,j\in J_k}$ la matriz inversa de $\left\{E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log_e p(x|\theta)\right)\right\}$. Entonces la <u>densidad inicial ALI</u> definida en $\theta=\theta_0$ es la solución (si existe) de las ecuaciones:

$$\frac{\partial}{\partial\theta_p}\log_e p(\theta) = -\sum_i\sum_j E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_p}\log_e p(x|\theta)\,\frac{\partial}{\partial\theta_j}\log_e p(x|\theta)\right\} g_{ij} \quad \forall p\in J_k$$

La obtención de distribuciones de referencia por estos métodos no siempre es posible, ya sea porque el método no está definido para la familia por no cumplir ésta con las hipótesis, o bien porque el método no lleva a una distribución inicial. Aparentemente todos los métodos funcionan únicamen-

te cuando la densidad es normal, y el único que parece funcionar en todos los casos es el de las distribuciones ALI, para las cuales Hartigan (1964) comenta que este método es bueno en el sentido de que estas distribuciones cumplen con las propiedades de invarianza para las inversiones. Por otro lado, la regla de Jeffreys dá distribuciones de referencia que también cumplen dichas propiedades, por lo que Hartigan (1964) la considera buena también, aunque los métodos no dan las mismas distribuciones para las mismas familias. En un artículo posterior, Hartigan (1965) menciona que si $\theta \in \mathbb{R}$, $L(d(x),\theta)$ es la función de pérdida para la decisión $d(x)$, y define a $d(x)$ insesgada en promedio (unbiased in the mean) si

$$E\{L(d(x),\theta)|\theta_0\} \geq E\{L(d(x),\theta_0)|\theta_0\} \quad \forall\, \theta,\theta_0 \in \mathbb{R}$$

entonces, puede demostrar que su método dá distribuciones iniciales que minimizan asintóticamente el sesgo de $d(x)$.

Sin embargo, si se tiene en cuenta la complejidad de obtener una densidad ALI, puede pensarse que las densidades obtenidas por la regla de Jeffreys son mejores en el sentido práctico.

Hartigan (1964) muestra que si $p(\theta)$ es una densidad obtenida mediante la regla de Jeffreys, entonces

$$p(T(\theta))\frac{\partial T(\theta)}{\partial \theta} = h(\theta)$$

donde T es una transformación biyectiva y diferenciable de S en S y de $\mathcal{H}$ en $\mathcal{H}$. Mientras que si $p(\theta)$ es una densidad ALI

$$p(T(\theta))\frac{\partial T(\theta)}{\partial \theta} = c\,p(\theta) \quad , \; c \in \mathbb{R}$$

Esto es, la densidad de Jeffreys es invariante por la izquierda y la ALI es relativamente invariante.

Dado que toda densidad invariante por la izquierda (o por la derecha) es relativamente invariante, el resultado de Hartigan (1964) no hace más que generalizar la regla de Jeffreys.

Stone (1965,1970) justifica el uso de densidades relativamente invarian - tes como distribuciones de referencia, mostrando que si $p(\theta|x)$ es una distribución final de referencia - la obtenida mediante el teorema de Bayes usando una distribución inicial de referencia - entonces, existe una sucesión de densidades iniciales que cumplen con

$$p_i(\theta) = p(\theta) \Big/ \int_{\Theta_i} p(\theta)\, d\sigma(\theta) \quad \forall \theta \in \Theta_i \,, \; i \in \mathbb{N}$$

con Θ_i compacto $\forall i \in \mathbb{N}$, y que convergen en probabilidad a $p(\theta|x)$, si y sólo sí $p(\theta)$ es invariante por la derecha, donde σ es la medida derecha de Haar con respecto a la cual $p(\theta)$ se define, y $\int |p_i(\theta) - p_i(\theta g)|\, d\sigma(\theta) \to 0$ uniformemente en cualquier subconjunto compacto de Γ, donde $g \in \Gamma$.

En el caso en que $\Theta = \mathbb{R}^n$ y la medida dominante que se utiliza es la medida de Lebesgue, puede mostrarse que invarianza por la derecha e invarianza por la izquierda son conceptos equivalentes (Halmos, 1974).

Esto significa que la densidad obtenida a partir de la regla de Jeffreys puede ser justificada, en estos casos, por el argumento de Stone, lo cual es un punto a favor de su uso.

Villegas (1971) obtiene un resultado similar. Demuestra que si el grupo de transformaciones posee sólo una medida invariante de Haar, entonces esta medida debe ser utilizada para expresar ignorancia. La justificación de este hecho, lo hace mediante consideraciones de invarianza y suponiendo que el grupo sólo consta de transformaciones lineales.

III.5 NOVICK Y HALL

Como se mencionó en II.2, existen diferentes formas de asignar distribu - ciones iniciales que representen las creencias del tomador de decisiones. Una de ellas es mediante el uso de distribuciones conjugadas, concepto que fué introducido formalmente por Barnard (1954) y usado ampliamente por - Raiff & Schlaifer (1961).

Novick & Hall (1965) desarrollan un método para encontrar distribuciones de referencia utilizando distribuciones conjugadas [22] (concepto que se definirá más adelante) e interpretando a las distribuciones propias (en las que la función integra uno) como aquellas que permiten hacer inferencia y predicción, mientras que las distribuciones impropias no lo permiten. Entonces especificando que observaciones son necesarias mínimamente para - permitir inferencia, una caracterización de ignorancia puede ser posible. Esta idea será explicada más adelante en forma detallada, pero antes es - necesario considerar algunas definiciones y sus consecuencias colaterales en el problema de las distribuciones de referencia.

Se dirá que $p(\theta)$, $\theta \in \Theta$, es una densidad conjugada natural (Natural Conjugate Density, NCD) relativa a la muestra $\{X_1, \ldots, X_n\}$ si $p(\theta) \propto p(X_n|\theta)$ $\forall \theta \in \Theta$, donde

22.- Bernardo (1979) menciona que Haldane (1948) también utiliza distribuciones conjugadas para encontrar reglas de indiferencia, desgraciadamente no pudo conseguirse ese trabajo.

$p(\underline{x}_n|\theta)$ denota la función de verosimilitud de la muestra.

Una de las propiedades más importantesde las NCD es que son cerradas bajo el muestreo, es decir si $p(\theta)$ es NCD entonces $p(\theta|x)$ también lo es; hecho fácilmente demostrable a partir de la definición de $p(\theta|x)$ (Raiffa & Schlaifer, 1961).

Esta definición generaliza la dada por Raiffa & Schlaifer, pues no necesariamente la función debe integrar uno, ni tampoco es requerida la existencia de una estadística suficiente para $\theta \in \Theta$.

Defínase ahora para cada familia de distribuciones a C_θ como el conjunto de densidades $p(\theta), \theta \in \Theta$ tales que

$$p(\theta) \propto \prod_{i=1}^{n} [p(x_i|\theta)]^{r_i} \quad \text{con} \quad r_i \in [-1,1] \quad \forall i \in J_n$$

Esta familia C_θ consta de NCD en las que muestras "fraccionales" pueden ser consideradas así como también muestras "negativas". Las primeras son útiles matemáticamente por ejemplo, en el sentido de que puede hacerse tender a cero el tamaño de mustra y obtener la distribución uniforme como límite de una sucesión de densidades NCD. Las segundas, pueden ser vistas como supresión de información. Es sencillo mostrar que si $p(\theta) \in C_\theta$ entonces $p(\phi(\theta)) \in C_{\phi(\theta)}$ siempre que ϕ sea diferenciable y su derivada sea proporcional a alguna verosimilitud.

Con estos conceptos Novick y Hall establecen la siguiente regla de indiferencia: "Para cada distribución ρ definida en Θ y para cada punto extremo $\theta_0 \in \Theta$ de ρ, es decir puntos en los que la distribución es impropia,[23] generar una partición en dos conjuntos S_{θ_0} y S'_{θ_0} del espacio muestral. Rechazar como posible distribución de referencia aquella que combinada con una muestra de S_{θ_0} (S'_{θ_0}) no lleve a una distribución final propia (impropia) $\forall \theta \in \Theta$. Toda distribución no eliminada, es una posible distribución de referencia".

En otras palabras, la regla de indiferencia establece el uso de distribuciones iniciales impropias que lleven a distribuciones finales propias obtenidas con la mínima muestra necesaria.

Como ejemplo, supóngase que X tiene una densidad exponencial con media θ, esto es $p(x|\theta) = \theta^{-1} e^{-x/\theta}$ si $x \in \mathbb{R}^+$ y $\theta \in \mathbb{R}^+$ y cero en otro caso. Es fácil ver que esta función es impropia en $\theta = 0$ y $\theta = \infty$, pues

$$\int_a^b \theta^{-1} e^{-x/\theta} = \log_e \theta \, e^{-x/\theta} \Big|_a^b - \int_a^b \log_e \theta \, \frac{x}{\theta^2} e^{-x/\theta} \, d\theta$$

y si $\theta = 0$ ó $\theta = \infty$, la integral no converge.

Si se considera a $p(\theta) = \theta^{-1}$ como posible distribución de referencia, ésta es impropia en $\theta = 0$ y $\theta = \infty$, por lo que para esta distribución sólo hay que considerar dos particiones: S_0, S'_0 y S_∞, S'_∞ en las que S_0 y S_∞ contie

23.- Entendiendo impropia en θ_0 si $\int_{\theta_0}^{c} p(\theta) d\theta$ diverge, con $c \in \mathbb{R}$.

nen únicamente muestras de tamaño positivo (aunque sean fraccionales), - mientras que S_0' y S_∞' contienen no positivos (esto es negativas y la - nula), ya que si X es una observación se tiene

$$p(\theta|x) = \frac{\theta^{-2} e^{-x/\theta}}{\int_0^\infty \theta^{-2} e^{-x/\theta} d\theta} = \theta^{-2} x e^{-x/\theta}$$

y $\int_0^\infty \theta^{-2} x e^{-x/\theta} d\theta = 1$

de donde $X \in S_0$ y $X \in S_\infty$

Ahora para justificar el uso de $p(\theta) = \theta^{-1}$ se utilizan familias conjugadas. Supóngase $p(\theta) \in C_\theta$, entonces $p(\theta) \propto \theta^a e^{b/\theta}$ con $a, b \in \mathbb{R}$, por definición de C_θ entonces

$$p(\theta|x) \propto \theta^{a-1} e^{(b-x)/\theta}$$

para que $p(\theta|x)$ sea propia se debe tener que $a \leq -1$ y $b \leq 0$ y para que la - distribución inicial se impropia en $\theta = \infty$, $b > 0$ y $a > -1$, por lo que los únicos valores posibles son $a = -1$ y $b = 0$, de donde la distribución de referencia está dada por $p(\theta) = \theta^{-1}$.

El uso de familias de distribuciones conjugadas, permite encontrar una - única distribución que cumple con la regla de indiferencia de Novick y - Hall, esto no significa que sea la única forma de hacerlo.

En 1969, Novick extiende la regla al caso multivariado y establece más - claramente la idea original. Para el caso de un parámetro real, la regla puede resumirse de la siguiente manera:

Dada una familia conjugada C_θ con parámteros (n, x) [24], en donde n es el tamaño de muestra de la variable X, la distribución de referencia es aquella en que $(n, x) = (0, 0)$, es decir en la que no se tiene información muestral inicial.

Por ejemplo, si θ es el parámetro de una Bernoulli, la verosimilitud es

$$p(x|\theta) = \theta^x (1-\theta)^{n-x}$$

y la familia conjugada asociada es la familia Beta (Raiffa & Schlaifer, 1961)

$$p(\theta) \propto \frac{\theta^{z-1} (1-\theta)^{m-z-1}}{B(z, m-z)} \quad \text{con } z \in \mathbb{R}, m \in \mathbb{R}$$

si $z = m = 0 \Rightarrow p(\theta) \propto \theta^{-1} (1-\theta)^{-1}$ será la distribución de referencia, pues es impropia y lleva a una distribución final propia.

Para el caso de varios parámetros el procedimiento es similar, basta especificar una distribución conjunta conjugada natural, la cual lleva a distribuciones de referencia -en el sentido mencionado por Novick y Hall (1965)- en los valores nulos.

Raiffa y Schlaifer (1961) y De Groot (1970) tratan este procedimiento para encontrar distribuciones de referencia, pero no en forma tan general como Novick y Hall (1965) y Novick (1969). Por otro lado, hay que hacer notar que las distribuciones de referencia obtenidas mediante el método tratado en esta sección son únicas sólo en el caso de que pertenezcan a C_θ -

24.- Recuérdese que la definición de NCD es relativa a la muestra.

por lo que si no existe una familia conjugada que satisfaga los requeri - mentos del interesado, a pesar de poder aplicar el método de Novick y Hall (1965), éste no le llevará a una distribución de referencia única, lo que podría presentar problemas al momento de elegir una distribución.

III.6 JAYNES

Partiendo de la premisa de que en dos problmeas en los que se tenga la misma información inicial, deben asignarse probabilidades iniciales iguales, Jaynes (1968) desarrolla un método para encontrar distribuciones de referencia. Este método obtiene las mismas distribuciones de referencia que se encuentran mediante la regla de Jeffreys -al menos en los casos más comunes- y justifica el uso de $p(\theta,\sigma)\propto\sigma^{-1}$ como distribución de referencia en el caso normal.

Como ejemplo del funcionamiento del método, supóngase que se tiene la siguiente densidad contínua

$$p(x|\theta,\sigma) = g\left(\frac{x-\theta}{\sigma}\right)\frac{1}{\sigma} \quad \text{con } \theta\in\mathbb{R} \text{ y } \sigma\in\mathbb{R}^+.$$

Si se requiere hacer inferencias sobre θ y σ es necesario especificar una densidad inicial, $p(\theta,\sigma)$. Si además lo único que se sabe sobre θ y σ es que son parámetros de localización y de escala respectivamente, se desea que $p(\theta,\sigma)$ sea una distribución de referencia.

Si un cambio en la escala o en la localización hace aparecer distinto el problema, significaría que se conoce algo sobre θ y σ, por lo que debe tenerse que

$$\theta' = \theta + a \qquad y \ \sigma' = b\sigma \text{ , con } \theta'\in\mathbb{R} \text{ y } \sigma'\in\mathbb{R}^+ \ ^{25}$$

para que el estado de conocimiento no cambie.

25.- Este tipo de invarianza había sido dado por Jeffreyes (1946) aunque no en forma tan clara. Debido a eso, es que este ejemplo fué incluído.

Si $p'(\theta',\sigma')$ es la distribución inicial de θ' y σ', dado el cambio de variable se tendrá $p'(\theta',\sigma') = p(\theta,\sigma)b^{-1}$. Ahora bien, si se desean hacer inferencias sobre θ' y σ', de acuerdo a la premisa establecida al principio de la sección, deberá tenerse que $p'(\theta',\sigma') = p(\theta,\sigma)$, de donde $p(\theta,\sigma) = \frac{1}{b}\, p(\theta+a, b\sigma)$

cuya solución general es

$$p(\theta,\sigma) = \frac{c}{\sigma} \quad , c \in \mathbb{R}$$

por lo que la distribución de referencia conjunta para θ y σ es

$$p(\theta,\sigma) \propto \sigma^{-1}$$

Esta distribución se había ya encontrado mediante la regla de Jeffreys, sólo que ese caso se había supuesto independencia y la regla se aplicaba por separado a cada parámetro, cosa que en este caso no se hace, por lo que puede decirse que este método justifica el uso de $p(\theta,\sigma) \propto \sigma^{-1}$ para el caso normal. En general, en los modelos de regresión, puede mostrarse que la distribución de referencia para los parámetros del modelo es proporcional a σ^{-1} .

El método de Jaynes (1968) puede ser expresado de la siguiente manera: Para expresar lo que se entiende por "ignorancia" es preciso establecer un conjunto de transformaciones que lleven un problema a otro equivalente, y la premisa de consistencia mencionada en el inicio de la sección, impondrá restricciones sobre la forma que deberá tener la distribución inicial de referencia.

Como comentario final, muchos autores adjudican a Jaynes el uso de teoría

de la información para encontrar distribuciones de referencia; sin embargo esto es erróneo, pues el concepto de entropía máxima que utiliza para encontrar distribuciones iniciales, parte del hecho de que es necesario tener cierta información inicial. El concepto de máxima entropía lo que hace es encontrar distribuciones iniciales que expresen esta información y no contradiga información futura.

Como Jaynes menciona es necesario resolver primero el problema de "igno - rancia completa", para después utilizar el principio de entropia máxima. (Jaynes, 1968 p 236, 1978 p 21).

Otro autor que utiliza el principio de invarianza que usa Jaynes, es Vi - llegas (1977a, 1977b). En el primer artículo considera sólo el caso de la asignación de distribuciones de referencia para los parámetros de la distribución Normal. La diferencia en las distribuciones de referencia que - obtienen, es que mientras Jaynes (1968) utiliza las transformaciones

$$\theta' = \theta + a \quad , \quad \sigma' = \sigma b \quad y \quad x' - \theta' = b(x - \theta)$$

Villegas (1977a) usa

$$\theta' = a + b\theta \quad , \quad \sigma' = b\sigma \quad y \quad (x' - \theta')/\sigma' = (x - \theta)/\sigma$$

lo que lleva a que la distribución de referencia encontrada por Jaynes sea

$$p(\theta, \sigma) \propto \sigma^{-1}$$

y la encontrada por Villegas

$$p(\theta, \sigma) \propto \sigma^{-2}$$

Sin embargo, Villegas menciona que si θ y σ son considerados independientes, entonces el procedimiento se utiliza para cada parámetro, resultando que

$$p(\theta,\sigma) \propto \sigma^{-1}$$

Este razonamiento es similar al hecho por Jeffreys (1946) y el cual se vió no era muy convincente. Por otro lado, las transformaciones dadas por Jaynes son más "naturales" que las de Villegas, si se toma en cuenta que θ es un parámetro de localización y σ de precisión.

En el segudo artículo (1977b), Villegas trabaja el proceso Poisson y el modelo Multinomial. En el primero, la transformación que propone es la misma que Jaynes y por tanto obtiene la misma distribución de referencia, $p(\theta) \propto \theta^{-1}$. En el segundo modelo, encuentra condiciones que la distribución inicial de referencia debe de cumplir y encuentra que debe de ser $p(\pi_1,\dots,\pi_I) \propto \prod_{i=1}^{I} \pi_i^{-1}$, en donde π_i es la probabilidad de la clase i .

Para encontrar esta distribución, establece el siguiente principio que debe ser cumplido por las distribuciones de referencia definidas para más de un parámetro. El principio es llamado Principio de Compatibilidad.

Se dirá que la distribución $p(\alpha)$ para el parámetro α , es compatible con

$p(\alpha,\beta)$ (α,β parámetros) si las inferencias que sobre α se hagan son las mismas usando $p(\alpha)$ y el modelo marginal y usando $p(\alpha,\beta)$ y el modelo completo y haciendo inferencias marginales. Esto es, el proceso de marginalización puede ser hecho antes o después de encontrar la distribución final - sin alterar las inferencias.

Esto intuitivamente parece obvio, sin embargo se verá en el próximo capítulo, como el proceso de marginalización crea problemas.

III.7 BOX Y TIAO

Como se mencionó anteriormente, algunos autores denominan a las distribuciones de referencia como "distribuciones que dejan hablar a los datos". El método que proponen Box y Tiao (1973, sec. 1.3), justifica de alguna manera esta expresión. Este método propone el uso de "distribuciones impropias localmente uniformes", en el sentido de que representarán el comportamiento local de la distribución inicial en la región en que la verosimilitud es apreciable, pero no sobre todo el rango admisible. En general, se utilizarán distribuciones iniciales que no varíen mucho dentro de la región en donde la verosimilitud es apreciable y que no tome valores grandes fuera de esta región. A partir de esta idea, Box y Tiao (1973) presentan un argumento para seleccionar una métrica particular en términos de la cual, una distribución inicial localmente uniforme, pueda ser considerada como de referencia y que sea invariante ante transformaciones uno a uno de los parámetros. El planteamiento general del método es el siguiente

Supóngase que $\theta \in \mathbb{R}$ y que existe una transformación ϕ uno a uno de θ tal que, en términos de ϕ que la verosimilitud correspondiente a θ sea trasladada por los datos (data translated). Esto quiere decir que la verosimilitud para $\phi(\theta)$ está completamente determinada a priori, salvo por su parámetro de localización que dependerá de los datos que van a obtenerse. Matemáticamente, $p(x|\theta)$ es trasladada por los datos, en términos de ϕ, si -

puede expresarse como

$$p(x|\theta) = g(\phi(\theta) - f(x))$$

donde g es una función conocida e independiente de X y f es una función sólo de X .

De esta forma, un estado de indiferencia puede ser expresado asociando una distribución uniforme para $\phi(\theta)$, esto es $p(\phi(\theta)) \propto cte$, de donde por el teorema de cambio de variable, una distribución de referencia para θ se - obtiene mediante la relación

$$p(\theta) \propto \left| \frac{d\phi}{d\theta} \right|$$

Sin embargo, no siempre es posible encontrar una transformación ϕ tal - que $p(x|\theta)$ sea una verosimilitud trasladada por los datos; pero en estos - casos es factible una transformación que aproximadamente lleve a $p(x|\theta)$ a una verosimilitud trasladada por los datos. Este procedimiento se pre- senta a continuación :

Supóngase que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución - - $p(x|\theta)$. Si esta distribución cumple con ciertas condiciones de regulari - dad, puede mostrarse (Johnson 1967, 1970) que cuando n crece, la verosi- militud $p(x|\theta)$ es aproximadamente normal y transformaciones uno a uno tam - bién lo son.

Ahora, sea $\hat{\theta}$ el estimador máximo verosímil de θ, desarrollando en se - ries de Taylor $\log_e p(x|\theta)$ alrededor de $\hat{\theta}$ y considerando sólo términos de orden menor o igual a dos se tiene

$$\log_e p(x|\theta) \doteq \log_e p(x|\hat{\theta}) + \left(\frac{\partial \log_e p(x|\theta)}{\partial \theta}\right)_{\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2}\left(\frac{\partial^2 \log_e p(x|\theta)}{\partial \theta^2}\right)_{\hat{\theta}} (\theta - \hat{\theta})^2$$

pero como $\hat{\theta}$ es el estimador máximo verosímil de θ, $\left(\frac{\partial \log_e p(x|\theta)}{\partial \theta}\right)_{\hat{\theta}} = 0$, de donde

$$\log_e p(x|\theta) \doteq \log_e p(x|\hat{\theta}) + \frac{1}{2}\left(\frac{\partial^2 \log_e p(x|\theta)}{\partial \theta^2}\right)_{\hat{\theta}} (\theta - \hat{\theta})^2$$

$$\Rightarrow \log_e p(x|\theta) \doteq \log_e p(x|\hat{\theta}) - \frac{n}{2}(\theta - \hat{\theta})^2 \left(-\frac{1}{n}\frac{\partial^2 \log_e p(x|\theta)}{\partial \theta^2}\right)_{\hat{\theta}} \qquad (1)$$

Considérese el logaritmo de la densidad normal

$$\log_e p(x|\mu,\sigma^2) = cte - \frac{1}{2\sigma^2}(x-\mu)^2 \qquad (2)$$

Denotando con $J(\hat{\theta}) \alpha \left(-\frac{1}{n}\frac{\partial^2 \log_e p(x|\theta)}{\partial \theta^2}\right)_{\hat{\theta}}$[26], comparando (1) y (2) y usando el hecho de que $p(x|\theta)$ se aproxima a una normal cuando n crece, es natural pensar que la desviación estándar de $p(x|\theta)$ es

$$n^{-\frac{1}{2}} \; J^{-\frac{1}{2}}(\hat{\theta})$$

26.- Suponiendo que $J(\hat{\theta})$ es sólo función de $\hat{\theta}$, lo cual es válido si $p(x|\theta)$ es de la familia exponencial.

Supóngase ahora que ϕ es una transformación uno a uno de θ, entonces

$$J(\phi) = \left(-\frac{1}{n}\frac{\partial^2 \log_e p(x|\phi)}{\partial \phi}\right)_{\hat{\phi}} = -\frac{1}{n}\left(\frac{\partial^2 \log_e p(x|\theta)}{\partial \theta^2}\right)_{\hat{\theta}} \Big/ \left(\frac{d\phi}{d\theta}\right)^2_{\hat{\theta}}$$

$$\Rightarrow J(\phi) = J(\hat{\theta}) \Big/ \left(\frac{d\phi}{d\theta}\right)^2_{\hat{\theta}}$$

Si ϕ es tal que

$$1 \Big/ \left|\frac{d\phi}{d\theta}\right|_{\hat{\theta}} \propto J^{-1/2}(\hat{\theta})$$

entonces $J(\hat{\phi})$ será independiente de $\hat{\phi}$, pues $J(\hat{\phi}) \propto J^{-1/2}(\hat{\theta})\, J^{\frac{1}{2}}(\hat{\theta}) = 1$

lo que implica que $J(\hat{\phi}) \propto cte$, de donde

$$\log_e p(x|\phi) \doteq \log_e p(x|\hat{\phi}) - \frac{n}{2}(\phi - \hat{\phi})^2$$

lo que significa que la verosimilitud será aproximadamente "trasladada - por los datos" en términos de ϕ. De aquí que una métrica para la cual una distribución inicial localmente uniforme sirva como distribución de referencia, puede ser obtenida de

$$\frac{d\phi}{d\theta} \propto J^{1/2}(\theta) \quad \text{o bien} \quad \phi \propto \int_{-\infty}^{\theta} J^{1/2}(t)\, dt$$

lo que a su vez implica que una distribución de referencia para θ es

$$p(\theta) \propto \left|\frac{d\phi}{d\theta}\right| \propto J^{1/2}(\theta)$$

Este procedimiento sólo es válido si $J(\hat{\theta})$ es función de $\hat{\theta}$ únicamente, en caso de que esto no sea cierto, Box & Tiao (1973) proponen como método alternativo a la regla de Jeffreys. De hecho, puede mostrarse que en el

caso de la familia exponencial que los dos métodos coinciden y por lo tanto las mismas distribuciones de referencia son obtenidas en esos casos. Por otro lado, el método propuesto por Box & Tiao puede extenderse al caso en que $\theta \in \mathbb{R}^n$, definiendo a la verosimilitud trasladada por los datos en términos de una transformación ϕ uno a uno de $\mathbb{R}^n$ en $\mathbb{R}^n$ como

$$p(x|\theta) = g(\phi - f(x))$$

donde g es una función independiente de x en $\mathbb{R}^n$ y f es una función de $\mathbb{R}^n$ en $\mathbb{R}^n$ que depende de x únicamente. De esta forma, una distribución de referencia para θ es encontrada a partir de

$$p(\theta) \propto |J|$$

donde $|J|$ es el valor absoluto del jacobiano de la transformación.

Sin embargo esta regla para el caso multivariado tiene los mismos problemas que la regla de Jeffreys para estos casos, por lo que no es adecuado su uso.

III.8 PICCINATO

Varios autores han desarrollado métodos para encontrar distribuciones de referencia, basándose en la idea de que la "no información" significa que el conocimiento final depende únicamente de la información muestral y que las opiniones iniciales no influyen en la obtención de la distribución final (Novick & Hall, 1965; Novick, 1969; Box & Tiao, 1973).

El método que desarrolla Piccinato (1973) parte de la misma premisa y presenta ciertas analogías con el método de Novick & Hall (1975). El procedimiento se basa en la existencia de una funiconal φ tal que $\varphi(p(x,\theta)) = \theta$, y de aquí encontrar una distribución inicial tal que la distribución final derivada de ella, tenga como "punto representativo" (concepto que se definirá posteriormente) a $\varphi(p_n)$ en donde p_n es la distribución empírica de una muestra aleatoria de $p(x|\theta)$. El método es el siguiente:

Sea $F = \{ p(x|\theta); \theta \in \Theta \}$ y sea $x_1, \ldots, x_n$ una muestra aleatoria de una función $p \in F$. Para esta muestra aleatoria puede construirse la distribución empírica p_n, la cual converge uniformemente en probabilidad a p[27]. Debido a ésto, p_n puede ser considerada, en cierto sentido, como un estimador de p. Denótese con P_D al conjunto que contiene tanto a F como a -

27.- Teorema de Glivenko-Cantelli. Véase por ejemplo Ash (1972) p 358

la clase de distribuciones empíricas asociadas a muestras aleatorias de elementos de F. Supóngase que existe una funcional $\varphi: P_D \to \Theta$, esto es

$$\varphi\left(p(x|\theta)\right) = \theta \quad \forall\, p \in F \quad y \quad \varphi\left(p_n(x)\right) = \theta_e$$

Ya que $p_n(x)$ existe independientemente de cualquier información inicial, θ_e puede ser considerado como un estimador de θ, y es obtenido sólo con la información muestral.

Sea P el conjunto de distribuciones iniciales definidas en Θ y considérese a la función $L(\theta, p, x)$ con $\theta \in \Theta$, $p \in P$ y $x \in \mathcal{X}$, que explicará la pérdida en que se incurre si se utiliza a θ en la representación de la distribución final derivada de p y los datos x. Si existe $\theta_r \in \Theta$ ∋

$$L(\theta_r, p, x) \leq L(\theta, p, x) \quad \forall\, \theta \in \Theta$$

se dirá que θ_r es un "punto representativo" de $p(\theta|x)$. Esto es, un "punto representativo" minimiza la pérdida definida por $L(\theta, p, x)$.

Con estos conceptos, Piccinato (1973) define a una distribución de referencia como aquella que cumple con

$$\theta_r = \theta_e \qquad \text{para todo conjunto de datos } x \in \mathcal{X}.$$

Es decir, $p_0 \in P$ es una distribución de referencia si el punto representativo de la distribución final obtenida a partir de p_0, es igual a $\varphi(p_0)$ y esta relación se cumple para toda $x \in \mathcal{X}$.

Esta distribución generará entonces, una distribución final que sólo depende de la información X.

En los ejemplos que Piccinato desarrolla en su exposición, la función de pérdida $L(\theta, \rho, x)$ es alguna de las dos siguientes

$$L(\tau, \rho, x) = \int_{\Theta} |\theta - \tau|^r \rho(\theta|x)\, d\theta \quad , \quad r \in \mathbb{N}$$

y

$$L(\tau, \rho, x) = 1 - \int_{\tau-\delta}^{\tau+\delta} \rho(\theta|x)\, d\theta \quad , \quad \delta \in \mathbb{R}^+ \cup \{0\}$$

Si en la primera se hace $r=1$, el punto representativo es la mediana; si $r=2$ la media es el punto que minimiza $L(\tau, \rho, x)$. Mientras que si $\delta=0$ en la segunda, el punto representativo resulta ser la moda.

Esto no significa que dichas funciones deban ser usadas siempre, pues la elección de $L(\tau, \rho, x)$ dependerá de la persona que estudie el problema; sin embargo, el uso de estas funciones facilita en muchos casos el problema y dan resultados satisfactorios.

Por otro lado, para calcular $L(\tau, \rho, x)$ se necesita conocer la forma analítica de $\rho(\theta|x)$, de aquí que usar distribuciones que sean conjugadas ayudará en la resolución del problema. Como ejemplo del uso del método, se tomará uno del artículo original de Piccinato (1973).

Supóngase que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución Poisson, con parámetro θ . Se sabe que la familia conjugada asociada es la Gama (De Groot, 1970). Entonces

$$\rho(\theta) = k\, e^{-\alpha\theta}\, \theta^{\beta-1} \quad \text{con } \alpha, \beta \in \mathbb{R}^+$$

y $$\rho(x|\theta) = e^{-\theta}\, \frac{\theta^x}{x!}$$

por lo que

$$\rho(\theta|x) = k\, e^{-\theta(\alpha+\Sigma x_i)}\, \theta^{\beta+\Sigma x_i - 1}$$

Si $\varphi(\rho(x|\theta)) = \int x\, \rho(x|\theta)\, dx$ entonces

$$\theta_e = \bar{X} \quad \text{y} \quad \theta_r = \frac{\beta+\Sigma x_i}{\alpha+n}$$

de donde $\theta_e = \theta_r \Leftrightarrow \alpha = \beta = 0$, de aquí resulta que la distribución de referencia para θ es

$$\rho(\theta) \propto \theta^{-1}$$

Si φ es tal que θ_r es la moda, esto es

$$\theta_r = \begin{cases} 0 & \text{si } \Sigma x_i + \beta < 1 \\ \dfrac{\beta+\Sigma x_i - 1}{\alpha+n} & \text{si } \beta+\Sigma x_i \geq 1 \end{cases}$$

entonces $\alpha = 0$, $\beta = 1 \Leftrightarrow \theta_e = \theta_r$, por lo que

$$\rho(\theta) \propto \text{constante}$$

Finalmente, si $\varphi(\rho)$ es la varianza, la condición $\theta_e = \theta_r$ lleva a la ecuación

$$\frac{1}{n}\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = \frac{\beta+\Sigma x_i}{\alpha+n}$$

la cual no dá condiciones sobre α y β que sean independientes de los datos experimentales, y por tanto no puede encontrarse una distribución de

referencia.

En resumen, la forma de la distribución inicial de referencia, dependerá de que punto $\theta \in \Theta$ es considerado como "representativo" para $p(\theta|x)$, lo cual a su vez dependerá de que $L(\theta, p, x)$ es usada para representar las pérdidas.

En un artículo posterior, Piccinato (1978), desarrolla otro método similar al presentado, con la diferencia de que se basa en el problema de predicción y no en el de estimación.

El problema de predicción queda planteado como sigue

Dado un $\theta \in \Theta$, se observa $x \in \mathcal{X}$ de acuerdo a $p(x|\theta)$ y se desea predecir $y \in Y$ que está regida por $p(y|\theta)$, en donde se supone que $p(x,y|\theta) = p(x|\theta)p(y|\theta)$ A partir de esto, la distribución predictiva de y dado que se observó x está dada por

$$p(y|x) = \int_{\Theta} p(y|\theta)\, p(\theta|x)\, d\theta$$

Ahora se define a una transformación $\varphi : \mathcal{X} \times \mathcal{F} \to \mathbb{R}^+ \cup \{0\}$, que medirá la "distancia" entre puntos de $\mathcal{X}$ y elementos de $\mathcal{F}$. Esta transformación es similar a $L(\theta, p, x)$ aunque con distinto significado, pues Piccinato (1978) menciona que dos formas obvias de φ son

$$\varphi(\xi, p) = \int_{\mathbb{R}} |\xi - x|^k \, p(x|\theta)\, dx, \quad k \in \mathbb{R}^+, \ \xi \in \mathcal{X} \ \text{y} \ p \in \mathcal{F}$$

$$\text{y} \quad \varphi(\xi, p) = 1 - p(\xi - h \leq x \leq \xi + h \,|\, \theta) \quad h \in \mathbb{R}^+ \cup \{0\}, \ \xi \in \mathcal{X} \ \text{y} \ p \in \mathcal{F}$$

Análogo al caso anterior, se dice que $\xi_0 \in \mathcal{X}$ es un "punto representativo" de $p(x|\theta)$ con respecto a φ si

$$\varphi(\xi_0, p) \leq \varphi(\xi, p) \quad \forall \ \xi \in \mathcal{X}$$

Por otro lado se dirá que $\{p(y|x) : x \in \mathcal{X}\}$ es una familia conservativa con respecto a φ, si ξ es un punto representativo de $p(y|\xi)$, esto es, si $\forall \ \xi, \eta \in \mathcal{X} \quad \varphi(\xi, p(y|\xi)) \leq \varphi(\eta, p(y|\xi))$

Finalmente, se define a una distribución de referencia, como aquella que genera una distribución predictiva conservativa (con respecto a φ).

Dada esta definición, al usar $\varphi(\xi, p(y|\xi))$ como distancia, el punto representativo (si $k=2$) resulta ser la esperanza condicional de y dado X, esto es

$$\int_Y y\, p(y|x)\, dy = \xi_0$$

o bien ($h=0$) la moda condicional, si es que usan las expresiones dadas arriba para $\varphi(\xi, p(y|\xi))$.

Nuevamente, para poder encontrar $p(y|x)$ es necesario conocer la forma de $p(\theta|x)$, por lo que el uso de familias conjugadas es justificable. El

procedimiento para encontrar la distribución de referencia es similar al caso anterior, y de nuevo, la forma de $p(\theta)$ dependerá del punto representativo elegido en $\mathcal{L}$, lo que equivale a la elección de $\varphi(\beta, \rho)$.

La analogía con el método de Novick & Hall, radica en el uso de familias conjugadas y en que la distribución de referenica es encontrada a partir de condiciones impuestas en los parámetros iniciales y finales. Sin embargo la justificación de cada método es distinta, pareciendo la de Novick & Hall más natural.

III.9 DISTRIBUCIONES DE REFERENCIA Y MEDIDAS DE INFORMACION

El uso de medidas de información para encontrar distribuciones iniciales se ha extendido al caso de distribuciones de referencia. Como se recordará, Lindley (1956) definió la medida de información esperada por un experimento y a partir de ella encontró una justificación a la regla de Jeffreys con estos argumentos (ver III.3). Posteriormente, Jaynes (1968) utiliza el concepto de entropía máxima para encontrar distribuciones iniciales no conflictivas con los datos (III.6). Fundamentalmente, las medidas de información han sido derivadas del trabajo de Shannon (1948) y justificadas por Lindley (1956), Lee (1964), Good (1966) y Bernardo (1979).

La definición más utilizada, es la establecida por Lindley, y es la siguiente

La cantidad de información que un experimento proporciona, cuando $p(\theta)$ es el conocimiento inicial que se tiene sobre Θ y X es el resultado del experimento, está dada por

$$\int_{\Theta} p(\theta|x) \log p(\theta|x) d\theta - \int_{\Theta} p(\theta) \log p(\theta) d\theta$$

por lo que la cantidad de información esperada es

$$\int_{X} \int_{\Theta} p(x)\, p(\theta|x) \log p(\theta|x) d\theta - \int_{\Theta} p(\theta) \log p(\theta)\, d\theta$$

Dos de los tres métodos que se presentarán en esta sección se basan en esta definición y son debidos a Bernardo (1975, 1979). El tercero, debido a Zellner (1977), usa otra medida de información, aunque es similar. La analogía de los tres casos es que dada una medida de información I, la distribución inicial de referencia será aquella que maximiza I, en cierto conjunto $\mathcal{P}$ de distribuciones iniciales; la diferencia es el tipo de medida de información usada.

El primer método que se tratará será el de Zellner y posteriormente los de Bernardo. Al final de la sección se compararán estos métodos.

Zellner (1977) considera la siguiente medida de información:

$$I = \int_{\Theta}\int_{\mathcal{X}} p(x|\theta)p(\theta)\log p(x|\theta)\,dx\,d\theta - \int_{\Theta} p(\theta)\log p(\theta)\,d\theta$$

que es equivalente a

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

Esta medida difiere en la dada por Lindley (1956) en el uso de la función de verosimilitud en lugar de la distribución final, hecho que no justifica muy claramente Zellner. Sin embargo, a partir de esta medida, define a la distribución de referencia como aquella que maximiza I y demuestra (Teorema 1, p213) que dicha distribución está dada por

$$p(\theta) = c\,\exp\left\{\int_{\mathcal{X}} p(x|\theta)\log p(x|\theta)\,dx\right\}$$

en donde C no es más que una constante de normalización, esto es

$$C = \int_{\Theta} \exp\left\{ \int_{\mathcal{X}} p(x|\theta) \log p(x|\theta)\, dx \right\} d\theta$$

Antes de discutir las consecuencias de esta definición, se verá el enfoque de Bernardo, para así discutirlos juntos y compararlos.

El desarrollo presentado en Bernardo (1975) es generalizado en un trabajo posterior (Bernardo, 1979) y será esta generalización la que se presente, puesto que es más clara y formaliza las ideas presentadas en 1975.

La medida de la cantidad de información esperada proporcionada por un experimento que utiliza Bernardo (1975, 1979) es la establecida por Lindley (1956) y está dada por

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

suponiendo que $p(\theta)$ es el conocimiento inicial que se tiene sobre $\theta \in \Theta$.

Sea $\mathcal{C}$ el conjunto de distribuciones iniciales que sean compatibles con cualquier conocimiento inicial que se está dispuesto a considerar y denótese por $I(n)$ a la cantidad de información esperada proporcionada por n repeticiones del experimento. Supóngase además que $\mathcal{C}$ es compacto con respecto a cualquier topología en donde el límite en distribución tenga sentido. Sea $\theta \in \Theta$, entonces $p(\theta)$ es la distribución de referencia para

θ si

$$p(\theta|x) \propto p(\theta)\, p(x|\theta)$$

cuando $p(\theta|x) = \lim_{\infty} p_n(\theta|x)$ y $p_n(\theta|x) \propto p_n(\theta)\, p(x|\theta)$ donde $p_n(x)$ es el máximo de $I(n)$ en C, $\forall n \in \mathbb{N}$.

En algunos casos, cuando ciertas condiciones de regularidad se dán, que garanticen las operaciones realizadas, la obtención de la distribución de referencia es más sencilla, ya que

$$I(n) = \int_{\mathcal{X}} \int_{\Theta} p(x)\, p(\theta|x) \log p(\theta|x)\, d\theta\, dx - \int_{\Theta} p(\theta) \log p(\theta)\, d\theta$$

$$\Rightarrow \quad I(n) = \int_{\mathcal{X}} \int_{\Theta} p(\theta)\, p(x|\theta) \log p(\theta|x)\, d\theta\, dx - \int_{\Theta} p(\theta) \log p(\theta)\, d\theta$$

$$\Rightarrow \quad I(n) = \int_{\Theta} p(\theta) \left(\int_{\mathcal{X}} p(x|\theta) \log p(\theta|x)\, dx - \log p(\theta) \right) d\theta$$

así $$I(n) = \int_{\Theta} p(\theta) \left(\log \left\{ \exp \left(\int_{\mathcal{X}} p(x|\theta) \log p(\theta|x)\, dx \right) \right\} - \log p(\theta) \right) d\theta$$

entonces $$I(n) = \int_{\Theta} p(\theta) \log \left\{ \frac{\exp \left(\int_{\mathcal{X}} p(x|\theta) \log p(\theta|x)\, dx \right)}{p(\theta)} \right\} d\theta$$

y el máximo es $$p_n(\theta) \propto \exp \left(\int_{\mathcal{X}} p(x|\theta) \log p^*(\theta|x)\, dx \right)$$

siempre y cuando n sea grande y en donde $p^*(\theta|x)$ es la distribución final asintótica independiente de la inicial.

En el caso de la distribución de referencia encontrada por Zellner, puede demostrarse que es invariante ante transformaciones lineales de paráme -

tros de escala y de localización, hecho que es más general en el segundo caso, pues si φ es una transformación continua uno a uno de Θ, se tiene que

$$p_n(\varphi) \propto \exp\left(\int_{\mathcal{X}} p(x|\varphi) \log p^*(\varphi|x) dx\right)$$

$$\Rightarrow \quad p_n(\varphi) \propto \exp\left(\int_{\mathcal{X}} p(x|\theta) \log |J| \, p^*(\theta|x) \, dx\right)$$

donde $|J|$ es el jacobiano de la transformación; desarrollando el exponente se llega a

$$p_n(\varphi) \propto \exp\left(\int_{\mathcal{X}} p(x|\theta) \log p^*(\theta|x) dx\right) \exp\left(\log |J| \int_{\mathcal{X}} p(x|\theta) dx\right)$$

$$\therefore \quad p_n(\varphi) \propto p_n(\theta) |J|$$

Este resultado también es válido para la distribución de referencia encontrada por Zellner, aunque no lo haga claro y sólo justifique los casos en que se tengan parámetros de localización y de escala.

Zellner muestra que si θ es un parámetro de localización, la distribución de referencia asociada es uniforme, mientras que si se trata de una parámetro de escala, la distribución es proporcional a θ^{-1}.

En general, las distribuciones de referencia que obtiene Zellner con este

método, coinciden con las que se obtienen con la regla de Jeffreys, con la diferencia de que en el caso Normal obtiene σ^{-1} en lugar de la de Jeffreys, σ^{-2}.

La definición que da Zellner puede ser extendida al caso multivariado, de hecho en la exposición dada aquí, no se especificó la dimensión de θ, lo que hace general la presentación.

En el caso del trabajo de Bernardo, resultados similares son obtenidos, además demuestra que cuando las condiciones de regularidad para obtener normalidad son cumplidas y no existen parámetros de ruido, la regla de Jeffreys y su método coinciden. La demostración no es ni por mucho sencilla, por lo que no se incluye.

La gran limitación de ambos métodos, en el aspecto práctico al menos, es la dificultad de calcular ciertas integrales que aparecen como resultado del método, sin embargo, tanto Zellner como Bernardo dán diferentes ejemplos, aunque, como es el caso de Zellner, no los desarrollan ó bien el desarrollo supone matemáticas avanzadas.

En el siguiente capítulo se tratará más el método de Bernardo, con referencia a los problemas que aparecen en el uso de distribuciones de referencia.

III.10 BIBLIOGRAFIA DEL CAPITULO

Ash (1972)

Barnard (1954)

Bayes (1763)

Bernardo (1975,1979)

Box & Tiao (1973)

Choquet (1976)

DeGroot (1970)

Fernández (1978)

Fisher (1922, 1925)

Good (1966)

Haldane (1948)

Halmos (1974)

Hartigan (1964,1965)

Jaynes (1968,1978)

Jeffreys (1939/61, 1946)

Kendall (1947)

Lee (1964)

Lindley (1956, 1961)

Novick (1969)

Novick & Hall (1965)

Perks (1947)

Piccinato (1973, 1978)

Raiffa & Schlaifer (1961)

Shannon (1948)

Solomon (1947)

Stone (1965, 1970)

Villegas (1971, 1977a, 1977b)

Zellner (1977).

CAPITULO IV

Se comentarán algunos de los problemas que surgen con el uso de distribuciones de referencia, en general cuando no son impropias, y varias alternativas de su solución serán mostradas.

IV.1 PARADOJAS DE MARGINALIZACION

Algunos autores cuestionan el usar de distribuciones de referencia impropias en el análisis estadístico, pues han desarrollado algunos ejemplos que arrojan resultados incompatibles al usar distribuciones impropias. Las paradojas de marginalización son, tal vez, de las críticas más fuertes al uso de este tipo de distribuciones. Stone & Dawid (1972) y Dawid, Stone & Zidek (1973) han generado varios ejemplos en este sentido.

En general el problema de marginalización puede resumirse en lo siguiente

Supóngase que $\theta \in (H)$ es tal que $\theta = (\theta_1, \cdots, \theta_n)$ y considérese una transformación de θ, $T(\theta) = (\theta_{i_1}, \cdots, \theta_{i_k})$, con $k \leq n$; si se desean hacer inferencias sobre $f(\theta_{i_1}, \theta_{i_2}, \cdots, \theta_{i_k})$, habrá que calcular la distribución final marginal. El problema aparece cuando estas inferencias se hacen basándose sólo en la información muestral y una distribución de referencia impropia es usada.

El siguiente ejemplo, planteado en Stone & Dawid (1972), ilustra la situación anterior.

Supóngase que se tiene una muestra aleatoria tomada de una distribución normal con media μ y varianza σ^2, desconocidas y que se desean hacer inferencias sobre $\theta = \mu/\sigma$, considerando sólo la información muestral.

La distribución de referencia más aceptada sobre μ y σ es $p(\mu,\sigma) \propto \sigma^{-1}$ (Jeffreys, 1946; Jaynes, 1968; Box & Tiao, 1973). Usando el teorema de Bayes, se encuentra fácilmente

$$p(\mu,\sigma|x_1,\ldots,x_n) \propto \sigma^{-(n+1)} \exp\left\{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2\right\}$$

de donde

$$p(\mu,\sigma|x_1,\ldots,x_n) \propto \sigma^{-(n+1)} \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2 - \frac{\sum(x_i-\bar{x})^2}{2\sigma^2}\right\}$$

Si $\theta = \mu/\sigma$, la distribución final en términos de θ y σ es

$$p(\theta,\sigma|x_1,\ldots,x_n) \propto \exp\left\{-\frac{n}{2}\theta^2 + \frac{n\bar{x}}{\sigma}\theta - \frac{\sum x_i^2}{2\sigma^2}\right\} \sigma^{-n}$$

por lo que

$$p(\theta|x_1,\ldots,x_n) \propto \exp\left\{-\frac{n}{2}\theta^2\right\} \int_0^\infty \sigma^{-n} \exp\left\{\frac{n\bar{x}}{\sigma}\theta - \frac{\sum x_i^2}{2\sigma^2}\right\} d\sigma$$

haciendo $\omega = \sqrt{\frac{\sum x_i^2}{\sigma^2}}$ se tiene

$$p(\theta|x_1,\ldots,x_n) \propto \exp\left\{-\frac{n}{2}\theta^2\right\} \int_0^\infty \omega^{n-2} \exp\left\{r\theta\omega - \frac{1}{2}\omega^2\right\} d\omega$$

con $r = \dfrac{\sum x_i}{(\sum x_i^2)^{1/2}}$; lo cual es una función de r únicamente.

Por otro lado, Stone & Dawid (1972) demuestran que la distribución de r depende μ y σ en función de θ, y está dada por

$$p(r|\mu,\sigma) = p(r|\theta) \propto \exp\left\{-\frac{n}{2}\theta^2\right\} \left(1-\frac{r^2}{n}\right)^{\frac{1}{2}(n-3)} \int_0^\infty \omega^{n-1} \exp\left\{r\theta\omega - \frac{1}{2}\omega^2\right\} d\omega$$

por lo que para hacer inferencias sobre θ, habría que considerar una distribución inicial para θ y combinarla con $p(r|\theta)$ para obtener $p(\theta|r)$. Pero, consideradas como función de θ, $p(r|\theta)$ y $p(\theta|x_1,\ldots,x_n)$ no son proporcionales, por lo que no existe <u>ninguna</u> distribución inicial $p(\theta)$ que permita

llegar a las mismas inferencias obtenidas a partir de $p(\theta|x_1,\cdots,x_n)$.

En Dawid, Stone & Zidek (1973) se plantean más ejemplos parecidos a éste, esto es, se tiene un parámetro θ que es función de los parámetros originales, sobre el que se desea hacer inferencias. Al calcular $p(\theta|x_1,\cdots,x_n)$ se observa que depende únicamente de cierta función de $x_1,\cdots,x_n$ (generalmente una estadística suficiente), cuya distribución sólo depende de θ y que para la cual no existe distribución inicial que lleve a las mismas conclusiones.

Stone & Dawid (1972) mencionan que si en lugar de usar $p(\mu,\sigma) \propto \sigma^{-1}$ se usa $p(\mu,\sigma) \propto \sigma^{-2}$ la paradoja es evitada, lo cual es evidente ya que

$$p_0(\mu,\sigma|x_1,\cdots,x_n) \propto \sigma^{-(n+2)} \exp\left\{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2 - \frac{\Sigma(x_i-\bar{x})^2}{2\sigma^2}\right\}$$

por lo que $$p_0(\theta,\sigma|x_1,\cdots,x_n) \propto \sigma^{-(n+1)} \exp\left\{-\frac{n}{2}\theta^2 + \frac{n\bar{x}}{\sigma}\theta - \frac{\Sigma x_i^2}{2\sigma^2}\right\}$$

de donde $$p_0(\theta|x_1,\cdots,x_n) \propto \exp\left\{-\frac{n}{2}\theta^2\right\} \int_0^\infty \omega^{n-1} \exp\left\{r\theta\omega - \frac{1}{2}\omega^2\right\} d\omega$$

con ω y r definidas como antes; teniéndose que $p_0(\theta|x_1,\cdots,x_n) \propto p(r|\theta)$ consideradas como funicón de θ únicamente. Sin embargo, la solución que proponen resulta ser ¡una distribución impropia!

Otra solución es dada por Bernardo (1977, 1979) y se basa en el método presentado en el capítulo anterior.

Supóngase que $\theta=(\psi,\omega)$ y ψ es el parámetro de interés. Stone (1958), demostró que si la distribución final de θ es asintóticamente normal con precisión $nh(\hat{\theta})$, con $\hat{\theta}$ el estimador máximo verosímil de θ, entonces

$$I(n) = \tfrac{1}{2}\log_e \frac{n}{2\pi e} + \int_{(H)} p(\theta)\log_e \frac{h^{1/2}(\theta)}{p(\psi)}\,d\theta + o(1)$$

a partir de lo cual puede demostrarse (Bernardo, 1976) que si $h(\theta)$ cumple con $h^{1/2}(\theta) = g(\psi)f(\omega)$ entonces la información desconocida para ψ, se maximiza para todo $p(\omega|\psi)$ cuando $p(\psi) = g(\psi)$. De igual forma, suponiendo normalidad para $p(\omega|\psi,x_1,\ldots,x_n)$ con precisión $n\,h_o(\hat{\theta})$ y $h_o^{1/2}(\theta) = g_o(\omega)\,f_o(\theta_{\cdot}|\omega)$[28], el máximo se alcanza cuando $p(\omega|\psi) = g_o(\omega)$; de forma que la distribución inicial de referencia estará dada por

$$p(\psi,\omega) = g(\psi)g_o(\omega)$$

Para este caso específico, se puede demostrar que la distribución final de $\eta=(\theta,\sigma)$ es asintóticamente normal (Walker, 1969) con matriz de precisión $nH(\hat{\eta})$, donde H es la matriz de información de Fisher, de donde la distribución asintótica de θ es normal con precisión

$$(1+\tfrac{\theta^2}{2})^{-1}$$

y la de σ condicionada a θ también es normal con precisión

$$(2+\theta^2)\,\sigma^{-2}$$

28.- La notación $\theta_{\cdot}|\omega$ se usa para cubrir el caso en que $\theta=\{\psi,\omega,\beta,\ldots\}$

de donde

$$g(\theta) = (1+\tfrac{\theta^2}{2})^{-1/2} \quad y \quad g_0(\sigma) = \sigma^{-1}$$

por lo que las distribuciones de referencia son

$$p(\theta) \propto (1+\tfrac{\theta^2}{2})^{-1/2} \quad y \quad p(\sigma|\theta) \propto \sigma^{-1}$$

de donde

$$p(\theta,\sigma) = p(\theta)\,p(\sigma|\theta) \propto \sigma^{-1}\,(1+\tfrac{\theta^2}{2})^{-1/2}$$

de aquí

$$p(\theta|x_1,\ldots,x_n) \propto (1+\tfrac{\theta^2}{2})^{-1/2} \exp\left\{-\frac{n\theta^2}{2}\right\} \int_0^\infty \omega^{n-1} \exp\left\{-\tfrac{1}{2}\omega^2 + r\theta\omega\right\} d\omega$$

lo cual es proporcional a $p(r|\theta)$ evitando la paradoja.

Bernardo (1979) comenta que, aunque no tiene una demostración, este método evita todas las paradojas de marginalización.

IV.2 OTROS PROBLEMAS

Las paradojas de marginalización no son la única crítica a las distribu - ciones de referencia. Aún cuando se usan distribuciones propias, pueden - surgir dificultades. El siguiente problema fué planteado por Efron (1973) y se basa en un artículo de Stein (1959).

Supóngase que $\theta_1, \theta_2, \ldots, \theta_{100}$ son parámetros desconocidos y que se tienen observaciones independientes $x_1, x_2, \ldots, x_{100}$ tales que $x_i \sim N(\theta_i, 1)$. Supóngase además, que quiere expresarse un "conocimiento vago" sobre las θ_i's , lo cual se "logra" haciendo $\theta_i \sim N(0, \varepsilon)$ con $\varepsilon = 10^{10000}$. Esta distribución es - "casi" uniforme pero es propia. Considérese ahora a $\xi = \sum_{i=1}^{100} \theta_i^2$ y supónga- se que $\sum x_i^2 = 200$.

La distribución final de θ_i es normal con media $\frac{x_i \varepsilon}{1+\varepsilon}$ y varianza $\frac{\varepsilon}{1+\varepsilon}$, - para toda $i \in J_{100}$. Además son independientes, por lo que la distribución final de ξ es una χ^2 no central con media aproximádamente igual a 300. ¡Este resultado difiere del estimador insesgado de $\hat{\xi} = 100$!

Este ejemplo, introduce un problema más de fondo que se comentará en el si guiente capítulo, el comparar resultados bayesianos con resultados clási- cos.

Una solución a este problema la dá Bernardo (1979) quien menciona que lo

que debe hacerse es construir una distribución de referencia para ξ y no usar la establecida para $\theta_1, \ldots, \theta_{100}$ (nótese la relación con los problemas de marginalización).

La mayoría de los problemas que resultan al usar distribuciones de referencia, surgen cuando se quieren comparar los resultados obtenidos con la metdología bayesiana con los obtenidos mediante métodos clásicos.

Así el problema de Fieller-Creasy (1954) referente a hacer inferencias sobre el cociente de dos medias normales, surge el comparar un intervalo de confianza con un intervalo de alta densidad, lo cual no tiene sentido, pues parten de principios distintos, por mucho que numéricamente puedan parecerse, además de que tienen significados distintos.

IV.3 BIBLIOGRAFIA DEL CAPITULO

Bernardo (1976) (1977, 1979)

Box & Tiao (1973)

Creasy (1954)

Dawid, Stone & Zidek (1973)

Efron (1973)

Fieller (1954)

Jaynes (1968)

Jeffreys (1946)

Stein (1959)

Stone (1958)

Stone & Dawid (1972)

Walker (1969)

CAPITULO V

CONCLUSIONES.

La finalidad de este trabajo fué dar una revisión al problema de distribuciones de referencia, revisión que no fué exhaustiva debido a los problemas obvios de recolectar toda la bibliografía del tema. Por otro lado el nivel matemático usado en muchos de los métodos, limitó la exposición más de lo deseado, sin embargo como una primera visita al problema, lo hace interesante, quedando como siguiente paso un estudio más profundo que permita comparar todos los métodos y hacerles una crítica mas fuerte o bien, recalcar sus ventajas.

El uso de distribuciones de referencia, como se mencionó al principio, surge al mismo tiempo que la idea de probabilidad subjetiva. En muchos casos, su uso es necesario o conveniente, sin embargo la razón más importante, aparentemente, es buscar una reconciliación con la estadística clásica. Este es un punto muy criticable desde el punto de vista bayesiano, pues por un lado existen autores que demuestran mediante ejemplos la incoherencia de los métodos clásicos y defienden la postura bayesiana, basándose en el hecho de que es un procedimiento coherente que no lleva a contradicciones; mientras que por otro lado, otros autores justifican ciertos resultados argumentando que mediante ellos, los resultados clásicos se "recuperan". Esto es incoherente, desde el punto de vista bayesiano incluso. Una explicación que puede darse a este fenómeno,

si así puede llamársele, es que en la práctica, los resultados clásicos han funcionado y por tanto puede servir como experiencia. Sin embargo, debe de quedar claro que el uso de distribuciones de referencia no es para reproducir resultados clásicos, sino que deben de ser utilizadas como marcos de referencia en el análisis estadístico.

Otro punto que debe de quedar claro y que en cierta forma es mencionado por Bernardo (1977) es que no existe la distribución de referencia para un problema específico. La razón de ésto es que cada método define lo que es "ignorancia" o "referencia" y a partir de eso, encuentra una distribución de probabilbidad que lo exprese, dándose así, distintas soluciones que no siempre coinciden. La mayoría de los métodos pretenden dar la solución al problema de distribuciones de referencia, siendo que lo único que se logra es una solución mas, que puede coincidir con otras dadas anteriormente. Lo que esto plantea, es que si el problema es encontrar la distribución de referencia, entonces no existe solución, esto es, desde el punto de vista bayesiano, la elección de la distribución inicial depende completamente de la persona que va a tomar una decisión o va a hacer inferencias, por lo que la elección de una distribución de referencia y de cierto concepto de "no información" dependerá también de cada persona.

En este trabajo se presentaron diferentes métodos, que pueden ser clasificados en forma general en tres grupos, quitando el principio de razón insuficiente de Bayes-Laplace.

En el primero de ellos están los métodos basados en diferentes tipos de -

invarianza como Jeffreys (1946), Perks (1947), Hartigan (1964,1965), Stone (1965,1970), Jaynes (1968), Villegas (1971,1977a,1977b), Box & Tiao (1973) y Piccinato (1973,1977).

Los que utilizan familias conjugadas, forman el segundo grupo y son los métodos desarrollados por Novick & Hall (1965), Novick (1969), DeGroot (1970), y Raiffa & Schlaifer (1961).

Finalmente, los métodos de Lindley (1961), Zellner (1977) y Bernardo (1975,1979), que utilizan argumentos de teoría de la información, forman el tercer grupo.

La mayoría de los métodos que usan principio de invarianza, obtienen para el caso de $\theta \in \mathbb{R}$, la regla de Jeffreys, e incluso parte de ellos para cuando $\theta \in \mathbb{R}^n$. Sin embargo, puede demostrarse con un ejemplo, que esta regla lleva a resultados incoherentes.

Supóngase que se tiene una observación de una disribución Binomial, $b^+(n,p)$ y se desea encontrar la distribución de referencia para p, mediante la regla de Jeffreys.

La función de verosimilitud es

$$L(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

de donde

$$\log_e L(x|p) = \log_e \binom{n}{x} + x \log_e p + (n-x) \log_e (1-p)$$

$$\Rightarrow \frac{\partial \log_e L(x|p)}{\partial p} = \frac{x}{p} - \frac{n-x}{1-p}$$

y

$$\frac{\partial^2 \log_e L(x|p)}{\partial p^2} = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

entonces $$E\left\{\frac{\partial^2 \log_e L(x|p)}{\partial p^2}\right\} = -\frac{np}{p^2} - \frac{n-np}{(1-p)^2}$$

$$\Rightarrow -E\left\{\frac{\partial^2 \log_e L(x|p)}{\partial p^2}\right\} = n\left(\frac{1}{p(1-p)}\right)$$

por lo que la distribución de referencia para p es

$$P(p) \propto p^{-1/2}(1-p)^{-1/2}$$

Por otro lado , si se tiene una observación de una distribución binomial - negativa $b^-(r,p)$, la distribución de referencia para p, usando el mismo - método, se encuentra como sigue. Suponiendo que se necesitan x ensayos para obtener r éxitos, la verosimilitud es

$$L(x|p) = \binom{r+x-1}{x-1} p^r (1-p)^x$$

de donde $$\frac{\partial^2 \log_e L(x|p)}{\partial p^2} = -\frac{r}{p^2} - \frac{x}{(1-p)^2}$$

$$\Rightarrow -E\left\{\frac{\partial^2 \log_e L(x|p)}{\partial p^2}\right\} = \frac{r}{p^2} + \frac{r(1-p)}{p(1-p)^2}$$

$$\Rightarrow -E\left\{\frac{\partial^2 \log_e L(x|p)}{\partial p^2}\right\} = \frac{r}{p}\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{r}{p^2(1-p)}$$

por lo que

$$P(p) \propto p(1-p)^{-1/2}$$

Sin embargo, cuando en ambos casos se tienen el mismo número de ensayos- y de éxitos, las verosimilitudes son proporcionales y las inferencias que se obtienen a partir de ellas son distintas, lo cual viola el principio - de verosimilitud y por tanto es incoherente.

De igual forma puede mostrarse que los demás métodos son incoherentes pues sus resultados reproducen los de Jeffreys.

En el caso de los métodos basados en teoría de la información, el de Lindley (1961) es incoherente pues no es mas que la regla de Jeffreys y los restantes también, pues como el mismo Lindley (1979) menciona, las distribuciones de referencia para los ejemplos anteriores coinciden, además de que el mismo Bernardo (1979) demuestra, en el caso de que no existan parámetros de ruido (como es éste), que su método reproduce la regla de Jeffreys. En general, cualquier método que utilice la función de verosimilitud, en el sentido usado en los métodos anteriores, obtiene diferentes distribuciones de referencia en el ejemplo anterior.

Si se usan familias conjugadas para encontrar la distribución de referencia en este ejemplo, se sabe (DeGroot,1970), que la familia conjugada es la Beta, obteniéndose como distribución de referencia (cf.III.5)

$$p(p) \propto p^{-1} (1-p)^{-1}$$

la cual es única, independientemente del esquema de muestreo usado.

Esta comparación somera, lleva a pensar que el método que se basa en familias conjugadas es el único coherente. Sin embargo, las incoherencias de los demás métodos sólo se dan en ciertos casos (al igual que las incoherencias en los métodos clásicos), por lo que se siguen usando las distribuciones obtenidas mediante ellos

Estos problemas deben ser estudiados más a fondo, cosa que nadie ha hecho; generalmente cierto autor introduce su método sin tomar en cuenta los anteriores.Esto es, lo desarrollado por otros autores, sólo es mencionado, pero nunca se critica ni se defiende, simplemente se ignora, lo que lleva a -

considerar las distrubuciones de referncia en el sentido de distribuciones iniciales comentado al principio del capítulo.

Como último comentario, hay que hacer notar que varios métodos son muy elaborados desde el punto de vista matemático, lo que hace que su exposición no sea muy clara, y su aplicación muy difícil. Si se piensa en el aspecto práctico del problema (que es muy importante) muchos de los métodos no son aplicables a situaciones concretas, a menos que se manejen diferentes conceptos matemáticos, lo cual es un impedimento para la gente que no tenga esta orientación, y si lo que busca es un rango amplio de utilidad en estos métodos, quedan limitados, sin embargo, por estas cuestiones.

Estos comentarios motivan el desarrollar mas ampliamente el tema, asimismo, proporciona diferentes líneas de investigación, como pueden ser: El uso de medidas de Haar en los métodos estadísticos, cuál es su aplicación y cómo; la importancia del concepto de información y su teoría, como herramienta en el análisi estadístico; la teoría de grupos de transformaciones aplicadas a problemas concretos de estadística, etc.

V.1 BIBLIOGRAFIA DEL CAPITULO.

Bernardo (1975,1979)

Box & Tiao (1973)

DeGroot (1970)

Hartigan (1964,1965)

Jaynes (1968)

Jeffreys (1946)

Lindley (1961)

Novick (1969)

Novick & Hall (1965)

Perks (1947)

Piccinato (1973,1977)

Raiffa & Schlaifer (1961)

Stone (1965,1970)

Villegas (1971,1977a,1977b)

BIBLIOGRAFIA

AITCHISON, J & DUNSMORE, I.R. (1975) Statiscal Prediction Analysis. Cambridge University Press.

ANSCOMBE, F, J (1973) Sequential Medical Trials.JASA 58, 365- 383

ARMITAGE, P (1973) Sequential Medical Trials: Some Comments on F. J. Anscombe's papel.JASA 58, 384-387

ASH, R(1972) Probability and Real Analysis.Academic Press

BARNARD, G.A. (1952) The Frequency Justification of Certain Sequential - Tests- Bk. 39, 144-150

BARTHOLOMEW, D.J (1965) A Comparison of Some Bayesian And Frequentist Inferencies Bk 52, 19-35

BAYES, T(1763) An Essay Toward Solvig A Problem in The Doctrine of Chances Reeditado en Bk 45, 293-315

BERNARDO, J.M. (1975) Non-informative Prior Distributions: A Subjetivist Approac. Bull. Inter. Stat. Inst. 46, 94-97

----------(1976) The Use of Information in The Design and Analysis of - Scientific Experimentation. Tesis Doctoral. University of London.

----------(1977) Inferencia Bayesiana Sobre el Coeficiente de Variación: Una solución a la paradoja de marginalización. Trab. Est. 28, 23-30

----------(1977) Memoria sobre Conceptos, Métodos y Fuentes de la Bioestadística. Dpto. de Bioestadiística, Universidad de Valencia, España.

----------(1979) Reference Posterior Distributions for Bayesian Inference JRSSB 41,113-147 (con discusión)

BOX, G.E.P. & TIAO G.C. (1973) Bayesian Inference in Statistical Analysis Addison- Mesley

-CREASY, M.A (1954) Límits for the Ratio of The Means JRSSB 16, 186-194

CHOQUET (1976) Lectures on Analysis. Benjamín.

DAWID, A.P. (1976) Properties of Diagnostic Data Distributions Bio 32, - 647-655.

DAWID, A.P., STONE, M & ZIDEK, J.V. (1973) Marginalization Paradoxes in - Bayesian and Structural Inference JRSS B35, 189-233 (con discusión).

DE FINETTI, B (1937) La previsión, Ses Lois Logiques, Ses Sonrces Subjectives. Reeditado en Inglés en Studies in Subjective Probability (Kyburg & Smokler ed) 93,158, Wiley (1969)

----------(1961) The Bayesian Approach to The Rejection of outliers Proc. 4th. Berkely Symp. 1, 199-210 Uni. of California Press.

----------(1962) Does it Make Sense To Speak of Good Probability Appraiser? En: The Scientist Speculates: An Anthology of Partly Baked Ideas (Good, ed) 357-364 Basic Books.

----------(1970) Optimal Stadistical Decisions, Mc. Graw Hill

DEMPSTER, A.P. (1963) On a Paradox Concerning Inference About a Covariance Matrix A.MS 34, 1414-1418.

DICKEY, JM (1973) Scientific Reporting And Personal Probabilities: Student Hypothesis JRSSB 35, 285-305.

DUNSMORE, I.R (1966) A Bayesian Approach To Classification JRSSB 28, 568-577

----------(1968) A Bayesian Approac To Calibration JRSSB 30, 396-405.

----------(1969) Regulatin and Optimization JRSSB 31, 160-170

EFRON, B (1973) En Dawid, Stone & Sidek (1973) JRSS B 35, 219

ERICSON W.A. (1965) Optimun Stratified Sampling Using Prior Information. JASSA 60, 750-771

----------(1969) Subjetive Bayesian Models in Sampling Finite Populations JRSSB 31, 195-233.

FERNANDEZ, J. (1978) Acerca de la Teoría de Información y algunas de sus aplicaciones. Com.Int.23, Serie Monografías. Ciencias,UNAM.

FIELLER, E.C. (1954) Some Problems in Interval Estimation JRSSB 16, 175-185.

GEISSER, S (1965) Bayesian Estimation in Multivariate Analysis, AMS 36, 150-159

----------& Cornfiel, J. (1963) Posterior Distributions Formulativariate Normal Parameters JRSSB 25, 368-376

GOOD, I.J. (1966) A Derivation of the Probabilistic Explication of Information JRSSB 28, 578-581

----------(1971) The probabilistic Explanation of Information, Evidence, Surprise, Casulity, Explanation And Utility. En Foundations of Statiscal Inference (Godambe & Sprotteds) 108-141 (con discusión) Holt, Rinehart & Winston

----------(1972)

GUTTMAN, I (1967) The Use of The Concept of a Future Observation in Goodness of fit Problems JRSSB 29, 83-100.

HALDANE, J.B.S. (1948) The Precision Of Observed Valves of Small Frequencies Bk 35, 297-303

HALMOS, P (1974) Measure Theory, Springer-Verlag

HARTIGAN, J.A. (1964) Invariant prior Distributions AMS 35, 836-845

----------(1965) The Asymptotically Unbiased Prior Distributions AMS 36, 836-845

----------(1969) Linear Bayesian Methods. JRSSB 31, 446-454

HILL, B.M. (1969) Foundations for the Theory of Least Squares JRSSB 31, 89-97.

JAYNES, E.T. (1965) Prior Probabilities IEEE Trans. Systems, Science And Cybernetics, SCC-4, 227-291

----------(1978) Marginalization and Prior Probabilities. Studies of Bayesian Statistics (Zellner, ed) North-Holland

JEFFREYS, H. (1946) An Invariant form for the Prior Probability in Stimation Problems. Proc R. S. London A. 186, 453-461.

----------(1939/61) Theory of Probability. Clarendon Press

KENDALL, M (1947) En Perks (1947) J. Inst. Actuaries 73.

LAPLACE, P.S. (1820) Theorie Analytique des Probabilités. Courcier

LEE, P.M. (1964) On the Axioms of Information Theory AM 535, 415-418

LINDLEY, D.V. (1956) On a Measure of the Information Provided by an Experiment AMS 27, 986-1005

----------(1961) The use of Prior Probability Distributions in Statiscal Inference and Decisions, Proc 4ht Berkely Sijmp. 1, 436-468. Uni.of California Press.

----------(1965) Introduction to Probability and Statistics from a Bayesian Viewpoint. Cambridge University Press

----------(1968) The Coice of Variables in Multiple Regression JRSSB 30, 31-66

----------(1969) Bayesian Least Squares. Boll. Inst. Int. St. 43, 1952-1953

----------(1971) Bayesian Statistics, a Review. SIAM

----------(1971) Making Decisions. Wiley

NOVICK, M.R. (1969) Multiparameter Bayesian Indifference Procedures JRSSB 31, 29-64 (con discusión)

----------&Hall, W. J. (1965) A Bayesian Indifference Procedure JASA 60, 1104-1117

PERKS, W (1947) Some Observations on Inverse Probability Including a New Indifference Rule J. Inst. Actuaries 73, 285-334

PICCINATO, L (1973) Un método per Determinare Distribuzioni Iniziali Relativamente Non-Informative. Metron 31, 1-13

----------(1978) Predictive distributions and Non-Informative Priors. - Trans. 7th Prague Conf Information Theory

PRATT, J.W., RAIFFA, H. & SCHLAIFER, R. (1961) The Foundations of Decision Under Uncertainty: An Elementary Exposition. JASA 59, 353-375

PRESS, S.J. (1972) Applied Multivariate Analysis. Hott, Rinehart & Winston.

RAIFFA, H & SCHLAIFER, R (1961) Applied Statiscal Decision Theory MIT Press.

RAMSEY, F.P (1931) Trhth and Probability. Reeditado en Studies in Subjective Probability. (Kyburg & Smokler eds.) 61-92, Wiley 1964.

RENYI, A (1961) Calcul. des Probabilités, Avec Un Appendice Sur la Theorie de L'information. Dunod

SAVAGE, L.J. (1954) The Foundations of Statistics. Wiley

----------(1961) The Foundations of Statistics Reconsidered. Proc 4th - Berkeley Symp. 1, 575-586

----------(1971) The Elicitation of Personal Probabilities And Expecta - tions. JASA 66, 783-801

SHANNON, C. E. (1948) A Matematical Theory of Communication Bell System Tech. J. 27, 379-423, 623-656

SMITH A.F.M. (1973a) Bayes Estimates in One-Way and Two-Way Models Bk 60, 319-329.

----------(1973b) A General Bayesian Linear Model JRSSB 35, 67-75

----------(1977) Análisis de Sensitividad y Elección de Modelo desde una Perspectiva Bayesiana. Univ. de Valencia

SALOMON (1947) En Perks (1947) J. Inst. Acturies 78

STEIN, C (1956) Inadmissibility of the Usual Estimation for the Mean of a Multivariate Normal Distribution Proc 3rd Berkeley Symp 1, 197-206. Univ. of California

----------(1959) An Example of Wide Discrepancy Between Fiducial and Confidence Interval. AMS 30, 877-880

STONE, M (1958) Studies with a Measure of Information. Tesis Doctoral, - Universidad de Cambridge.

----------(1965) Righ Haar Measures for Convergence in Probability to Invariant Posterior Distributions AMS 36, 440-453.

----------(1970) Necessary and Sufficient Conditions for Convergence in probability to Invariant Posterior Distributions AMS 41, 1939-1953.

VILLEGAS, C (1971) On Haar Priors. En Foundations of Statistical Inference (Godambe & Sprott Eds) 409-414. Halt, Rinehart & Winston.

----------(1977a) Inner Statistical Inference JASA 72, 651-654

----------(1977b) On The Representation of Ignorance JASA 72, 651-654

WETHERILL, G. B. (1961) Bayesian Sequential Analysis Bk 48, 281-292

WINKLER, R.L. (1967) The Assessment of Prior Distributions in Bayesian Analysis JASA 62, 776-800

----------(1968) "Good" probability Assessors J. Appl. Metereol. 7, 751-758

----------(1969) Scoring Rules and the Evaluation of Probability Assessors JASA 64, 1073-1078.

----------(1972) Introduction to Bayesian Inference and Decision. Holt, Rinehart & Winston

ZELLNER, A (1971) An Introduction to Bayesian Inference in Econometrics. Wiley

----------(1977)Maximal Data Information Prior Distributions. New Developments in the Applications of Bayesian Methods (Aykac & Brumat eds.)114-132, North Holland.