74
2Ej

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

INTRODUCCION A LA TEORIA DE LA DECISION ESTADISTICA

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I A

P R E S E N T A:

OLIVIA PEREZ PEREZ

DIRECTOR DE TESIS

M. en C. J. GABRIEL HUERTA GOMEZ

TESIS CON
FALLA DE ORIGEN

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

"Introducción a la Teoría de la Decisión Estadística"

realizado por Olivia Pérez Pérez

con número de cuenta 8729655-0 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	M. en C. J. Gabriel Huerta Gómez	Gabriel Huerta H.
Propietario	M. en C. José Antonio Flores Díaz	
Propietario	Act. Ma. del Pilar Alonso Reyes	
Suplente	Dr. José R. Mendoza Blanco	José Rolly Mendoza Blanco.
Suplente	M. en C. J. Salvador Zamora Muñoz	Zamora Muñoz J. S.

Claudia Carrillo Q.
Consejo Departamental de Matemáticas
Act. Claudia Carrillo Quiroz

A mis padres. Gracias por su paciencia y amor.

A Moisés. Quien lo es todo.

Quiero externar un sincero agradecimiento a mis sinodales, quienes participaron en la revisión de este trabajo:

M. en C. José Antonio Flores Díaz

Act. Ma. del Pilar Alonso Reyes

Dr. José Rodolfo Mendoza Blanco

M. en C. J. Salvador Zamora Muñoz.

Un agradecimiento especial para mi director de tesis, M. en C. J. Gabriel Huerta Gómez, por su paciencia, amistad, etc.

No podría dejar de agradecer a todos mis amigos, ellos representan un complemento indispensable de mi vida.

Dra. Patricia Arrieta, Olivia E., Oscar, Gabriel, Victor, Juan Luis, Guille, Gustavo, Paola, Leticia, Norma, Adalberto, Erick, etc.

I N D I C E

INTRODUCCIÓN

Los métodos estadísticos constituyen una herramienta de análisis cada vez más importante en las áreas de trabajo más diversas. Es por esto que resulta interesante ahondar en su estudio.

Como se sabe, la Estadística contribuye al desarrollo de teorías y técnicas apropiadas para hacer inferencia bajo condiciones de incertidumbre, por tal motivo, la *Estadística Bayesiana*, que podría considerarse como la *Teoría de la Decisión Estadística*, constituye una eficiente y necesaria alternativa al enfoque Clásico en la solución de problemas típicamente estadísticos, como son Estimación, Prueba de Hipótesis y Predicción.

La teoría de la decisión estadística y los métodos Bayesianos han sido desarrollados intensamente durante las últimas décadas por investigadores como Jeffreys, Barnard, Ramsey, De Finetti, Savage, Lindley, Anscombe y Stein, quienes han contribuido con un sinnúmero de resultados.

La Estadística Bayesiana se caracteriza fundamentalmente por abordar los problemas estadísticos como problemas de decisión. Este enfoque proporciona una forma natural de plantearlos, ya que a fin de cuentas lo que se hace en un problema de inferencia es "*decidir*" sobre qué valor se asigna a un parámetro desconocido, qué hipótesis se acepta o bien, por cuál tamaño de muestra se opta. Además, por contar con una base axiomática, la Estadística Bayesiana proporciona una serie de resultados consistentes entre sí.

Como se mencionó anteriormente la Estadística Bayesiana es una alternativa necesaria a la Clásica o Frecuentista, ya que el enfoque clásico se dedica a hacer inferencia únicamente a partir de la información muestral, olvidándose de dos aspectos que para algunos de los problemas pueden ser de extrema importancia. Tales aspectos son:

- La evaluación de las posibles consecuencias de las decisiones, que puede ser representada a través de una *función de pérdida* (o de *utilidad*) y

- El conocimiento previo (*Información a priori*) que se tenga del parámetro sobre el que se desea inferir.

Además de involucrar estos dos factores en el análisis de los problemas, si se cuenta con datos obtenidos en una muestra, la metodología Bayesiana combina el conocimiento *a priori*, representado a través de la función de distribución inicial; con la información muestral, función de verosimilitud, por medio del *Teorema de Bayes* para obtener un conocimiento final de las probabilidades, representado por una *distribución a posteriori*.

Cuando se supone conocimiento inicial "no informativo", los resultados bayesianos coinciden frecuentemente con las conclusiones numéricas obtenidas a través de procesos clásicos, aunque obviamente, sus interpretaciones son diferentes.

En este trabajo se proporciona un panorama general de las ideas, principios y conceptos detrás de la Teoría de la Decisión Estadística, así como la manera en que la misma funciona. Se puso especial interés en mantener un equilibrio entre la fundamentación teórica y la ilustración mediante ejemplos. Por su importancia en la práctica, se incluyen procedimientos y algoritmos que ayudan a la construcción tanto de la función de utilidad (o pérdida) como de la distribución a priori.

Los conceptos y procedimientos aquí planteados pretenden ser construidos paso a paso, esto es, se intenta que vayan surgiendo naturalmente y que la idea intuitiva sea clara.

El trabajo está pensado para servir de material auxiliar en un curso de Teoría de la Decisión Estadística y básicamente se encuentra estructurado de la siguiente manera.

En el primer capítulo, que podría considerarse como parte de la introducción, se exhiben algunos ejemplos en donde el uso (o quizás abuso) de la estadística clásica ha llevado a resultados poco congruentes. Cabe señalar que esto se hace con la finalidad de motivar la búsqueda de una alternativa estadística más consistente en sus resultados.

La Teoría de la Decisión proporciona un marco de trabajo fundamental para visualizar los problemas estadísticos, es por ello que en el capítulo 2 se da una introducción de la misma. En

particular, se estudia la estructura de un problema de decisión y se plantean diferentes criterios de solución.

El capítulo 3 está dedicado al estudio de los Axiomas de Coherencia, los cuales dan sustento a la Teoría de la Decisión y por consiguiente a la Estadística Bayesiana. Este soporte teórico permite definir formalmente los conceptos de probabilidad y de "utilidad" y una vez que se cuenta con estas definiciones se demuestra que la única solución *"coherente"* es la decisión de Bayes o Utilidad Esperada Mínima.

En el capítulo 4 se plantean los problemas de inferencia como problemas de decisión en ambiente de incertidumbre. Se resalta el hecho de que su planteamiento resulta natural en el marco de la Teoría de la Decisión, por lo cual su solución es teóricamente sencilla. Como parte de este capítulo también se analizan algunos métodos para la asignación de la distribución inicial.

Finalmente, ya con todas las herramientas disponibles, en el último capítulo se analizan problemas típicos de inferencia estadística desde el enfoque bayesiano. Se ejemplifica con casos particulares y cuando es posible se dan algunos resultados generales.

CAPÍTULO 1

ALGUNAS INCONSISTENCIAS EN LA ESTADÍSTICA CLÁSICA

La Estadística Clásica ha sido intensamente estudiada y utilizada, logrando satisfacer en muchos de los casos las expectativas de los investigadores. Sin embargo, debido a que la Teoría clásica o frecuentista no cuenta con una base axiomática, algunos de sus resultados no son necesariamente consistentes. La estadística Bayesiana, al fundamentarse en los llamados Axiomas de Coherencia no cae en tales inconsistencias.

Con la intención de ilustrar esto y al mismo tiempo de motivar el interés hacia una alternativa estadística, se presentan algunos ejemplos de problemas en donde se muestra que el uso indiscriminado de la Estadística Clásica desencadena en soluciones contradictorias o poco satisfactorias.

E1. Sea X una variable aleatoria con distribución Poisson(λ), se desea encontrar un estimador insesgado de $P[X=0]$.

El problema se reduce a estimar $e^{-\lambda}$, pues $P[X=0]=\frac{e^{-\lambda}\lambda^0}{0!}=e^{-\lambda}$. De aquí que se requiera buscar $h(X)$ tal que $E[h(X)]=e^{-\lambda}$.

Se tiene que $E[h(X)]=\sum_{x=0}^{\infty}h(x)\frac{e^{-\lambda}\lambda^x}{x!}=e^{-\lambda}$ si y sólo si

$$\sum_{x=0}^{\infty}h(x)\frac{\lambda^x}{x!}=1.$$

De la última igualdad se sigue que $\sum_{x=0}^{\infty}h(x)\frac{\lambda^x}{x!}$ es el Polinomio de Taylor de la constante 1, y por la unicidad de los coeficientes se concluye que

$$h(x)=\begin{cases}1 & si\ x=0\\ 0 & eoc\end{cases}$$

lo cual muestra la unicidad del estimador insesgado.

Es evidente que estimar $P[X=0]$ con 1 ó 0 no resulta congruente, ya que $e^{-\lambda}$ no puede tomar estos valores. Esto debido a que $e^{-\lambda}=1$ si y sólo si $\lambda=0$, pero por la definición de distribución Poisson, $\lambda>0$; y por otro lado, $e^{-\lambda}$ siempre será mayor que 0 para cualquier valor de λ.

E2. Sean X y Y variables aleatorias independientes con función de distribución Poisson(λ), sea Z otra variable aleatoria con la misma distribución de X y Y. Se desea estimar insesgadamente $P[X=0, Y=0]=e^{-2\lambda}$ con base en Z.

Se requiere encontrar $h(Z)$, tal que $E[h(Z)]=e^{-2\lambda}$; donde $h(z)>0$ para toda Z.

Por otro lado se tiene que $E[h(z)]=\sum_{z=0}^{\infty}h(z)\frac{e^{-\lambda}\lambda^z}{z!}$, por lo que $E[h(z)]=e^{-2\lambda}$ si y sólo si $\sum_{z=0}^{\infty}\frac{h(z)\lambda^z}{z!}=e^{-\lambda}$. Resulta evidente que la suma corresponde al polinomio de Taylor de $e^{-\lambda}$ y por la unicidad en los coeficientes, se tiene que

$$h(z)=\begin{cases}1 & si\ z\ es\ par\\ -1 & si\ z\ es\ impar.\end{cases}$$

Este resultado que arroja la estadística clásica presenta gran incongruencia con la definición de probabilidad, ya que $h(z)$ estima una probabilidad y por lo tanto no puede tomar valores negativos.

E3. (Berger, 1985). Una muestra aleatoria $\underline{X}=(X_1, X_2, ..., X_n)$ independiente e idénticamente distribuida (i.i.d.) va a ser tomada de una N(0,1) con la finalidad de realizar la prueba de hipótesis

$$H_0: \theta=0 \quad \text{versus} \quad H_1: \theta\neq 0$$

con un nivel de significancia $\alpha=0.05$.

Por el cociente de verosimilitudes generalizado (ver Mood, Graybill y Boes, 1974) se obtiene que la región de rechazo es

$$\mathcal{C}=\left\{(x_1, x_2, ..., x_n)=\underline{x}\ \middle|\ \sqrt{n}|\bar{x}|>1.96\right\}.$$

Es poco razonable pensar que la hipótesis nula, H_0, sea verdadera; es decir, que $\theta = 0$. Si $\theta = 10^{-10}$, aunque no es estrictamente cero representa una diferencia poco significativa de cero, en la mayoría de los casos prácticos. Si se toma un tamaño de muestra muy grande, por decir algo, $n = 10^{24}$, entonces con una probabilidad alta, $\bar{x}$ diferirá en 10^{-11} de la verdadera media $\theta = 10^{-10}$; ya que $\sigma_{\bar{X}} = \sqrt{\mathrm{Var}\,\frac{\sum_{i=1}^{10^{24}} X_i}{10^{24}}} = \sqrt{\frac{1}{(10^{24})^2}\sum_{i=1}^{10^{24}}(\mathrm{Var}\,x_i)} = 10^{-12}$. Pero para $\bar{x}$ en esta región, es claro que $10^{12}|\bar{x}| > 1.96$. De aquí que la prueba de hipótesis clásica rechazará H_0, aún cuando la verdadera media difiera mínimamente de cero.

Este mismo fenómeno se presenta sin importar qué tamaño de $\alpha > 0$ se elija, y sin importar tampoco, qué tan pequeña sea la diferencia $\varepsilon > 0$ entre cero y la verdadera media, pues un tamaño de muestra suficientemente grande hará que H_0 sea rechazada.

Es sabido desde el principio que la hipótesis nula de tipo simple casi nunca es verdadera, y esto siempre es confirmado por una muestra suficientemente grande. Sin embargo, el dilema podría estar en la manera de formular el problema, pues muchas veces podría resultar de mayor utilidad averiguar si el parámetro de interés se encuentra dentro de algún intervalo, en lugar de determinar si éste toma exactamente algún valor específico.

Por ejemplo, del ejercicio anterior se podría estar interesado en detectar una diferencia de al menos 10^{-3}; en tal caso, una hipótesis nula más adecuada sería $|\theta| \leq 10^{-3}$.

Hay ciertas situaciones en las cuales es razonable formular el problema como una prueba de hipótesis simple, pero aún así surgen serias preguntas concernientes a la "precisión" de la prueba clásica.

E4. Sea $\underline{X} = (X_1, X_2, ..., X_n)$ una muestra aleatoria donde X_i se distribuye Bernoulli con parámetro desconocido θ, ($Bllí(0)$). Se desea contrastar el siguiente conjunto de hipótesis,

$$H_0: \theta \leq \theta_0, \quad H_1: \theta_0 < \theta \leq \theta_1, \quad H_2: \theta > \theta_1; \quad \text{con } \theta_0 \leq \theta_1.$$

Intentar resolver este ejemplo como un problema clásico de prueba de hipótesis trae una serie de indefiniciones e interrogantes. Inmediatamente surge la pregunta de cuál es la hipótesis nula y cuál la alternativa.

En el caso en el que $\theta_0 = \theta_1$, se tiene un planteamiento de contraste de hipótesis conocido,

$$H_0: \theta \leq \theta_0 \text{ versus } H_1: \theta > \theta_0 .$$

Utilizando el cociente de verosimilitudes generalizadas se obtiene una región de rechazo, $\mathcal{C}^*$, para la hipótesis nula, de la forma

$$\mathcal{C}^* = \{(x_1, x_2, ..., x_n) = \underline{x} \mid \bar{x} > k\}, \text{ donde } P[\bar{X} > k \mid \theta = \theta_0] = \alpha ;$$

en donde $\mathcal{C}^*$ define una prueba uniformemente más potente (*p.u.m.p.*) de tamaño α (ver Mood, Graybill y Boes, 1974).

En el problema original parece lógico y natural determinar una estrategia en la que se acepte H_0 si $\bar{x} \leq k_0$, H_1 si $k_0 < \bar{x} \leq k_1$ y H_2 si $\bar{x} > k_1$. Sin embargo la pregunta sería entonces cómo determinar k_0 y k_1. Una forma de hacerlo sería calcular k_0 y k_1 tales que $P[\bar{x} > k_0 \mid \theta = \theta_0] = \alpha$ y $P[\bar{x} > k_1 \mid \theta = \theta_1] = \alpha$. Sin embargo ¿es éste un procedimiento "*óptimo*"?

Es muy importante hacer notar que la hipótesis nula y la potencia quedan indefinidas y como desde la perspectiva clásica se requiere establecerlas, las opciones para ello serían.

1) H_0 versus $H_1 \cup H_2$

2) $H_0 \cup H_1$ versus H_2.

Sin embargo, ningún par de hipótesis es equivalente a lo que se quiere contrastar, puesto que con ellas el planteamiento se reduce a

1) $\theta \leq \theta_0$ versus $\theta > \theta_0$

2) $\theta \leq \theta_1$ versus $\theta > \theta_1$.

De los resultados anteriores queda claro que $\mathcal{C}_1 = \{\underline{X} \mid \bar{x} > k_0\}$ define una *p.u.m.p.* para el primer caso y $\mathcal{C}_2 = \{\underline{X} \mid \bar{x} > k_1\}$ define una *p.u.m.p.* para el segundo. Pero aceptar

$$\begin{array}{lll} H_0 & \text{si y sólo si} & \bar{x} \le k_0 \\ H_1 & \text{si y sólo si} & k_0 < \bar{x} \le k_1 \text{ y} \\ H_2 & \text{si y sólo si} & \bar{x} > k_1. \end{array}$$

no define un procedimiento óptimo.

E5. (Problema de Fieller) Sea $\underline{X} = (X_1, X_2, ..., X_n)$ una muestra aleatoria i.i.d. de una $N(\mu_1, 1)$ y $\underline{Y} = (Y_1, Y_2, ..., Y_n)$ otra muestra i.i.d. de una $N(\mu_2, 1)$, además son independientes entre sí. Encontrar un intervalo de confianza (I_ρ) al $(1-\alpha)\times 100\%$ para $\rho = \dfrac{\mu_1}{\mu_2}$.

Antes que nada es deseable verificar que el cociente está bien definido; por lo tanto es necesario realizar el contraste de hipótesis

$$H_0: \mu_2 = 0, \text{ versus } H_1: \mu_2 \neq 0.$$

Mediante el cociente de verosimilitudes generalizadas se obtiene que H_0 se rechaza a nivel α' si y sólo si

$$n\bar{Y}^2 \ge \chi^{2,\,1-\alpha'}_{(1)}.$$

Ahora bien, para encontrar el intervalo de confianza para ρ (ver Mood, Graybill y Boes, 1974) se propone la pivotal $Z = \bar{X} - \rho\bar{Y}$. Es claro notar que $Z \sim N\left(0, \dfrac{1+\rho^2}{n}\right)$; por lo que $\dfrac{Z}{\sqrt{\frac{1+\rho^2}{n}}} \sim N(0,1)$ y finalmente $\dfrac{nZ^2}{1+\rho^2} \sim \chi^2_{(1)}$. De ahí que

$$P\left[\frac{n(\bar{X} - \rho\bar{Y})^2}{1+\rho^2} \le \chi^{2,\,1-\alpha}_{(1)}\right] = 1-\alpha.$$

De la expresión anterior se tiene que el intervalo de confianza está formado por los valores de ρ que cumplen con

$$f(\rho)=\frac{n\left(\bar{X}-\rho\bar{Y}\right)^2}{1+\rho^2}\leq k\text{, donde } k=\chi_{(1)}^{2,\,1-\alpha}.$$

Es importante verificar que la desigualdad $f(\rho)\leq\chi_{(1)}^{2,\,1-\alpha}$ determina realmente un intervalo.

Desarrollando tal desigualdad se obtiene

$$\rho^2(n\bar{Y}^2-k)-\rho(2n\bar{X}\bar{Y})+n\bar{X}^2-k\leq 0.$$

El conjunto de valores de ρ que cumplen con la desigualdad es en cada caso:

a. (ρ_1,ρ_2) si $n\bar{Y}^2-k>0$, ya que esto define una parábola cóncava hacia arriba.
b. $(-\infty,\rho_1)\bigcup(\rho_2,\infty)$ si $n\bar{Y}^2-k<0$, ya que esto define una parábola cóncava hacia abajo.
c. $(-\infty,\rho_1)$ ó (ρ_2,∞) si $n\bar{Y}^2-k=0$. Este caso queda prácticamente descartado pues $\bar{Y}^2=\frac{\chi_{(1)}^{2(1-\alpha)}}{n}$ con probabilidad cero.

Es claro, por lo tanto, que utilizando la cantidad pivotal $f(\rho)$ sólo se podría definir un intervalo de confianza en el caso en que $n\bar{Y}^2-k>0$, que es además la condición para tener evidencia de que $\mu_2\neq 0$ $(n\bar{Y}^2\geq\chi_{(1)}^{2(1-\alpha')}=k)$. Para determinar si existen raíces reales de la ecuación $\rho^2(n\bar{Y}^2-k)-\rho(2n\bar{X}\bar{Y})+n\bar{X}^2-k=0$ veamos el signo del discriminante.

$$\begin{aligned}\Delta&=4n^2\bar{X}^2\bar{Y}^2-4(n\bar{Y}^2-k)(n\bar{X}^2-k)\\&=4n\bar{Y}^2k+4n\bar{X}^2k-4k^2\\&=4k(n\bar{Y}^2+n\bar{X}^2-k)\\&=4k\left[(n\bar{Y}^2-k)+n\bar{X}^2\right]\end{aligned}$$

Suponiendo nuevamente que $n\bar{Y}^2-k>0$ se tiene que $\Delta>0$ y por lo tanto la ecuación tiene las raíces reales ρ_1 y ρ_2.

Es importante resaltar que en este problema existe una relación estrecha entre una prueba de hipótesis y un intervalo de confianza, pues se tendrá un intervalo de confianza para ρ única y exclusivamente en el caso en el que se tenga evidencia de que ρ está bien definido, lo cual pasa cuando $n\bar{Y}^2 \geq \chi^{2(1-\alpha')}_{(1)}$. De la función, $f(\rho)$, que determina el intervalo de confianza, se observan las siguientes características

i) $f(\rho)\geq 0$ para todo $\rho \in \mathbb{R}$, y además $f(\rho)=0$ si y sólo si $\rho = \frac{\bar{X}}{\bar{Y}}$.

ii) $f(\rho)$ es continua, entonces alcanza su máximo y su mínimo para cualquier intervalo cerrado.

iii) $\lim_{|\rho|\to\infty} f(\rho) = \lim_{|\rho|\to\infty} \frac{n(\bar{X}-\rho\bar{Y})^2}{1+\rho^2} = n\bar{Y}^2$ (que puede identificarse con la estadística de prueba para la hipótesis $H_0: \mu_2 = 0$).

iv) por i), ii) y iii) se tiene que $f(\rho)$ es acotada, lo cual indica que existe un $M \in \mathbb{R}^+$ tal que $f(\rho)\leq M$ para todo $\rho \in \mathbb{R}$.

Se puede probar que $\rho = \frac{\bar{X}}{\bar{Y}}$ y $\rho = -\frac{\bar{Y}}{\bar{X}}$ son un mínimo y un máximo global de $f(\rho)$, con $f\left(\frac{\bar{X}}{\bar{Y}}\right)=0$ y $f\left(-\frac{\bar{Y}}{\bar{X}}\right) = n(\bar{X}^2+\bar{Y}^2) = M$. Como M es positivo, puede ser pensado como un cuantil $1-\alpha$ de la $\chi^2_{(1)}$. Por lo tanto, se sigue que

$$I_\rho = \left\{\rho \in \mathbb{R} \middle| f(\rho)\leq M\right\} = \mathbb{R}.$$

De la última igualdad se concluye que existe un valor de α estrictamente positivo tal que la correspondiente región para ρ (con un nivel de confianza estrictamente menor que 1) coincide con toda la recta real. Este resultado es considerado por algunos como una descalificación del procedimiento clásico (Mendoza, 1988).

Cabe mencionar que el procedimiento de Fieller para el problema de cociente de medias ha sido criticado muy duramente ya que parece inadmisible que R sea un intervalo de confianza menor al 100%, sin embargo, el resultado tiene cierta justificación.

Con el valor de α que se cumple que el intervalo de confianza es la recta real no se tiene evidencia de que ρ esté bien definido, ya que si $n(\bar{X}^2+\bar{Y}^2)\leq \chi^{2,\,1-\alpha}_{(1)}$ en general no se cumple que $\bar{Y}^2 \geq \frac{\chi^{2,\,1-\alpha}_{(1)}}{n}$. Esto viene a ser una justificación a que R sea un intervalo con una confianza menor al 100%, ya que en tal caso se tiene que el cociente no está bien definido y por lo tanto no es un valor real.

CAPÍTULO 2

TEORÍA DE LA DECISIÓN

La Estadística Bayesiana se caracteriza por abordar los problemas estadísticos como problemas de decisión, es por ello que resulta necesario brindar un panorama general de la Teoría de la Decisión. Tal teoría describe el proceso lógico que debe seguir un decisor para elegir "razonablemente" la mejor forma de actuar.

En este capítulo se estudiará la estructura de un problema de decisión, así como los diferentes criterios para solucionarlo. El mayor énfasis del trabajo estará puesto en aquellos problemas que se plantean en un ambiente de incertidumbre.

2.1 ESTRUCTURA DE UN PROBLEMA DE DECISIÓN

Se dice que alguien se encuentra ante un problema de decisión cuando debe elegir entre dos o más formas de actuar. Los problemas de esta naturaleza son muy comunes en la vida cotidiana pues continuamente se debe decidir entre tomar una u otra decisión.

En este tipo de problemas se identifican inmediatamente dos elementos, uno que es el conjunto de decisiones y otro que es el conjunto de posibles consecuencias resultado de haber elegido determinada opción. Por lo tanto, dos conjuntos que intervienen en un problema de decisión son:

1. *Conjunto de todas las formas posibles de actuar o espacio de decisiones*, representado por D; donde $D = \{d_1, d_2, ..., d_n, ...\}$ y d_i denota a la decisión i.

Este conjunto contiene las alternativas para decidir sobre la forma de actuar, es por esto que debe ser un conjunto exhaustivo en donde los elementos sean mutuamente excluyentes. Es decir, este conjunto debe estar construido de tal manera que el problema se reduzca a elegir de manera óptima para el decisor, uno y sólo uno de sus elementos.

2. *Conjunto o espacio de consecuencias*, representado por C donde $C = \{c_1, c_2, ..., c_n, ...\}$ y cada c_i representa la posible consecuencia de tomar alguna decisión.

Como se verá más adelante, para simplificar el desarrollo teórico en la solución de problemas de decisión, se supondrá en muchos de los casos, que tanto el conjunto de alternativas, D, como el de consecuencias, C, son finitos.

Es natural pensar que para el decisor debe existir un orden de preferencias entre las posibles consecuencias, dicho orden puede cambiar de decisor en decisor. Esto da lugar a un elemento más en los problemas de decisión llamado *Relación de preferencia u Orden de preferencia.*

3. (<) *Relación de preferencia* entre las consecuencias dependiendo del decisor.

$c_i < c_j$ indica que c_i es menos preferible que c_j,

$c_i \leq c_j$ indica que c_i es menos o igualmente preferible a c_j,

y análogamente se definen

$c_i > c_j$ y $c_i \geq c_j$.

Es posible que las consecuencias de tomar una u otra decisión no estén determinadas, sino que estén sujetas a la aparición de sucesos inciertos, los cuales generan un ambiente de incertidumbre. La presencia o no de dicho ambiente origina la existencia de dos tipos de problemas de decisión:

- Problemas de decisión sin incertidumbre
- Problemas de decisión con incertidumbre.

Es importante hacer notar que la diferencia en estructura entre estas clases de problemas radica en que el segundo cuenta con un elemento más llamado "*espacio de sucesos inciertos*".

2.2 PROBLEMAS DE DECISIÓN SIN INCERTIDUMBRE

Un problema de decisión sin incertidumbre es aquél en el que se sabe con certeza lo que va a pasar como consecuencia de tomar cada decisión, es decir, las consecuencias no están sujetas a la aparición de sucesos inciertos.

Los elementos de un problema de decisión sin incertidumbre (suponiendo finitud) son los planteados en un problema de decisión general. Esto es, el espacio de decisiones, representado por $D = \{d_1, d_2, ..., d_n\}$; el espacio de consecuencias, representado por $C = \{c_1, c_2, ... c_n\}$ y la relación de preferencia entre las consecuencias ($<$).

El problema, en este caso, se reducirá a determinar la decisión d^* del espacio de decisiones que produzca para el decisor la mejor consecuencia. Esto es, la solución será d^* tal que:

$c^* \geq c_i$ para toda i ,

donde c^* es la consecuencia de haber elegido d^*.

La estructura de un problema de decisión de este tipo, puede ser presentada esquemáticamente a través de un diagrama llamado *Árbol de Decisión*. Dicho diagrama iniciará con un *nodo de decisión* representado por un cuadro, a través del cual se desprenden ramas donde cada *rama* representa una decisión, y cada consecuencia está asociada a una *hoja del árbol*. Ver figura 2.2.1.

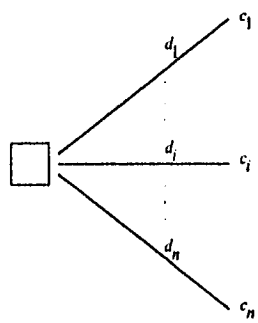

Figura 2.2.1

El proceso de solución visto a través del diagrama de árbol consistirá en *podar* el árbol para quedarse con una sola de sus ramas, aquélla de la que cuelgue la hoja con la mejor consecuencia.

Ejemplo 2.2.1: Un inversionista en la Bolsa de Valores se encuentra ante el problema de tomar una decisión de entre las tres siguientes:

d_1 = No invertir,
d_2 = Invertir en Telmex,
d_3 = Invertir en Celanese.

Con certeza sabe que las opciones le producirán las siguientes ganancias:

Decisión	Ganancia
d_1	$ 0
d_2	$ 100
d_3	$ 85

Es evidente que la mejor decisión es invertir en Telmex, ya que ésta es la inversión que produce la mayor ganancia (Ver figura 2.2.2).

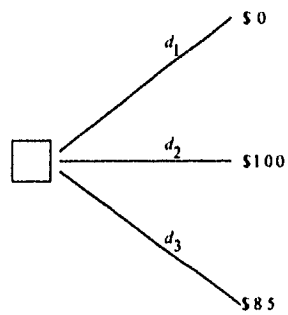

Figura 2.2.2

Existen problemas sin incertidumbre cuya solución no es tan trivial como lo planteado en el ejemplo 2.2.1. Sin embargo, la posible complejidad del problema se deberá a cuestiones técnicas de cálculo o a dificultades en establecer preferencias. Ejemplo de ello es la elección de la estrategia óptima en una partida de ajedrez. Las jugadas permisibles pueden ser en teoría listadas, al igual que las consecuencias de cada una de ellas. Sin embargo, en la práctica esto resulta difícil ya que las posibles combinaciones son casi innumerables.

2.3 PROBLEMAS DE DECISIÓN CON INCERTIDUMBRE

Un problema de decisión con incertidumbre es aquél que se plantea en un ambiente de desconocimiento acerca de lo que sucederá según se actúe de acuerdo a una u otra forma. Es decir, existe desconocimiento de la consecuencia resultante de tomar cierta decisión, ya que las consecuencias estarán condicionadas a la ocurrencia de determinados sucesos inciertos.

Los elementos de un problema de decisión con incertidumbre son los que conforman un problema de decisión general $(D,C,<)$, más el conjunto de todos los sucesos inciertos (Ω). Por simplicidad, se supondrá que todos estos conjuntos son finitos. De tal forma que

- $D = \{d_1,d_2,...,d_n\}$, es el espacio de decisiones;
- $\Omega = \{\varpi_1,\varpi_2,...,\varpi_n\}$, es el espacio de sucesos inciertos, donde $\varpi_i = \{\omega_{i1},\omega_{i2},...,\omega_{im_i}\}$ y ω_{ij} es igual al suceso incierto j asociado a la decisión i ;
- $C = \{c_1,c_2,...c_n\}$, es el espacio de consecuencias, donde $c_i = \{c_{i1},c_{i2},...,c_{im_i}\}$ y c_{ij} la consecuencia resultado de tomar la decisión i y haber ocurrido el suceso incierto j;
- $(<)$ es el orden de preferencia entre las consecuencias.

Al igual que en los problemas de decisión sin incertidumbre se pueden plantear esquemáticamente los problemas a través de un árbol de decisión. El diagrama inicia con un nodo de decisión (representado por un cuadro), las ramas que se originan de él representan las posibles decisiones y conducen a un *nodo de incertidumbre* o *nodo aleatorio* (representado por un círculo). De cada uno de estos nodos salen ramas que representan a los sucesos ω_{ij}. De la rama ω_{ij} colgará la consecuencia c_{ij}. Ver figura 2.3.1.

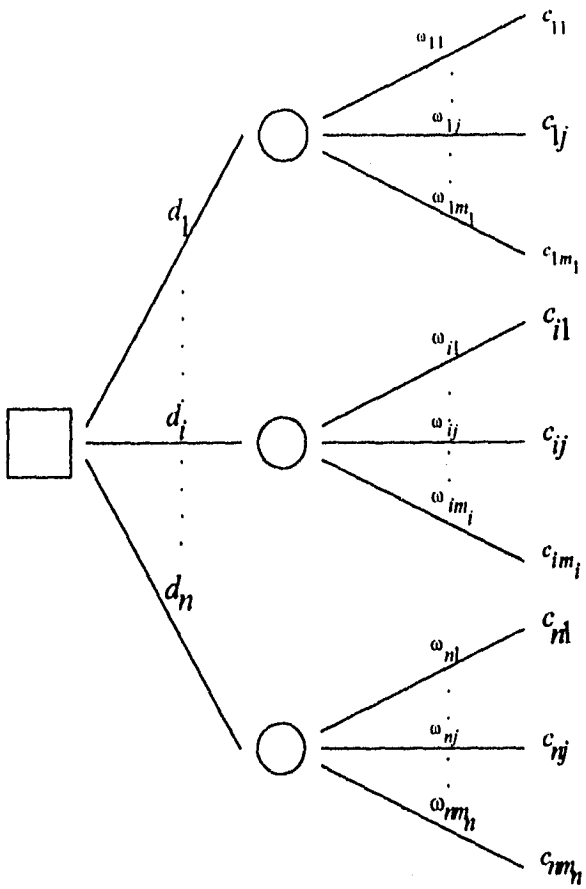

Figura 2.3.1.

Resolver el problema a través del diagrama del árbol, consiste en *podar sus ramas* buscando eliminar de alguna manera *racional* la incertidumbre.

Para estudiar la solución de un problema de decisión con incertidumbre bastará considerar aquellos casos en los que los eventos inciertos son comunes para cada $d_i \in D$, ya que los problemas que no cumplan esta condición pueden ser reestructurados considerando uniones de eventos. De esta manera, el árbol de la figura 2.3.1 queda como el de la figura 2.3.2.

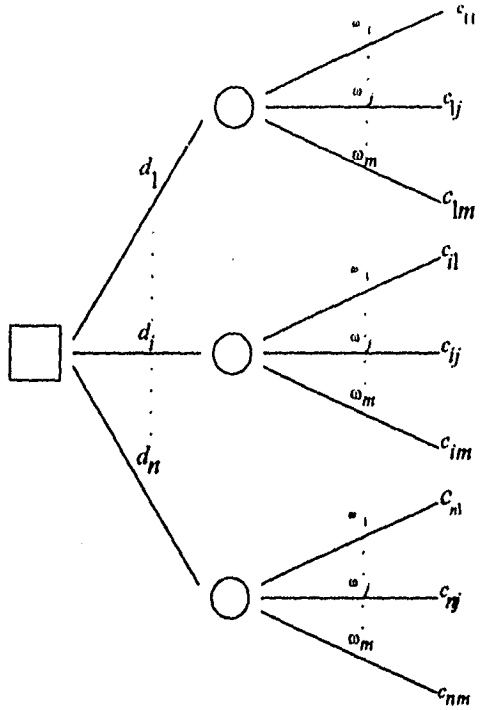

Figura 2.3.2

La simplificación anterior permitirá representar el problema de decisión en términos de una tabla como la siguiente.

		Sucesos inciertos				
		ω_1	ω_2	ω_3	.....	ω_m
Decisiones	d_1	c_{11}	c_{12}	c_{13}	.....	c_{1m}
	d_2	c_{21}	c_{22}	c_{23}		c_{2m}
	d_3	c_{31}	c_{32}	c_{33}		c_{3m}
	.	.	.	.		.
	.	.	.	.		.
	.	.	.	.		.
	d_n	c_{n1}	c_{n2}	c_{n3}	.....	c_{nm}

Tabla 2.3.1

De igual forma que en el problema de decisión sin incertidumbre, puede suceder que las consecuencias no sean numéricas, aunque en algunos casos resultará posible evaluarlas en términos numéricos reflejando las preferencias del decisor. Un concepto que permite efectuar esta cuantificación del orden de preferencias es el de *función de utilidad*.

Para cuantificar las preferencias que sobre las consecuencias tiene el decisor existe una *función de utilidad* $u: C \to \mathbf{R}$, tal que

$$c_1 < c_2 \text{ si y sólo si } u(c_1) < u(c_2),$$

donde, abusando de la notación, $<$ denota la relación de orden en C y en $\mathbf{R}$.

Es muy natural pensar que algunos decisores se inclinen por expresar sus preferencias en términos de pérdida y no en términos de utilidad, por lo cual para cuantificar preferencia existe también una *función de pérdida* $l: C \to \mathbf{R}$, que satisface lo siguiente:

$$c_1 < c_2 \text{ si y sólo si } l(c_1) > l(c_2).$$

Ciertamente, si l es una función de pérdida $u = -l$ define una función de utilidad y viceversa, si u es una función de utilidad $l = -u$ define una función de pérdida.

La función de utilidad está definida no sólo sobre las consecuencias, sino también sobre el producto cartesiano $D \times \Omega$, ya que las consecuencias son producto de la conjunción de tomar una decisión y de que ocurra cierto evento. Por lo que U: $D \times \Omega \to \mathbf{R}$ es una función tal que para cada $d \in D$ y $\omega \in \Omega$, $U(d,\omega)$ denota la utilidad que se obtiene al tomar la decisión d cuando ocurre el suceso ω.

En el caso de las pérdidas ocurre algo completamente análogo al de las utilidades, esto es, se define L: $D \times \Omega \to \mathbf{R}$, tal que para cada $d \in D$ y $\omega \in \Omega$, $L(d,\omega)$ denota la pérdida que se obtiene al tomar la decisión d cuando ocurre el suceso ω.

Si se tiene una función de utilidad asignada, la tabla 2.3.1 se puede expresar como la mostrada en 2.3.2 a través del uso de la función U.

Sucesos inciertos

Decisiones		ω_1	ω_2	ω_3	.....	ω_m
	d_1	U_{11}	U_{12}	U_{13}	.....	U_{1m}
	d_2	U_{21}	U_{22}	U_{23}		U_{2m}
	d_3	U_{31}	U_{32}	U_{33}		U_{3m}
	.	.	.	.		.
	.	.	.	.		.
	.	.	.	.		.
	d_n	U_{n1}	U_{n2}	U_{n3}	.....	U_{nm}

Tabla 2.3.2

donde $U_{ij} = U(d_i, \omega_j)$ para $i = 1,...,n$ y $j = 1,...,m$.

Ejemplo 2.3.1: Suponga que el problema del ejemplo 2.2.1 está sujeto necesariamente a la aparición de uno de estos tres eventos: que las acciones se vayan a la alza (con $P(\omega_1) = 0.4$), permanezcan igual (con $P(\omega_2) = 0.3$) o se vayan a la baja (con $P(\omega_3) = 0.3$). Las consecuencias se muestran en la Tabla 2.3.3. Plantear el árbol de decisión correspondiente.

	ω_1	ω_2	ω_3
d_1	0	0	0
d_2	100	0	-200
d_3	200	0	-500

Tabla 2.3.3

con

d_1 = No invertir, ω_1 = Las acciones se van a la alza

d_2 = Invertir en Telmex, ω_2 = Las acciones permanecen igual

d_3 = Invertir en Celanese. ω_3 = Las acciones se van a la baja

El árbol de decisión de este problema se muestra en la figura 2.3.3.

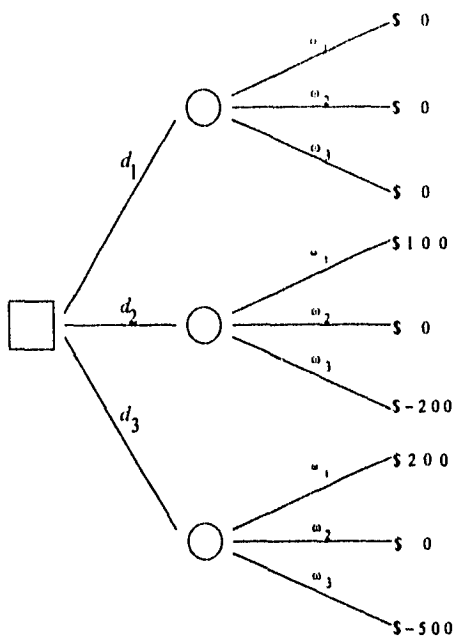

Figura 2.3.3

2.4 CRITERIOS DE SOLUCIÓN DE UN PROBLEMA DE DECISIÓN

Ahora que se conoce la estructura de un problema de decisión planteado en ambiente de incertidumbre, el siguiente paso es determinar una solución satisfactoria. El problema fundamental radica en que no se puede elegir la decisión óptima directamente, debido a que las consecuencias de cada decisión están sujetas a la aparición de sucesos inciertos. Entonces, un paso fundamental en la solución de problemas de decisión en ambiente de incertidumbre consiste en eliminar la incertidumbre de alguna manera razonable.

Existen diferentes criterios para la solución de un problema de decisión con incertidumbre, la diferencia entre ellos se presenta en la manera en que los diferentes procesos eliminan la incertidumbre. Los criterios que se estudiarán en esta sección son: *Maximin (Minimax), Maxmax (Minmin), Consecuencia más probable* y el criterio de la *Utilidad (pérdida) esperada máxima (mínima)*. En el capítulo siguiente (sección 3.4) se demuestra que el único criterio consistente con la base axiomática de la Teoría de la Decisión es el de la Utilidad (pérdida) esperada máxima (mínima)

2.4.1 CRITERIO PESIMISTA (MAXIMIN, MINIMAX)

Consiste en eliminar la incertidumbre en un problema de decisión suponiendo que cada decisión va a llevar a la peor consecuencia. Hecho esto, se escoge la decisión que produzca la consecuencia más favorable, lo anterior, con la finalidad de maximizar la utilidad garantizada.

Cuando se están evaluando las consecuencias con una función de utilidad, el criterio pesimista se conoce como MAXIMIN y para el caso de una función de pérdida el criterio recibe el nombre de MINIMAX.

CRITERIO MAXIMIN

Asignada una función de *utilidad* al problema de decisión, el modo de actuar "*pesimista*" conduce al criterio *MAXIMIN*. Este criterio consiste en minimizar la utilidad para cada decisión fija y posteriormente buscar la decisión que produzca la máxima utilidad mínima. Esto es, d^* es Maximin con valor Maximin igual a $U^*(d^*)$, si y sólo si

$$U^*(d^*) = \underset{d \in D}{Max} U^*(d), \text{ donde } U^*(d) = \underset{\omega \in \Omega}{Min} U(d,\omega).$$

Lo anterior se visualiza en la tabla 2.4.1.1.

Sucesos inciertos

		ω_1	ω_2	ω_3	.....	ω_m	
	d_1	U_{11}	U_{12}	U_{13}	.....	U_{1m}	$U_1^* = \underset{j=1,2,..,m}{Min} U_{1j}$
	d_2	U_{21}	U_{22}	U_{23}		U_{2m}	$U_2^* = \underset{j=1,2,..,m}{Min} U_{2j}$
Decisiones	d_3	U_{31}	U_{32}	U_{33}		U_{3m}	$U_3^* = \underset{j=1,2,..,m}{Min} U_{3j}$
	.	.	.	.		.	
	.	.	.	.		.	
	.	.	.	.		.	
	d_n	U_{n1}	U_{n2}	U_{n3}	.....	U_{nm}	$U_n^* = \underset{j=1,2,..,m}{Min} U_{nj}$

Tabla 2.4.1.1

d^* es Maximin, si y sólo si $U^* = \underset{i=1,2,..,n}{Max} U_i^*$.

Visto en un diagrama de árbol, lo que se está haciendo es podar las ramas de sucesos inciertos como si fuera a ocurrir lo "peor", esto es, sustituyendo el nodo de incertidumbre por el valor de la peor consecuencia (la utilidad mínima) y posteriormente resolver como un problema de decisión sin incertidumbre (quedarse con aquella decisión cuya utilidad sea máxima).

Al resolver el problema del ejemplo 2.3.1, queda un árbol sin incertidumbre como el de la figura 2.4.1.1, se toma el $Max\{0,-200,-500\}=0$ y se tiene que la decisión es

$d^* = d_1$ = No invertir, con valor Maximin igual a 0.

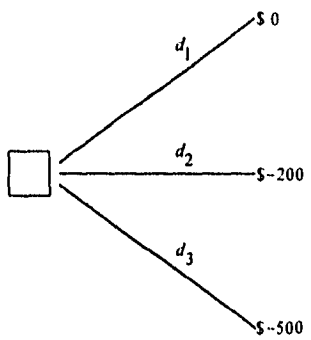

Figura 2.4.1.1

CRITERIO MINIMAX

Asignada una función de *pérdida* al problema de decisión, el modo de actuar *"pesimista"* conduce al criterio *MINIMAX*. Este criterio consiste en maximizar la pérdida para cada decisión fija y posteriormente buscar la decisión que produzca la mínima pérdida máxima. Esto es, d^* es Minimax con valor minimax igual a $L^*(d^*)$ si y sólo si

$$L^*(d^*) = \min_{d \in D} L^*(d), \text{ donde } L^*(d) = \max_{\omega \in \Omega} L(d,\omega).$$

Lo anterior se visualiza en la tabla 2.4.1.2.

Sucesos inciertos

Decisiones		ω_1	ω_2	ω_3	.....	ω_m	
	d_1	L_{11}	L_{12}	L_{13}	.....	L_{1m}	$L_1^* = \underset{j=1,2,...,m}{Max} L_{1j}$
	d_2	L_{21}	L_{22}	L_{23}		L_{2m}	$L_2^* = \underset{j=1,2,...,m}{Max} L_{2j}$
	d_3	L_{31}	L_{32}	L_{33}		L_{3m}	$L_3^* = \underset{j=1,2,...,m}{Max} L_{3j}$
	.	.	.	.		.	
	.	.	.	.		.	
	.	.	.	.		.	
	d_n	L_{n1}	L_{n2}	L_{n3}	.....	L_{nm}	$L_n^* = \underset{j=1,2,...,m}{Max} L_{nj}$

Tabla 2.4.1.2

d^* es minimax, si y sólo si $L^* = \underset{i=1,2,...,n}{Min} L_i^*$.

Visto en un diagrama de árbol, lo que se está haciendo es podar las ramas de sucesos inciertos como si fuera a ocurrir lo "peor", esto es, sustituyendo el nodo de incertidumbre por el valor de la peor consecuencia (la pérdida máxima) y posteriormente resolver como un problema de decisión sin incertidumbre (quedarse con aquella decisión cuya pérdida sea mínima). Al resolver el problema del ejemplo 2.3.1 suponiendo que se asoció una función de pérdida, se tiene un árbol de decisión como el de la figura 2.4.1.2. Por lo que al podar el árbol se obtiene el árbol de decisión de la figura 2.4.1.3.

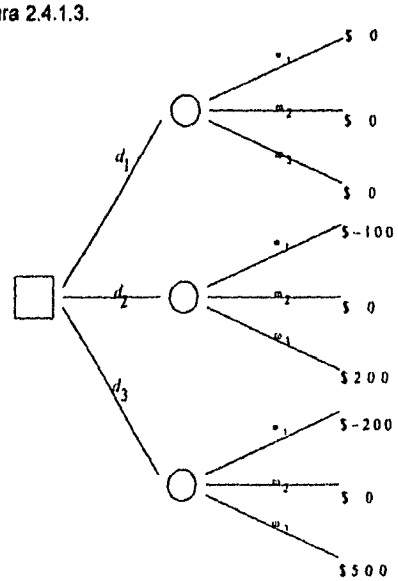

Figura 2.4.1.2

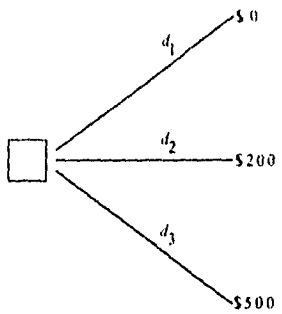

Figura 2.4.1.3

Se toma el $Min\{0,200,500\}=0$ y se tiene que la decisión es

$d^* = d_1$ = No invertir, con valor Minimax igual a 0.

2.4.2 CRITERIO OPTIMISTA (MAXMAX, MINMIN)

Consiste en eliminar la incertidumbre en un problema de decisión suponiendo que cada decisión va a llevar a la mejor consecuencia. Hecho esto, se escoge la decisión que produzca la consecuencia más favorable.

Cuando se están evaluando las consecuencias con una función de utilidad, el criterio optimista se conoce como MAXMAX y para el caso de una función de pérdida el criterio recibe el nombre de MINMIN.

CRITERIO MAXMAX

Asignada una función de *utilidad* al problema de decisión, el modo de actuar "*optimista*" conduce al criterio *MAXMAX*. Este criterio consiste en maximizar la utilidad para cada decisión fija y posteriormente buscar la decisión que produzca la máxima utilidad máxima. Esto es, d^* es Maxmax con valor Maxmax igual a $U^*(d^*)$, si y sólo si

$$U^*(d^*)= \underset{d\in D}{Max} U^*(d), \text{ donde } U^*(d)= \underset{\omega\in\Omega}{Max} U(d,\omega).$$

Lo anterior se visualiza en la tabla 2.4.2.1.

Sucesos inciertos

		ω_1	ω_2	ω_3	.....	ω_m	
	d_1	U_{11}	U_{12}	U_{13}	.....	U_{1m}	$U_1^* = \underset{j=1,2,..,m}{Max}\, U_{1j}$
	d_2	U_{21}	U_{22}	U_{23}		U_{2m}	$U_2^* = \underset{j=1,2,..,m}{Max}\, U_{2j}$
Decisiones	d_3	U_{31}	U_{32}	U_{33}		U_{3m}	$U_3^* = \underset{j=1,2,..,m}{Max}\, U_{3j}$
	.	.	.	.		.	
	.	.	.	.		.	
	.	.	.	.		.	
	d_n	U_{n1}	U_{n2}	U_{n3}	.....	U_{nm}	$U_n^* = \underset{j=1,2,..,m}{Max}\, U_{nj}$

Tabla 2.4.2.1

d^* es Maxmax, si y sólo si $U^* = \underset{i=1,2,..,n}{Max}\, U_i^*$.

En un diagrama de árbol, se necesita podar las ramas de sucesos inciertos como si fuera a ocurrir lo "mejor", esto es, sustituyendo el nodo de incertidumbre por el valor de la mejor consecuencia (la utilidad máxima) y posteriormente resolver como un problema de decisión sin incertidumbre (quedarse con aquella decisión cuya utilidad sea máxima).

Al resolver el problema del ejemplo 2.3.1, queda un árbol sin incertidumbre como el de la figura 2.4.2.1, se toma el $Max\{0,100,200\}=200$ y se tiene que la decisión es

$d^* = d_3$ = Invertir en Celanese, con valor Maxmax igual a 200.

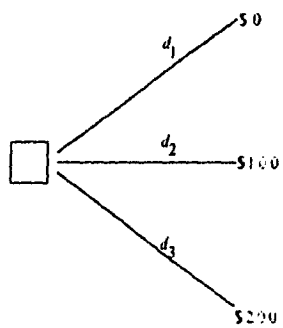

Figura 2.4.2.1

CRITERIO MINMIN

Asignada una función de *pérdida* al problema de decisión, el modo de actuar "*optimista*" conduce al criterio *MINMIN*. Este criterio consiste en minimizar la pérdida para cada decisión fija y posteriormente buscar la decisión que produzca la mínima pérdida mínima. Esto es, d^* es Minmin con valor Minmin igual a $L^*(d^*)$ si y sólo si

$$L^*(d^*) = \underset{d \in D}{Min}\, L^*(d), \text{ donde } L^*(d) = \underset{\omega \in \Omega}{Min}\, L(d,\omega).$$

Lo anterior se visualiza en la tabla 2.4.2.2.

		Sucesos inciertos					
		ω_1	ω_2	ω_3	.....	ω_m	
Decisiones	d_1	L_{11}	L_{12}	L_{13}	.....	L_{1m}	$L_1^* = \underset{j=1,2,..,m}{Min}\, L_{1j}$
	d_2	L_{21}	L_{22}	L_{23}		L_{2m}	$L_2^* = \underset{j=1,2,..,m}{Min}\, L_{2j}$
	d_3	L_{31}	L_{32}	L_{33}		L_{3m}	$L_3^* = \underset{j=1,2,..,m}{Min}\, L_{3j}$
	.	.	.	.		.	
	.	.	.	.		.	
	.	.	.	.		.	
	d_n	L_{n1}	L_{n2}	L_{n3}	.....	L_{nm}	$L_n^* = \underset{j=1,2,..,m}{Min}\, L_{nj}$

Tabla 2.4.2.2

d^* es Minmin, si y sólo si $L^* = \underset{i=1,2,..,n}{Min}\, L_i^*$.

Visto en un diagrama de árbol, lo que se está haciendo es podar las ramas de sucesos inciertos como si fuera a ocurrir lo "mejor", esto es, sustituyendo el nodo de incertidumbre por el valor de la mejor consecuencia (la pérdida mínima) y posteriormente resolver como un problema de decisión sin incertidumbre (quedarnos con aquella decisión cuya pérdida sea mínima). Al resolver el problema del ejemplo 2.3.1, se le debe asociar una función de pérdida, con esto se tiene un árbol de decisión como el de la figura 2.4.1.2. Por lo que al podar el árbol se obtiene el árbol de decisión de la figura 2.4.2.2, se toma el $Min\{0,-100,-200\} = -200$ y se tiene que la decisión es

$d^* = d_3$ = Invertir en Celanese, con valor Minmin igual a -200.

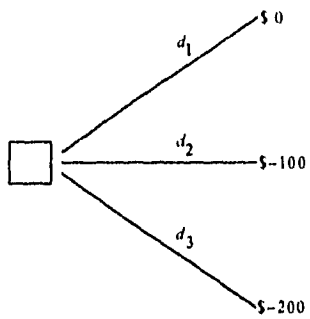

Figura 2.4.2.2

RESUMEN

La manera de solucionar un problema de decisión, ya sea actuando de manera pesimista u optimista se puede sintetizar de la siguiente manera:

Dados $(D,\Omega,C,<)$ se puede actuar de manera "OPTIMISTA" o "PESIMISTA".

Si la relación de preferencias se encuentra expresada a través de una función de utilidad $U(d,\omega)$, se es "PESIMISTA" siempre y cuando se aplique MAXIMIN y se es "OPTIMISTA" si se aplica MAXMAX.

Si la relación de preferencias se encuentra expresada a través de una función de pérdida $L(d,\omega)$, se es "PESIMISTA" siempre y cuando se aplique MINIMAX y se es "OPTIMISTA" si y sólo si se aplica MINMIN.

Los criterios de solución estudiados hasta el momento, se pueden aplicar sin importar cuál sea la cardinalidad de los conjuntos D, Ω y C, basta con que existan los mínimos y máximos respectivos. Además son aplicables aún cuando las consecuencias no sean numéricas.

Estos criterios, tienen una característica muy importante y es el hecho de que *no incorporan probabilidades*, lo cual puede llevar a tomar decisiones poco acertadas. Por ejemplo, en el problema de la Bolsa (con una función de utilidad asociado a él) cuando se actuó de forma optimista, se decidió invertir en Celanese porque de esa manera se podía obtener la utilidad más alta. Sin embargo, la probabilidad de no ganar dicha utilidad es de 0.6,

que es una probabilidad que puede ser considerablemente alta. En general, en un problema de decisión resuelto con el criterio Maxmax, se puede optar por la decisión que reporta la mejor consecuencia aunque la probabilidad de que ocurra el evento que la genera sea pequeña. Situaciones análogas ocurren con los demás criterios vistos en esta sección.

2.4.3 CRITERIO DE LA CONSECUENCIA MÁS PROBABLE

El *criterio de la consecuencia más probable (c.m.p.)* consiste en eliminar la incertidumbre de un problema de decisión suponiendo que cada decisión llevará a la consecuencia que tiene asignada la mayor probabilidad. Hecho esto, se resuelve como un problema de decisión sin incertidumbre, esto es, se elige la decisión que produzca la consecuencia más favorable.

Si se tiene asignada una función de utilidad, $U(d,\omega)$, d^* es la decisión vía la consecuencia más probable con valor de la consecuencia más probable igual a $U^*(d^*)$, si y sólo si

$$U^*(d^*)= \underset{d\in D}{Max}\, U^*(d)$$

donde

$U^*(d)= U(d,\omega_0)$ con $\omega_0 \in \Omega$ tal que $P(\omega_0)\geq P(\omega)$ para todo $\omega \in \Omega$.

Si se tiene asignada una función de pérdida, $L(d,\omega)$, d^* es la decisión vía la consecuencia más probable con valor de la consecuencia más probable igual a $L^*(d^*)$, si y sólo si

$$L^*(d^*)= \underset{d\in D}{Min}\, L^*(d)$$

donde

$L^*(d)= L(d,\omega_0)$ con $\omega_0 \in \Omega$ tal que $P(\omega_0)\geq P(\omega)$ para todo $\omega \in \Omega$.

Visto en un diagrama de árbol, se necesita podar las ramas, de tal manera que se sustituya cada nodo de incertidumbre por el valor de la consecuencia (utilidad o pérdida) más probable. Posteriormente se debe proceder a resolver como un problema sin incertidumbre. Al resolver el problema de la Bolsa, cuyo árbol de decisión aparece en la Figura 2.3.3, se obtiene un árbol podado como el de la figura 2.4.3.1. Por lo que la decisión es invertir en Celanese con valor de la consecuencia más probable igual a 200.

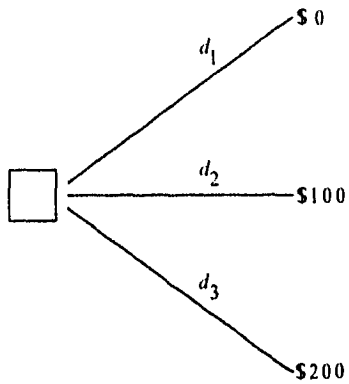

Figura 2.4.3.1

Esta manera de proceder para determinar la solución a un problema de decisión en ambiente de incertidumbre parece ser una buena alternativa, ya que incorpora probabilidades. Sin embargo, existen algunas dificultades que provocan que no quede definida su aplicación o que se llegue a soluciones poco satisfactorias. Ejemplos de esto son:

• Si $\Omega = \mathbb{R}$, la medida de probabilidad sobre Ω seguramente será una función de densidad de probabilidad (f.d.p.g.) $\Pi(\omega)$ definida sobre un σ-álgebra en Ω. En este caso $P(\omega) = 0$ para toda $\omega \in \Omega$ y como consecuencia *no existe el evento más probable*. Sin embargo, en algunas situaciones es posible definir algo parecido al suceso más probable, la moda podría fungir como el evento con mayor verosimilitud.

Por ejemplo, supóngase que se tiene una función de densidad como la de la figura 2.4.3.2. Todos los puntos tienen la misma probabilidad (cero), pero se podría pensar que "ω_0 es el más probable".

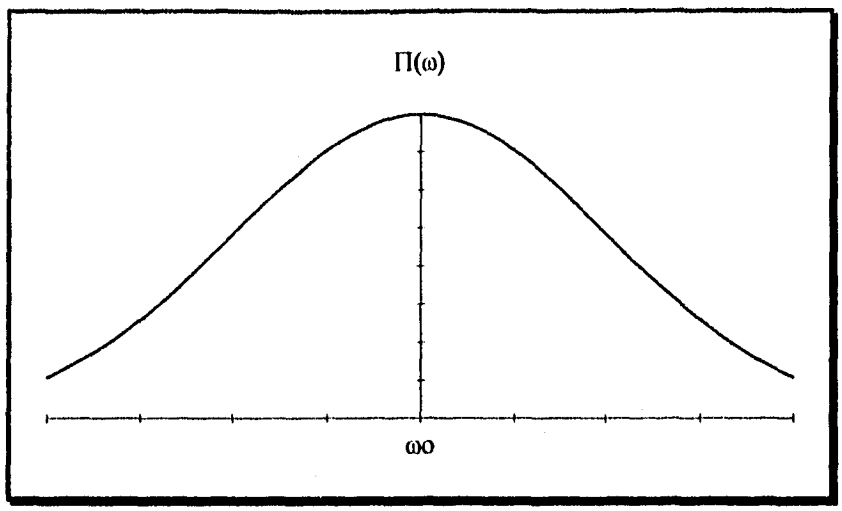

Figura 2.4.3.2

Si la situación se presenta como la de la figura 2.4.3.3 ω_0 y ω_1 se podrían considerar como los *más probables.*

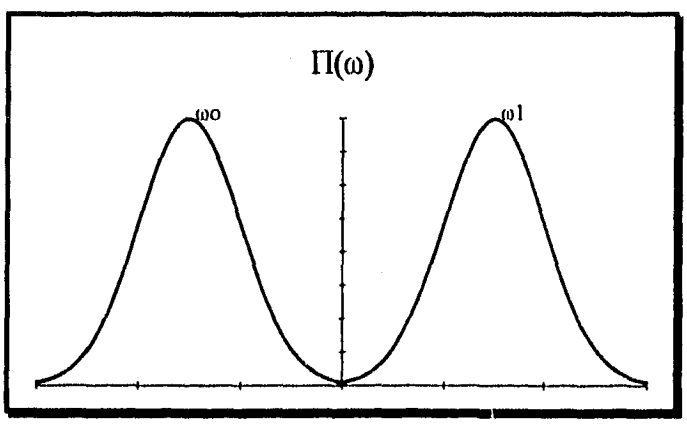

Figura 2.4.3.3

Por supuesto, si en un problema de decisión $\Omega = \{\omega_1, \omega_2, ..., \omega_m\}$ y $P(\omega_1) = P(\omega_2) = ... = P(\omega_m) = \frac{1}{m}$, el criterio de la consecuencia más probable queda indefinido.

• En el problema de la Bolsa supóngase que $P(\omega_1) = 0.3$, $P(\omega_2) = 0.4$ y $P(\omega_3) = 0.3$. El criterio de la consecuencia más probable origina un árbol podado como el de la figura 2.4.3.4. De ahí que todas las decisiones sean igualmente preferibles.

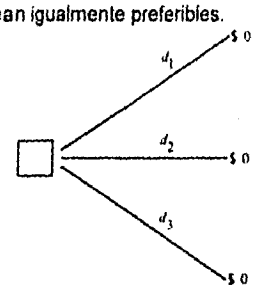

Figura 2.4.3.4

• Si ahora, para el problema de la Bolsa se tiene un árbol como el de la figura 2.4.3.5 y si los eventos de interés tienen probabilidades $P(\omega_1) = 0.3$, $P(\omega_2) = 0.4$ y $P(\omega_3) = 0.3$, el criterio de la consecuencia más probable establece que todas las decisiones son igualmente preferibles. Sin embargo, es claro que d_3 es la "mejor opción".

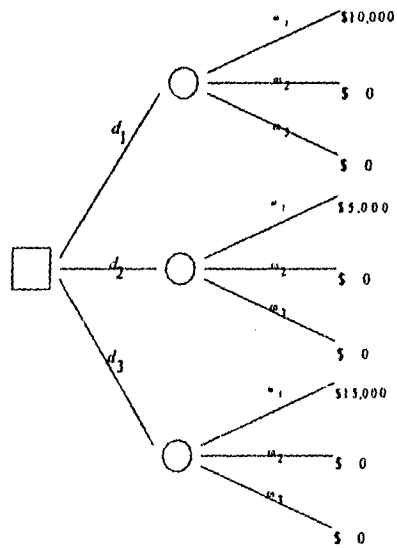

Figura 2.4.3.5

Por supuesto, si en un problema de decisión $\Omega = \{\omega_1, \omega_2, ..., \omega_m\}$ y $P(\omega_1) = P(\omega_2) = ... = P(\omega_m) = \frac{1}{m}$, el criterio de la consecuencia más probable queda indefinido.

• En el problema de la Bolsa supóngase que $P(\omega_1) = 0.3$, $P(\omega_2) = 0.4$ y $P(\omega_3) = 0.3$. El criterio de la consecuencia más probable origina un árbol podado como el de la figura 2.4.3.4. De ahí que todas las decisiones sean igualmente preferibles.

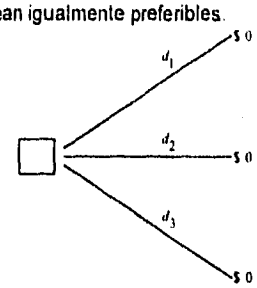

Figura 2.4.3.4

• Si ahora, para el problema de la Bolsa se tiene un árbol como el de la figura 2.4.3.5 y si los eventos de interés tienen probabilidades $P(\omega_1) = 0.3$, $P(\omega_2) = 0.4$ y $P(\omega_3) = 0.3$, el criterio de la consecuencia más probable establece que todas las decisiones son igualmente preferibles. Sin embargo, es claro que d_3 es la "mejor opción".

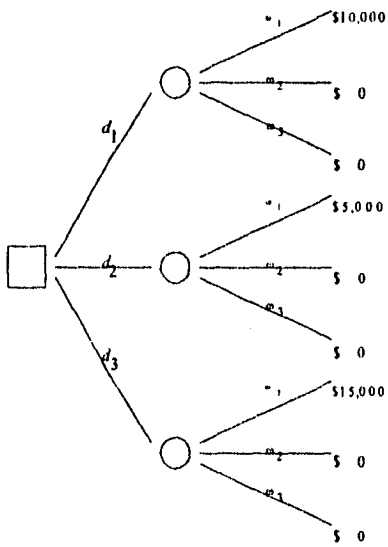

Figura 2.4.3.5

Como se ha visto, esta manera de proceder lleva frecuentemente a soluciones poco satisfactorias. Esto se debe a que en la etapa de "*podado*" del árbol, el criterio de la consecuencia más probable hace caso omiso del orden de preferencia entre consecuencias que no están asociadas al evento más probable.

2.4.4 CRITERIO DE LA UTILIDAD (PÉRDIDA) ESPERADA MÁXIMA (MÍNIMA)

Este es un criterio de solución que toma en cuenta tanto las probabilidades de las consecuencias como las preferencias del decisor sobre las mismas. Este criterio consiste en eliminar la incertidumbre de un problema de decisión asignando como consecuencia a cada decisión la utilidad (pérdida) esperada. La decisión óptima es aquella que *maximiza (minimiza) la utilidad (pérdida) esperada,* dicha decisión recibe el nombre de *Decisión de Bayes*. El valor esperado (máximo o mínimo, según el caso) se conoce como *valor de Bayes*.

Formalmente se dice que d^* es decisión de Bayes si y sólo si

$$E_{\Pi(\omega)}(U(d^*,\omega)) = \underset{d \in D}{Max}\, E_{\Pi(\omega)}(U(d,\omega)),$$

donde para cada $d \in D$, $E_{\Pi(\omega)}(U(d,\omega))$ es el valor esperado de $U(d,\omega)$ respecto a $\Pi(\omega)$, una medida de probabilidad sobre un σ –álgebra en Ω. El valor de Bayes es $E_{\Pi(\omega)}(U(d^*,\omega))$.

En términos de "Tabla" el criterio consiste en calcular

$$u_i = \sum_{j=1}^{n} u_{ij} P(\omega_j) \text{ para } i = 1,2,....n$$

y determinar d_k, tal que

$$u_k = \underset{j=1,2,...n}{\text{Max}}\, u_j.$$

Por lo que el valor de Bayes es u_k.

Planteado el problema por un árbol de decisión como el de la figura 2.3.2, el criterio o principio de la utilidad esperada máxima consiste en sustituir el nodo de incertidumbre para cada decisión por $E_{\Pi(\omega)}(U(d,\omega))$. De aquí, se obtiene un problema de decisión sin incertidumbre, por lo que la decisión será aquella con el valor esperado máximo.

El criterio es análogo en el caso de tener asociada al problema de decisión con incertidumbre una función de pérdida.

Resolver el problema de la bolsa (Ejemplo 2.3.1) mediante el principio de la utilidad esperada máxima implica determinar las utilidades esperadas para cada decisión. Por lo que se tiene,

$$E_{\Pi(\omega)}U(d_1,\omega)= 0(.4)+0(.3)+0(.3)= 0,$$
$$E_{\Pi(\omega)}U(d_2,\omega)= 100(.4)+0(.3)-200(.3)= -20,$$
$$E_{\Pi(\omega)}U(d_3,\omega)= 200(.4)+0(.3)-500(.3)= -70.$$

De aquí que se obtenga un árbol podado como el de la figura 2.4.4.1 y la decisión óptima sea *no invertir*, con valor de Bayes igual a cero.

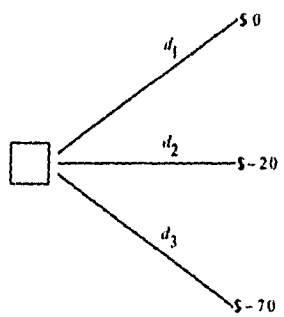

Figura 2.4.4.1

Después de haber estudiado algunos criterios para elegir la *decisión óptima* en un problema de decisión planteado en ambiente de incertidumbre, surge la duda sobre qué criterio adoptar. En la sección 4 del capítulo 3, se demuestra que si el decisor está de acuerdo con una base axiomática (referida en el mismo capítulo) para la teoría de la decisión, el criterio que deberá usar es el de la utilidad (pérdida) esperada máxima (mínima). En esta sección únicamente se exhibe un ejemplo en el que se ilustra que el criterio de la pérdida esperada mínima es el que lleva a una solución más satisfactoria.

Ejemplo 2.4.4.1: Considere el árbol de decisión de la figura 2.4.4.2 y suponga que tiene asignada una función de pérdida.

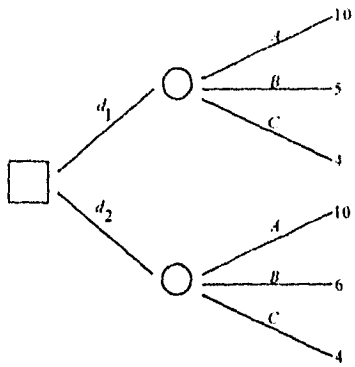

Figura 2.4.4.2

Si se es *pesimista*, lo peor que puede pasar para ambas decisiones es perder 10, por lo tanto, para un decisor que utiliza el criterio *Minimax* es exactamente igual elegir d_1 o d_2.

Si se es *optimista* (criterio *Minmin*), el decisor podrá elegir indiscriminadamente entre d_1 y d_2, ya que lo mejor que puede pasar para ambas decisiones es perder 4.

Si A o C fueran, respectivamente, los eventos con mayor probabilidad, el criterio de *la consecuencia más probable* lleva a la misma conclusión que los criterios *Minimax y Minmin.*

Supóngase ahora que $P(A)=p$, $P(B)=q$ y $P(C)=1-p-q$, con $p,q>0$; entonces

$$E_{\Pi(\omega)}(d_1,\omega)=10p+5q+4(1-p-q),$$
$$E_{\Pi(\omega)}(d_2,\omega)=10p+6q+4(1-p-q)=E_{\Pi(\omega)}(d_1,\omega)+q,$$

de aquí que

$$E_{\Pi(\omega)}L(d_1,\omega)<E_{\Pi(\omega)}L(d_2,\omega),$$

por lo que el *criterio de Bayes* determina como solución óptima a d_1.

A pesar de que los criterios Minimax y Minmin concluyen que ambas decisiones son igualmente preferibles, para cualquier decisor es evidente que d_1 *es más preferible* a d_2. Esto

debido a que si ocurren los eventos A y C las pérdidas serán iguales para ambas decisiones pero en caso de ocurrir B, d_1 produce una pérdida menor que d_2. En este caso se dice que d_1 *domina* d_2 y que por lo tanto d_2 es una decisión *inadmisible*. El único criterio con una solución razonable en casos como este es el de la *pérdida esperada mínima* (o *criterio de Bayes*).

DEFINICIÓN 2.4.4.1 *(Decisión inadmisible): Sea* $U(d,\omega)$ *una función de utilidad del problema de decisión* (D,Ω,C), $d_0 \in D$ *es inadmisible si y sólo si existe* $d_1 \in D$ *tal que*

$$U(d_0,\omega) \leq U(d_1,\omega) \text{ para todo } \omega \in \Omega$$

y

$$U(d_0,\omega_0) < U(d_1,\omega_0) \text{ para algún } \omega_0 \in \Omega.$$

En este caso se dice que d_1 *domina a* d_0.

Cabe aclarar que cuando se tiene una función de pérdida asociada al problema la definición es similar.

DEFINICIÓN 2.4.4.2 *(Decisión admisible)*: *Una decisión* $d_0 \in D$ *es admisible si y sólo si no es inadmisible.*

TEOREMA 2.4.4.1: *Si* d^* *es la decisión de Bayes del problema de decisión* (D,Ω,C), *donde* $\Omega = \{\omega_1,\omega_2,...,\omega_m\}$ *y* $P(\omega_i) > 0$ *para* $i = 1,2,...,m$, *entonces* d^* *es una decisión admisible.*

Demostración: Si d^* es la decisión de Bayes, entonces

$$E_{\Pi(\omega)}(U(d^*,\omega)) \geq E_{\Pi(\omega)}(U(d,\omega)) \text{ para toda } d \in D.$$

Supóngase que d^* es inadmisible (no admisible), entonces existe d_0 tal que

$$U(d^*,\omega) \leq U(d_0,\omega) \text{ para todo } \omega \in \Omega$$

y

$$U(d^*,\omega_0) < U(d_0,\omega_0) \text{ para algún } \omega_0 \in \Omega.$$

Como $P(\omega_i) > 0$ para toda $i = 1,2,..,m$

$$U(d^*,\omega)P(\omega) \leq U(d_0,\omega)P(\omega) \text{ para todo } \omega \in \Omega$$

y

$$U(d^*,\omega_0)P(\omega_0) < U(d_0,\omega_0)P(\omega_0) \text{ para algún } \omega_0 \in \Omega.$$

En consecuencia

$$E_{\Pi(\omega)}(U(d^*,\omega)) < E_{\Pi(\omega)}(U(d_0,\omega)),$$

lo cual es una contradicción con el hecho de que d^* es la decisión de Bayes. Por lo tanto, d^* es admisible■

En un problema de decisión, en el que algún $\omega_A \in \Omega$ tenga probabilidad cero, puede ser que la decisión de Bayes no sea admisible, tal es el caso del siguiente ejemplo.

Ejemplo 2.4.4.2: Sea el problema de decisión con una función de pérdida expresada en la siguiente tabla

	ω_1	ω_2	ω_3
d_1	1	5	0
d_2	1	0	0

con $P(\omega_1)=\frac{1}{3}$, $P(\omega_2)=0$ y $P(\omega_3)=\frac{2}{3}$.

Ya que $E_{\Pi(\omega)}(L(d_1,\omega))=E_{\Pi(\omega)}(L(d_2,\omega))=\frac{1}{3}$ ambas son soluciones de Bayes, sin embargo, d_1 es inadmisible. Por tanto, se tiene que d_1 es una decisión de Bayes inadmisible.

En el caso en que D y C no son finitos puede suceder que la solución de Bayes no sólo sea inadmisible, sino que ni siquiera exista, esto se ilustra en el siguiente problema.

Ejemplo 2.4.4.3 (DeGroot, 1970): Considere un problema de decisión en el que $\Omega = \{\omega_1,\omega_2,...\omega_m,....\}$ y que la función de pérdida L asociada a él está dada por la tabla 2.4.4.1.

Se probará que d^* es la única decisión admisible y que sin embargo d^* no es una decisión de Bayes para cualquier distribución del parámetro ω. Más aún, se probará que en este problema, si $P(\omega)>0$ para todo $\omega \in \Omega$, la decisión de Bayes no existe.

	ω_1	ω_2	ω_3	ω_4	ω_5	.....
d^*	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	.....
d_1	0	1	1	1	1	.....
d_2	0	0	1	1	1	.....
d_3	0	0	0	1	1	.....
d_4	0	0	0	0	1	.....
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Tabla 2.4.4.1

Para toda $i = 1,2,...$, $L(d_i,\omega_{i+1})=1$ y $L(d^*,\omega)=\frac{1}{2}$ para todo $\omega \in \Omega$,
en particular $L(d^*,\omega_{i+1})=\frac{1}{2}$.
En consecuencia no existe $d_k \in D$ diferente a d^* tal que

$$L(d_k,\omega)\leq L(d^*,\omega) \text{ para toda } \omega \in \Omega,$$

y por tanto d^* *es admisible.*

Por otro lado

$$L(d_i,\omega_j)=\begin{cases}0 & si\ j \leq i\\ 1 & si\ j > i.\end{cases}$$

Entonces

$$L(d_{i+1},\omega_j)= L(d_i,\omega_j) \quad \text{si } j \neq i+1$$
$$L(d_{i-1},\omega_{i+1})< L(d_i,\omega_{i+1}).$$

En consecuencia d_i *es inadmisible para todo* i =1,2,....
Con lo que se tiene que *la única decisión admisible es* d^*.
Ahora, $E_{\Pi(\omega)}(L(d^*,\omega))=\sum_{i=1}^{\infty}\frac{1}{2}P(\omega_i)=\frac{1}{2}\sum_{i=1}^{\infty}P(\omega_i)=\frac{1}{2}$,
de aquí que para algún $k_0 \in \mathbf{N}$

$$\frac{1}{2}=E_{\Pi(\omega)}(L(d^*,\omega))> E_{\Pi(\omega)}(L(d_{k_0},\omega))=\sum_{i=k_0+1}^{\infty}P(\omega_i)$$

ya que $\sum_{j=1}^{\infty} P(\omega_j)=1$.

Por lo que se tiene que d^* no es una decisión de Bayes.

Además $E_{\Pi(\omega)}(L(d_i,\omega))=\sum_{j=i+1}^{\infty} P(\omega_j)$ por lo que si $P(\omega)>0$ para todo $\omega \in \Omega$, entonces $P(\omega_{i+1})>0$, de donde

$$E_{\Pi(\omega)}(L(d_i,\omega))> E_{\Pi(\omega)}(L(d_{i+1},\omega)) \text{ para todo } i=1,2,\ldots,$$

por lo tanto *no existe la decisión de Bayes.*

De este problema también se concluye que si existe k_0 tal que $P(\omega_{k_0})>0$ y $P(\omega_k)=0$ para toda $k > k_0$, entonces d_l *será una decisión de Bayes inadmisible para toda* $l > k_0$, con valor de Bayes igual a 0.

2.5 PROBLEMAS DE DECISIÓN SECUENCIAL

Aunque siempre es posible (en el caso finito) asociar un árbol de decisión a los problemas con incertidumbre, no siempre resulta fácil estructurarlo, ya que los nodos de incertidumbre pueden llevar a su vez a nuevos nodos de incertidumbre. Sin embargo, la complejidad será solamente técnica. Un problema de este tipo se ilustra en el siguiente ejemplo.

Ejemplo 2.5.1 (Smith, 1988): El día que cumple 20 años un paciente ingresa al hospital con síntomas que sugieren que padece la enfermedad A (con $P(A)=0.4$) o bien la enfermedad B (con $P(B)=0.6$). Si el paciente no se trata, sin importar la enfermedad (A o B) morirá ese mismo día con probabilidad 0.8 o bien sobrevivirá con probabilidad 0.2. El responsable de urgencias tiene 3 opciones excluyentes:

d_1, no administrar tratamiento alguno;
d_2, administrar la droga S;
d_3, remitir a cirugía.

Tanto la droga como la cirugía son peligrosas. Sin importar la enfermedad, el paciente puede morir en la operación con probabilidad de 0.5 y también sin importar la enfermedad, la droga puede causarle una reacción fatal con probabilidad 0.2. Si el paciente sobrevive a los

efectos adversos de la droga y tenía la enfermedad A puede que se cure (Probabilidad 0.5) o quede igual. Si tenía la enfermedad B seguro queda igual. Si el paciente sobrevive a la operación y tenía la enfermedad A se cura con probabilidad de 0.8 y queda igual con probabilidad de 0.2. Si tenía la enfermedad B se cura con probabilidad de 0.4 y queda igual con probabilidad de 0.6. Sano o curado el paciente tiene una esperanza de vida de 50 años. Resolver el problema de decisión que tiene el responsable de urgencias pensando que las consecuencias se miden por los años de esperanza de vida del paciente.

Se recomienda construir el árbol de decisión asociado al problema, para que de esta manera se tenga clara su estructura y poder resolver el problema a partir de él.

Del nodo de decisión se desprenden 3 ramas, ya que D está formado por 3 decisiones (d_1, d_2 y d_3). Los sucesos inciertos involucrados en el problema dependen de la decisión que se tome, por esto, los nodos de incertidumbre no serán los mismos en todas las decisiones.

Si se elige d_1 (no administrar tratamiento alguno) sólo puede ocurrir que el paciente muera con probabilidad de 0.8 o sobreviva con probabilidad de 0.2 y en este caso se tendría una esperanza de vida de 50 años; de aquí que para d_1 se tenga una rama como la de la Figura 2.5.1-1.

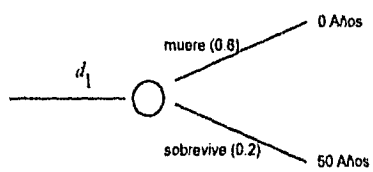

Figura 2.5.1-1

Si se elige d_2 (administrar la droga S) puede suceder que se presente un efecto fatal con probabilidad de 0.2, lo cual obviamente lleva a una esperanza de vida de 0 años. Si el paciente sobrevive, puede ser que tenga la enfermedad A o la enfermedad B con probabilidad de 0.4 y 0.6 respectivamente. Si padece la enfermedad A se llega a un nuevo nodo de incertidumbre, ya que puede suceder que se cure con probabilidad de 0.5 (esperanza de vida igual a 50 años) o quede igual; si queda igual se genera un nodo de incertidumbre como el de d_1, ya que puede suceder entonces que el paciente muera (con probabilidad de 0.8) o sobreviva (con probabilidad de 0.2) con esperanza de vida de 0 y 50 años respectivamente. Si el paciente padece la enfermedad B sabe con certidumbre que quedará igual, y que por lo tanto, puede

suceder que muera con probabilidad de 0.8 y esperanza de vida de 0 años o sobreviva con probabilidad de 0.2 y esperanza de vida de 50 años. Ver Figura 2.5.1-2.

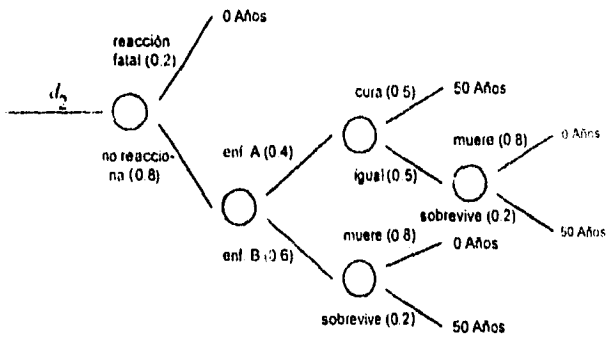

Figura 2.5.1-2

Para la decisión d_3 se realiza un seguimiento análogo en los posibles sucesos inciertos y se obtiene lo mostrado en la Figura 2.5.1-3.

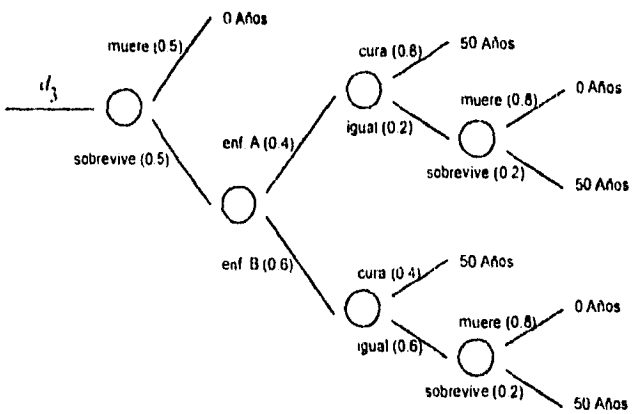

Figura 2.5.1-3

Con las tres ramas de decisión seguidas de su respectivo nodo de incertidumbre y consecuencias, queda constituido el árbol de decisión, el cual se presenta en la figura 2.5.1.

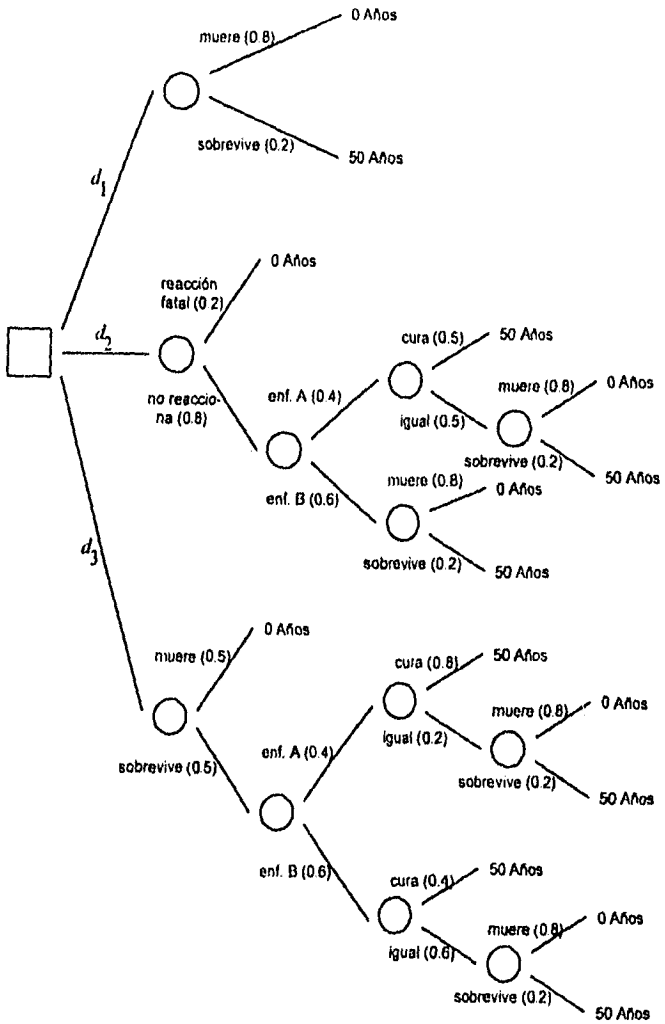

Figura 2.5.1

Resolviendo vía el criterio *Maximin*, se obtiene el árbol de la figura 2.5.2, de donde se observa que cualquier decisión es *Maximin* con valor *Maximin* igual a 0 años.

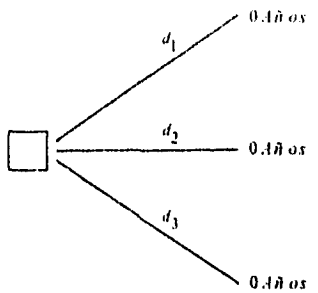

Figura 2.5.2

Por el criterio *Maxmax* resulta que el árbol podado queda como el de la figura 2.5.3, de donde se obtiene que cualquier decisión es *Maxmax* con valor *Maxmax* de 50 años.

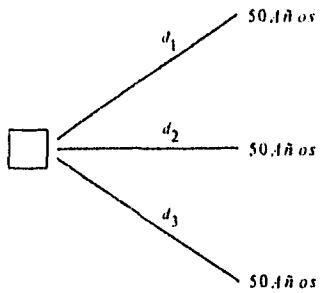

Figura 2.5.3

Para resolver el problema vía la consecuencia más probable y vía la utilidad esperada, el árbol se reducirá a uno como el de la figura 2.3.1 tomando intersecciones de eventos.

Considere los siguientes eventos inciertos

- ω_1, morir por no tener tratamiento;
- ω_2, vivir a pesar de no tener tratamiento;
- ω_3, morir por reacción fatal debido a la droga;
- ω_4, no tener reacción a la droga, tener enfermedad A y curarse;
- ω_5, no tener reacción a la droga, tener enfermedad A, quedarse igual y morir por falta de tratamiento;
- ω_6, no tener reacción a la droga, tener enfermedad A, quedarse igual y sobrevivir sin un nuevo tratamiento,

ω_7, no tener reacción a la droga, tener enfermedad B y morir por falta de un nuevo tratamiento;

ω_8, no tener reacción a la droga, tener enfermedad B, y sobrevivir sin un nuevo tratamiento;

ω_9, morir en la cirugía;

ω_{10}, sobrevivir a la cirugía, tener la enfermedad A y curarse por la cirugía;

ω_{11}, sobrevivir a la cirugía, tener la enfermedad A, quedar igual y morir por falta de un nuevo tratamiento;

ω_{12}, sobrevivir a la cirugía, tener la enfermedad A, quedar igual y sobrevivir aún sin un nuevo tratamiento;

ω_{13}, sobrevivir a la cirugía, tener la enfermedad B y curarse por la cirugía;

ω_{14}, sobrevivir a la cirugía, tener la enfermedad B, quedar igual y morir por falta de un nuevo tratamiento;

ω_{15}, sobrevivir a la cirugía, tener la enfermedad B, quedar igual y sobrevivir aún sin un nuevo tratamiento.

Con estos nuevos eventos, el problema de decisión original se puede representar con el árbol de la figura 2.5.4.

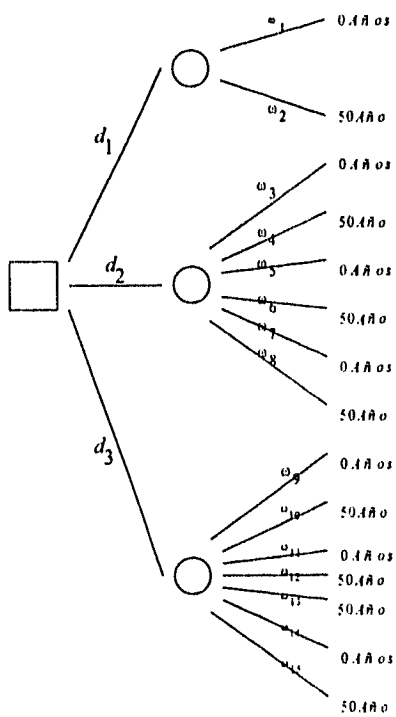

Figura 2.5.4

Para poder implantar los criterios de la consecuencia más probable y de la utilidad esperada máxima al problema de decisión, es necesario determinar las probabilidades de los eventos ω_i, $i = 1,2,...,15$. Se calculará la probabilidad de ω_{12}. El resto de las probabilidades se obtienen de manera análoga.

ω_{12} es igual a la intersección de los eventos {sobrevivir cirugía}, {enfermedad A}, {quedar igual} y {sobrevivir sin un nuevo tratamiento}. Sin embargo, en el contexto del problema, el evento {sobrevivir sin un nuevo tratamiento} podría considerarse igual al evento {sobrevivir sin tratamiento}, entonces

$P(\omega_{12})$= P({sobrevivir cirugía}$\cap${enfermedad A}$\cap${quedar igual}$\cap${sobrevivir sin tratamiento}).

{sobrevivir sin tratamiento} es independiente a {sobrevivir cirugía} {enfermedad A} y {quedar igual}, por lo que

$P(\omega_{12})$= P({sobrevivir cirugía}$\cap${enfermedad A}$\cap${quedar igual})P({sobrevivir sin tratamiento}),

pero

P({sobrevivir cirugía}$\cap${enfermedad A}$\cap${quedar igual})
= P({quedar igual} | {enfermedad A}$\cap${sobrevivir cirugía}) P({enfermedad A}$\cap${sobrevivir cirugía}),

entonces $P(\omega_{12})$ es igual a

P({quedar igual} | {enfermedad A}$\cap${sobrevivir cirugía}) P({enfermedad A}$\cap${sobrevivir cirugía}) P({sobrevivir sin tratamiento}).

{enfermedad A} y {sobrevivir cirugía} se pueden considerar independientes por el contexto del problema, de aquí que $P(\omega_{12})$ es igual a

P({quedar igual} | {enfermedad A}$\cap${sobrevivir cirugía}) P({enfermedad A}) P({sobrevivir cirugía}) P({sobrevivir sin tratamiento})=(.2)(.4)(.5)(.2)=.008

Como se puede notar, en este caso, todo se reduce a multiplicar las probabilidades que aparecen en una misma rama del árbol inicial (figura 2.5.1). Por lo que las probabilidades de los sucesos ω_i son

$P(\omega_1)= 0.8$ $P(\omega_9)= 0.5$
$P(\omega_2)= 0.2$ $P(\omega_{10})= 0.16$
$P(\omega_3)= 0.2$ $P(\omega_{11})= 0.032$
$P(\omega_4)= 0.16$ $P(\omega_{12})= 0.008$
$P(\omega_5)= 0.128$ $P(\omega_{13})= 0.12$
$P(\omega_6)= 0.032$ $P(\omega_{14})= 0.144$
$P(\omega_7)= 0.384$ $P(\omega_{15})= 0.036$
$P(\omega_8)= 0.096$

Podando el árbol vía el criterio de la *consecuencia más probable* se obtiene el árbol de la figura 2.5.5, lo que origina que cualquier decisión sea óptima vía este criterio, con valor de la c.m.p. igual a 0.

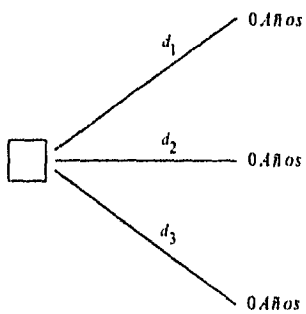

Figura 2.5.5

Por el criterio de Bayes, se tiene

$E_{\Pi(\theta)}U(d_1,\omega)= 0(0.8)+50(0.2)= 10$
$E_{\Pi(\theta)}U(d_2,\omega)= 0(0.2)+50(0.16)+0(0.128)+50(0.032)+0(0.384)+50(0.096)= 14.4$
$E_{\Pi(\theta)}U(d_3,\omega)=0(0.5)+50(0.16)+0(0.032)+50(0.008)+50(0.12)+0(0.144)+50(0.036)=16.2$

En consecuencia, *la decisión de Bayes es intervenir quirúrgicamente al paciente* (d_3) con valor de Bayes igual a 16.2 años.

En este problema el criterio de la "*Utilidad Esperada Máxima*" es el único que establece una distinción entre las 3 opciones involucradas e intuitivamente proporciona una decisión razonable.

Este problema puede ser resuelto sin tener que reducir el árbol de la forma en que se hizo ya que se puede resolver a partir de la estructura inicial (figura 2.5.1) "*podando*" de derecha a izquierda y sustituyendo cada nodo por *la utilidad esperada* que le corresponda.

Los problemas de decisión vistos hasta este momento presentan diferentes opciones (decisiones), donde cada una, dependiendo de sucesos inciertos lleva a posibles consecuencias. Sin embargo, existen problemas con una estructura aparentemente más compleja. Tal es el caso de un problema en el que se toma una decisión y dependiendo del suceso incierto que se presente a partir de tomar dicha decisión, se deberá tomar una nueva decisión (complementaria de la primera) y así sucesivamente. Ver figura 2.5.6.

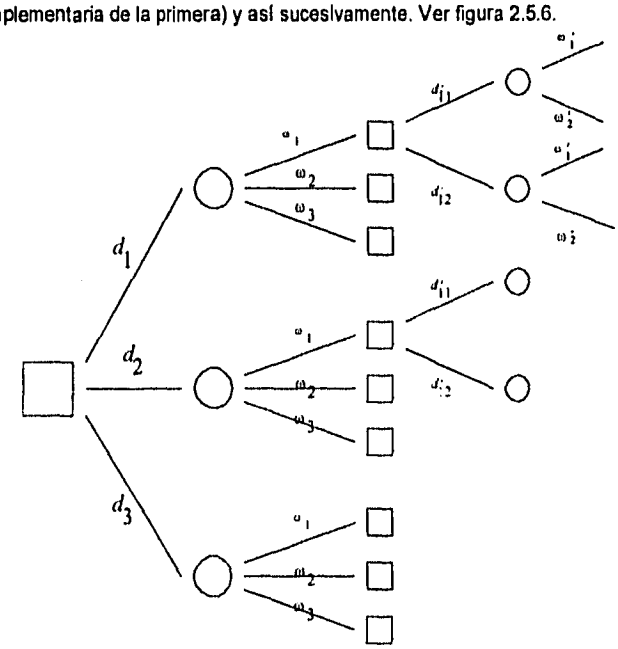

Figura 2.5.6

La manera de resolver estos problemas es podar cada árbol de derecha a izquierda vía algún criterio de los estudiados. En el caso de "podarlo" vía la utilidad (pérdida) esperada máxima (mínima), el valor de los nodos deberá irse sustituyendo consecutivamente por los valores esperados obtenidos en cada parte del proceso.

La solución de un problema de este estilo se muestra en el siguiente ejemplo.

Ejemplo 2.5.2 (Bernardo, 1981)**:** Se dispone de dos urnas denotadas por I y II. La urna I contiene 3 bolas rojas y 2 blancas, la urna II contiene 5 rojas y 4 blancas. El juego consiste en sacar un máximo de dos bolas. Por cada bola extraída se gana 100 pesos si es roja, y se pierden 50 si es blanca. Por la primera extracción no hay que pagar nada. Por la segunda extracción hay que pagar 50 pesos si se trata de la urna I y 25 si se trata de la urna II. Determinar la estrategia óptima para un jugador.

En este problema el conjunto de decisiones está formado por el conjunto de estrategias de juego. Primero se debe decidir entre sacar una bola de la urna I (d_1), sacarla de la urna II (d_2) o no sacar nada (d_3). Cuando no se saca nada la estrategia termina ahí y corresponde a la de no jugar. Cuando sí se extrae una bola se debe decidir de cuál urna, una vez extraída, se observa su color y a continuación se debe tomar otra decisión, sacar una segunda bola de la urna I (d_1'), de la urna II (d_2') o no sacar ninguna bola (d_3'). En cualquiera de las tres opciones, la estrategia termina ahí, ya que por las condiciones del juego, a lo más se pueden sacar 2 bolas. Por lo tanto, el árbol de decisión generado por este juego corresponde al de la figura 2.5.7.

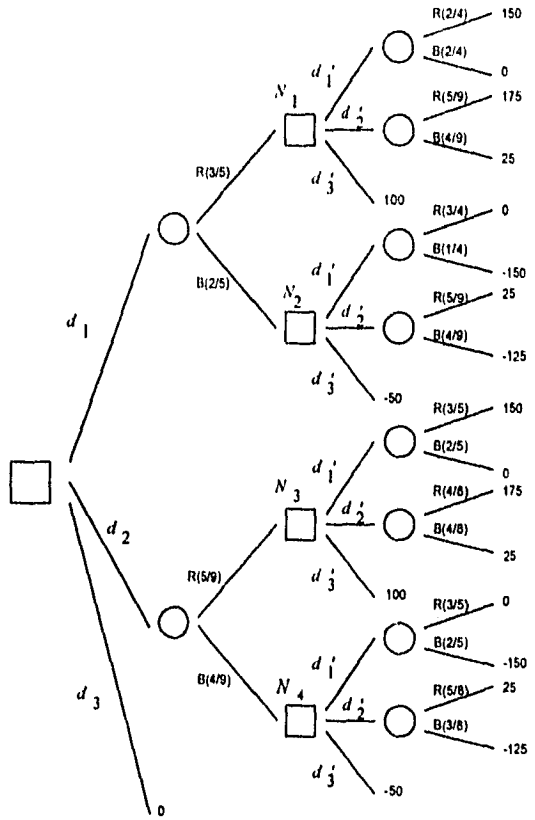

Figura 2.5.7

Como el árbol debe ser podado de derecha a izquierda, la primera parte de la solución del problema consiste en resolver como problemas de decisión independientes los que corresponden a los nodos de decisión N_1, N_2, N_3 y N_4.

Para N_1 se tiene:

$E_1(d_1')=150(\frac{2}{4})=75$

$E_1(d_2')=175(\frac{5}{9})+25(\frac{4}{9})=\frac{325}{3}=108.33$

$E_1(d_3')=100$.

Y así, análogamente se tiene,

para N_2	para N_3	y para N_4
$E_2(d_1')=-\frac{75}{2}=-37.5$	$E_3(d_1')=90$	$E_4(d_1')=-60$
$E_2(d_2')=-\frac{475}{9}=-52.77$	$E_3(d_2')=100$	$E_4(d_2')=-\frac{125}{4}=-31.25$
$E_2(d_3')=-50,$	$E_3(d_3')=100,$	$E_5(d_3')=-50.$

con lo que se obtiene que la decisión es d_2' para N_1, d_1' para N_2, d_2' o d_3' para N_3 y d_2' para N_4. Sustituyendo estos valores esperados en los nodos de decisión correspondientes se obtiene un árbol de decisión como el de la figura 2.5.8. Es importante recordar que la decisión final deberá ser la unión secuencial de las soluciones que fueron obtenidas en cada paso del proceso de solución.

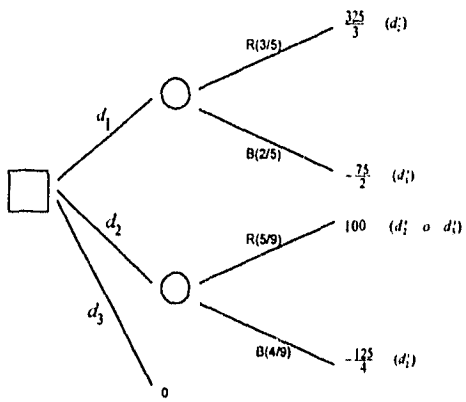

Figura 2.5.8

Al resolver el problema de decisión representado por el árbol de la figura 2.5.8 se obtiene que $E(d_1)$=50, $E(d_2)$=41.66 y $E(d_3)$=0, de aquí que la mejor opción es aquélla asociada a d_1.

De lo anterior se concluye que la estrategia óptima es extraer una primera bola de la urna I, si sale roja se obtiene una segunda de la urna II, y si sale blanca de la urna I. La estrategia tiene un valor esperado (valor de Bayes) de 50 pesos.

2.6 INTERPRETACIONES DE LA PROBABILIDAD

Como se ha visto en este capítulo algunos criterios de solución para el problema de decisión definido por $(D,\Omega,C,<)$; requieren una medida de probabilidad sobre un σ –álgebra

definido en Ω. Es por esto que resulta importante discutir acerca de las diferentes interpretaciones y formas de asignar una probabilidad.

Existen fundamentalmente tres enfoques de la probabilidad: *Clásico, Frecuentista* y *Subjetivo*.

En un principio, la probabilidad estaba asociada a los juegos de azar. Uno de los más comunes es el de lanzamiento de monedas, en donde los eventos de interés son águila y sol.

Si la moneda es *honesta* P(águila)=$\frac{1}{2}$ y P(sol)=$\frac{1}{2}$, es decir, los dos eventos tienen la misma probabilidad; esta equidad de probabilidades da lugar a la definición clásica de probabilidad.

DEFINICIÓN 2.6.1 (*Probabilidad clásica* (Mood, Graybill and Boes, 1974): *Si un experimento aleatorio puede resultar en n eventos mutuamente excluyentes e igualmente verosímiles y si n_A de ellos tienen el atributo A, entonces la probabilidad de A es la fracción*

$$\frac{n_A}{n}.$$

Este concepto de probabilidad es insuficiente ya que se limita a ser aplicable a un número muy reducido de casos: aquéllos en los que haya equiprobabilidad y en donde el espacio de sucesos inciertos es finito. Esto se ilustra en los siguientes ejemplos.

Ejemplo 2.6.1: Calcular la probabilidad de extraer un número par dentro del total de números naturales. Intuitivamente esta probabilidad debería ser $\frac{1}{2}$, ya que de 1 a $2N$ existen N números pares, y aplicando la definición 2.6.1, se tiene que:

$$P_N(\textit{Sacar un par entre } 1 \textit{ y } 2N) = \frac{N}{2N} = \frac{1}{2}.$$

Desafortunadamente, como el conjunto de los Naturales es infinito, no es posible aplicar la definición 2.6.1 para calcular la probabilidad del evento de interés. Sólo se podría obtener la probabilidad de sacar un número par en un subconjunto finito de los naturales.

Ejemplo 2.6.2: Se desea calcular la probabilidad de obtener un águila al lanzar una moneda "no honesta".

Debido a que la moneda está "cargada" el enfoque clásico no es aplicable. Para definir $P(\text{águila})$ se puede pensar en lanzar la moneda "*varias veces*" (n veces) y aproximar $P(\text{águila}) = \frac{No.\ de\ águilas}{n}$, donde n es el número total de lanzamientos realizados. Esta forma de proceder origina la *definición frecuentista de la probabilidad.*

DEFINICIÓN 2.6.2 (*Probabilidad frecuentista*): *Si un experimento se puede repetir un número ilimitado de veces en las mismas condiciones y para un número n de repeticiones del experimento el atributo A ocurrió n_A veces, entonces la probabilidad del evento A se define como*

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

Volviendo al ejemplo, cuando la moneda es "honesta", la definición clásica es aplicable. Cuando no y se tiene a la mano la moneda, es posible lanzarla un número grande de veces, contar el número de águilas sobre el total de lanzamientos y usar la definición frecuentista. Sin embargo, si la moneda no es honesta y no se tiene acceso a la misma, es imposible asignar la probabilidad de águila mediante el enfoque frecuentista.

Aún sin lanzar la moneda es muy factible que un decisor pueda hacer afirmaciones sobre la probabilidad de águila si se le ofrecen opciones (*loterías*) del estilo

$l_1 = \{1000 \text{ con } P(A),\ 0 \text{ con } P(S)\}$,
$l_2 = \{1000 \text{ con } P = 0.5,\ 0 \text{ con } P = 0.5\}$;

ya que si elige l_1 estará diciendo que la probabilidad de águila es mayor a 0.5 y en caso de elegir l_2, lo que el decisor estaría pensando es que la probabilidad de águila es menor a 0.5.

Además, si estas loterías se ofrecen a distintos individuos, lo más seguro es que evalúen la probabilidad de águila en forma diferente; lo cual sugiere que no existen probabilidades absolutas sino subjetivas.

Se podría pensar que la probabilidad de águila asignada con el enfoque frecuentista es "absoluta" ya que si se lanza 100 veces la moneda y se cuentan las águilas que salieron, la probabilidad "verdadera" será $\frac{No.\ de\ águilas}{100}$. Sin embargo, esta probabilidad también es subjetiva, ya que se está considerando que 100 lanzamientos son suficientes para aproximar el $\lim_{n\to\infty}\frac{n_A}{n}$, esto es, se está siendo subjetivo al evaluar el número "razonable" de repeticiones; además de que se supone que los lanzamientos de la moneda siempre se realizan en condiciones idénticas. La definición clásica también requiere de un aspecto subjetivo para determinar probabilidades, ya que es necesario suponer que los eventos son igualmente verosímiles.

El enfoque clásico también asigna probabilidades condicionales, ya que determina la probabilidad de obtener águila dado que la moneda es honesta.

Por todo lo anterior, el enfoque de probabilidad que parece resultar más atractivo en los problemas de decisión es el de *probabilidad subjetiva*. La formalización de este enfoque se estudiará en el capítulo siguiente, debido a que para ello se requiere de los Axiomas de Coherencia de la Teoría de la Decisión.

Es importante hacer notar que por difícil que sea, las personas son capaces de cuantificar de manera consciente o inconsciente su incertidumbre en términos de probabilidades. Aún cuando no se pueda hablar de equiprobabilidad ni de repeticiones de experimentos.

CAPÍTULO 3

TRATAMIENTO AXIOMÁTICO DE LA TEORÍA DE LA DECISIÓN

En el capítulo anterior se definió la estructura de un problema de decisión y se estudiaron diferentes criterios para determinar la manera óptima de proceder, todo esto, en el contexto de la *Teoría de la Decisión*. Ahora, en este capítulo, se fundamenta tal teoría mediante una base axiomática formada por los llamados *Axiomas de Coherencia*, los cuales permiten definir formalmente los conceptos de *probabilidad* y *utilidad*. Se demuestra que el único criterio consistente con los principios axiomáticos es el de la *Utilidad (pérdida) esperada Máxima (mínima)*.

Como se ha mencionado anteriormente, el objetivo de resolver un problema de decisión en ambiente de incertidumbre es determinar aquella forma de actuar que resulte óptima para el decisor y que además sea "*razonable*" en algún sentido. La racionalidad, en un problema de decisión, se debe hacer presente en la forma de elegir acciones en situaciones de incertidumbre y representar creencias con respecto a los eventos relevantes que se ven involucrados en él. Dicha racionalidad se basa en los Axiomas de Coherencia, los cuales gobiernan los patrones para determinar relaciones de preferencia entre opciones.

Es importante señalar que los axiomas de coherencia son prescriptivos y no descriptivos, esto es, indican en cierta manera la forma en la que un decisor debería expresar sus preferencias mas no describe la manera en la que generalmente actúa en la vida práctica.

En conclusión, los Axiomas de Coherencia son una respuesta natural a la intención de tener un comportamiento coherente en el momento de tomar decisiones en ambiente de incertidumbre. Tales axiomas son abordados en este trabajo de manera intuitiva, algunas referencias a un tratamiento más formal son Rubin (1987), Lindley (1982), Lane and Sudderth (1983) y Bernardo y Smith (1994).

3.1 AXIOMAS DE COHERENCIA

Con frecuencia, y de manera natural, un decisor expresa sus preferencias entre las distintas opciones que se le presentan. Una opción es una situación en la que se obtiene la consecuencia c_1 si sucede A_1, la c_2 si sucede A_2, ..., y la c_k si sucede A_k. Dicha opción también es llamada lotería y se denota por

$$l=\{c_1\ A_1, c_2\ A_2, \ldots, c_k\ A_k\}.$$

De esta definición de opción se concluye inmediatamente que cada uno de los elementos del espacio de decisiones son opciones, ya que las decisiones de un problema de decisión pueden ser escritas como

$$\begin{aligned} d_1&=\{c_{11}\ \omega_{11}, c_{12}\ \omega_{12}, \ldots, c_{1m_1}\ \omega_{1m_1}\}\\ d_2&=\{c_{21}\ \omega_{21}, c_{22}\ \omega_{22}, \ldots, c_{2m_2}\ \omega_{2m_2}\}\\ &\ldots\ldots\ldots\ldots\\ d_n&=\{c_{k1}\ \omega_{k1}, c_{k2}\ \omega_{k2}, \ldots, c_{nm_n}\ \omega_{nm_n}\}. \end{aligned}$$

De igual manera las consecuencias son casos particulares de opciones, pues se pueden escribir como

$$c=\{c\ \ \Omega\}.$$

Para simplificar la notación c será usada tanto para denotar a la consecuencia misma como para referirse a ella como un caso particular de decisión.

El concepto de preferencia es muy común y natural en la vida cotidiana. Cada individuo generalmente está capacitado para manifestar sus preferencias, aunque ello en algunas ocasiones pueda resultar complicado. Para expresar tales preferencias, se define la relación de orden entre opciones, esto es, $l_1 > l_2$ si se prefiere la opción l_1 a la opción l_2, $l_1 \geq l_2$ si la opción l_1 resulta más o igualmente preferible a la opción l_2 y $l_1 \sim l_2$ si las opciones l_1 y l_2 son igualmente preferibles (de modo análogo se interpreta $l_1 < l_2$ y $l_1 \leq l_2$). Como las consecuencias son casos particulares de las opciones se puede utilizar esta misma relación de orden entre ellas.

Si se aspira a hacer una elección racional entre opciones alternativas, se debería al menos estar en condiciones de expresar preferencias entre las opciones que se presentan.

Además, para el conjunto de consecuencias resulta razonable pensar que se puede encontrar una consecuencia suprema c^* y una consecuencia ínfima c_* pertenecientes o no a C tales que para cualquier $c \in C$ sucede que $c_* \leq c \leq c^*$; esto se resume en el siguiente axioma

AXIOMA 1 (*Comparabilidad*): *Para cada par de opciones l_1 y l_2, es cierta una y sólo una de las siguientes relaciones: $l_1 > l_2$, $l_1 < l_2$ o $l_1 \sim l_2$. Además existen dos consecuencias c^* y c_*, tales que $c_* < c^*$ y para toda $c \in C$ sucede que $c_* \leq c \leq c^*$.*

De este axioma se implica la comparabilidad entre consecuencias y verosimilitudes. La comparabilidad entre consecuencias es inmediata debido a que éstas son casos particulares de opciones y por lo tanto les es aplicable el axioma. Algo similar ocurre con las verosimilitudes, pues cualquier suceso A perteneciente a Ω se puede identificar como una opción.

Si se comparan las opciones $l_1=\{c^*, A, c_*, A^c\}$ y $l_2=\{c^*, B, c_*, B^c\}$ se están comparando las verosimilitudes de los eventos A y B pertenecientes a Ω. Esta comparabilidad entre verosimilitudes se estudia más detalladamente en la sección (3.2).

Resulta natural pensar que el decisor debe ser transitivo al expresar sus preferencias, ya que no es razonable que se prefiere la opción l_1 a l_2 y l_2 a l_3, y que además se afirme que l_3 es más preferible que l_1. Esta condición de transitividad se plantea en el siguiente axioma

AXIOMA 2 (*Transitividad*): *Si $l_1 > l_2$ y $l_2 > l_3$, entonces $l_1 > l_3$. Análogamente, si $l_1 \sim l_2$ y $l_2 \sim l_3$, entonces $l_1 \sim l_3$ y si $l_1 < l_2$ y $l_2 < l_3$, entonces $l_1 < l_3$.*

Este axioma se justifica intuitivamente por reducción al absurdo.

Supóngase que el decisor prefiere l_1 a l_2 y l_2 a l_3, entonces estaría dispuesto a pagar una cantidad x_1 (en especie o dinero) para pasar de l_3 a l_2, y por el mismo razonamiento estaría dispuesto a pagar otra cantidad x_2 por pasar de l_2 a l_1. Si además se supone que el decisor es intransitivo y prefiere l_3 a l_1, entonces debería estar dispuesto a pagar otra cantidad x_3 para pasar de l_1 a l_3. Por lo tanto el decisor habría vuelto a la opción inicial después de haber pagado una cantidad de dinero equivalente a $x_1 + x_2 + x_3$. De acuerdo con lo anterior, el decisor podría repetir este proceso intercambiando opciones hasta quedarse sin recursos, y

obviamente, sin haber obtenido ningún beneficio adicional. Con la discusión anterior se concluye que es conveniente ser transitivo al manifestar preferencias.

Es lógico pensar que existen opciones l_1 y l_2 tales que l_1 es preferible a l_2 cuando se presenta el evento A, pero que no lo es cuando éste no se presenta. Sin embargo, hay opciones que son preferibles tanto cuando se presenta el evento como cuando no se presenta. Por ejemplo, si un decisor prefiere la opción l_1 a la l_2 cuando llueve y cuando no llueve, entonces se dice simplemente que "la opción l_1 es preferible a la opción l_2".

AXIOMA 3 (*Sustitución y Dominancia*): *Si $l_1 > l_2$ cuando sucede A y $l_1 > l_2$ cuando sucede A^c, entonces $l_1 > l_2$. De igual manera, si $l_1 \sim l_2$ cuando sucede A y $l_1 \sim l_2$ cuando sucede A^c, entonces $l_1 \sim l_2$.*

Este axioma plantea que si en general se tienen dos opciones l_1 y l_2 definidas como $\{c_{11}|A_1, c_{12}|A_2, \ldots, c_{1k}|A_k\}$ y $\{c_{21}|A_1, c_{22}|A_2, \ldots, c_{2k}|A_k\}$ respectivamente, y se cumple que las consecuencias de la primera opción *dominan* a las de la segunda, esto es, $c_{1i} \geq c_{2i}$ para toda i, entonces, $l_1 \geq l_2$. Si además existe una j tal que $c_{1j} > c_{2j}$, sucederá que $l_1 > l_2$.

Como resultado de este axioma se tiene que si dos opciones l_1 y l_2 son igualmente preferibles dado A y A^c, entonces son equivalentes y en consecuencia, podrán ser intercambiadas. Derivado de este principio de sustitución se puede reemplazar en una opción una consecuencia por otra opción que sea equivalente a dicha consecuencia. Esto es, si se tiene $l_2=\{c_1|A, c_2|A^c\}$ y $c_1 \sim l_1$, se cumple que $l_2 \sim l_3$, donde $l_3=\{l_1|A, c_2|A^c\}$.

Por otra parte, se ha hecho hincapié en que el decisor que pretenda resolver razonablemente sus problemas de decisión debe medir (de manera cuantitativa) la información inicial con que cuenta, así como sus preferencias.

Este valor no podrá ser expresado explícitamente si no se cuenta tanto con una unidad de medida como con un mecanismo que facilite su obtención. Con esta finalidad se plantea que es posible imaginar métodos para seleccionar puntos al azar dentro del cuadrado unitario (no es necesario que estos mecanismos se construyan físicamente). Es importante considerar que los mecanismos que generan puntos aleatorios cumplan con que todos los puntos tengan la misma verosimilitud de ser seleccionados.

Dichos mecanismos son fáciles de construir, por ejemplo, si se piensa en una ruleta de circunferencia igual a la unidad, se toma la longitud del arco que separa el origen predeterminado del punto señalado por la aguja, obteniendo de esta manera un número aleatorio $X \in [0,1]$. Con una repetición del experimento se obtendrá otro punto $Y \in [0,1]$. Los números "x" y "y" son las coordenadas del punto $z=(x,y)$.

En este experimento surge un problema, si la aguja coincide con el punto señalado como origen, entonces, ¿qué valor tomar? Para resolver este dilema se propone lanzar una moneda "honesta" para decidir si se considera como longitud 0 ó 1. Otro ejemplo sería tomar las observaciones mediante la distribución uniforme en el intervalo $[0,1]$ a través de un programa computacional de simulación. A partir de lo discutido anteriormente se llega al siguiente axioma

AXIOMA 4 (*Suceso de referencia, experimento auxiliar*): *El decisor puede imaginar un procedimiento para generar un punto aleatorio Z en el cuadrado unitario contenido en $\mathbb{R}^2$, esto es, un punto $z=(x,y)$ en $I=[0,1]\times[0,1]$ tal que para cualquier par de regiones R_1, R_2 de I, el suceso $\{Z \in R_1\}$ le resulta menos verosímil que el suceso $\{Z \in R_2\}$ si y sólo si, el área de R_1 es menor que la de R_2.*

Para simplificar la notación R será utilizado tanto para referirse a la *región* del cuadrado como al *suceso* de que el punto aleatorio se localice en tal región.

Como es evidente, el axioma construye un mecanismo con el cual se puede medir mediante comparaciones, que tan deseables son las opciones. Las opciones son de la forma

$$l_R=(c^* \quad R, \; c, \quad R^c)$$

donde $R \subset [0,1]^2=[0,1]\times[0,1]$.

Dicha opción significa que se obtiene la mejor consecuencia si sucede el evento $\{Z \in R\}$, pero en caso de ocurrir el suceso $\{Z \in R^c\}$ se obtiene la peor consecuencia.

La medida que dé el decisor a partir de expresar sus preferencias entre opciones debe ser precisa. Esta precisión se puede garantizar al considerar que el conjunto $\{l_R \quad R \subset I\}$ es denso con respecto a la relación (~). Es decir, que se pueda asegurar que para cada l_i existe una región R contenida en I tal que se llega a la relación $l_i \sim l_R$. La densidad del conjunto

formado por los l_R se sustenta en la densidad de los números reales. El concepto de precisión se aborda en el siguiente postulado

AXIOMA 5 (*Axioma de la medida precisa de preferencia*): *Para toda opción l_i existe una región R contenida en I tal que $l_i \sim l_R$*

Intuitivamente existe R tal que $l_i \sim l_R$, ya que si l_i se compara con opciones $l_{R(x)}$, con $R(x)$ un rectángulo de lado x y altura 1, por la densidad en los reales debe existir un $x_o \in [0,1]$ tal que $l_i \sim l_{R(x_o)}$.

Estos axiomas son útiles en la construcción formal de las definiciones de probabilidad y utilidad.

3.2 DEFINICIÓN DE PROBABILIDAD

En el capítulo anterior se habló de la necesidad de utilizar el enfoque subjetivo para abordar el concepto de probabilidad. Mediante este enfoque la probabilidad será considerada como un grado de creencia.

Anteriormente no se pudo dar una definición formal de probabilidad porque era necesario introducir los Axiomas de Coherencia; ahora que se han presentado se cuenta con los elementos necesarios para definirla. La definición se obtiene a partir de establecer preferencias entre las verosimilitudes de los sucesos, considerándolos casos particulares de opciones.

Sea $d_i \in D$ definido como opción, entonces $d_i=\{c_{i1}\ B_1, c_{i2}\ B_2, \ldots, c_{ik}\ B_k\}$, donde c_{ij} denota la consecuencia de adoptar la decisión d_i dado que ocurrió el evento B_j.

Considere el conjunto D_1 definido como la unión del conjunto D y el conjunto de las opciones de la forma $c=\{c\ \ \Omega\}$ que corresponden a las consecuencias.

De igual manera, cualquier suceso A que pertenece a Ω se puede identificar como una opción d_A, con $d_A=\{c^*\ \ A,\ c_*\ \ A^c\}$, donde c^* y c_* son las consecuencias extremas del axioma 1. La opción d_A se puede considerar como "el evento A".

Se define D_2 como el conjunto unión de D_1 con el conjunto de opciones de la forma d_A con $A \in \Omega$. Sean d_A y d_B elementos de D_2, por el Axioma 1 ocurre una de las siguientes situaciones

i) $d_A < d_B$
ii) $d_A > d_B$
iii) $d_A \sim d_B$.

i) implica que B es más creíble que A, ii) que B es menos creíble que A y por supuesto, iii) indica que A y B son igualmente verosímiles o creíbles.

Un suceso $R \subset I$ referido en el axioma 4 también se escribe como una opción. Llámese a esta opción d_R, entonces, $d_R = \{c^* \quad R, c, \quad R^c\}$.

Si ahora se define D_3 como la unión del conjunto de las opciones d_R con D_2 y se toman dos elementos de éste, d_R y d_Q, se tiene que por el axioma 1 se cumplirá alguna de las tres relaciones siguientes:

iv) $d_R < d_Q$
v) $d_R > d_Q$
vi) $d_R \sim d_Q$.

iv) implica que el área de R es menor que el área de Q, esto es, $A(R) < A(Q)$, lo cual indica que de generar un punto aleatorio Z (con las características del axioma 4), sería más creíble que este cayera en Q que en R. En otras palabras, se diría que "Q es más creíble que R". v) indica que Q es menos creíble que R ($A(R) > A(Q)$) y vi) significa que R y Q son igualmente creíbles ($A(R) = A(Q)$).

Ahora, por el axioma 1, se tiene que d_A (donde $A \in \Omega$) y d_R (donde $R \subset I$) son opciones comparables y en consecuencia ocurre una de las tres siguientes afirmaciones

vii) $d_A < d_R$
viii) $d_A > d_R$
ix) $d_A \sim d_R$.

Estas relaciones sugieren que si se presenta vi) cualquier medida de la incertidumbre que se asocie al evento A debe ser menor que el área de R. Si ocurre viii) cualquier medida de la

incertidumbre que se asocie al evento A debe ser mayor que el área de R. En el caso de tener la condición a) el área de R cuantifica la incertidumbre del evento A. Esto sugiere la siguiente definición.

DEFINICIÓN 3.2.1. *Sea $A \subset \Omega$ un suceso incierto relevante. La probabilidad de A en las condiciones H, que se denota por $P(A|H)$ se define como el área de una región $R \subset I = [0,1] \times [0,1]$ tal que $d_A \sim d_R$, esto es, $\{c^*\ A,\ c,\ A^c\} \sim \{c^*\ R,\ c,\ R^c\}$.*

Por simplicidad de notación se expresará, cuando no haya confusión, con $P(A)$ la probabilidad condicional de A dado H, en lugar de $P(A|H)$.

Por el axioma 5, existe R tal que $d_A \sim d_R$, y por tanto la probabilidad está definida para cualquier suceso A.

Además de probar la existencia, es necesario verificar que la definición establecida da un valor único a la probabilidad de A. Si existen R' y R contenidas en I tales que

$d_A \sim d_{R'}$ y

$d_A \sim d_R$,

por el axioma 2 (transitividad) se tiene que $d_R \sim d_{R'}$ y por el axioma 4

$$Area(R') = Area(R) = P(A);$$

de esta forma se concluye que la probabilidad de un evento es única.

La definición sugiere una nueva forma para denotar una opción en general. Dicha notación es $l = \{c_1\ P_1, c_2\ P_2, \ldots, c_k\ P_k\}$, la cual conduce a la consecuencia c_1 con probabilidad P_1, c_2 con probabilidad P_2, ... , y c_k con probabilidad P_k.

La definición 3.2.1 satisface los llamados axiomas de Kolmogorov.

TEOREMA 3.2.1 *Cualesquiera que sean las condiciones de referencia H, la medida de probabilidad verifica*

(i) Para todo suceso A, $0 \le P(A|H) \le 1$ y $P(H|H) = 1$

(ii) Si A y B son dos sucesos excluyentes dado H,

$P(A \cup B\ H) = P(A|H) + P(B|H)$

(iii) Para todo par de sucesos A y B,

$P(A \cap B\ H) = P(A|H)P(B|A,H) = P(B|H)P(A|B,H)$.

La demostración de este teorema aparece en Bernardo (1981).

3.3 DEFINICIÓN DE UTILIDAD

Es necesario formalizar la definición de utilidad, ya que ésta mide las preferencias que el decisor tiene sobre las posibles consecuencias de tomar una u otra decisión. De manera análoga a la definición 3.2.1 surge la definición de utilidad, la cual está basada en los axiomas de coherencia.

La utilidad de cada consecuencia, debido a su construcción, será un número del intervalo [0,1], y las utilidades de las consecuencias extremas c^*, c_* serán 1 y 0 respectivamente.

Sea $c \in C$, entonces $c=\{c \quad \Omega\} \in D_3$. Por el axioma de la medición precisa, existe $R \subset I$ tal que $\{c \quad \Omega\} \sim \{c^* \quad R, \ c_* \quad R^c\}$. Esta relación permite definir una utilidad para cada c.

DEFINICIÓN 3.3.1. *Sea $c \in C$, la utilidad canónica de la consecuencia c que se denota por $u(c)$ se define como el área de una región $R \subset I=[0,1]^2$, tal que $\{c \quad \Omega\} \sim \{c^* \quad R, \ c_* \quad R^c\}$. De esta forma, para toda consecuencia c se tiene que $c \sim \{c^* \quad u(c), \ c_* \quad 1-u(c)\}$, esto debido a que $Area(R)=u(c)$.*

Para garantizar que $u(c)$ define una función de utilidad, es necesario demostrar que si $c_1 < c_2$, entonces $u(c_1) < u(c_2)$. Esto se demostrará por contradicción.

Si $c_1 < c_2$, entonces $\{c_1 \quad \Omega\} < \{c_2 \quad \Omega\}$,

por la definición de utilidad y el axioma 3 (sustitución), se tiene que

$$\{c^* \quad R, \ c_* \quad R^c\} < \{c^* \quad Q, \ c_* \quad Q^c\} \qquad (1)$$

donde R y Q son tales que $u(c_1)=Area(R)$ y $u(c_2)=Area(Q)$.

Por otro lado, si $u(c_1) \geq u(c_2)$, implica que $Area(R) \geq Area(Q)$, y como consecuencia del experimento auxiliar (axioma 4) se tiene que

$$\{c^* \quad R, \ c_* \quad R^c\} \geq \{c^* \quad Q, \ c_* \quad Q^c\},$$

lo que contradice la relación (1) y por lo tanto $u(c_1) < u(c_2)$.

La existencia y unicidad de la utilidad son también resultado de los axiomas de coherencia, y su demostración se lleva a cabo de manera análoga a lo hecho para la definición de probabilidad.

3.4 PRINCIPIO DE UTILIDAD (PÉRDIDA) ESPERADA MÁXIMA (MÍNIMA)

En las dos secciones anteriores se definieron mediante los axiomas de coherencia los conceptos de probabilidad y utilidad. Estos dos conceptos permiten al decisor asignar una medida a la verosimilitud de los sucesos inciertos y proporcionar también una medida de sus preferencias entre las diferentes consecuencias.

La intención ahora, es poder asignar un "número" a cada decisión, el cual mida de alguna manera las preferencias entre decisiones. Esto es, se establecerá un orden entre las decisiones. En esta sección se demuestra que el orden en D consistente con los axiomas de coherencia está dado por el principio de Utilidad (pérdida) esperada máxima (mínima).

En conclusión, si al plantearse un problema de decisión, el decisor está de acuerdo con los Axiomas de Coherencia, entonces existirá una medida de la probabilidad sobre los sucesos de interés y una utilidad sobre las consecuencias. Como resultado de lo anterior, se tiene que la manera racional de resolver el problema es buscar aquella opción que maximiza (minimiza) la utilidad (pérdida) esperada.

Por claridad, en este trabajo se presenta el caso en el que D y Ω son finitos, sin embargo los resultados se pueden generalizar a problemas donde tales elementos son infinitos. En Bernardo y Smith (1994) se discute el caso más general.

Sea $d_i=\{c_{i1}\ A_{i1}, c_{i2}\ A_{i2}, \ldots, c_{im_i}\ A_{im_i}\}$.

Por definición de utilidad canónica se tiene

$$\{c_{ij}\ \Omega\} \sim \{c^*\ \ R_{ij}, c_*\ \ R_{ij}^c\} \text{ con } R_{ij} \subset I.$$

Por el axioma de sustitución (axioma 3) aplicado sobre las consecuencias, se tiene

$$d_i \sim \{\{c^*\ R_{i1}, c_*\ R_{i1}^c\} \mid A_{i1}, \{c^*\ R_{i2}, c_*\ R_{i2}^c\} \mid A_{i2}, \ldots, \{c^*\ R_{im_i}, c_*\ R_{im_i}^c\} \mid A_{im_i}\}.$$

Entonces

$$d_i \sim \{c^*\ R_{i1} \cap A_{i1}, c_*\ R_{i1}^c \cap A_{i1}, c^*\ R_{i2} \cap A_{i2}, c_*\ R_{i2}^c \cap A_{i2}, \ldots, c^*\ R_{im_i} \cap A_{im_i}, c_*\ R_{im_i}^C \cap A_{im_i}\},$$

por lo que finalmente

$$d_i \sim \{c^* \quad \bigcup_{j=1}^{m_i} R_{ij} \cap A_{ij},\ c_* \quad \bigcup_{j=1}^{m_i} R_{ij}^c \cap A_{ij}\} \qquad (1)$$

$$i = 1,2,...,k$$

Ahora si $d_1 < d_2$, entonces por (1)

$$\{c^* \quad \bigcup_{j=1}^{m_1} R_{1j} \cap A_{1j},\ c_* \quad \bigcup_{j=1}^{m_1} R_{1j}^c \cap A_{1j}\} < \{c^* \quad \bigcup_{j=1}^{m_2} R_{2j} \cap A_{2j},\ c_* \quad \bigcup_{j=1}^{m_2} R_{2j}^c \cap A_{2j}\} \qquad (2)$$

Por la definición de probabilidad (3.2.1) ocurre (2) si y sólo si

$$P(\bigcup_{j=1}^{m_1} R_{1j} \cap A_{1j}) < P(\bigcup_{j=1}^{m_2} R_{2j} \cap A_{2j}) \qquad (3)$$

Como los eventos A_{ij} son excluyentes, los elementos en la unión también lo son, y por lo tanto, por el teorema 3.2.1 se tiene

$$P(\bigcup_{j=1}^{m_i}(A_{ij} \cap R_{ij})) = \sum_{j=1}^{m_i} P(A_{ij} \cap R_{ij}). \qquad (4)$$

Por el axioma 4 (experimento auxiliar) y el teorema 3.2.1

$$\sum_{j=1}^{m_i} P(A_{ij} \cap R_{ij}) = \sum_{j=1}^{m_i} P(A_{ij})P(R_{ij}). \qquad (5)$$

Al sustituir (5) en (4) se tiene

$$P(\bigcup_{j=1}^{m_i}(A_{ij} \cap R_{ij})) = \sum_{j=1}^{m_i} P(A_{ij})P(R_{ij}).$$

Además por el axioma 4 se tiene que $P(R_{ij}) = Area(R_{ij})$.

Por la definición de utilidad, se tiene que

$$Area(R_{ij}) = u(c_{ij}), \text{ donde } c_{ij} \sim \{c^* \quad R_{ij},\ c_* \quad R_{ij}^c\}.$$

Por lo tanto, al sustituir en (3)

$$\sum_{j=1}^{m_1} P(A_{1j})u(c_{1j}) < \sum_{j=1}^{m_2} P(A_{2j})u(c_{2j}).$$

De ahí que

$$d_1 < d_2 \Leftrightarrow \sum_{j=1}^{m_1} P(A_{1j})u(c_{1j}) < \sum_{j=1}^{m_2} P(A_{2j})u(c_{2j}),$$

esto es, $d_1 < d_2$ si y sólo si la utilidad esperada de d_1 es menor a la utilidad esperada de d_2. Entonces, por el axioma 2 (transitividad), la solución al problema será aquella decisión que *maximice la utilidad esperada.*

De lo anterior se concluye que el criterio para resolver un problema de decisión consistente con los axiomas de coherencia es el conocido como *Principio de utilidad (pérdida) esperada máxima (mínima)*. La decisión obtenida por este criterio es la referida como *Decisión de Bayes*. Esto se formaliza en el siguiente teorema.

TEOREMA 3.4.1 *(Principio de la utilidad esperada máxima (P.U.E.M.) o Criterio de Bayes) Considérese un problema de decisión definido por* $(D,\Omega,C,\prec)$, *con*

$$D=\{d_1,d_2,\ldots,d_k\} \text{ y } d_i=\{c_{i1}\ \omega_{i1}, c_{i2}\ \omega_{i2}, \ldots, c_{im_i}\ \omega_{im_i}\}.$$

Sea $P(\omega_{ij}|d_i,H)$ *la probabilidad de que suceda* ω_{ij} *si se elige* d_i *en las condiciones* H *y sea* $u(c_{ij})$ *la utilidad de la consecuencia que a ello da lugar. Entonces la utilidad esperada de la decisión* d_i *es*

$$E_p(U(d_i,\omega))=\sum_{j=1}^{m_i} u(c_{ij})P(\omega_{ij}|d_i,H) \qquad (1)$$

y la decisión óptima es aquella d^* *tal que*

$$E_p(U(d^*,\omega))=max_D E_p(U(d,\omega)).$$

Si Ω *es continuo y están determinadas* $P(\omega)$ *y* $U(d,\omega)$, *la decisión óptima es aquella* d^* *que maximice* $E_p(U(d,\omega))$, *donde*

$$E_p(U(d,\omega))=\int_{\Omega} U(d,\omega)P(\omega)d\omega .$$

La decisión proporcionada por el P.U.E.M. se conoce como Decisión de Bayes y $E_p(U(d^*,\omega))$ es conocido como Valor de Bayes.

El análisis de un problema de decisión puede llevarse a cabo en términos de una función de pérdida, esto es, en lugar de medir las preferencias de un decisor mediante utilidades, éstas se miden a través de pérdidas. El criterio de maximizar la utilidad esperada, se convierte entonces en el criterio que minimiza la pérdida esperada.

3.5 TEORÍA DE LA UTILIDAD

Se ha demostrado que la manera racional de resolver un problema de decisión es a través del principio de la Utilidad (pérdida) esperada máxima (mínima). Los axiomas garantizan

la existencia de una función de utilidad y una medida de probabilidad definida sobre Ω, por lo que el problema de decisión queda representado por $(D, \Omega, P(\omega), U(d\omega))$. Para implantar en la práctica el Principio de la Utilidad Esperada Máxima es necesario construir (exhibir explícitamente) dicha función y dicha medida, de tal manera que se expresen las preferencias y el conocimiento que tenga el decisor del problema. En esta sección se discutirá lo referente a la construcción de la función de utilidad y en el siguiente capítulo la asignación de la medida de probabilidad.

Asignar una utilidad a las consecuencias no es un proceso fácil, ya que generalmente se presentan dos clases de problemas:

- El valor de las consecuencia no tiene una escala obvia de medida. Por ejemplo, es difícil medir el valor del prestigio, de la reputación, del tiempo, etc.

- Aún cuando las consecuencias están expresadas en una escala numérica la escala no refleja el verdadero valor de tomar una decisión. Por ejemplo, tratándose de dinero, en algunos casos resulta ser totalmente inadecuado plantear: $u(c)=c$, o en general $u(c)=ac+b$. La utilidad del dinero será estudiada en la Subsección 3.5.1.

Para llevar a cabo la evaluación de las consecuencias se plantean dos algoritmos, los cuales están encaminados básicamente a evaluar consecuencias en los casos en los que la escala no es obvia o no existe la suficiente homogeneidad entre las consecuencias para poder determinar una unidad común. Ambos algoritmos están basados en la definición de la utilidad canónica (Definición. 3.3.1), por lo cual, se establecen los valores de las utilidades a través de una comparación directa entre opciones.

ALGORITMO 1 PARA LA CONSTRUCCIÓN DE LA FUNCIÓN DE UTILIDAD.

PASO 1: Se determinan c^* y c_* (las consecuencias extremas del Axioma 1) y se establece
$u(c^*)=1$ y $u(c_*)=0$

PASO 2: Para c_1, tal que $c_* \le c_1 \le c^*$ se busca α_1 tal que
$\{c_1 \quad \Omega\} \sim \{c^* \quad A, c_* \quad A^c\}$ donde $P(A)=\alpha_1$.
Con lo que se define
$u(c_1)=u(c^*)P(A)+u(c_*)P(A^c)$

por lo tanto,

$u(c_1)=(1)\alpha_1+(0)(1-\alpha_1)=\alpha_1$

PASO 3: Para c_2, tal que $c_1 \le c_2 \le c^*$, se busca α_2 tal que

$\{c_2 \quad \Omega\} \sim \{c^* \quad B,\ c_1 \quad B^c\}$, donde $P(B)=\alpha_2$,

con lo que

$u(c_2)=u(c^*)P(B)+u(c_1)P(B^c)=\alpha_2+\alpha_1(1-\alpha_2)=\alpha_1+\alpha_2(1-\alpha_1)$.

$u(c_2)=\alpha_1+\alpha_2(1-\alpha_1)$.

PASO 4: Para c_3, tal que $c_* \le c_3 \le c_1$, se busca α_3 tal que

$\{c_3 \quad \Omega\} \sim \{c_1 \quad F,\ c_* \quad F^c\}$, donde $P(F)=\alpha_3$,

con lo que

$u(c_3)=u(c_1)P(F)+u(c_*)P(F^c)=\alpha_1\alpha_3+0(1-\alpha_2)=\alpha_1\alpha_3$.

PASO 5: Hacer verificaciones periódicas de la consistencia en la construcción de la función de utilidad. Esta verificación se hace comparando nuevas combinaciones de consecuencias cuyas utilidades han sido encontradas por medio de la técnica descrita.

Por ejemplo, como $c_* \le c_1 \le c_2$ se busca α' tal que:

$\{c_1 \quad \Omega\} \sim \{c_2 \quad G,\ c_* \quad G^c\}$ donde $P(G)=\alpha'$,

con lo que se define:

$u(c_1)=u(c_2)P(G)+u(c_*)P(G^c)$

$u(c_1)=u(c_2)\alpha'+(0)(1-\alpha')=u(c_2)\alpha'$.

Por otro lado se sabe que $u(c_1)=\alpha_1$, por lo que

$\alpha_1 \cong \alpha'(\alpha_1+\alpha_2(1-\alpha_1))$.

De aquí que $\alpha' \cong \dfrac{\alpha_1}{\alpha_1+\alpha_2(1-\alpha_1)}$. Si esto no se cumple, habrá que revisar las asignaciones anteriores hasta encontrar relaciones "aproximadamente" consistentes.

Si se tienen cuatro consecuencias, por ejemplo $C=\{c_*, c_1, c_2, c^*\}$ donde $c_* < c_1 < c_2 < c^*$, es necesario hacer 2 verificaciones de consistencia, una para c_1 y otra para c_2.

Si se tienen 5 consecuencias $C=\{c_*, c_1, c_2, c_3, c^*\}$, entonces será necesario hacer 7 verificaciones: 2 para c_1, 3 para c_2 y 2 para c_3.

Como es evidente, el número de verificaciones que debe hacerse crece a medida que crece la cardinalidad de C. En general, si se tiene $C=\{c_*, c_1, c_2, ..., c_n, c^*\}$, el número de verificaciones necesarias para cada c_k es $k(n-k+1)-1$, lo que lleva a un número total de verificaciones de $\sum_{k=1}^{n} k(n-k+1)-n = \frac{n(n-1)(n+4)}{6}$.

Este algoritmo para asignar la utilidad a cada consecuencia es sumamente complicado y engorroso por el número de verificaciones que se deben realizar.

Existe otro método equivalente para la asignación de utilidades; sin embargo, dicho algoritmo sólo es utilizable en los casos en donde el conjunto C es numérico y continuo, como es el caso en donde las utilidades están dadas en unidades monetarias.

ALGORITMO 2 PARA LA CONSTRUCCIÓN DE LA FUNCIÓN DE UTILIDAD.

PASO 1: Se determinan c^* y c_* (las consecuencias extremas del Axioma 1) y se establece $u(c^*)=1$ y $u(c_*)=0$.

PASO 2: Encontrar c_1, tal que
$\{c_1 \ \ \Omega\} \sim \{c^* \ \ \frac{1}{2}, c_* \ \ \frac{1}{2}\}$, de tal manera que
$u(c_1)=u(c^*)\frac{1}{2}+u(c_*)\frac{1}{2}$.
Por lo tanto,
$u(c_1)=\frac{1}{2}$

PASO 3: Encontrar c_2, tal que
$\{c_2 \ \ \Omega\} \sim \{c_1 \ \ \frac{1}{2}, c_* \ \ \frac{1}{2}\}$, de tal manera que
$u(c_2)=u(c_1)\frac{1}{2}+u(c_*)=\frac{1}{2}(\frac{1}{2})+0(\frac{1}{2})=\frac{1}{4}$.

PASO 4: Encontrar c_3, tal que
$\{c_3 \ \ \Omega\} \sim \{c^* \ \ \frac{1}{2}, c_1 \ \ \frac{1}{2}\}$, de tal manera
$u(c_3)=u(c^*)\frac{1}{2}+u(c_1)\frac{1}{2}=1(\frac{1}{2})+\frac{1}{2}(\frac{1}{2})=\frac{3}{4}$.

PASO 5: Continuar el proceso anterior encontrando los puntos con utilidades $\frac{i}{2^n}$ y verificar la consistencia.

Por ejemplo, si se tienen cinco consecuencias, tales que $u(c_*)=0$, $u(c_2)=\frac{1}{4}$, $u(c_1)=\frac{1}{2}$, $u(c_3)=\frac{3}{4}$ y $u(c^*)=1$.

Buscar c_4 tal que $\{c_4 \ \ \Omega\} \sim \{c_3 \ \ \frac{1}{2}, c_2 \ \ \frac{1}{2}\}$, de tal manera que

$$u(c_4)=u(c_3)\tfrac{1}{2}+u(c_2)\tfrac{1}{2}=\tfrac{3}{4}(\tfrac{1}{2})+\tfrac{1}{2}(\tfrac{1}{4})=\tfrac{1}{2}.$$

Por ser u monótona, c_4 tiene que ser c_1.

Si se renombran las consecuencias de tal manera que se tenga $c_* < c_1 < c_2 <,...,< c_n < c^*$, y se obtiene que

$$u(c_{i+1})-u(c_i)=u(c_{j+1})-u(c_j) \text{ para } i,j = 1,2,...,n-1, \qquad (1)$$

entonces n será impar y el número de verificaciones que se deberá hacer para cada c_k (N_k=número de verificaciones necesarias para c_k) es el $min\{k, n-k+1\}-1=min\{k-1, n-k\}$, lo que lleva a un número total de verificaciones de:

$$\sum_{k=1}^{n} N_K = \sum_{k=1}^{n} min\{k-1, n-k\} = \sum_{k=1}^{\frac{n}{2}+1}(k-1)+\sum_{k=\frac{n}{2}+1}^{n}(n-k)=\left(\frac{n-1}{2}\right)^2. \qquad (2)$$

En caso de no cumplirse la condición (1), se tendrá que

$$\sum_{k=1}^{n} N_K \le \left[\left(\frac{n-1}{2}\right)^2\right].$$

Es evidente que este algoritmo (2) reduce el número de verificaciones con respecto al algoritmo 1, ya que es inmediato demostrar que

$$\left(\frac{n-1}{2}\right)^2 < \frac{n(n-1)(n+4)}{6}, \text{ para } n>1.$$

En el caso de $n = 1$, el número de verificaciones es 0 para ambos algoritmos.

Los dos algoritmos generan una función de utilidad cuyo rango es el $[0,1]$. Sin embargo, se puede realizar un cambio de escala si se conoce la utilidad de las consecuencias extremas en esta nueva escala. Si $u(c)$ es una función de utilidad canónica, entonces $u'(c)=au(c)+b$ $(a>0)$ es también una función de utilidad (cuyo rango es $[b,a+b]$) que lleva a la misma solución cuando el problema es resuelto vía el P.U.E.M.

Es muy importante hacer notar que cuando C es finito y no muy "grande", se trata de asignar utilidades a todas las consecuencias. Sin embargo, cuando C es "grande" o infinito, la construcción de u se hace mucho más difícil. En esos casos (y siendo C un conjunto de números), lo conveniente es calcular la utilidad para algunas consecuencias por medio de uno de los dos algoritmos, y para las consecuencias restantes, estimar el valor correspondiente a partir de los puntos ya evaluados.

En el caso de C infinito y numérico, la solución es igual a lo ya planteado, se asignan utilidades a algunas consecuencias, hasta que queda claro quién es la curva de utilidad, esto es, hasta que se vislumbra el comportamiento de la función de utilidad. Una situación común que se presenta, es aquella en la que C es un intervalo de la recta real. El ejemplo más importante de esto es cuando las consecuencias son monetarias, y por lo tanto U será una función de utilidad del dinero (Subsección 3.5.1).

Ejemplo 3.5.1: Se desea determinar una posible función de utilidad de un conjunto de consecuencias C donde $C=[0,5000]$.
Se utilizará el algoritmo número 2 para calcular la función de utilidad para algunos puntos del conjunto C.

PASO 1: Se sabe que $c_*=0$ y $c^*=5000$, entonces,
$u(c^*)=u(5000)=1$ y $u(c_*)=u(0)=0$.

PASO 2: Se busca c_1, tal que
$\{c_1 \quad \Omega\} \sim \{5000 \quad \frac{1}{2}, 0 \quad \frac{1}{2}\}$.
Se toma por ejemplo 2500 y se pregunta cuál es la relación de preferencia (<,>,~) entre las loterías
$l_1=\{2500\ \Omega\}$ y $l_2=\{5000\ \frac{1}{2}, 0\ \frac{1}{2}\}$.

Si la relación es ~, entonces $c_1=2500$.
Si la relación es >, entonces $c_1<2500$.

Si la relación es <, entonces $c_1 > 2500$.

Supóngase que $c_1 < 2500$, entonces se toma una cantidad entre 0 y 2500, por ejemplo, 1250 y se pregunta nuevamente cuál es la relación entre las loterías l_3={1250 Ω} y l_2={2500 $\frac{1}{2}$, 0 $\frac{1}{2}$}.
Supóngase que ocurre <, entonces $c_1 > 1250$ y que después de algunas comparaciones se llega a que $c_1 = 1500$, por lo que $u(1500) = \frac{1}{2}$.

PASO 3: Buscar ahora c_2, tal que
{c_2 Ω} ~ {1500 $\frac{1}{2}$, 0 $\frac{1}{2}$}. Supóngase que resulta que c_2=700, por lo que $u(c_2) = u(700) = \frac{1}{4}$.

PASO 4: Se busca c_3, tal que
{c_3 Ω} ~ {5000 $\frac{1}{2}$, 1500 $\frac{1}{2}$}. Supóngase que c_3=3000, por lo que $u(c_3) = u(3000) = \frac{3}{4}$.

PASO 5: Verificar la consistencia. Se busca c_4 tal que
{c_4 Ω} ~ {3000 $\frac{1}{2}$, 700 $\frac{1}{2}$}.

Si c_4 es por ejemplo, 1700, entonces hay una inconsistencia, ya que c_4 debe ser igual a c_1 y c_1 es igual a 1500.

Si se supone que c_1 y c_2 están bien, entonces se busca un valor para c_3 que haga c_4=c_1, esto es,

{1500 Ω} ~ {c_3 $\frac{1}{2}$, 700 $\frac{1}{2}$}.
Supóngase que c_3=2500.

Los valores asignados se grafican (Fig. 3.5.1), lo que permite tener cierto conocimiento de la función de utilidad.

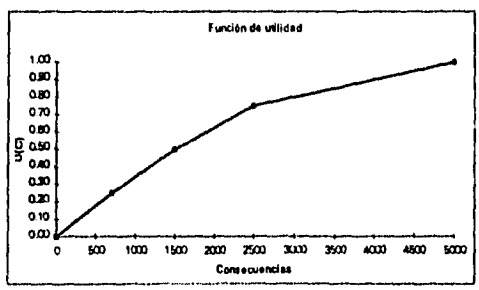

Figura 3.5.1

Existen problemas de decisión en los cuales las consecuencias están dadas como vectores, esto es $c = (c_1, c_2, ..., c_m)$. Casos como éste son frecuentes en situaciones médicas, ya que por ejemplo, algunos tratamientos podrían tener efectos sobre "el paciente" además de tenerlos sobre la "enfermedad". La construcción de $u(c)$, por lo tanto, resulta muy difícil. Debido a esta dificultad es frecuente aceptar que

$$u(c) = \sum_{i=1}^{m} k_i u_i(c_i)$$

El problema se reduce entonces a encontrar utilidades unidimensionales para cada tipo de consecuencia y determinar las constantes k_i para poder realizar la combinación lineal de las utilidades. Estas constantes pueden ser encontradas de la misma manera que se ha realizado la construcción de la función de utilidad. Esta simplificación no siempre es razonable, especialmente cuando hay una interacción significativa entre las coordenadas de c. Por ejemplo, dos reacciones de una droga pueden ser casi inofensivas cuando se presentan por separado, pero tener un efecto realmente peligroso cuando ocurren juntas. Un modelo probablemente satisfactorio para este caso es:

$$u(c) = \sum_{i=1}^{m} k_i u_i(c_i) + \sum_{i=1}^{m} \sum_{j=i+1}^{m-1} k_{ij} u_i(c_i) u_j(c_j).$$

3.5.1 UTILIDAD DEL DINERO

En muchos problemas de decisión es natural plantear el valor de las consecuencias en términos de dinero, tal es el caso de lo concerniente a problemas económicos y financieros. Además de estos, existe otra clase de problemas cuya apreciación de las consecuencias no parece mostrar una equivalencia monetaria, sin embargo, a tales consecuencias se les puede asignar un valor monetario. Se podría, por ejemplo, no estar de acuerdo en que el morir pudiera tener consecuencias monetarias, no obstante, se suscribe diariamente una cantidad considerable de seguros de vida. De igual manera, podría objetarse la asignación económica al placer de asistir a un concierto, sin embargo, se paga un boleto para tener acceso a ello.

Lo anterior sugiere que existe un cierto equivalente económico para diversas clases de consecuencias, por lo cual, el problema de asignar utilidades se verá muchas veces reducido a determinar una función que describa la "utilidad del dinero".

Para poder asignar correctamente una función de utilidad monetaria se debe tener cierto cuidado, ya que por ejemplo, puede resultar erróneo considerar la utilidad de una consecuencia directamente igual a la cantidad monetaria o como función lineal de ella. Esto se ilustra en el siguiente ejemplo.

Ejemplo 3.5.1.1 Suponga que un estudiante debe decidir entre aceptar o no un trabajo con un sueldo mensual de $\$K_1$. Sus ingresos mensuales son de $\$I$ y valora en $\$c$ el esfuerzo que le demanda dicho trabajo.

Intuitivamente se observa que si I es grande respecto a K_1, el estudiante no estará muy dispuesto a aceptar el trabajo. Si I es pequeño con respecto a K_1 es muy factible que decida trabajar.

Sean las dos posibles decisiones

d_1= Aceptar, d_2= No aceptar, entonces

$d_1 < d_2$ si y sólo si $u(I + K_1 - c) < u(I)$
$d_2 < d_1$ si y sólo si $u(I) < u(I + K_1 - c)$.

Si $u(d) = ad + b$,

entonces, se tiene que

$d_1 < d_2$ si y sólo si $a(I + K_1 - c) + b < aI + b$
$d_2 < d_1$ si y sólo si $aI + b < a(I + K_1 - c) + b$.

Por lo que

$d_1 < d_2$ si y sólo si $a(K_1 - c) < 0$
$d_2 < d_1$ si y sólo si $a(K_1 - c) > 0$.

Estas relaciones indican que se puede discriminar entre d_1 y d_2 sin importar cuánto vale I, lo cual resulta absurdo.

Este ejemplo, sugiere que la función de utilidad del dinero debe tomar en cuenta el capital con el que se dispone, ya que no es igualmente importante recibir $1,000 cuando se posee una

fortuna que recibirlos cuando no se dispone de absolutamente nada. Esto indica que la función de utilidad debe ser de la forma

$$u(c)=u(c+k)-u(k),$$

donde k es el capital inicial. Esta diferencia recibe el nombre de *Utilidad marginal del dinero*.

El incremento de la utilidad producto de una determinada ganancia c es positivo, pero el efecto de éste es menor en cuanto más dinero se tenga. Esto es, la diferencia $u(c+k)-u(k)$ decrece a medida que k crece.

Otro ejemplo que permite inferir algunas propiedades razonables que debe tener una función de utilidad del dinero, es la llamada Paradoja de San Petersburgo.

Ejemplo 3.5.1.2: Suponga que se dispone de una moneda "honesta" (P(Águila)=P(Sol)=$\frac{1}{2}$) y se plantea el juego siguiente: La moneda será lanzada repetidamente hasta que aparezca la primera águila, en ese momento el juego finaliza. Entonces se pagará al jugador como recompensa $X=2^k$, donde k es el número de lanzamientos requeridos para obtener águila por primera vez. ¿Cuál es el valor esperado del juego? o ¿Cuánto debería estar dispuesto a pagar el jugador para participar?

Si se considera que $u(c)=c$. Entonces,

$$E\big(U(juego)\big)=\sum_{k=1}^{\infty}u(2^k)P(X=2^k)=\sum_{k=1}^{\infty}\left(2^k\right)\left(\frac{1}{2^k}\right)=\sum_{k=1}^{\infty}1=\infty.$$

Esto lleva a concluir que uno estaría dispuesto a pagar cualquier cantidad de dinero por grande que sea para participar en el juego, ya que como recompensa se espera una utilidad infinita. Paradójicamente, resulta inmediato observar que sólo se está dispuesto a pagar una cantidad muy pequeña por participar, ya que simplemente, para obtener más de \$16 se requiere que ocurra un evento con probabilidad $\frac{1}{16}=0.0625$. Es evidente que a medida que la utilidad crece, la probabilidad de obtenerla decrece.

La teoría de la utilidad corrige esta paradoja suponiendo que el verdadero valor de jugar es

$$\sum_{k=1}^{\infty}u(2^k-c_0)(\tfrac{1}{2^k}),\ \text{donde}\ u(0)=0\ \text{y}$$

$c_0>0$ es el costo por participar en dicho juego.

Considérese

$$u(2^k - c_0) = \begin{cases} \log(2^k - c_0) & si\ 2^k - c_0 > 0 \\ 0 & si\ 2^k - c_0 = 0 \\ -\log(c_0 - 2^k) & si\ 2^k - c_0 < 0. \end{cases}$$

De aquí que

$$E(U(juego)) = \sum_{k=k_0+1}^{\infty} \log(2^k - c_0)\left(\frac{1}{2^k}\right) - \sum_{k=1}^{k_0-1} \log(c_0 - 2^k)\left(\frac{1}{2^k}\right) < \infty,$$

$$\text{donde } k_0 = \left[\frac{\log c_0}{\log 2}\right].$$

En general, utilizar una función de utilidad cóncava resuelve la paradoja. Además se sabe que la concavidad de la curva representa un decremento en la utilidad marginal a medida que el dinero crece.

También es razonable pedir que la utilidad sea una función acotada dado que en la práctica no se puede aspirar a obtener una ganancia infinita.

Del análisis de otros problemas de asignación de utilidad surgen características como las siguientes. Sea $u(c)$ una función de utilidad donde c está expresado en unidades monetarias, entonces

i) $u(c)$ es *aproximadamente* ***lineal*** *para valores pequeños de* c. Como una regla general cuando c es pequeño comparado con los ingresos del decisor se considera $u(c)$ casi lineal, ya que en este caso la consecuencia no tiene un gran impacto en el decisor. Esto es, casi se puede ser indiferente a la utilidad o pérdida que produce una consecuencia de esta naturaleza.

ii) $u(c)$ *es cóncava para* $c > 0$. Esta característica se debe básicamente a que la utilidad marginal del dinero es típicamente decreciente. La utilidad marginal puede ser pensada como $u'(c) = \frac{du(c)}{dc}$. Si $u'(c)$ es decreciente, entonces $u''(c) = \frac{du'(c)}{dc} < 0$, lo cual implica que $u(c)$ es una función cóncava.

iii) $u(c)$ *es acotada*. Esto es, existe $B < \infty$ tal que para todo $c \in C$ $|u(c)| \leq B$.

iv) $u(c)$ con frecuencia es diferente para $c > 0$ y $c < 0$ (tanto en signo como en forma). Esto se debe a que muchas veces el impacto que produce una pérdida es diferente al que puede producir una ganancia. Por esta razón es deseable construir $u(c)$ separadamente para $c > 0$ y $c < 0$.

Debido a que muchas funciones de utilidad de la vida real pueden ser perfectamente asociadas a curvas de ciertas funciones, surge el concepto de la *Teoría de la Utilidad Paramétrica.*

La Teoría de la Utilidad Paramétrica propone asignar algún modelo paramétrico que resulte conveniente para representar la utilidad cuando las consecuencias son numéricas. Una función muy utilizada es

$$u(c) = a\log(bc+1) \quad a,b > 0 \qquad (1)$$
con $u(0) = 0$.

Muchas veces, es preferible la aproximación paramétrica a realizar un proceso laborioso para la construcción de u. De acuerdo con esta teoría sólo se deben elegir a y b en la función (1), de tal manera que la utilidad quede razonablemente expresada. Esto se puede hacer obteniendo $u(c)$ para dos puntos (diferentes a $c = 0$) por medio de alguno de los dos algoritmos estudiados en la sección 3.5 y entonces resolver para a y b un par de ecuaciones.

La función de utilidad (1) es cóncava, además si la constante b se elige pequeña, la función resultará casi lineal para valores pequeños de c.

Otras familias paramétricas de funciones de utilidad cóncavas empleadas para evaluar consecuencias son:

i) $u(c) = \sqrt{c}$	(no acotada)
ii) $u_{a,b}(c) = (\sqrt{ac}+b)\, I_{[0,k]}(c)$	(acotada)
iii) $u(c) = \log(c)$	(no acotada)
iv) $u_a(c) = \log(ac) \quad a > 0$	(no acotada)
$u_a(c) = \log(ac) I_{[k_1,k_2]}(c) \quad a > 0$	(acotada)

v) $u_{a,b}(c)=\log(ac^b)=\log a+b\log c$	(no acotada)
vi) $u(c)=\log(c+b)$	(no acotada)
vii) $u_a(c)=a\sqrt{c}$	(no acotada)
viii) $u_{a,b}(c)=a\sqrt{c}+b$	(no acotada)
ix) $u(c)=1-\frac{1}{c}$	(acotada)
x) $u_{a,b}(c)=a+b\frac{1}{c}\quad b<0$	(acotada)
xi) $u(c)_{a,b}=(c+b)^a\quad 0<a<1,\ b>0$	(no acotada)
xii) $u(c)_{a,b}=(c+b)^{-a}\quad a,b>0$	(no acotada)

Otras funciones de utilidad son tratadas en Keeney y Raiffa (1976).

A partir de lo estudiado sobre la función de utilidad, es claro ver que no existe un procedimiento general aplicable a todo conjunto de consecuencias para determinar una función de utilidad. Sin embargo, las ideas básicas sobre lo que se debe hacer al asignar dicha función son fundamentalmente las mismas en los diferentes procedimientos. Estas ideas llevan a dividir el procedimiento en cinco pasos, aunque en la práctica esta división no siempre es muy clara. Tales pasos son:

1. *Preparar al decisor*: En esta parte se debe de explicar al decisor cómo funciona el paradigma de decisión, cuáles son los procedimientos, etc. Esto con la finalidad de tener al decisor preparado y consciente de la importancia que tiene manifestar adecuadamente sus preferencias entre las posibles consecuencias.

2. *Identificación de características cualitativas de la función de decisión*: En esta parte se exploran características de la función de utilidad tales como concavidad o convexidad y acotamiento. Esto es, durante esta fase es apropiado, por ejemplo, preguntar:
   Si c_i es más grande que c_j, ¿es siempre preferible c_i a c_j?, si la respuesta es sí, entonces la función u es monótona creciente.

3. *Identificación de características cuantitativas de la función de utilidad*: Después de que han sido determinadas las características cualitativas de la función de utilidad, es necesario valorarla cuantitativamente para algunos puntos de C.

Se puede evaluar la utilidad de los puntos por medio de los algoritmos descritos en la sección anterior.

4. *Elección de la función de utilidad*: En esta parte se analiza si existe una función de utilidad conocida que cumpla con las características tanto cualitativas como cuantitativas determinadas previamente por el decisor. Si tal función existe ¿cómo determinarla? y si no existe ¿cómo obtener un conjunto consistente de valores para constituir con éste la función de utilidad?

   Un método para contestar la primera pregunta consiste en encontrar una familia paramétrica de funciones de utilidad (como las listadas anteriormente) que cumpla con las características cualitativas expresadas por el decisor y posteriormente calcular los parámetros particulares a través de las condiciones cuantitativas del decisor.

   En caso de que no exista una familia paramétrica adecuada, se deberá calcular mediante los algoritmos 1 y 2 el valor de la utilidad de un número suficiente de puntos, de tal manera que se pueda vislumbrar el comportamiento de la función.

3.5.2 RIESGO

Se podría pensar que el de P.U.E.M (Bayes) no toma en cuenta el riesgo ya que por ejemplo, si se tuviera un problema de decisión planteado de la siguiente manera

d_1 = Ganar $ 100 con probabilidad de 1,
d_2 = Ganar $ 200 con probabilidad de 0.5 o 0 con la misma probabilidad,
d_3 = Ganar $ 1,000 con probabilidad de 0.1 o 0 con probabilidad de 0.9,
d_4 = Ganar $ 200 con probabilidad de 0.9 o perder 800 con probabilidad de 0.1,

el P.U.E.M. no discrimina entre estas decisiones debido a que el valor esperado de todas es $100. Sin embargo, es claro que para un decisor generalmente sí existirá gran diferencia entre ellas. Por ejemplo, un decisor al que le gusta correr riesgos, preferiría quizás d_3; en cambio, otro decisor que prefiere lo seguro optaría por d_1. Aquel decisor que prefiere d_3 podría ser considerado un "tomador de riesgos", mientras que el decisor que se queda con d_1, es un decisor "adverso a riesgos".

Esta diferencias entre decisiones y decisores debe estar expresada en la función de utilidad, que como consecuencia tendrá un efecto en el P.U.E.M.

DEFINICIÓN 3.5.2.1 *Un decisor se considera adverso al riesgo si prefiere la consecuencia esperada (valor esperado) con probabilidad 1 de una lotería a la lotería misma. En otras palabras, un decisor se considera adverso al riesgo si para cualquier lotería se tiene que*

$$u[E(c)] > E[u(c)].$$

TEOREMA 3.5.2.1 *Un tomador de decisiones se considera adverso al riesgo si y sólo si su función de utilidad es cóncava.*

Demostración: Considere una lotería que asigna c_1 con probabilidad p o c_2 con probabilidad $1-p$, $0 < p < 1$. La consecuencia esperada es $\bar{c} = pc_1 + (1-p)c_2$, por la definición 3.5.2.1, se tiene que

$$u[pc_1 + (1-p)c_2] > pu(c_1) + (1-p)u(c_2),$$

por lo tanto, u es un función cóncava.

Para probar la implicación contraria, considere que una lotería proporciona c_i con probabilidad p_i, para $i = 1,...,m$, donde $p_i \neq 1$. Ya que la función es estrictamente cóncava, se sabe que

$$u[\sum_{i=1}^{m} p_i c_i] > \sum_{i=1}^{m} p_i u(c_i).$$

Dada esta desigualdad, por la definición 3.5.2.1, se tiene que esta función de utilidad implica aversión al riesgo■

DEFINICIÓN 3.5.2.2 *Un decisor se considera tomador de riesgo si prefiere cualquier lotería a la consecuencias esperada (valor esperado) de la misma. En otras palabras, un decisor se considera tomador de riesgo si para cualquier lotería se tiene que*

$$u[E(c)] < E[u(c)].$$

TEOREMA 3.5.2.2 *Un tomador de decisiones se considera tomador de riesgos si y sólo si su función de utilidad es convexa.*

La demostración es análoga a la del teorema 3.5.2.1.■

CAPÍTULO 4

PROBLEMAS DE DECISIÓN ESTADÍSTICA

Los capítulos anteriores están dedicados al estudio de la Teoría de la Decisión; en ellos se abordan desde ideas básicas de la Decisión, hasta la fundamentación axiomática de la misma.

En este capítulo, en primera instancia, se presentan los problemas de inferencia estadística como problemas de decisión. Durante este proceso se verá involucrado de manera importante el Teorema de Bayes, el cual sirve para actualizar el conocimiento inicial mediante información muestral. El paradigma concerniente a esta manera de plantear y resolver problemas estadísticos da origen a la *Estadística Bayesiana*.

En el planteamiento de los problemas estadísticos como problemas de decisión nos enfrentamos al problema de determinar una función de densidad o de distribución que cuantifique el conocimiento inicial que el decisor tiene sobre el evento aleatorio. En este capítulo se estudian algunas formas de cuantificar la incertidumbre.

4.1 INFERENCIA ESTADÍSTICA EN EL MARCO DE LA TEORÍA DE LA DECISIÓN

Debido a su naturaleza, los problemas estadísticos pueden ser abordados como problemas de decisión con sus respectivos elementos $(D, \Theta, p(\theta), U(d,\vartheta))$. Generalmente, cuando se habla de problemas estadísticos se habla de funciones de pérdida y no de funciones de utilidad, tomando esto como convención, los problemas de decisión estadística aquí planteados estarán definidos a partir de este momento por $(D, \Theta, p(\theta), L(d,\vartheta))$.

Comúnmente Θ corresponde al espacio parametral, esto es, al conjunto de valores que puede tomar el parámetro desconocido θ de la función de distribución o de densidad, $f(X|\theta)$ (o $P(X|\theta)$), que describe el comportamiento de una variable o vector aleatorio X. Con $p(\theta)$ se representa la probabilidad o en el caso de ser θ continua la densidad asociada al evento de que el parámetro de interés tome el valor θ.

Típicamente la Teoría de la Decisión Estadística concierne a la toma de decisiones en presencia de conocimiento estadístico, el cual ayuda a eliminar la incertidumbre; es por esto que resulta natural que $p(\theta)$ involucre además del conocimiento inicial del decisor información dada por n experimentos, $\underline{X}=(X_1, X_2, ..., X_n)$. Cuando esto sucede, $p(\theta)$ es conocida como distribución *final* o *a posteriori* y se denota por $P(\theta|X_1, X_2, ..., X_n) = P(\theta|\underline{X})$. Por otra parte, la función $p(\theta)$ puede representar únicamente el conocimiento que sobre el parámetro tenga el decisor sin involucrar información muestral, en este caso $p(\theta)$ recibe el nombre de distribución *a priori* o *inicial* y se denota por $P(\theta)$.

Cuando se ha determinado la distribución a priori $P(\theta)$ y se conoce $f(\underline{X}|\theta)$ que es la función de densidad de $X_1, X_2, ..., X_n$ dado θ, entonces la distribución a posteriori $P(\theta|\underline{X})$ puede ser obtenida vía el *Teorema de Bayes*, mediante la siguiente igualdad

$$P(\theta|\underline{X}) = \frac{f(\underline{X}|\theta)P(\theta)}{P(\underline{X})}.$$

$P(\underline{X})$ no depende del parámetro desconocido θ ya que en el caso discreto se tiene $P(\underline{X}) = \sum_{\theta_j \in \Theta} f(\underline{X}|\theta_j)P(\theta_j)$ y en el caso continuo $P(\underline{X}) = \int_{\Theta} f(\underline{X}|\theta)P(\theta)d\theta$. Debido a esta propiedad, $P(\theta|\underline{X})$ vista como función de θ puede pensarse como

$$P(\theta|\underline{X}) \propto f(\underline{X}|\theta)P(\theta).$$

El tener presente que la distribución a posteriori es proporcional al producto de la función de verosimilitud por la distribución inicial puede, en algunos de los casos, facilitar la determinación de la final, ya que dicho producto representa al núcleo de la distribución a posteriori y si este núcleo corresponde a una función conocida, la determinación completa de la a posteriori es inmediata. Esto será estudiado a detalle en la sección 4.2.2.

Determinar la solución de un problema estadístico cuando éste ha sido planteado como problema de decisión es una tarea teóricamente sencilla, pues de acuerdo a lo estudiado en el capítulo 3 se tiene que la solución consistente con los axiomas de coherencia es la decisión de Bayes; por lo que la solución será aquella d^* que minimiza la pérdida esperada.

En caso de no contar con información muestral d^* es tal que

$$E_{P(\theta)}(L(d^*,\theta)) = min_D E_{P(\theta)}(L(d,\theta)) \qquad (1)$$

donde

$E_{P(\theta)}(L(d,\theta))=\int_{\Theta} L(d,\theta)P(\theta)d\theta$ para el caso continuo y

$E_{P(\theta)}(L(d,\theta))=\sum_{\theta_j\in\Theta} L(d,\theta_j)P(\theta_j)$ en el caso discreto.

Cuando si se cuenta con información muestral y por lo tanto con una distribución a posteriori, d^* es tal que

$$E_{P(\theta|\underline{X})}(L(d^*,\theta))=min_d E_{P(\theta|\underline{X})}(L(d,\theta)) \qquad (2)$$

donde

$E_{P(\theta|\underline{X})}(L(d,\theta))=\int_{\Theta} L(d,\theta)P(\theta|\underline{X})d\theta$ para el caso continuo y

$E_{P(\theta|\underline{X})}(L(d,\theta))=\sum_{\theta_j\in\Theta} L(d,\theta_j)P(\theta_j|\underline{X})$ en el caso discreto.

Ejemplo 4.1.1: Sea X una variable aleatoria con f.d.p.g. $f(X|\theta)$; suponga además que θ tiene una función de densidad $p(\theta)$ (inicial o final) cuyo primer y segundo momento son finitos. Se desea estimar puntualmente el parámetro θ considerando un pérdida cuadrática $L(d,\theta)=k(d-\theta)^2$ con $k>0$.

De acuerdo a la ecuación (1) se tiene que d^* es tal que $E_{p(\theta)}(L(d^*,\theta))=min_d E_{p(\theta)}(L(d,\theta))$. De aquí que

$$E_{p(\theta)}(L(d,\theta))=E_{p(\theta)}(k(d-\theta)^2)$$
$$=k(d^2-2dE_{p(\theta)}(\theta)+E_{p(\theta)}(\theta^2))$$
$$\frac{\partial E_{p(\theta)}(L(d,\theta))}{\partial d}=2k(d-E_{p(\theta)}(\theta))$$
$$\frac{\partial E_{p(\theta)}(L(d,\theta))}{\partial d}=0 \text{ si y sólo si } d=E_{p(\theta)}(\theta);$$

como $\frac{\partial^2 E_{p(\theta)}(L(d,\theta))}{\partial d^2}=2k>0$ se tiene un mínimo en $d=E_{p(\theta)}(\theta)$.

Por lo tanto

$$d^*=E_{p(\theta)}(\theta)$$

ESTA TESIS NO DEBE
SALIR DE LA BIBLIOTECA

es la solución con valor de Bayes igual a la $Var_{p(\theta)}(\theta)$; donde $E_{p(\theta)}(\theta)$ y $Var_{p(\theta)}(\theta)$ son respectivamente la media y la varianza de la función de densidad $p(\theta)$.

Ejemplo 4.1.2: En las condiciones del ejemplo 4.1.1 se desea estimar nuevamente el parámetro θ pero considerando ahora una función de pérdida $L(d,\theta)=k|d-\theta|$ con $k>0$.

$$E_{p(\theta)}(L|d,\theta|)=k\left(\int_{\theta<d}(d-\theta)p(\theta)d\theta+\int_{\theta>d}(\theta-d)p(\theta)d\theta\right)$$

$$=k\left(\left(d\int_{\theta<d}p(\theta)d\theta-d\int_{\theta>d}p(\theta)d\theta\right)-\left(\int_{\theta<d}\theta p(\theta)d\theta-\int_{\theta>d}\theta p(\theta)d\theta\right)\right)$$

$$=k\left(d-E_{p(\theta)}(\theta)-2d\int_{\theta>d}p(\theta)d\theta+2\int_{\theta>d}\theta p(\theta)d\theta\right)$$

Por el Teorema Fundamental del Cálculo

$$\frac{\partial E_{p(\theta)}(L|d,\theta|)}{\partial d}=k\left(1-2d(-p(d))-2\int_{\theta>d}p(\theta)d\theta-2dp(d)\right)$$

$$=k\left(1-2\int_{\theta>d}p(\theta)d\theta\right)$$

$$\frac{\partial E_{p(\theta)}(L|d,\theta|)}{\partial d}=0 \quad \text{si y sólo si} \quad \int_{\theta>d}p(\theta)d\theta=\frac{1}{2};$$

de aquí que d^* es la mediana de la función de densidad $p(\theta)$.

Como $\frac{\partial^2 E_{p(\theta)}(L|d^*,\theta|)}{\partial d^2}=2kp(d^*)>0$ se tiene que la solución d^* vía el criterio de Bayes es la mediana de la función de densidad $p(\theta)$.

En los ejemplos 4.1.1 y 4.1.2 se puede observar que el resultado no depende directamente de la forma de $f(x|\theta)$; lo único que se debe satisfacer es que $E_{p(\theta)}(\theta)<\infty$ y que $E_{p(\theta)}(\theta^2)<\infty$. Es importante también hacer notar que para los casos en los que se tiene una función de pérdida cuadrática y una pérdida de error absoluto, los estimadores de Bayes son respectivamente la media y la mediana, ya sea de la función de densidad final o inicial dependiendo de que se haya o no, incorporado datos muestrales. En caso de contar con dicha información muestral la distribución a posteriori se obtiene mediante el Teorema de Bayes.

Ejemplo 4.1.3: Sea $\underline{X}=(X_1, X_2, ..., X_n)$ una muestra aleatoria de una distribución Exponencial$(X|\theta)$. Suponga que $P(\theta)$ es una densidad Gamma, $Ga(\theta|\alpha,\beta)$. Se desea estimar puntualmente el parámetro θ considerando una función de pérdida $L(d,\theta)=k(d-\theta)^2$ con $k>0$.

Por lo expuesto en el ejemplo 4.1.1 se sabe que la solución, cuando se incorporan datos, será $d^* = E_{P(\theta|\underline{X})}(\theta)$, por tanto, se debe calcular la distribución final y obtener la media, así como la varianza si se desea conocer el Valor de Bayes.

Si $X_i \sim \text{Exp}(\theta)$, $f(X_i|\theta)=\theta e^{-\theta x_i}$ con $\theta>0$. Por lo tanto

$f(\underline{X}|\theta)=\theta^n e^{-\theta\sum_{i=1}^{n} X_i}$ y además $P(\theta)=Ga(\theta|\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$ con $\alpha, \beta>0$.

Por el *Teorema de Bayes* se tiene que

$$P(\theta|\underline{X})=\frac{P(\underline{X}|\theta)P(\theta)}{P(\underline{X})}$$

y debido a que $P(\underline{X})$ no depende de θ

$$P(\theta|\underline{X}) \propto \theta^n e^{-\theta\sum_{i=1}^{n} X_i}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

$$\propto \theta^{n+\alpha-1}e^{-(\sum_{i=1}^{n} X_i+\beta)\theta}.$$

Es claro que $\theta^{\alpha+n-1}e^{-(\beta+\sum_{i=1}^{n} X_i)\theta}$ corresponde al núcleo de una función Gamma, por lo que $P(\theta|\underline{X})=Ga(\alpha',\beta')$ donde $\alpha'=\alpha+n$ y $\beta'=\beta+\sum_{i=1}^{n} x_i$.

Por lo que la decisión es

$$d^* = E_{P(\theta|Z_n)}(\theta)=\frac{\alpha'}{\beta'}=\frac{\alpha+n}{\beta+\sum_{i=1}^{n} x_i} \text{ con valor de Bayes } \frac{(\alpha+n)}{(\beta+\sum_{i=1}^{n} x_i)^2}.$$

En el caso de que no se hubiera contado con información muestral se tendría que la decisión de Bayes es $d^*=\frac{\alpha}{\beta}$ con valor esperado $d^*=\frac{\alpha}{\beta^2}$.

4.1.1 PRINCIPIO DE VEROSIMILITUD

Dentro del estudio de la estadística resulta de gran relevancia el Principio de Verosimilitud, ya que éste hace explícita la idea natural e intuitiva de que con respecto a la información experimental, sólo los valores observados deben ser relevantes para conclusiones o evidencias sobre el parámetro de interés. A partir de tal principio, se puede establecer una gran controversia sobre qué paradigma estadístico seguir, pues existen por ejemplo situaciones en donde la Estadística Clásica viola el Principio de Verosimilitud (Berger, 1985).

PRINCIPIO DE VEROSIMILITUD (Berger, 1985): Al hacer inferencias o tomar decisiones sobre el parámetro θ después de que se observó una muestra $\underline{X}$, toda la información experimental relevante está contenida en la función de verosimilitud, $f(\underline{X}|\theta)$. Más aún, dos funciones de verosimilitud contienen la misma información sobre θ si son proporcionales (como función de θ).

RESULTADO 4.1.1: Dada una distribución inicial $P(\theta)$, si las funciones de verosimilitud $f_1(\underline{Y}|\theta)$ y $f_2(\underline{Z}|\theta)$ son proporcionales como funciones de θ, entonces las respectivas distribuciones finales para θ son idénticas.

Demostración:

Sin perder generalidad la demostración se hará para el caso continuo.

Sea $P(\theta)$ una distribución inicial sobre θ. Si se obtiene una muestra $\underline{Y}=(Y_1,...,Y_n)$ de observaciones independientes con densidad conjunta $f_1(\underline{Y}|\theta)$, entonces por el Teorema de Bayes

$$P_1(\theta|\underline{Y})=\frac{P(\theta)f_1(\underline{Y}|\theta)}{\int_{\theta}P(\theta)f_1(\underline{Y}|\theta)d\theta}.$$

Sea $\underline{Z}=(Z_1,...,Z_m)$ una muestra de observaciones independientes con densidad conjunta $f_2(\underline{Z}|\theta)$ y supóngase que

$$f_2(\underline{Z}|\theta)=K(\underline{Y},\underline{Z})f_1(\underline{Y}|\theta) \text{ para todo } \theta\in\Theta,$$

de manera que las verosimilitudes sean proporcionales vistas como funciones de θ. Entonces

$$P_2(\theta|\underline{Z}) = \frac{P(\theta)f_2(\underline{Z}|\theta)}{\int_\theta P(\theta)f_2(\underline{Z}|\theta)d\theta} = \frac{P(\theta)K(\underline{Y},\underline{Z})f_1(\underline{Y}|\theta)}{\int_\theta P(\theta)K(\underline{Y},\underline{Z})f_1(\underline{Y}|\theta)d\theta} = \frac{P(\theta)f_1(\underline{Y}|\theta)}{\int_\theta P(\theta)f_1(\underline{Y}|\theta)d\theta} = P_1(\theta|\underline{Y}) \blacksquare$$

Es importante hacer notar que en el Paradigma Bayesiano el Principio de Verosimilitud no "sólo se sigue", sino que se deriva automáticamente del teorema de Bayes. Además, tal paradigma acepta la función de verosimilitud como un resumen completo de la información provista por los datos sobre el parámetro de interés. Para más detalle se puede consultar Bimbaum (1962), Basu (1975) y Berger y Wolpert (1984).

4.2 DETERMINACIÓN DE LA DISTRIBUCIÓN DE PROBABILIDAD INICIAL.

Lo discutido anteriormente, así como lo ilustrado en los ejemplos refuerza la necesidad de realizar una buena asignación de la función de densidad a priori. $P(\theta)$ es generalmente una función difícil de precisar.

Cuando el decisor se encuentra en la posibilidad de expresar su conocimiento inicial acerca del parámetro desconocido, es factible, aunque no es una tarea sencilla, representar dicho conocimiento mediante una función de distribución o de densidad *a priori informativa.*

4.2.1 DISTRIBUCIÓN DE PROBABILIDAD INICIAL INFORMATIVA.

La definición de probabilidad es constructiva (sección 3.2.), sin embargo, en la mayoría de los casos no resulta ser la mejor opción para asignar probabilidades ya que es un proceso largo y engorroso.

Una manera natural de asignar probabilidades es mediante la comparación de eventos. Para determinar la probabilidad del evento A el decisor puede comparar A con A^c y expresar por ejemplo que es doblemente creíble que ocurra A^c a que ocurra A. En este caso se tendría que $P(A)=\frac{1}{3}$ y $P(A^c)=\frac{2}{3}$

Otra manera semejante de determinar probabilidades es mediante *juegos*. Se plantea un juego en el que se gana x si ocurre A y se pierde $1-x$ si ocurre A^c, con $0 \leq x \leq 1$. Para que el

juego sea justo (la esperanza del juego sea cero) x debe ser tal que $u(x)P(A)-u(1-x)(1-P(A))=0$, por lo cual

$$P(A)=\frac{u(1-x)}{u(1-x)+u(x)}.$$

Para el caso de problemas de inferencia estadística, cuando Θ es un conjunto de cardinalidad finita el problema se reduce a determinar las probabilidades de cada uno de sus elementos; sin embargo a medida que la cardinalidad de Θ se incrementa o de que se trata de un conjunto infinito éste procedimiento resulta prácticamente imposible.

En los casos en los que no es viable asignar probabilidades a cada uno de los elementos de Θ lo que procede es encontrar una función que aproxime $P(\theta)$. A continuación se describen algunos de los métodos empleados para ello.

a) MÉTODO DE HISTOGRAMAS

Cuando θ es una variable aleatoria continua, en particular cuando Θ es un intervalo de $\mathbf{R}$, una aproximación natural a $P(\theta)$ es a través de histogramas; a partir de estos se puede ajustar una densidad, posiblemente suavizando el histograma. El número y el tamaño de los intervalos usados no está sujeto a ninguna regla, sino que depende de las necesidades del problema específico; algunas veces será suficiente con un número pequeño de intervalos, mientras que en otros casos se requerirá un mayor detalle y precisión, por lo que se construirán más intervalos de menor longitud. Una de las limitaciones graves de este método es que sólo sirve para asignar probabilidades a un conjunto acotado, esto es, la distribución obtenida a través de histogramas no tiene colas.

Es importante enfatizar que este método tiene los problemas usuales de tratar de visualizar una densidad a través de histogramas.

b) MÉTODO DE VEROSIMILITUDES RELATIVAS

Este método se usa también en subconjuntos de la recta real y consiste simplemente en comparar las "*verosimilitudes*" de varios puntos de Θ. Por ejemplo, suponga que se tiene un espacio parametral $\Theta=[0,1]$ en donde el punto $\theta=\frac{1}{2}$ es el más creíble mientras que $\theta=0$ y $\theta=1$ son los menos verosímiles. Si se asigna a $\theta=0$ un valor a priori de 1 y se dice que $\theta=\frac{1}{2}$

es cuatro veces más creíble, $\theta = \frac{1}{4}$ y $\theta = \frac{3}{4}$ dos veces y medio más creíble y $\theta = \frac{3}{8}$ y $\theta = \frac{5}{8}$ tres veces y medio más creíble con respecto a $\theta = 0$, entonces les corresponde los valores a priori de 4, 2.5 y 3.5. respectivamente. La función de densidad a priori así obtenida no integra 1, sin embargo, es fácil encontrar una constante c tal que $cP(\theta)$ sea una función con masa igual a la unidad. En realidad, no es necesario determinar esta constante, ya que es claro que cualquier d que minimice $E_{P(\theta)}(L(d,\theta))$ minimizará $E_{cP(\theta)}(L(d,\theta))$.

La función de densidad obtenida para el ejemplo mencionado se ilustra en la figura 4.2.1.1.

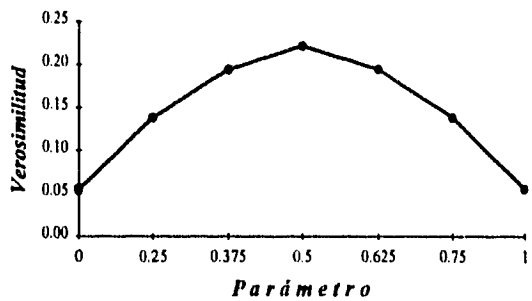

Figura 4.2.1.1

Con este procedimiento sucede algo similar a lo que ocurre con el método de histogramas, y es el hecho de que ambos son aplicables a regiones finitas, por lo que hay problemas cuando se trata de utilizar en conjuntos no acotados o infinitos. En tales casos se debe determinar un intervalo central, aproximar $P(\theta)$ para dicha región y posteriormente decidir cómo evaluarla fuera de la región finita; obviamente esto representa una dificultad extra. En el Análisis de Robustez Bayesiana (cuyo estudio escapa de los objetivos de este trabajo) se presentan alternativas al problema (ver Berger, 1984).

c) ASOCIACIÓN DE UNA FORMA PARAMÉTRICA DADA

Esta es una de las maneras más usadas para determinar la densidad a priori, consiste en suponer que $P(\theta)$ tiene una forma paramétrica particular, por lo cual el problema se reduce a seleccionar los parámetros de tal manera que la función resultante refleje lo más cercanamente posible las creencias a priori del decisor. En resumen, se busca seleccionar un elemento de

$$F=\left\{P(0)\middle|\, P(0)=g(0|\lambda),\ \lambda\in\Lambda\right\}.$$

El método es de gran utilidad cuando mediante el análisis de histogramas o gráficas se ha vislumbrado una familia paramétrica o en el caso contrario, cuando sólo se cuenta con un panorama generalizado de la información a priori.

Una de las formas de estimar más fácilmente los parámetros a priori es mediante el cálculo de los momentos a priori. Por ejemplo, si se piensa que la función de distribución es una Gamma, $Ga(\alpha,\beta)$, entonces únicamente se debe decidir sobre los valores de la media (μ) y de la varianza (σ^2) ya que es sabido que $\mu=\frac{\alpha}{\beta}$ y $\sigma^2=\frac{\alpha}{\beta^2}$. A partir de estas relaciones se tiene que $\alpha=\frac{\mu^2}{\sigma^2}$ y $\beta=\frac{\mu}{\sigma^2}$.

El problema como en todas las técnicas vistas se hace presente en los conjuntos no acotados, pues las colas de una densidad pueden tener gran impacto en sus momentos.

Otra alternativa para estimar los parámetros de la distribución a priori es mediante la determinación subjetiva de algunos de sus cuantiles. Una vez encontrados los cuantiles, se seleccionan los parámetros adecuados tal que los cuantiles de la distribución resultante sean lo más próximos a aquéllos estimados subjetivamente . Un α-cuantil de una distribución continua es un punto $z(\alpha)$ tal que una variable aleatoria con esta distribución tiene probabilidad α de ser menor o igual a $z(\alpha)$.

Debido a que el decisor generalmente se encuentra en la posibilidad de estimar probabilidades de regiones, éste parece ser un método viable; además, existen tablas de cuantiles de las densidades más utilizadas así como paquetes estadísticos (por ejemplo S-plus) que los calculan.

Ejemplo 4.2.1.1 (Berger, 1985): Considere que el espacio parametral es $\Theta=(-\infty,\infty)$ y que se sospecha que la función a priori pertenece a la familia normal. Se determina subjetivamente que la $P(\theta\leq 0)=0.5$ y que los cuartiles (.25 y .75) son -1 y 1. Ya que para una distribución normal la media y la mediana son iguales, es claro que la media es $\mu=0$. Usando tablas de probabilidades de una normal, se concluye que la varianza de la distribución a priori

debe ser $\sigma^2 = 2.198109$ pues $P(0 < \frac{1}{(2.198109)^{1/2}}) = .75$ y $P(0 < \frac{-1}{(2.198109)^{1/2}}) = .25$. De aquí que $P(0)$ será una densidad $N(0, 2.19)$.

Para encontrar el valor de los parámetros de la función a priori generalmente se necesita un número pequeño de cuantiles, tal y como se ilustra en el ejemplo, sin embargo esto puede acarrear un nuevo problema, ya que puede suceder que cuantiles no involucrados en la determinación de la a priori no sean coherentes con ella. Una situación como esta implicaría en algunos de los casos que el decisor no ha sido coherente al expresar su conocimiento inicial, y en otros, que no se ha elegido la familia paramétrica adecuada. Esto sugiere que se debe poner especial atención para elegir la familia de distribuciones más apropiada.

En la determinación de la distribución a priori resulta conveniente utilizar de manera combinada los métodos antes mencionados. En algunos casos la densidad bosquejada mediante histogramas puede resultar sugerente, de ahí que sea factible suponer una familia paramétrica y a través de los momentos a priori o del método conocido como de verosimilitudes relativas calcular los parámetros de interés.

Ejemplo 4.2.1.2: Sea $\underline{X} = (X_1, X_2, ..., X_n)$ una muestra aleatoria con función de distribución $Bili(0)$ y $\Theta = [0, 1]$. Al estimar subjetivamente se obtuvo el histograma ilustrado en la figura 4.2.1.2.

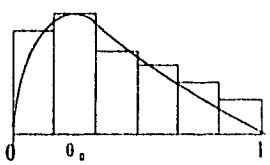

Figura 4.2.1.2.

El histograma mostrado así como el intervalo en el que puede tomar valores θ, hacen razonable suponer una densidad beta, $Be(\alpha, \beta)$, esto es,

$$P(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{(0,1)}(\theta).$$

Para definir completamente $P(\theta)$ se requiere determinar los correspondientes valores para α y β. A partir de la figura se observa que θ_0 es la moda de la función de densidad y se puede

demostrar que la moda correspondiente a $P(\theta|\alpha,\beta)$ es $\frac{\alpha-1}{\alpha+\beta-2}$ y la media es $\frac{\alpha}{\alpha+\beta}$. Si además se estima subjetivamente que θ_1 es la media obtenida del histograma, se tiene que

$$P(\theta)= Be\left(\frac{\theta_1(2\theta_0-1)}{\theta_0-\theta_1}, \frac{(1-\theta_1)(2\theta_0-1)}{\theta_0-\theta_1}\right).$$

4.2.2 FAMILIAS CONJUGADAS PARAMÉTRICAS

Ejemplo 4.2.2.1: En las condiciones del ejemplo 4.2.1.2 suponga que se desea actualizar la distribución a priori con información obtenida a través de una muestra aleatoria.

Hasta el momento se tiene $\underline{X} = (X_1, X_2, ..., X_n)$, una m.a. con $X \sim Blli(\theta)$, $\Theta = [0, 1]$, y $P(\theta)$ una $Be(\alpha, \beta)$ con α y β ya determinados. Utilizando el Teorema de Bayes se tiene que

$$\begin{aligned} P(\theta|\underline{X}) &\propto P(\underline{X}|\theta)P(\theta) \\ &\propto \theta^{Y}(1-\theta)^{n-Y}\theta^{\alpha-1}(1-\theta)^{\beta-1}I_{(0,1)}(\theta) \\ &\propto \theta^{\alpha+Y-1}(1-\theta)^{\beta+n-Y-1}I_{(0,1)}(\theta). \end{aligned} \qquad (1)$$

donde $Y = \sum_{i=1}^{n} X_i$.

De la expresión (1) queda claro que $P(\theta|\underline{X})$ tiene la forma de una distribución Beta con parámetros $\alpha' = \alpha + Y$ y $\beta' = \beta + n - Y$. Lo anterior demuestra que bajo un muestreo Bernoulli y una inicial (a priori) Beta, la final (a posteriori) vuelve a ser una Beta pero con parámetros actualizados.

Algo análogo se muestra en el ejemplo 4.1.3., pues de él se concluye que bajo un muestreo Exponencial y una inicial Gamma, la final vuelve a ser una Gamma.

Los dos ejemplos mencionados ilustran un concepto de gran utilidad en la asignación de distribuciones iniciales y cálculo de distribuciones finales; este concepto es el de Familias conjugadas paramétricas. En general, toda familia paramétrica se dice que es conjugada si es cerrada bajo la aplicación del Teorema de Bayes con respecto a una verosimilitud específica. En tal caso se dice además que la familia es *cerrada bajo muestreo*.

DEFINICIÓN 4.2.2.1 *(Familia Conjugada Paramétrica): Sea* $\underline{X} = (X_1, X_2, ..., X_n)$ *una m.a. con* $X \sim f(X|\theta)$, *con* $\theta \in \Theta \subset \mathbb{R}^k$ *y* $\mathbf{F} = \{P(\theta) | P(\theta) = g(\theta|\lambda), \lambda \in \Lambda\}$, $\mathbf{F}$ *es una familia*

conjugada paramétrica para $\mathrm{H} = \{f(X|\theta)\ \theta \in \Theta\}$ *si y sólo si para todo* $P(\theta)$ *que pertenece a* F *y* $f(X|\theta)$ *que pertenece a* H *se tiene que* $P(\theta|\underline{X})$ *también pertenece a* F.

El tomar una distribución inicial de alguna familia conjugada paramétrica puede facilitar la determinación de la función de distribución a priori pues representa considerables ventajas. Para dejar completamente definida la inicial se debe estimar de manera subjetiva los valores de los parámetros; lo cual se puede hacer mediante los métodos descritos anteriormente; además el cálculo de la distribución final es inmediato ya que sólo hace falta actualizar los parámetros con la información muestral.

A continuación se presentan algunas familias conjugadas con sus respectivas reglas de actualización de los parámetros. Las demostraciones se excluyen debido a que sólo se trata de cálculos algebraicos.

Tabla 4.2.2.1

VEROSIMILITUD	PARÁMETRO DE INTERÉS	INICIAL	FINAL
$Blli(X\|\theta)$	θ	$Be(\theta\|\alpha,\beta)$	$P(\theta\|\underline{X}) = Be(\alpha+Y, \beta+n-Y)$
$Binomial(X\|m\theta)$	θ	$Be(\theta\|\alpha,\beta)$	$P(\theta\|\underline{X}) = Be(\alpha+Y, \beta+nm-Y)$
$Geometrica(X\|\theta)$	θ	$Be(\theta\|\alpha,\beta)$	$P(\theta\|\underline{X}) = Be(\alpha+n, \beta+Y)$
$Exp(X\|\theta)$	θ	$Ga(\theta\|\alpha,\beta)$	$P(\theta\|\underline{X}) = Ga(\alpha+n, \beta+Y)$
$Polsson(X\|\theta)$	θ	$Ga(\theta\|\alpha,\beta)$	$P(\theta\|\underline{X}) = Ga(\alpha+Y, \beta+n)$
$N(X\|\mu,\sigma^2)$ μ conocida	$\tau = \frac{1}{\sigma^2}$	$Ga(\tau\|\alpha,\beta)$	$P(\tau\|\underline{X}) = Ga\left(\alpha+\frac{n}{2}, \beta+\frac{ns^2}{2}\right)$
$N(X\|\mu,\sigma^2)$ σ^2 conocido	μ	$N(\mu\|\mu_0,\sigma_0^2)$	$P(\mu\|\underline{X}) = N\left(\frac{Y\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)$.

con $Y = \sum_{i=1}^{n} x_i$, y $s^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$.

La relevancia de cada familia conjugada está ligada directamente a la flexibilidad y versatilidad para poder expresar la incertidumbre que el decisor tiene sobre el parámetro de interés.

En general el poder precisar una función de distribución $P(\theta)$ es una tarea compleja, sin embargo, el concepto de familias conjugadas parece resultar de gran utilidad; es más, Diaconis e Ylvisaker (1984) afirman que la incertidumbre sobre θ *siempre* se puede representar como una *mezcla de distribuciones conjugadas*, esto es, como una combinación lineal convexa de ellas. Aunque esto no garantiza la simplicidad en la práctica.

TEOREMA 4.2.2.1: *Sea* $\underline{X}=(X_1, X_2, ..., X_n)$ *una m.a. con* $X \sim f(X|\theta)$, $\Theta \subset R^k$ *y* $\mathbf{F}=\{P(\theta)|\ P(\theta)=g(\theta|\lambda),\ \lambda \in \Lambda\}$ *una familia conjugada paramétrica bajo tal muestreo,* entonces $\mathbf{F}^*=\left\{P^*(\theta)\middle|\ P^*(\theta)=\sum_{i=1}^{m} k_i P_i(\theta);\ P_i(\theta)=g(\theta|\lambda_i),\ \sum_{i=1}^{m} k_i=1,\ k_i \geq 0 \text{ y } \lambda \in \Lambda\right\}$ *es también una familia conjugada paramétrica bajo el mismo muestreo.*

Demostración:

Sea $P^*(\theta)$ que pertenece a $\mathbf{F}^*$, entonces $P^*(\theta)=\sum_{i=1}^{m} k_i P_i(\theta)$. Por el Teorema de Bayes se tiene que:

$$\begin{aligned} P^*(\theta|\underline{X}) &\propto P^*(\theta) f(\underline{X}|\theta) \\ &= \sum_{i=1}^{m} k_i P_i(\theta) f(\underline{X}|\theta) \\ &= \sum_{i=1}^{m} k_i P_i(\underline{X}) \frac{P_i(\theta) f(\underline{X}|\theta)}{P_i(\underline{X})} \\ P^*(\theta|\underline{X}) &\propto \sum_{i=1}^{m} k_i P_i(\underline{X}) P_i(\theta|\underline{X}) \end{aligned} \tag{2}$$

pero como $P_i(\theta)=g(\theta|\lambda_i)$ pertenece a la familia conjugada con respecto a $f(\underline{X}|\theta)$,

$$P_i(\theta|\underline{X}) = g(\theta|\lambda_i'), \tag{3}$$

donde $g(\theta|\lambda_i')$ es la distribución final asociada a la inicial $g(\theta|\lambda_i)$ con λ_i' el parámetro actualizado.

De aquí que combinando (2) y (3) se tiene

$$P^*(\theta|\underline{X}) \propto \sum_{i=1}^{m} k_i' g(\theta|\lambda_i'), \tag{4}$$

con $k_i' = k_i P_i(\underline{X})$ y $P_i(\underline{X})$ una constante que puede ser calculada fácilmente (por ser $\mathbf{F}$ conjugada) mediante el Teorema de Bayes .

De (4) se deduce que

$$P^*(\theta|\underline{X})=\sum_{i=1}^{m} k_i'' g(\theta|\lambda_i') \text{ con } k_i''=\frac{k_i'}{\sum_{i=1}^{n} k_i'}.$$

La expresión anterior muestra claramente que $P^*(\theta|Z_n)$ es una combinación lineal convexa de elementos de F, por lo que $P^*(\theta|\underline{X})$ también pertenece a F* y en consecuencia, F* es una familia conjugada ■

En conclusión, la familia de mezcla de conjugadas es también una familia conjugada y dicha mezcla es de gran utilidad para representar distribuciones iniciales multimodales. La demostración anterior exhibe las reglas de actualización de los parámetros.

Si $P^*(\theta)=\sum_{i=1}^{n} k_i P_i(\theta)$, entonces $P^*(\theta|\underline{X})=\sum_{i=1}^{n} c_i P_i(\theta|\underline{X})$, donde $P_i(\theta|\underline{X})$ es la distribución actualizada a través de la información muestral (ver tabla de familias conjugadas para algunos casos específicos) y $c_i=\frac{k_i P_i(\underline{X})}{\sum_{i=1}^{m} k_i P_i(\underline{X})}$ (claramente $\sum_{i=1}^{m} c_i=1$).

ESTADÍSTICA SUFICIENTE

Otro aspecto que es importante resaltar y que se muestra tanto en el problema 4.1.3 como en el 4.2.2.1 es el hecho de que $P(\theta|\underline{X})$ depende de la muestra únicamente a través de $Y=\sum_{i=1}^{n} X_i$. Lo anterior implica que desde el punto de vista bayesiano Y es suficiente para θ.

DEFINICIÓN 4.2.2.2 *(Estadística Suficiente): Sea T_n una estadística de la muestra $\underline{X}$, se dice que T_n es suficiente si para cualquier distribución inicial se tiene*

$$P(\theta|\underline{X})=P(\theta|T_n)$$

con

$$P(\theta|T_n)\propto f(T_n|\theta)P(\theta)$$

donde $f(T_n|\theta)$ es la f.d.p.g. de T_n condicional a θ.

Existen algunos resultados interesantes sobre estadísticas suficientes y uno de ellos es el del Teorema de Factorización.

TEOREMA 4.2.2.2: *Una estadística T es suficiente si y sólo si $f(\underline{X}|\theta)$ puede ser factorizado como*

$$f(\underline{X}|\theta)=U(\underline{X})V(T(\underline{X}),\theta) \qquad (1)$$

para todo $\underline{X}$ en el espacio muestral, donde U es una función positiva que no depende de θ y V es no negativa y depende de $\underline{X}$ solamente a través de T.

Demostración:

$\Rightarrow$) Suponga que T es una estadística suficiente, entonces $P(\theta|\underline{X})=h(T(\underline{X}),\theta)\geq 0$ donde h es una función que involucra solamente a $T(\underline{X})$ y a θ. Por el Teorema de Bayes se tiene que

$$P(\theta|\underline{X})=\frac{f(\underline{X}|\theta)P(\theta)}{\int_\theta f(\underline{X}|\theta)P(\theta)d\theta}, \text{ por lo que } f(\underline{X}|\theta)=\left(\int_\theta f(\underline{X}|\theta)P(\theta)d\theta\right)\frac{P(\theta|\underline{X})}{P(\theta)}$$

y por ser T suficiente se sigue que

$$f(\underline{X}|\theta)=\left(\int_\theta f(\underline{X}|\theta)P(\theta)d\theta\right)\frac{h(T(\underline{X}),\theta)}{P(\theta)}$$

que es una factorización como la indicada en (1), con $U(\underline{X})=\int_\theta f(\underline{X}|\theta)P(\theta)d\theta>0$ y $V(\underline{X},\theta)=\frac{h(T(\underline{X}),\theta)}{P(\theta)}\geq 0$.

$\Leftarrow$) Suponga que se cumple la factorización (1), entonces por el Teorema de Bayes

$$P(\theta|\underline{X})=\frac{f(\underline{X}|\theta)P(\theta)}{\int_\theta f(\underline{X}|\theta)P(\theta)d\theta}$$

$$=\frac{U(\underline{X})V(T(\underline{X},\theta)P(\theta)}{\int_\theta U(\underline{X})V(T(\underline{X}),\theta)P(\theta)d\theta}.$$

Como U es una función que no depende de θ, entonces

$$P(\theta|\underline{X})=\frac{V(T(\underline{X}),\theta)P(\theta)}{\int_\theta V(T(\underline{X}),\theta)P(\theta)d\theta}. \qquad (2)$$

Nuevamente utilizando el hecho de que U es una función que no depende de θ, se tiene que

$$P(\theta|\underline{X})=\frac{U(T(\underline{X}))V(T(\underline{X}),\theta)P(\theta)}{\int_\theta U(T(\underline{X}))V(T(\underline{X}),\theta)P(\theta)d\theta}$$

$$=\frac{f(T(\underline{X})|\theta)P(\theta)}{\int_\theta f(T(\underline{X})|\theta)P(\theta)d\theta}=P(\theta|T(\underline{X}))$$

Debido a que $P(\theta|\underline{X})=P(\theta|T(\underline{X}))$ se sigue que $T(\underline{X})$ es una estadística suficiente ■

El concepto de estadística suficiente fue introducido por Fisher (1922) y lo aquí expuesto es equivalente al teorema de factorización de Neyman. Con esto se tiene que si una estadística es suficiente desde el punto de vista clásico o frecuentista, también lo será desde el punto de vista bayesiano.

Ejemplo 4.2.2.2: Sea $\underline{X}=(X_1, X_2,..., X_n)$ una m.a. donde X es una v.a. cuya distribución pertenece a la familia exponencial generalizada, esto es

$$f(X|\theta)=h(x)w(\theta)\exp\left(\sum_{i=1}^{k}c_j(\theta)U_j(x)\right),\ \theta\in\mathbb{R}^k;$$

en donde el recorrido de X no depende de θ y $h(x), w(\theta), c_j(\theta), U_j(x),\ j=1,...,k$ son funciones totalmente especificadas. Encontrar $T(\underline{X})\in\mathbb{R}^k$ tal que $T(\underline{X})$ es una estadística suficiente para θ.

$$f(\underline{X}|\theta)=\prod_{i=1}^{n}\left(h(x_i)w(\theta)\exp\left(\sum_{i=1}^{k}c_j(\theta)U_j(x_i)\right)\right)$$

$$=\left(\prod_{i=1}^{n}h(x_i)\right)w^n(\theta)\exp\left(\sum_{i=1}^{k}c_j(\theta)U_j(x_1)+\sum_{i=1}^{k}c_j(\theta)U_j(x_2)+...+\sum_{i=1}^{k}c_j(\theta)U_j(x_n)\right)$$

$$f(\underline{X}|\theta)=\left(\prod_{i=1}^{n}h(x_i)\right)w^n(\theta)\exp\left(c_1(\theta)\sum_{i=1}^{n}U_1(x_i)+c_2(\theta)\sum_{i=1}^{n}U_2(x_i)+...+c_k(\theta)\sum_{i=1}^{n}U_k(x_i)\right)\quad(3)$$

De la ecuación (3) es claro que se tiene una factorización como la del Teorema 4.2.2.2., donde $U(\underline{X})=\left(\prod_{i=1}^{n}h(x_i)\right)>0$ y $V(T(\underline{X})|\theta)=w^n(\theta)\exp\left(\sum_{j=1}^{k}\left(c_j(\theta)\sum_{i=1}^{n}U_j(x_i)\right)\right)\geq 0$, que depende de $\underline{X}$ únicamente a través de la estadística

$$T(\underline{X})=\left(\sum_{i=1}^{n}U_1(x_i),\ \sum_{i=1}^{n}U_2(x_i),\ ...,\ \sum_{i=1}^{n}U_k(x_i)\right).$$

Por tanto, $T(\underline{X})$ es una estadística suficiente y de dimensión fija (igual a la dimensión de Θ) para el parámetro θ de la familia exponencial generalizada. Esto es un resultado de suma importancia que se plasma en el siguiente teorema.

TEOREMA 4.2.2.3: *Sea $\underline{X}=(X_1, X_2,..., X_n)$ una m.a. donde X es una v.a. cuya distribución pertenece a la familia exponencial generalizada, esto es*

$$f(X|\theta)= h(x)w(\theta)\exp\left(\sum_{j=1}^{k} c_j(\theta)U_j(x)\right),\ \theta \in R^k;$$

en donde el recorrido de X *no depende de* θ *y* $h(x), w(\theta), c_j(\theta), U_j(x),\ j=1,...,k$ *son funciones totalmente especificadas. Entonces,*

$$T(\underline{X})=\left(\sum_{i=1}^{n} U_1(x_i), \sum_{i=1}^{n} U_2(x_i), ..., \sum_{i=1}^{n} U_k(x_i)\right)$$

es una estadística suficiente de dimensión fija para θ.

Algunos ejemplos de distribuciones pertenecientes a esta familia son la Bernoulli, la Poisson, la Normal (con media desconocida y varianza conocida o desconocida), la Gamma, la Beta y la Normal Multivariada con vector de medias y matriz de varianzas y covarianzas desconocidas.

Existe una relación directa entre las familias conjugadas y el concepto de suficiencia ya que si se tiene una estadística suficiente de dimensión fija, siempre es posible construir una familia conjugada paramétrica. Se puede demostrar que las únicas funciones de densidad que admiten una estadística suficiente de dimensión fija son las pertenecientes a la familia exponencial.

TEOREMA 4.2.2.4: *Sea* X *una variable aleatoria con f.d.p.g.* $f(X|\theta)$ *y* $\theta \in \Theta$, *si existe una estadística suficiente de dimensión fija para* θ *y para todo* n *ocurre que* $\int_{\Theta} f(\underline{X}_n|\theta)d\theta < \infty$ *entonces existe una familia conjugada para* $f(X|\theta)$.

Demostración

Sea $T(\underline{X}_n)$ una estadística suficiente de dimensión fija para θ, entonces por el teorema de factorización se tiene que existe V_n tal que

$$f(\underline{X}_n|\theta) \propto V_n(T(\underline{X}_n),\theta), \tag{4}$$

por lo que $\int_{\Theta} f(\underline{X}_n|\theta)d\theta \propto \int_{\Theta} V_n(T(\underline{X}_n),\theta)d\theta$. De aquí se tiene que

$$\int_{\Theta} V_n(T(\underline{X}_n),\theta)d\theta < \infty,$$

y por lo tanto existe una f.d.p.g.

$$g(\theta|T(\underline{X}_n),n) \propto V_n(T(\underline{X}_n),\theta). \tag{5}$$

A continuación se demuestra que la familia **G** de todas las f.d.p.g. de la forma expresada en (5), es una familia conjugada para $f(X|\theta)$.

Sea $g(\theta|t_1,n_1)$ un elemento arbitrario de **G** y $\underline{X}_{n_0}=(x_1, x_2, ..., x_{n_0})$ la muestra; entonces por el teorema de Bayes

$$g(\theta|\underline{X}_{n_0}) \propto g(\theta|t_1,n_1) f(\underline{X}_{n_0}|\theta). \qquad (6).$$

Al sustituir (4) y (5) en el lado derecho de la expresión (6), se tiene

$$g(\theta|\underline{X}_{n_0}) \propto f(\underline{X}_{n_1}|\theta) f(\underline{X}_{n_0}|\theta) = f(\underline{X}_{n_0+n_1}|\theta)$$

y por ser t una estadística de dimensión fija

$$g(\theta|\underline{X}_{n_0}) \propto V'\left(T(\underline{X}_{n_1+n_0}),\theta\right).$$

De lo anterior se concluye que **G** es una familia conjugada para $f(X|\theta)$■

Esta demostración es muy ilustrativa ya que a partir de ella se deduce la manera natural de construir la familia conjugada paramétrica para una $f(X|\theta)$ dada cuando existe la estadística suficiente de dimensión fija. La familia conjugada paramétrica es entonces

$$\mathbf{G} = \left\{g(\theta|t,n) \propto V_n(t,\theta)\right\}$$

en donde V_n es igual que en el Teorema de Factorización. Esta familia es conocida como *familia conjugada básica.*

Ejemplo 4.2.2.3: Sea $\underline{X}=(X_1, X_2, ..., X_n)$ una muestra aleatoria con f.d.g.p. $Blli(\theta)$. Determinar la familia conjugada básica.

$$f(\underline{X}|\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i},$$

y por el Teorema de factorización, se tiene que $U(\underline{X})=1$ y

$$V'(T(\underline{X})\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

por lo que $T(\underline{X}) = \sum_{i=1}^{n} x_i$ y por consiguiente

$$G = \left\{ Be(\alpha,\beta) \middle| \alpha = \sum_{i=1}^{n} x_i + 1,\ \beta = n - \sum_{i=1}^{n} x_i + 1 \right\}.$$

En este caso $G = \{Be(\alpha,\beta) | \alpha,\beta \in \mathbf{N}\}$, esto es, la familia conjugada básica es la familia de distribuciones Beta con parámetros en los naturales.

4.2.3 DISTRIBUCIÓN DE PROBABILIDAD INICIAL NO INFORMATIVA.

Como ya se ha mencionado, existen situaciones en las cuales la determinación de una distribución a priori informativa es imposible, ya sea por que se carece de información a priori, o debido a que el decisor se encuentra imposibilitado para manifestar su conocimiento inicial. En otros casos, es deseable no considerar información inicial para llegar a resultados comparables con la estadística clásica o frecuentista. El Análisis Bayesiano, sin embargo, requiere de una función de densidad a priori; motivo por el cual existe toda una escuela encargada del estudio de la determinación de una función *a priori no informativa*; esto es, una función que represente el desconocimiento del comportamiento del parámetro de interés; una distribución que no favorezca ciertos valores de θ.

El plantear una *a priori no informativa* es un problema difícil, de hecho es un problema abierto a la discusión. Existen algunas alternativas de solución, cada una con sus puntos a favor y sus puntos en contra; en esta sección se tratarán las más comunes.

PRINCIPIO DE LA RAZÓN INSUFICIENTE

Si no se sabe nada sobre el espacio parametral Θ, entonces no hay razón alguna para asignarle una probabilidad más alta a algunos de estos sucesos inciertos.

En el caso en el que $\Theta = \{\theta_1, ..., \theta_k\}$, la probabilidad de cada evento bajo este principio estará dada por $P(\theta_i) = \frac{1}{k}$ $i = 1,...,k$; esto es, la inicial no informativa será la distribución uniforme. Sin embargo, en situaciones en donde el número de sucesos inciertos no es finito, la aplicación de este principio ya no es tan obvia.

Ejemplo 4.2.3.1 (Berger, 1985): Suponga que el parámetro de interés es el de la media de una normal, con el espacio parametral $\Theta = (-\infty,\infty)$. Si se desea una a priori no informativa,

el principio de la razón insuficiente llevaría a asignar igual peso a todos los posibles valores de Θ. Desafortunadamente si se elige $P(\theta)=c>0$, entonces $\int_{\Theta} P(\theta)d\theta=\infty$, esto es, no se tiene una densidad propia. No obstante, se puede trabajar con ella. La elección de c no es importante, por lo tanto la distribución a priori típica para este problema es $P(\theta)=1$. Esta es frecuentemente llamada la densidad no informativa y fue introducida y usada por Laplace (1812).

Como en el ejemplo de arriba sucederá frecuentemente que la no informativa natural es una a priori impropia. Se conoce como impropia aquella densidad que tiene una masa infinita.

Un problema más grave que el de tener una masa infinita es el de falta de invarianza ante reparametrizaciones. Esto se ilustra en el siguiente ejemplo.

Ejemplo 4.2.3.2 Sea θ la proporción de individuos con cierta característica. Entonces la inicial no informativa de a cuerdo con el principio de razón insuficiente es

$$P(\theta)=\begin{cases}1 & 0<\theta<1\\ 0 & e.o.c.\end{cases}.$$

Supóngase ahora que el interés está puesto sobre $\varphi=-\log\theta$. Resulta obvio que si no se tiene información sobre θ, tampoco se tiene sobre φ. En este caso el principio de la razón insuficiente llevaría a asignar la distribución

$$P(\varphi)=\begin{cases}c & 0<\varphi<\infty\\ 0 & e.o.c.\end{cases}.$$

Por otro lado la distribución para φ inducida por $P(\theta)$ es

$$P_1(\varphi)=P(\theta(\varphi))|J_0(\varphi)|=\begin{cases}e^{-\varphi} & 0<\varphi<\infty\\ 0 & e.o.c.\end{cases}$$

la cual es claramente informativa y por tanto diferente a $P(\varphi)$.

En torno a esto se puede argumentar que uno usualmente elige la parametrización que resulta intuitivamente más razonable; este argumento es, en general, difícil de defender.

PRINCIPIO DE INVARIANZA (REGLA DE JEFFRYES)

La carencia de invarianza de la a priori (constante) derivada del principio de razón insuficiente ha conducido a la búsqueda de iniciales no informativas las cuales sean invariantes ante transformaciones del parámetro, ya que por ejemplo Jeffreys dice (Lee, 1989): "cualquier parametrización arbitraria del modelo debe de llevar a los mismos resultados". Se han dado muchas sugerencias para resolver este problema. El método más ampliamente usado es el de Jeffreys (1961).

Sea X una variable aleatoria con f.d.p.g. $P(X|\theta)$ y $\theta \in \Theta \subset \mathbf{R}^r$. Basándose en la noción de invarianza, Jeffreys propuso tomar la distribución inicial no informativa para el parámetro θ como

$$P(\theta) \propto \left|\det\{I(\theta)\}\right|^{1/2}, \quad \theta \in \Theta,$$

donde (bajo condiciones de regularidad) $I(\theta) = -E_{X|\theta}\left[\dfrac{\partial^2 \log P(x|\theta)}{\partial\theta'\partial\theta}\right]$ es la matriz (pxp) de información de Fisher.

La idea intuitiva de basarse en la información de Fisher es que $I(\theta)$ se interpreta generalmente como un indicador de información provista por el modelo (o por las observaciones) acerca del valor del parámetro θ. Además parece razonable suponer que los valores de θ para los cuales $\left|\det\{I(\theta)\}\right|^{1/2}$ es grande deben ser más probables bajo una distribución inicial no informativa. Dicho de otra manera, favorecer los valores de θ para los cuales $\left|\det\{I(\theta)\}\right|^{1/2}$ es grande es equivalente a minimizar la influencia de la distribución inicial y por lo tanto ésta será lo más no informativa posible.

Ejemplo 4.2.3.2 Encontrar la distribución inicial no informativa de Jeffreys para el parámetro θ de una distribución Binomial-Negativa $(x|\theta,r)$ con función de probabilidad dada por

$$P(x|\theta) = \frac{(r+x-1)!}{x!(r-1)!}\theta^r(1-\theta)^x \quad x = 0,1,\ldots$$

$$\log P(x|\theta) = \log\frac{(r+x-1)!}{x!(r-1)!} + r\log\theta + x\log(1-\theta)$$

$$\frac{\partial \log P(x|\theta)}{\partial \theta} = \frac{X}{\theta} - \frac{r}{1-\theta}$$

$$\frac{\partial^2 \log P(x|\theta)}{\partial \theta^2} = -\frac{X}{\theta^2} - \frac{r}{(1-\theta)^2}$$

$$I(\theta) = -E_{x|\theta}\left[\frac{\partial^2 \log P(x|\theta)}{\partial \theta^2}\right] = -E_{x|\theta}\left[-\frac{x}{\theta^2} - \frac{r}{(1-\theta)^2}\right]$$

$$= \frac{E_{x|\theta}(x)}{\theta^2} + \frac{r}{(1-\theta)^2}$$

y como $E_{x|\theta}(x) = \frac{r\theta}{1-\theta}$

$$I(\theta) = \frac{r\theta/1-\theta}{\theta^2} + \frac{r}{(1-\theta)^2} = \frac{r}{\theta(1-\theta)^2} + \frac{r}{(1-\theta)^2}$$

$$= \frac{r}{\theta(1-\theta)^2} = r\theta^{-1}(1-\theta)^{-2}$$

por lo tanto

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

Esta es una distribución claramente impropia (pues tiene masa infinita) que podría pensarse como una "*distribución beta*", $Be(1/2, 0)$; sin embargo, en la distribución beta se requiere $\alpha > 0$, $\beta > 0$. Cabe aclarar que a pesar de que tal a priori es impropia suele ser usada, pues la final sí es propia y resulta ser una distribución beta especificada por $Be\left(\sum x_i + 1/2,\ nr\right)$.

Ejemplo 4.2.3.3 Encontrar la distribución inicial no informativa de Jeffreys para el parámetro $\theta = (\mu, \sigma^2)$ de una distribución normal con media y varianza desconocidas.

Se tiene que $\log P(x|\theta) = k - \frac{1}{2}\log\sigma^2 - \frac{1}{2}(x-\mu)^2/\sigma^2$ y por tanto

$$\frac{\partial^2 \log P(x|\theta)}{\partial \theta' \partial \theta} = \begin{pmatrix} -1/\sigma^2 & -(x-\mu)/(\sigma^2)^2 \\ -(x-\mu)/(\sigma^2)^2 & 1/2(\sigma^2)^2 - (x-\mu)^2/(\sigma^2)^3 \end{pmatrix}.$$

Además como se tiene que $E(x) = \mu$ y $E(x-\mu)^2 = \sigma^2$ se sigue que

$$I(\theta)=\begin{pmatrix}(\sigma^2)^{-1} & 0\\ 0 & -\frac{1}{2}(\sigma^2)^{-2}\end{pmatrix}$$

$$\text{y } \det I(\theta)=-\tfrac{1}{2}(\sigma^2)^{-3},$$

por lo tanto se concluye que

$$P(\theta)\propto(\sigma^2)^{-3/2}.$$

La distribución de Jeffreys como se ilustra en los dos ejemplos anteriores resulta en muchos de los casos impropia, sin embargo, respeta el principio de invarianza.

TEOREMA 4.2.3.1 *Si* $\varphi=\varphi(\theta)$ *es una transformación uno a uno de* θ, *entonces*

$$P\varphi(\varphi)=P(\theta(\varphi))\left|J_\theta(\varphi)\right|,\quad \varphi\in\Phi$$

donde $\Phi=\varphi(\Theta)$ *y* $P_\varphi(\varphi)\propto\left|\det\{I_\varphi(\varphi)\}\right|^{1/2}$ *es la distribución inictal no informativa de Jeffreys (y* $I_\varphi(\varphi)$ *la información de Fisher) para el parámetro* φ.

Demostración:

La distribución no informativa de Jeffreys para θ es $P(\theta)\propto\left|\det\{I_\theta(\theta)\}\right|^{1/2}$ donde $I_\theta(\theta)=-E_{X|\theta}\left[\dfrac{\partial^2\log P(x|\theta)}{\partial\theta'\partial\theta}\right]$. Sea ahora $\varphi=\varphi(\theta)$ una reparametrización del modelo con inversa $\theta(\cdot)=\varphi^{-1}(\cdot)$.

La distribución no informativa de Jeffreys para φ es

$$P_\varphi(\varphi)\propto\left|\det\{I_\varphi(\varphi)\}\right|^{1/2}=\left|\det\left\{-E_{x|\varphi}\left[\frac{\partial^2\log P(x|\varphi)}{\partial\varphi'\partial\varphi}\right]\right\}\right|^{1/2} \tag{1}$$

$$\frac{\partial\log P(X|\varphi)}{\partial\varphi}=\frac{\partial\log P(X|\theta(\varphi))}{\partial\varphi}=\frac{\partial\log P(X|\theta)}{\partial\theta}\frac{\partial\theta}{\partial\varphi}$$

entonces

$$\frac{\partial^2 \log P(X|\varphi)}{\partial\varphi'\partial\varphi} = \left(\frac{\partial^2 \log P(X|\theta)}{\partial\theta'\partial\theta}\frac{\partial\theta}{\partial\varphi}\right)\frac{\partial\theta}{\partial\varphi} + \frac{\partial \log P(X|\theta)}{\partial\theta}\frac{\partial^2\theta}{\partial\varphi'\partial\varphi}$$

$$= \frac{\partial^2 \log P(X|\theta)}{\partial\theta'\partial\theta}\left(\frac{\partial\theta}{\partial\varphi}\right)^2 + \frac{\partial \log P(X|\theta)}{\partial\theta}\frac{\partial^2\theta}{\partial\varphi'\partial\varphi}. \qquad (2)$$

Obteniendo el valor esperado de (2) se tiene

$$E_{X\theta}\left(\frac{\partial^2 \log P(X|\varphi)}{\partial\varphi'\partial\varphi}\right) = E_{X|\theta}\left(\frac{\partial^2 \log P(X|\theta)}{\partial\theta'\partial\theta}\right)\left(\frac{\partial\theta}{\partial\varphi}\right)^2 + E_{X|\theta}\left(\frac{\partial \log P(X|\theta)}{\partial\theta}\right)\left(\frac{\partial^2\theta}{\partial\varphi'\partial\varphi}\right)$$

$$= -I_\theta(\theta)\left(\frac{\partial\theta}{\partial\varphi}\right)^2 + \frac{\partial^2\theta}{\partial\varphi'\partial\varphi}\int\frac{\partial \log P(X|\theta)}{\partial\theta}P(X|\theta)dX$$

$$= -I_\theta(\theta)\left(\frac{\partial\theta}{\partial\varphi}\right)^2 + \frac{\partial^2\theta}{\partial\varphi'\partial\varphi}\int\frac{\partial P(X|\theta)}{P(X|\theta)}P(X|\theta)dX.$$

Pero se tiene que

$$\int\frac{\partial P(X|\theta)}{P(X|\theta)}P(X|\theta)dX = \int\frac{\partial P(X|\theta)}{\partial\theta}dX$$

y bajo ciertas condiciones de regularidad, las cuales suponen que se puede intercambiar el orden entre derivación e integración

$$\int\frac{\partial P(X|\theta)}{\partial\theta}dX = \frac{\partial}{\partial\theta}\int P(X|\theta) = \frac{\partial}{\partial\theta}(1) = 0.$$

Por lo tanto

$$E_{X|\varphi}\left(\frac{\partial^2 \log P(X|\theta)}{\partial\varphi'\partial\varphi}\right) = -I_\theta(\theta)\left(\frac{\partial\theta}{\partial\varphi}\right)^2. \qquad (3)$$

Sustituyendo (3) en (1)

$$P(\varphi) \propto \left|\det\left\{-\left[-I_\theta(\theta)\left(\frac{\partial\theta}{\partial\varphi}\right)^2\right]\right\}\right|^{1/2} = \left|\det\left\{I_\theta(\theta)\left(\frac{\partial\theta}{\partial\varphi}\right)^2\right\}\right|^{1/2}$$

y como el determinante de un producto es el producto de los determinantes

$$P(\varphi) \propto \left|\det I(\theta)\right|^{1/2} \left|\det\left(\frac{\partial\theta}{\partial\varphi}\right)^{2}\right|^{1/2}$$

$$P(\varphi) \propto P(\theta)\left|\det\left(\frac{\partial\theta}{\partial\varphi}\right)\right| = P(\theta(\varphi))\left|J_{\theta}(\varphi)\right|$$

y por consiguiente

$$P(\varphi) \propto P(\theta(\varphi))\left|J_{\theta}(\varphi)\right| \blacksquare$$

FAMILIAS CONJUGADAS NO INFORMATIVAS.

En la sección 4.2.2 se habló de la conveniencia que representa en algunos casos, el usar como distribución inicial un miembro de la familia conjugada al modelo. Por lo tanto, una manera natural de obtener una distribución a priori no informativa es partir de una familia conjugada y elegir la distribución de tal manera que los parámetros (iniciales) no influyan en la a posteriori, esto es, que la distribución final esté dominada por la información muestral. Para ilustrar esto se presentan los dos ejemplos siguientes.

Ejemplo 4.2.3.4 Sea $\underline{X}=(X_1, X_2, ..., X_n)$ una muestra aleatoria con función de distribución $Blli(\theta)$, $\Theta=[0, 1]$. Se sabe que la distribución inicial conjugada es una distribución beta con parámetros α, β; esto es, $Be(\alpha, \beta)$. La distribución final es por tanto $Be(\alpha+\sum x_i, \beta+n-\sum x_i)$.

Si se hace tender $\alpha \to 0$ y $\beta \to 0$, se tiene que

$$P(\theta|\underline{X}) \propto \theta^{\sum x_i - 1}(1-\theta)^{n-\sum x_i - 1}$$

y por consiguiente

$$P(\theta) \propto \theta^{-1}(1-\theta)^{-1}.$$

Por lo tanto la distribución no informativa usando conjugadas es una "*distribución beta impropia*", $Be(0, 0)$. A pesar de que esta distribución inicial es impropia, la final generalmente es una distribución propia pues $P(\theta|\underline{X}) = Be(\sum x_i,\ n-\sum x_i)$.

Ejemplo 4.2.3.5 Sea $\underline{X}=(X_1, X_2, ..., X_n)$ una muestra aleatoria con función de distribución $N(\mu, 1)$. Se sabe que la distribución inicial conjugada es una distribución normal,

$N(\mu_0, \sigma_0^2)$. La distribución final es por tanto (ver sección 4.2.2) $N\left(\frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}, \frac{\sigma_0^2}{n\sigma_0^2+1}\right)$ con $Y=\sum_{i=1}^{n} x_i$.

Se puede observar que $\frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1} = \frac{Y+\frac{\mu_0}{\sigma_0^2}}{n+\frac{1}{\sigma_0^2}}$ y que $\frac{\sigma_0^2}{n\sigma_0^2+1} = \frac{1}{n+\frac{1}{\sigma_0^2}}$, por lo tanto, haciendo $\sigma_0^2 \to \infty$, los parámetros de la distribución final dependerán únicamente de la información de la muestra. En consecuencia se tendrá que

$$P(\mu|\underline{X})=N\left(\bar{x}, \frac{1}{n}\right) \text{ y}$$

$$P(\mu)\propto 1.$$

Nuevamente, se tiene una a priori no informativa impropia que lleva a una final de masa finita.

Se pueden presentar muchos otros ejemplos, y en general la siguiente definición será de utilidad para determinar iniciales no informativas vía familias conjugadas.

DEFINICIÓN 4.2.3.1 *(Inicial no informativa-familias conjugadas) Sea* $\underline{X}=(X_1, X_2, ..., X_n)$ *una m.a. con* $X \sim f(X|\theta)$, $\Theta \subset R^k$ *y* $F=\{P(\theta \mid \lambda), \lambda \in \Lambda\}$ *familia conjugada paramétrica, de esta manera* $P(\theta \mid \underline{X}) = P(\theta \mid \lambda')$ *en donde* $\lambda' = h(\lambda, Z_n)$. *Si* $\lambda_0 \in \hat{c}\Lambda$ *(frontera de* Λ*) es tal que* λ' *sólo depende de* $\underline{X}$, *entonces* $P(\theta \mid \lambda_0)$ *es la inicial no informativa vía familias conjugadas.*

4.2.4 INICIALES DE MÁXIMA ENTROPÍA

Existen muchas situaciones en las cuales se cuenta con información parcial a priori sobre el parámetro θ. En estos casos no resulta apropiado utilizar una distribución no informativa ya que se perdería la información con que se cuenta. Por otro lado, tal información tampoco es suficiente para especificar completamente $P(\theta)$. Por ejemplo, suponga que se especifica la media a priori, entonces una solución razonable será elegir de entre todas las distribuciones con esta media, aquélla que sea lo menos informativa posible. Un método útil para tratar este

problema es a través del concepto de *Entropía*. Este concepto es más natural para distribuciones discretas, por lo tanto se comenzará con el caso discreto.

DEFINICIÓN 4.2.4.1 *(Entropía) Sea Θ un conjunto discreto y P la densidad de probabilidad sobre Θ. La entropía de P, denotada por $\mathrm{H}(P)$, se define como*

$$\mathrm{H}(P)=-\sum_{\theta}\left(\log P(\theta_i)\right)P(\theta_i)$$

(Si $P(\theta_i)=0$, entonces se define $\left(\log P(\theta_i)\right)P(\theta_i)=0$).

El concepto de entropía tiene una estrecha relación con la cantidad de información disponible, por lo tanto puede ser pensada como una medida de la incertidumbre inherente en la distribución de probabilidad (ver Rosenkrantz, 1977).

Ejemplo 4.2.4.1. (Berger, 1985). Considere $\Theta=\{\theta_1,\theta_2,...,\theta_n\}$. Si $P(\theta_k)=1$ y $P(\theta_i)=0$ para todo $i\neq k$, entonces claramente la distribución de probabilidad describe exactamente la ocurrencia del parámetro θ_k, en otras palabras, se tiene información perfecta. La "incertidumbre" es cero. Correspondientemente

$$\mathrm{H}(P)=-\sum_{\theta}\left(\log P(\theta_i)\right)P(\theta_i)=-\log(1)=0.$$

En la determinación de distribuciones a priori la medida de entropía es de gran importancia pues es deseable obtener la distribución que cumpla con las restricciones iniciales y que además sea la menos informativa, esto es, la de máxima entropía.

Ejemplo 4.2.4.2. Sea $\Theta=\{\theta_1,\theta_2\}$ $P(\theta_1)=p$ y $P(\theta_2)=1-p$, determinar el valor de p de tal manera que la que la distribución resultante tenga máxima entropía.

$$\mathrm{H}(P)=-\left[(\log p)p+\log(1-p)(1-p)\right],$$

entonces $\mathrm{H}'(P)=-\left[(\log p)-\log(1-p)\right]=\log\left(\frac{1-p}{p}\right)$ e igualando la derivada a cero se tiene que $\frac{1-p}{p}=1$. Por lo tanto $p=\frac{1}{2}$ es un punto crítico. Además $\mathrm{H}''(P)=\frac{-1}{p(1-P)}\leq 0$ y en particular, para $p=\frac{1}{2}$ es estrictamente negativa, en consecuencia, este valor maximiza la entropía (ver figura 4.2.2.1).

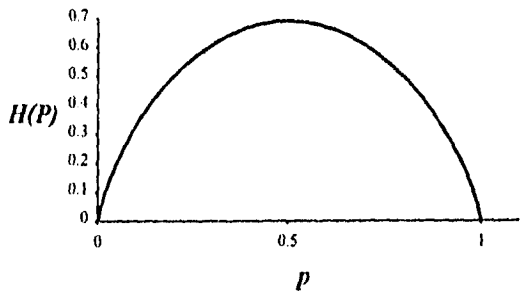

Figura 4.2.2.1

De lo anterior se concluye que la distribución de máxima entropía es $P(\theta_i)=\frac{1}{2}$ para $i=1,2$, que es la distribución uniforme. Cabe mencionar que esta distribución coincide con la obtenida mediante el principio de la razón insuficiente o principio de indiferencia.

De manera análoga a lo realizado para el caso $\Theta=\{\theta_1,\theta_2\}$, se puede demostrar que si se tiene $\Theta=\{\theta_1,\theta_2,...,\theta_n\}$, la distribución de máxima entropía será $P(\theta_i)=\frac{1}{n}$ para $i=1,2,...,n$.

Se sabe que para el caso discreto la entropía de una distribución de probabilidad está dada por $\mathbf{H}(P)=-\sum_{\Theta}(\log P(\theta_i))P(\theta_i)$. De esta expresión es claro que la entropía puede ser también definida como

$$\mathbf{H}(P)=-E_{P(\theta)}(\log P(\theta)).$$

Esta última expresión sugiere la definición de entropía en el caso continuo, pues el operador esperanza puede ser aplicado tanto en el caso discreto como en el caso continuo.

DEFINICIÓN 4.2.4.2 ***(Entropía)*** *Sea θ una variable aleatoria continua con función de densidad $P(\theta)$, entonces la entropía de $P(\theta)$ se define como*

$$\mathbf{H}(P)=-E_{P(\theta)}(\log P(\theta))=-\int_{\Theta}(\log P(\theta))P(\theta)d\theta.$$

Ejemplo 4.2.4.3. Sea $\theta \sim U(a,b)$. Determinar la entropía de $P(\theta)$.

$$\mathrm{H}(U)=\mathrm{H}\left(\frac{1}{b-a}I_{(a,b)}(x)\right)=-E\left(\log\frac{1}{b-a}I_{(a,b)}(x)\right)$$

$$=-\log\frac{1}{b-a}=\log(b-a)$$

por lo tanto $\mathrm{H}(U)=\log(b-a)$.

Ejemplo 4.2.4.4. Se desea determinar la entropía de $P(\theta)$, en donde $\theta \sim N(\mu,\sigma^2)$. Utilizando la definición de entropía se tiene que

$$\mathrm{H}(P_\theta)=\mathrm{H}\left(\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(\theta-\mu)^2}{2\sigma^2}}\right)$$

$$=-\int\left(\log\frac{1}{\sigma\sqrt{2\pi}}-\frac{(\theta-\mu)^2}{2\sigma^2}\right)P(\theta)d\theta$$

$$=\int_{\mathbb{R}}\frac{(\theta-\mu)^2}{2\sigma^2}P(\theta)d\theta-\int_{\mathbb{R}}\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)P(\theta)d\theta=\tfrac{1}{2}-\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)=\tfrac{1}{2}+\log\left(\sigma\sqrt{2\pi}\right)$$

por lo que se concluye que

$$\mathrm{H}\left(N(\mu,\sigma^2)\right)=\frac{1}{2}+\log\left(\sigma\sqrt{2\pi}\right).$$

Para este caso particular de la normal se puede observar que no existe la distribución de máxima entropía pues

$$\lim_{\sigma\to\infty}\mathrm{H}\left(N(\mu,\sigma^2)\right)=\infty.$$

En general, no existe la distribución inicial de máxima entropía sin embargo cuando el soporte de θ es finito o cuando se conocen la media y la varianza a priori se pueden obtener algunos resultados.

TEOREMA 4.2.4.1 *Sean* $p(\theta)$ *y* $q(\theta)$ *dos funciones de densidad para* θ *en* $\Theta\subset\mathbb{R}$, *si* $\int_\theta \log(p(\theta))p(\theta)d\theta$ *y* $\int_\theta \log(q(\theta))p(\theta)d\theta$ *existen, entonces*

$$-\int_\theta \log(p(\theta))p(\theta)d\theta \le -\int_\theta \log(q(\theta))p(\theta)d\theta\ .$$

Demostración:

La desigualdad de Jensen (Mood, Graybill and Boes, 1974) dice que si $g: \mathrm{R} \to \mathrm{R}$ es una función continua convexa y θ una v.a. con media $E(\theta)$ entonces $E(g(\theta)) \geq g(E(\theta))$. Una función es cóncava si y sólo si su negativa es convexa, la función logaritmo natural es cóncava, por tanto su negativa es convexa, entonces por la desigualdad de Jensen

$$E_{P(\theta)}\left[-\log\left(\frac{q(\theta)}{p(\theta)}\right)\right] \geq -\log E_{P(\theta)}\left[\frac{q(\theta)}{p(\theta)}\right],$$

por lo que $E_{P(\theta)}\left[\log\left(\frac{q(\theta)}{p(\theta)}\right)\right] \leq \log E_{P(\theta)}\left[\frac{q(\theta)}{p(\theta)}\right]$, lo cual sucede si y sólo si $E_{P(\theta)}\left[\log\left(\frac{q(\theta)}{p(\theta)}\right)\right] \leq \log \int_{\theta} \frac{q(\theta)}{p(\theta)} p(\theta)\, d\theta = \log(1)$, y de aquí que

$$E_{P(\theta)}\left[\log\left(\frac{q(\theta)}{p(\theta)}\right)\right] \leq 0.$$

Por propiedades de la función logaritmo y debido a la aditividad de la función esperanza

$$\int_{\theta} \log(q(\theta)) p(\theta)\, d\theta - \int_{\theta} \log(p(\theta)) p(\theta)\, d\theta \leq 0,$$

por lo que finalmente $-\int_{\theta} \log(p(\theta)) p(\theta)\, d\theta \leq -\int_{\theta} \log(q(\theta)) p(\theta)\, d\theta$ ■

COROLARIO 4.2.4.1 *Sea θ una variable aleatoria continua con soporte en $[a,b]$ y función de densidad $P(\theta)$, entonces*

$$H(P(\theta)) \leq H(U(a,b)) = \log(b-a).$$

Demostración:

Por el teorema 4.2.4.1

$$\begin{aligned} -\int_a^b \log(p(\theta)) p(\theta)\, d\theta &\leq -\int_a^b \log\left(\frac{1}{b-a}\right) p(\theta)\, d\theta \\ &= -\log\left(\frac{1}{b-a}\right) \int_a^b p(\theta)\, d\theta = -\log\left(\frac{1}{b-a}\right) \\ &= \log(b-a). \end{aligned}$$

De la última expresión se tiene que $H(P(\theta)) \leq \log(b-a)$ y en el problema 4.2.4.3. se concluyó que $\mathrm{H}(U) = \log(b-a)$, por lo tanto $H(P(\theta)) \leq H(U(a,b))$ ■

COROLARIO 4.2.4.2 *Sea* θ *una variable aleatoria continua con soporte en* $\mathbf{R}$, *si* $E_{P(\theta)}(\theta)=\mu$ *y* $Var_{P(\theta)}(\theta)=\sigma^2$, *entonces*

$$H(P(\theta)) \leq H(N(\mu,\sigma^2))$$

Demostración:

Por el teorema 4.2.4.1

$$-\int_{\mathbf{R}} \log(p(\theta))p(\theta)d\theta \leq -\int_{\mathbf{R}} \log(N(\mu,\sigma^2))p(\theta)d\theta$$

$$= -\int_{\mathbf{R}} \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\theta-\mu)^2}{2\sigma^2}}\right) p(\theta)d\theta$$

$$= -\int \left(\log\frac{1}{\sigma\sqrt{2\pi}} - \frac{(\theta-\mu)^2}{2\sigma^2}\right) P(\theta)d\theta$$

$$= \int_{\mathbf{R}} \frac{(\theta-\mu)^2}{2\sigma^2} P(\theta)d\theta - \int_{\mathbf{R}} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) P(\theta)d\theta = \frac{1}{2} - \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) = \frac{1}{2} + \log\left(\sigma\sqrt{2\pi}\right)$$

y del problema 4.2.2.4 se concluyó que $\mathbf{H}(N(\mu,\sigma^2)) = \frac{1}{2} + \log\left(\sigma\sqrt{2\pi}\right)$, por lo tanto

$H(P(\theta)) \leq H(N(\mu,\sigma^2))$ ■

La utilidad de estos dos corolarios en la determinación de distribuciones iniciales es muy clara. En conclusión se tiene que si θ es una variable continua en $[a,b]$, entonces la distribución que cumple con el principio de máxima entropía es la $U(a,b)$. Por otra parte, si $\theta \in \mathbf{R}$ con media y varianza a priori μ y σ^2 respectivamente, la distribución $N(\mu,\sigma^2)$ es la de máxima entropía.

En muchos problemas de las ciencias físicas se dispone de información expresada en momentos de la a priori, en tales casos, el uso de distribuciones de máxima entropía reporta excelentes resultados.

4.3 ALTERNATIVAS PARA EL CÁLCULO Y EXPLORACIÓN DE DISTRIBUCIONES FINALES.

En la mayoría de los problemas estadísticos, la inferencia se hace a partir de la distribución final, la cual resume la información a priori y la información muestral. El Teorema de Bayes proporciona la manera de determinar tal distribución a posteriori, sin embargo en muchos de los casos será difícil encontrar una expresión analítica para ella. Es por eso que resulta necesario encontrar formas alternas de calcularla y explorarla.

4.3.1 APROXIMACIÓN ASINTÓTICA NORMAL PARA LA DISTRIBUCIÓN FINAL.

Sea $\underline{X}=(X_1,...,X_n)$ una muestra de observaciones independientes de $P(X|\theta)$ con $\theta \in \mathbf{R}^k$ y $P(\theta)$ una distribución inicial sobre θ. La distribución final de θ está dada entonces por $P(\theta|\underline{X}) \propto P(\theta)P(\underline{X}|\theta)$, lo cual pasa si y sólo si

$$P(\theta|\underline{X}) \propto \exp\{\log P(\theta)+\log P(\underline{X}|\theta)\}.$$

Si ahora se considera la expansión de Taylor para $\log P(\theta)$ y $\log P(\underline{X}|\theta)$ alrededor de sus respectivos máximos m_0 y $\hat{\theta}$, se tiene

$$\log P(\theta)=\log P(m_0)-\tfrac{1}{2}(\theta-m_0)'H_0(\theta-m_0)+R_0$$

$$\log P(\underline{X}|\theta)=\log P(\underline{X}|\hat{\theta})-\tfrac{1}{2}(\theta-\hat{\theta})'H(\theta-\hat{\theta})+R;$$

en donde $H_0=\left(-\dfrac{\partial^2 \log P(\theta)}{\partial\theta_i\partial\theta_j}\right)_{\theta=m_0}$, $H=\left(-\dfrac{\partial^2 \log P(\underline{X}|\theta))}{\partial\theta_i\partial\theta_j}\right)_{\theta=\hat{\theta}}$ y R_0 y R denotan los residuos de segundo orden de las correspondientes expresiones.

Bajo ciertas condiciones de regularidad, que garantizan que R_0 y R son pequeños si n es grande, se tiene que aproximadamente

$$P(\theta|\underline{X}) \propto \exp\left\{-\tfrac{1}{2}(\theta-m_0)'H_0(\theta-m_0)-\tfrac{1}{2}(\theta-\hat{\theta})'H(\theta-\hat{\theta})\right\}$$

y por lo tanto

$$P(\theta|\underline{X}) \dot{\propto} \exp\left\{-\tfrac{1}{2}(\theta-m_n)'H_n(\theta-m_n)\right\},$$

en donde $H_n=H_0+H$ y $m_n=H_n^{-1}(H_0m_0+H\hat{\theta})$.

En conclusión, $P(\theta|\underline{X})$ puede aproximarse, si n es "grande", a través de la distribución normal multivariada

$$N_k\left(\theta \mid m_n,\ H_n^{-1}\right)$$

APROXIMACIÓN ASINTÓTICA ALTERNATIVA

A continuación se presenta una aproximación asintótica a los parámetros de la distribución final que no toma en cuenta el conocimiento inicial sobre el parámetro θ. Por la Ley fuerte de los Grandes Números

$$\lim_{n\to\infty}\left\{\frac{1}{n}\left(\frac{-\partial^2 \log P(\underline{X}|\theta))}{\partial\theta_i\partial\theta_j}\right)\right\} = \lim_{n\to\infty}\left\{\frac{1}{n}\sum_{r=1}^{n}\frac{-\partial^2 \log P(X_r|\theta))}{\partial\theta_i\partial\theta_j}\right\}$$

$$\cong \int P(X|\theta)\left(\frac{-\partial^2 \log P(X|\theta))}{\partial\theta_i\partial\theta_j}\right)dX \quad (\forall i=1,\ldots,k).$$

De esta manera, si n es "grande", $H_n \approx nI(\hat{\theta})$, en donde $I(\theta)$ denota la matriz de información de Fisher por unidad muestral.

Si n es "grande", la precisión de la distribución inicial tenderá a ser pequeña comparada con la precisión obtenida de los datos. Así, la distribución final $P(\theta|\underline{X})$ puede aproximarse, si n es "grande", a través de la distribución normal multivariada

$$N_k\left(\theta \mid \hat{\theta},\ n^{-1}I(\hat{\theta})^{-1}\right)$$

4.3.2 ALGORITMO COMPUTACIONAL PARA LA EXPLORACIÓN DE DISTRIBUCIONES FINALES.

En la Teoría de la Decisión Estadística, una vez que se tiene la distribución final es necesario obtener una colección de distribuciones y momentos marginales, ya que a partir de ellos es que se podrá hacer inferencia sobre los parámetros. Esta transición de distribución conjunta a distribuciones marginales (y cálculo de momentos) implica cálculo de integrales de $P(\theta|\underline{X})$. En muchos de los casos, la solución analítica es casi imposible.

Otro problema frecuente es que resulta difícil encontrar una expresión analítica para la distribución final, y en muchas de estas aplicaciones específicas no es fácil verificar si la distribución asintótica es adecuada (sobre todo si n es "pequeña").

Esta serie de problemas, en conjunción con el desarrollo computacional, han motivado el desarrollo de métodos eficientes de integración. Uno de los más conocidos y útiles es el "Muestreo de Gibbs".

MUESTREO DE GIBBS (Gibbs sampler)

El "Muestreo de Gibbs" es una técnica para generar (indirectamente) variables aleatorias de una distribución (marginal o conjunta) sin tener que calcular la densidad. Este método está basado en propiedades elementales de Cadenas de Markov. A pesar de que la mayoría de las aplicaciones de la técnica ha sido en modelos Bayesianos, ha representado también gran utilidad en la estadística clásica.

Sea $f(x,y_1,...,y_p)$, considere que se desea obtener la media de la densidad marginal

$$f(x)=\int...\int f(x,y_1,...,y_p)dy_1...dy_p .$$

La manera natural de resolver este problema es calcular $f(x)$ y a partir de ahí obtener la media; sin embargo pensemos en el caso en el que es extremadamente difícil calcular la integral, ya sea para obtener la marginal, la media o ambas. En tal caso, el "Muestreo de Gibbs" proporciona una forma alternativa para determinar $f(x)$ (y/o sus momentos). A diferencia de los métodos numéricos tradicionales el "Muestreo de Gibbs", más que tratar de aproximar directamente $f(x)$, permite generar una muestra $X_1,...X_m$ de $f(x)$ sin requerir la función de densidad. Si se simula una muestra suficientemente grande, la media, varianza, o cualquier otra característica de $f(x)$ puede ser aproximada con el grado de exactitud deseado.

Una vez que se tiene $X_1,...X_m$, $E(X)$ puede ser estimada por $\overline{X}=\sum_{i=1}^{m} X_i$, ya que es bien sabido que $\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} xf(x)dx = E(X)$. Así si se toma m suficientemente grande, cualquier característica poblacional, aún la densidad misma, puede ser aproximada.

Para ilustrar el funcionamiento del método, considere el par de variables aleatorias (X,Y). El "Muestreo de Gibbs" genera una muestra de $f(x)$, a partir de las distribuciones condicionales $f(x|y)$ y $f(y|x)$, las cuales se suponen conocidas y se pueden simular. Esto se hace generando una "secuencia Gibbs" de variables aleatorias

$$Y_0', X_0', Y_1', X_1', ..., Y_k', X_k'.$$

El valor inicial de $Y_0' = y_0'$ es especificado y el resto se obtiene mediante iteraciones generando, alternadamente, valores de

$$X_j' \sim f\left(x|Y_j' = y_j'\right)$$
$$Y_{j+1}' \sim f\left(y|X_j' = x_j'\right)$$

Bajo ciertas condiciones (ver Schervish y Carlin, 1990) la sucesión X_k' proviene de $f(x)$. Por ejemplo, Gelfand y Smith (1990) sugieren generar m secuencias de Gibbs independientes y de longitud k; y usar el valor de X_k' de cada secuencia. Si k se elige suficientemente grande, se tiene una muestra (aproximada) de variables aleatorias independientes e idénticamente distribuidas de acuerdo a $f(x)$.

El "Muestreo de Gibbs" puede también ser usado para estimar la densidad $f(x)$. Promediando las densidades condicionales finales de cada secuencia Gibbs se obtiene

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} f\left(x|y_i\right),$$

en donde $y_1,...,y_m$ son los valores obtenidos de la variable Y final de cada secuencia. La teoría detrás del cálculo es que el valor esperado de la densidad condicional es

$$E\left[f\left(x|Y\right)\right] = \int f\left(x|y\right)f(y)dy = f(x).$$

Las cantidades $f\left(x|y_1\right),...,f\left(x|y_m\right)$, calculadas usando los valores simulados $y_1,...y_m$, tienen más información sobre $f(x)$ que $x_1,...x_m$ solas, y por lo tanto proporcionarán mejores estimadores.

En conclusión, el "Muestreo de Gibbs" genera una cadena de Markov de variables aleatorias, la cual converge a la distribución conjunta (distribución de equilibrio) de las variables bajo consideración. Las condiciones necesarias para la convergencia del método son discutidas a detalle por Schervish y Carlin (1990) y Tierney (1994).

La utilidad del "Muestreo de Gibbs" se incrementa a medida que crece la dimensión del problema. Esto es porque el método evita calcular integrales complicadas. Más aún, los cálculos de integrales múltiples pueden ser reemplazados por una serie de simulaciones de variables aleatorias univariadas.

Cabe señalar que el "Muestreo de Gibbs" ha sido extensamente usado en problemas prácticos tanto Bayesianos como clásicos. En el ámbito Bayesiano , el "Muestreo de Gibbs" es básicamente usado para generar distribuciones a posteriori, mientras que para los clásicos el uso se ha centrado en el cálculo de funciones de verosimilitud y características de los estimadores máximo verosímiles.

CAPÍTULO 5

INFERENCIA ESTADÍSTICA

En los capítulos anteríores se ha presentado toda la herramienta teórica para abordar problemas específicos de inferencia desde la perspectiva de la Teoría de la Decisión Estadística (paradigma Bayesiano). Por lo tanto, su planteamiento, así como su solución resultan naturales.

En este último capítulo se estudian algunos resultados generales sobre problemas típicos de inferencia, como son estimación puntual, estimación por intervalos y prueba de hipótesis. Además, se ilustra mediante ejemplos su tratamiento.

Es importante no perder de vista que el objetivo principal de muchas aplicaciones específicas es el de estar en la posibilidad de predecir valores futuros de la variable aleatoria, generalmente denominada por X. Este es el motivo por el cual el problema de predicción será también tema de estudio en este capítulo.

5.1 ESTIMACIÓN PUNTUAL

El problema de estimación puntual es un problema de decisión estadístico definido por $(D, \Theta, p(\theta), L(d,\theta))$, en el cual, la decisión consiste del valor con el que se estimará algún parámetro θ perteneciente al conjunto Θ. Por lo tanto, el espacio de decisiones coincide con el espacio parametral, esto es, $D = \Theta$.

La decisión será entonces $d = \hat{\theta}$, por lo cual, la función de pérdida, $L(d,\theta)$, deberá reflejar de alguna manera la discrepancia entre el valor θ y el valor estimado, $\hat{\theta}$. Es por ello que frecuentemente, $L(d,\theta)$ tiene la forma

$$L(\theta,d) = k\Lambda(\theta - d)$$

o equivalentemente

$$L(\theta,\hat{\theta}) = k\Lambda(\theta-\hat{\theta}), \qquad (1)$$

en donde Λ es una función no negativa del error $(\theta-\hat{\theta})$, de tal suerte que $\Lambda(0)=0$ y k es una constante positiva que pondera la relación del error con respecto a los posibles valores de θ. Ejemplos de este tipo de funciones de pérdida son la pérdida cuadrática y la pérdida proporcional al error absoluto (ejemplos 4.1.1. y 4.1.2. respectivamente).

En el capítulo 3 se demostró que la solución consistente con los axiomas de coherencia es la decisión de Bayes, y como en este tipo de problemas la decisión óptima es un estimador, se puede hablar del Estimador Bayesiano.

DEFINICIÓN 5.1.1. *(Estimador Bayesiano) Un estimador Bayesiano de* θ *con respecto a la función de pérdida* $L(d,\theta) = L(\theta,\hat{\theta})$ *y la distribución* $p(\theta)$ *es cualquier* $\hat{\theta}^* \in \Theta$ *que minimiza*

$$E_{p(\theta)}\left[L(\theta,\hat{\theta})\right] = \int_{\Theta} L(\theta,\hat{\theta})p(\theta)d\theta .$$

EJEMPLO 5.1.1. Sea X una v.a. Bernoulli(θ). Se desea conocer el estimador (puntual) Bayesiano para θ cuando se utiliza una distribución inicial conjugada, la información de una muestra aleatoria $\underline{X} = (X_1,\ldots,X_n)$ y la función de pérdida

$$L(\theta,\hat{\theta}) = \frac{\left(\frac{\theta-\hat{\theta}}{\theta}\right)^2}{(1-\theta)^2}.$$

En la tabla 4.2.2.1 de familias conjugadas se observa que la conjugada natural es la distribución beta, $Be(\alpha_0,\beta_0)$, por lo que la distribución final será otra beta con parámetros actualizados (α,β), en donde $\alpha = \alpha_0 + \sum_{i=1}^{n} X_i$ y $\beta = \beta_0 + n - \sum_{i=1}^{n} X_i$. Por lo tanto

$$p(\theta|\underline{X}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Ahora, calculando la pérdida esperada se tiene

$$E_{P'(\theta|X)}\left[L(\theta,\hat{\theta})\right] = k\int_{\Theta}\frac{(\theta-\hat{\theta})^2}{\theta^2}\left(\frac{1}{(1-\theta)^2}\right)\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = k\int_{\Theta}(\hat{\theta}-\theta)^2\theta^{(\alpha-2)-1}(1-\theta)^{(\beta-2)-1}d\theta$$
$$= k'E_{P'(\theta|X)}\left[(\theta-\hat{\theta})^2\right]$$

en donde $P'(\theta|X) = Be(\theta|\alpha-2,\beta-2)$. Evidentemente, para que la esperanza exista se requiere que $\alpha > 2$ y $\beta > 2$.

De la última igualdad es claro que el $\hat{\theta}^*$ que minimiza $E_{P'(\theta|X)}\left[L(\theta,\hat{\theta})\right]$ es el mismo que minimiza $k'E_{P'(\theta|X)}\left[(\theta-\hat{\theta})^2\right]$, y en el ejemplo 4.1.1 se demostró que cuando se tiene una pérdida cuadrática la decisión de Bayes es la media de la distribución. Por tanto el estimador Bayesiano es $\hat{\theta}^* = E_{P'(\theta|X)}[\theta] = \frac{\alpha-2}{\alpha+\beta-4}$.

En el problema de estimación como en cualquier otro problema de decisión cada investigador puede elegir la función de pérdida que mejor satisfaga sus intereses; sin embargo, debido a la naturaleza de este problema, algunas de las pérdidas con las que salen expresiones algebraicamente cerradas son la pérdida cuadrática, la pérdida proporcional al error absoluto, y la divergencia logarítmica de Kullback-Leibler.

PÉRDIDA CUADRÁTICA

A pesar de que la función de pérdida cuadrática, $L(\theta,\hat{\theta}) = k(\theta-\hat{\theta})^2$, no es ni acotada ni cóncava es utilizada con alguna frecuencia en problemas de decisión. Una de las razones para ello es la relación estrecha entre el Valor de Bayes correspondiente, $kE_{P(\theta)}\left[(\theta-\hat{\theta})^2\right]$ y la teoría clásica de mínimos cuadrados. Además, los cálculos necesarios para llevar al cabo el proceso de estimación son generalmente sencillos.

Una justificación que pudiera ser más convincente para el uso de esta función de pérdida es la siguiente. Sea $L(\theta,\hat{\theta})$ definida como en (1) y supóngase que es diferenciable al menos de orden 2. Si se aproxima con una serie de Taylor con términos hasta de segundo orden (bajo la idea de que debido a que la diferencia $(\theta-\hat{\theta})$ debe ser pequeña los términos de orden superior serán despreciables) se tiene que:

$$L(\theta,\hat{\theta}) \approx c_0 + c_1(\theta-\hat{\theta}) + c_2(\theta-\hat{\theta})^2.$$

Como $\Lambda(\theta)=0$ se implica que $c_0=0$ y debido a que $L(\theta,\hat{\theta})$ debe ser una función no negativa se requiere que $c_1=0$. Por lo tanto, la aproximación se reduce a $k(\theta-\hat{\theta})^2$, con $k>0$. En conclusión, se dice que en un contexto más general, la pérdida cuadrática puede ser considerada como una buena aproximación a una función más apropiada.

Cabe recordar que cuando se considera adecuado el uso de la pérdida cuadrática, el estimador Bayesiano para un valor real θ es (ver ejemplo 4.1.1 para demostración)

$$\hat{\theta}^* = E_{p(\theta)}(\theta).$$

EJEMPLO 5.1.2. Sea $\underline{X}=(X_1,\ldots,X_n)$ una muestra aleatoria con X v.a. Bernoulli(θ). Se desea estimar θ puntualmente cuando se considera una distribución inicial conjugada y una función de pérdida cuadrática.

Del ejemplo 5.1.1 se observa que la distribución inicial es una beta, $Be(\alpha_0,\beta_0)$, y que por consiguiente la distribución final será $Be(\alpha,\beta)$, en donde $\alpha=\alpha_0+\sum_{i=1}^{n}X_i$ y $\beta=\beta_0+n-\sum_{i=1}^{n}X_i$.

Es bien conocido que la media de la distribución $Be(\alpha,\beta)$ es $\frac{\alpha}{\alpha+\beta}$, por lo tanto el estimador Bayesiano está dado por

$$\hat{\theta}=\frac{\alpha_0+\sum_{i=1}^{n}X_i}{\alpha_0+\beta_0+n}.$$

Si se desea ser no informativo se hacen desaparecer los parámetros de la inicial, esto es, α_0 y $\beta_0 \to 0$, y con ello se obtiene que

$$\hat{\theta}^* = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x},$$

que desde el punto de vista clásico es el mejor estimador linealmente insesgado de varianza mínima (y además máximo verosímil), esto es, es el estimador óptimo (BLUE por sus siglas en inglés).

EJEMPLO 5.1.3. Sea $\underline{X}=(X_1,...,X_n)$ una muestra aleatoria en donde X se distribuye $N(\mu,1)$. Se desea estimar μ puntualmente cuando se considera una distribución inicial conjugada y una función de pérdida cuadrática.

Se sabe que la distribución inicial conjugada es una distribución normal, $N(\mu_0, \sigma_0^2)$ y la distribución final es por tanto $N\left(\frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}, \frac{\sigma_0^2}{n\sigma_0^2+1}\right)$ con $Y=\sum_{i=1}^{n} x_i$.

Por lo tanto el estimador Bayesiano es

$$\hat{\mu}^* = \frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}.$$

Si se hace $\sigma_0^2 \to \infty$, el estimador se reduce (ver ejemplo 4.2.3.4) a $\hat{\mu}^* = \bar{x}$. Nuevamente, el estimador Bayesiano coincide con el clásico cuando se supone una a priori no informativa.

Resultados análogos se pueden observar del ejemplo 4.1.3.

FUNCIÓN DE PÉRDIDA LINEAL

Algunas veces resulta natural utilizar una función de pérdida lineal, sobre todo cuando θ pertenece a un intervalo pequeño de la recta real. Esta pérdida debe reflejar además, la discrepancia entre θ y su estimador. Tras estas dos consideraciones surge la función de pérdida lineal, la cual está dada por

$$L(\theta,\hat{\theta}) = \begin{cases} k_0(\theta-\hat{\theta}) & si\ \theta-\hat{\theta} \geq 0 \\ k_1(\hat{\theta}-\theta) & si\ \theta-\hat{\theta} < 0. \end{cases}$$

Las constantes k_0 y k_1 son positivas y de alguna manera ponderan la sub y sobre estimación respectivamente. Si $k_0 = k_1 = k$, se tiene la función conocida como *pérdida proporcional al error absoluto*,

$$L(\theta,\hat{\theta}) = k|(\theta - \hat{\theta})| \propto |(\theta - \hat{\theta})|.$$

Cabe recordar que cuando se considera apropiado el uso de la pérdida proporcional al error absoluto, el estimador Bayesiano para un valor real θ es la mediana de la distribución correspondiente (ver ejemplo 4.1.2 para demostración).

EJEMPLO 5.1.4. En el contexto del problema 5.1.3 considere ahora una función de pérdida dada por $L(\mu,\hat{\mu}) = k|(\mu - \hat{\mu})|$.

El estimador de Bayes para este caso será la mediana de la distribución final, pero cuando se tiene una distribución normal, por ser ésta simétrica, la mediana coincide con la media, por lo tanto, el estimador será el mismo que cuando se supone una pérdida cuadrática, esto es $\hat{\mu}^* = \frac{Y\sigma_0^2 + \mu_0}{n\sigma_0^2 + 1}$ o $\hat{\mu}^* = \bar{x}$, cuando se supone una inicial no informativa vía familias conjugadas.

En general la media y la mediana de una distribución de probabilidad son distintas, en tales casos la pérdida cuadrática y la pérdida proporcional al error absoluto darán estimaciones diferentes.

Se podrían seguir analizando casos particulares de estimación considerando diferentes funciones de pérdida, sin embargo, son de mayor utilidad los resultados generales que se presentan a continuación.

PROPOSICIÓN 5.1.1

A. (*Función de pérdida cuadrática*). *Sea* $\Theta = \mathbb{R}^k$ *y* $L(\theta,\hat{\theta}) = (\theta - \hat{\theta})' H(\theta - \hat{\theta})$, *en donde* H *es una matriz conocida, simétrica y definida positiva. El estimador Bayesiano es cualquier vector* $\hat{\theta}^*$ *que satisface*

$$H\hat{\theta}^* = H \cdot E(\theta).$$

Si la matriz H *es invertible*

$$\hat{\theta}^* = E_{p(\theta)}(\theta).$$

B. *Sea* $\Theta = \mathbb{R}$ y $L(\theta,\hat{\theta}) = k_1(\hat{\theta}-\theta)I_{(\theta\le\hat{\theta})}(\theta) + k_2(\theta-\hat{\theta})I_{(\theta>\hat{\theta})}(\theta)$, *entonces el estimador de Bayes,* $\hat{\theta}^*$, *es tal que*

$$p(\theta \le \hat{\theta}^*) = \frac{k_2}{k_1+k_2}.$$

En particular si $k_1 = k_2 = k$ *se sigue que* $L(\theta,\hat{\theta}) = k|(\theta-\hat{\theta})|$ *y por lo tanto el estimador será igual a la mediana de* θ.

Demostración:

A) La pérdida esperada es

$$E_{p(\theta)}(L(\theta,\hat{\theta})) = \int(\theta-\hat{\theta})'H(\theta-\hat{\theta})p(\theta)d\theta,$$

derivando con respecto a $\hat{\theta}$ (bajo ciertas condiciones de regularidad) e igualando a cero se sigue que $2H\int(\theta-\hat{\theta}^*)p(\theta)d\theta = 0$, de aquí que

$$H\hat{\theta}^* = H\cdot E(\theta).$$

De la expresión anterior, es claro que si H^{-1} existe, entonces

$$\hat{\theta}^* = E_{p(\theta)}(\theta).$$

Debido a que H es definida positiva, $\hat{\theta}^*$ efectivamente minimiza $E_{p(\theta)}(L(\theta,\hat{\theta}))$.

B) La pérdida esperada está dada por

$$E_{p(\theta)}(L(\theta,\hat{\theta})) = k_1\int_{\theta\le\hat{\theta}}(\hat{\theta}-\theta)p(\theta)d\theta + k_2\int_{\theta<\hat{\theta}}(\theta-\hat{\theta})p(\theta)d\theta.$$

Al derivar con respecto a $\hat{\theta}$ (bajo ciertas condiciones de regularidad) e igualando a cero se sigue que $k_1\int_{\theta\le\hat{\theta}^*}p(\theta)d\theta = k_2\int_{\hat{\theta}<\theta}p(\theta)d\theta$, por lo que al sumar $k_2\int_{\theta\le\hat{\theta}^*}p(\theta)d\theta$ a ambos lados de la igualdad se tiene

$$k_1P(\theta\le\hat{\theta}^*) + k_2P(\theta\le\hat{\theta}^*) = k_2.$$

Por lo tanto el estimador de Bayes es tal que $p(\theta\le\hat{\theta}^*) = \frac{k_2}{k_1+k_2}$. Evidentemente si $k_1 = k_2$, $\frac{k_2}{k_1+k_2} = \frac{1}{2}$ y en consecuencia $\hat{\theta}^*$ es la mediana de θ ■

DIVERGENCIA LOGARÍTMICA DE KULLBACK-LEIBLER

Otra función de pérdida que se utiliza en el problema de estimación estadística es la llamada divergencia logarítmica de Kullback-Leibler (Kullback, Leibler, 1951), la cual más que medir la discrepancia entre θ y $\hat{\theta}$, mide la discrepancia entre los modelos de densidad correspondientes.

DEFINICIÓN 5.1.2: *Sea* $X \sim P(x|\theta)$, *con* $\theta \in \Theta$, *la función de pérdida dada por la divergencia logarítmica de Kullback-Leibler está definida por*

$$L(\theta,\hat{\theta}) = \int P(x|\theta)\log\frac{P(x|\theta)}{P(x|\hat{\theta})}dx.$$

Es claro ver que no es una función simétrica, pero que cumple con una de las características más deseadas en una función de pérdida y es el hecho de que es cóncava (Gutiérrez, 1991).

PROPOSICIÓN 5.1.2: Dada una muestra $\underline{X}=(X_1, ..., X_n)$ de la densidad exponencial $P(x|\theta)=a(x)\exp\{x'\theta - M(\theta)\}$, y suponiendo una distribución inicial para θ en la familia conjugada natural, el estimador Bayesiano correspondiente a la función de pérdida

$$L(\theta,\hat{\theta}) = \int P(x|\theta)\log\frac{P(x|\theta)}{P(x|\hat{\theta})}dx$$

es la moda de la distribución final de θ.

Demostración:

Es fácil demostrar que la distribución inicial perteneciente a la familia conjugada natural es $P(\theta|t_0,n_0)\propto \exp\{t_0'\theta - n_0 M(\theta)\}$, por lo que la distribución final será entonces

$$P(\theta|\underline{X})\propto \exp\{t_1'\theta - n_1 M(\theta)\},$$

en donde $t_1 = t + t_0$ y $n_1 = n + n_0$, con $t = \sum_{i=1}^{n} x_i$ igual a la estadística suficiente.

El estimador de Bayes es aquel $\hat{\theta}$ que minimiza la pérdida esperada, la cual está dada por

$$E_{P(\theta|\underline{X})}L(\theta,\hat{\theta}) = \int_{\theta}\left(\int P(x|\theta)\log\frac{P(x|\theta)}{P(x|\hat{\theta})}dx\right)P(\theta|\underline{X})d\theta$$

$$= \int_{\theta}\left(\int P(x|\theta)\log P(x|\theta)\,dx\right)P(\theta|\underline{X})d\theta - \int_{\theta}\left(\int P(x|\theta)\log P(x|\hat{\theta})dx\right)P(\theta|\underline{X})d\theta$$

$$= \int_{\theta}\left(\int P(x|\theta)\log P(x|\theta)\,dx\right)P(\theta|\underline{X})d\theta - \int_{\theta}\left(\int\left(\log a(x)+x'\hat{\theta}-M(\hat{\theta})\right)P(x|\theta)dx\right)P(\theta|\underline{X})d\theta$$

El primer término al igual que $\int_{\theta}\left(\int\log a(x)P(x|\theta)dx\right)P(\theta|\underline{X})d\theta$ no depende de $\hat{\theta}$,

por lo cual minimizar $E_{P(\theta|\underline{X})}L(\theta,\hat{\theta})$ es equivalente a maximizar

$$\int_{\theta}\left(\int\left(x'\hat{\theta}-M(\hat{\theta})\right)P(x|\theta)dx\right)P(\theta|\underline{X})d\theta = \int_{\theta}\left(E_{P(x|\theta)}(x')\hat{\theta}-M(\hat{\theta})\right)P(\theta|\underline{X})d\theta \qquad (1).$$

En Gutiérrez (1991) se demuestra que

$$E_{P(x|\theta)}(x) = \frac{\partial M(\theta)}{\partial\theta} \qquad (2)$$

$$E_{P(\theta|\underline{X})}\left[\frac{\partial M(\theta)}{\partial\theta}\right] = \frac{t_1}{n_1}. \qquad (3)$$

Sustituyendo (2) en (1)

$$\int_{\theta}\left(\left(\frac{\partial M(\theta)}{\partial\theta}\right)'\hat{\theta}-M(\hat{\theta})\right)P(\theta|\underline{X})d\theta = E_{P(\theta|\underline{X})}\left[\left(\frac{\partial M(\theta)}{\partial\theta}\right)'\right]\hat{\theta}-M(\hat{\theta}).$$

Ahora sustituyendo (3) en la expresión anterior se tiene

$$\frac{t_1'}{n_1}\hat{\theta}-M(\hat{\theta}) = \frac{1}{n_1}\left(t_1'\hat{\theta}-n_1M(\hat{\theta})\right)$$

y maximizar esta expresión con respecto a $\hat{\theta}$ (y debido a que $\frac{1}{n_1}>0$) es equivalente a maximizar $\left(t_1'\theta-n_1M(\theta)\right)$ con respecto a θ.

Ahora, como la función exponencial es monótona creciente se sigue que

$$\min_{\theta}E_{P(\theta|\underline{X})}(L(\theta,\hat{\theta})) \text{ es equivalente a } \max_{\theta}P(\theta|\underline{X});$$

por lo tanto el estimador Bayesiano será la moda de la distribución final.

En este ejemplo, es claro que si se supone una inicial vía familia conjugadas, el estimador Bayesiano será igual al estimador máximo verosímil de la estadística clásica.

5.2 ESTIMACIÓN POR REGIONES (Intervalos cuando $\theta \in \mathbf{R}$)

La distribución final o a posteriori contiene toda la información relevante sobre el parámetro de interés, por lo cual no sólo se cuenta con un estimador puntual sino que se conoce el comportamiento del parámetro mismo. Es por ello que muchos problemas referentes al parámetro pueden ser resueltos. Uno de estos problemas de interés es el de determinar un intervalo, y en general una región de probabilidad, que contenga el valor del parámetro con el cual se condicionó el modelo. Este problema es conocido como "estimación por regiones (intervalos)". Formalmente, una región de probabilidad se define de la siguiente manera.

DEFINICIÓN 5.2.1. ***(Región de probabilidad)*** *Una región* $C \subseteq \Theta$ *es llamada región de probabilidad del* $(1-\alpha)\times 100\%$ $(0<\alpha<1)$, *si es tal que*

$$\int_C p(\theta)d\theta = P(\theta \in C)=1-\alpha\,.$$

Obviamente cuando $\theta \in \mathbf{R}$, $C \subseteq \Theta$ será un intervalo de la recta real.

Para un α particular puede haber varias regiones que tengan probabilidad igual a $1-\alpha$; el problema será entonces determinar con cuál región quedarse. Intuitivamente, la respuesta sería optar por aquella región cuyo "tamaño" sea el menor. Más formalmente, el problema puede ser planteado como un problema de decisión.

El espacio de decisiones es $D = \{C \mid P(\theta \in C)=1-\alpha\}$. Una función de pérdida que refleja esa idea intuitiva de que entre más grande sea la región más pérdida se debe tener es

$$L(\theta, C) = k\|C\| - I_C(\theta), \quad k>0\,,$$

en donde $\|C\|$ representa la norma de C.

Bajo este contexto y cuando $p(\theta)$ es continua, la solución Bayesiana es elegir la región C^*, cuyo tamaño sea mínimo. Esta región coincide con la región que contiene los puntos de mayor densidad. De esta manera la solución es frecuentemente llamada la región de máxima densidad final. En la literatura es también conocido con nombres como "Región Bayesiana de Confianza" y "Región de Credibilidad".

DEFINICIÓN 5.2.2 (*Región de máxima densidad*): *Una región* $C \subseteq \Theta$ *es llamada región de máxima densidad* $(1-\alpha)\times 100\%$ *con respecto a* $p(0)$ *si*

i. $P(C)=1-\alpha$

ii. $p(0_1) \geq p(0_2)$ *para todo* $0_1 \in C$ y $0_2 \notin C$, *excepto para un subconjunto de probabilidad cero.*

Ejemplo 5.2.1: Como en el ejemplo 5.1.3, sea $\underline{X}=(X_1,...,X_n)$ una muestra aleatoria en donde X se distribuye $N(\mu,1)$. Se desea encontrar la región de máxima densidad para μ cuando la distribución inicial a priori es $N(\mu_0, \sigma_0^2)$.

Del ejemplo 5.1.3. se observa que la distribución final es

$$N\left(\frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}, \frac{\sigma_0^2}{n\sigma_0^2+1}\right) = N(\mu',\sigma'^2) \text{ con } Y=\sum_{i=1}^{n} x_i .$$

Es bien conocido que la distribución normal es simétrica con respecto a su moda y además es una función de densidad continua. Por lo tanto, el intervalo de máxima densidad estará centrado en la media (que coincide con la moda) de la distribución final.

Si se toma $Z=\frac{\mu-\mu'}{\sigma'}$, entonces $Z \sim N(0,1)$. Por lo tanto $\zeta(1-\frac{\alpha}{2})$ será el cuantil que deja a la derecha un área de $\frac{\alpha}{2}$. De aquí que la región de máxima densidad para μ sea

$$\left(\mu'-\zeta(1-\tfrac{\alpha}{2})\sigma',\ \mu'+\zeta(1-\tfrac{\alpha}{2})\sigma'\right),$$

lo que en la notación original conduce a:

$$\left(\frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}-\zeta(1-\tfrac{\alpha}{2})\left(\frac{\sigma_0^2}{n\sigma_0^2+1}\right)^{1/2},\ \frac{Y\sigma_0^2+\mu_0}{n\sigma_0^2+1}+\zeta(1-\tfrac{\alpha}{2})\left(\frac{\sigma_0^2}{n\sigma_0^2+1}\right)^{1/2}\right).$$

Es de particular interés observar que si se hace $\sigma_0^2 \to \infty$, esto es, se supone una a priori no informativa vía familia conjugadas, el intervalo de máxima densidad se reduce a

$$\left(\bar{x}-\frac{\zeta(1-\frac{\alpha}{2})}{\sqrt{n}},\ \bar{x}+\frac{\zeta(1-\frac{\alpha}{2})}{\sqrt{n}}\right),$$

el cual es idéntico al intervalo de Confianza (al 95%) de la Estadística Clásica.

En este problema, debido a que se tiene una función a posteriori Normal, es fácil determinar la región de máxima densidad ya que la función es simétrica con respecto a la moda y adicionalmente se tienen cuantiles tabulados. En el caso más general lo sugerido, cuando el parámetro es un real, es trazar una recta horizontal que corte a la función de densidad. Esta recta se va subiendo o bajando hasta que el área contenida en la región delimitada por la función de densidad y las rectas que pasan por los puntos de corte y que son perpendiculares a la recta trazada sea igual a $1-\alpha$. El intervalo o región de máxima densidad final estará entonces determinada por los puntos de intersección entre la recta horizontal y la función de densidad a posteriori (ver Figura 5.2.1).

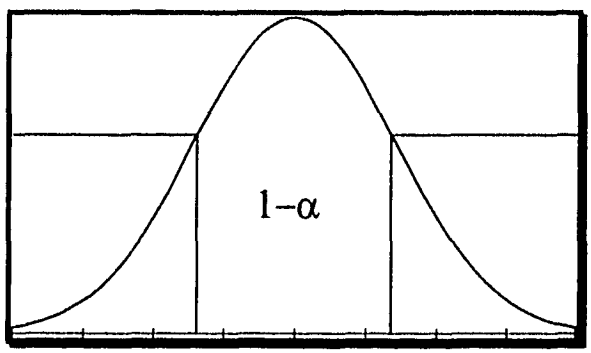

Figura 5.2.1

5.3 CONTRASTE DE HIPÓTESIS

El contraste de hipótesis es un problema de inferencia estadística , el cual consiste en apoyar como verdadera alguna de las hipótesis propuestas. Generalmente se plantean dos hipótesis, sin embargo puede existir problemas en donde sea necesario contrastar más de dos hipótesis. El problema se planlea de la siguiente manera.

Sea $\underline{X}=(X_1,X_2,...,X_n)$ una m.a., tal que X tiene f.d.p.g. $P(X|\theta)$, con $\theta \in \Theta$. Se desea contrastar

$$H_0: \theta \in \Theta_0 \quad \text{vs} \quad H_1: \theta \in \Theta_1,$$

en donde $\Theta_0 \cup \Theta_1 = \Theta$ y $\Theta_0 \cap \Theta_1 = \phi$, con $\Theta_i \neq \phi$.

Es evidente que el problema de contraste de hipótesis puede ser abordado como un problema de decisión con dos alternativas. Desde este punto de vista,

- $D = \{d_0, d_1\}$, en donde
  d_0 es correcta cuando H_0 es cierta (elegir d_0 es equivalente a rechazar H_1).
  d_1 es correcta cuando H_1 es cierta (elegir d_1 es equivalente a rechazar H_0).

- Una función de pérdida comúnmente usada, está dada por

$L(d_i,\Theta_j)$	H_0	H_1
d_0	0	k_{01}
d_1	k_{10}	0

Tabla 5.3.1

en donde k_{01}, $k_{10} > 0$ podrían ser pensados como los errores tipo I y II de la estadística clásica respectivamente (ver Mood, Graybill and Boes, 1974).

En el caso de que $k_{01} = k_{10} = 1$ se tiene una pérdida conocida como "*todo o nada*".

- La distribución sobre los eventos de interés queda definida por
  $$P(\Theta_0) = p \text{ y } P(\Theta_1) = 1-p \text{ con } 0 < p < 1.$$

Por lo que al incorporar la información muestral, las probabilidades a posteriori quedan determinadas por:

$$P(\Theta_0|\underline{X}) = \frac{P(\Theta_0)P(\underline{X}|\Theta_0)}{P(\Theta_0)P(\underline{X}|\Theta_0) + P(\Theta_1)P(\underline{X}|\Theta_1)}$$

$$P(\Theta_1|\underline{X}) = \frac{P(\Theta_1)P(\underline{X}|\Theta_1)}{P(\Theta_0)P(\underline{X}|\Theta_0) + P(\Theta_1)P(\underline{X}|\Theta_1)}.$$

Una vez planteado el problema de esta manera, la solución Bayesiana será optar por la decisión que minimice la pérdida esperada. Por simplicidad de notación sea $E_{P(\theta)}(L(d_i,\theta)) = E_{P(\theta)}(L(d_i))$.

Calculando la pérdida esperada con respecto a la distribución final se tiene

$$E_{P(\Theta|\underline{X})}(L(d_0)) = k_{01}P(\Theta_1|\underline{X}) \text{ y } E_{P(\Theta|\underline{X})}(L(d_1)) = k_{10}P(\Theta_0|\underline{X}).$$

De acuerdo a la Teoría de la Decisión Estadística, la decisión será Rechazar H_0 cuando

$$E_{P(\Theta|\underline{X})}(L(d_0)) > E_{P(\Theta|\underline{X})}(L(d_1)).$$

La expresión es válida si y sólo si $\frac{P(\Theta_0|\underline{X})}{P(\Theta_1|\underline{X})} < \frac{k_{01}}{k_{10}}$, lo cual se reduce a decir que H_0 será rechazada cuando

$$P(\Theta_0|\underline{X}) < \frac{k_{01}}{k_{01}+k_{10}}. \quad (1)$$

Equivalentemente H_0 será rechazada cuando

$$\frac{P(\underline{X}|\Theta_0)}{P(\underline{X}|\Theta_1)} < \frac{k_{01}}{k_{10}}\left(\frac{P(\Theta_1)}{P(\Theta_0)}\right). \quad (2)$$

El cociente $B = \frac{P(\underline{X}|\Theta_0)}{P(\underline{X}|\Theta_1)}$ es conocido como Factor de Bayes y puede ser en algún sentido comparado con el Coeficiente de Verosimilitudes de la Teoría Clásica.

Dependiendo de la cardinalidad de Θ_i se generan distintos tipos de contrastes de hipótesis (simple contra simple, simple contra compuesta y compuesta contra compuesta). Una hipótesis H_i es conocida como "simple" si el conjunto Θ_i contiene únicamente un punto. Por otro lado, si Θ_i contiene más de un punto, H_i es conocida como hipótesis compuesta.

5.3.1 SIMPLE CONTRA SIMPLE

Sea $\underline{X}=(X_1,X_2,...,X_n)$ una m.a., tal que X tiene f.d.p.g. $P(X|\theta)$, con $\theta \in \Theta$. Es de interés contrastar

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta = \theta_1,$$

en donde $\theta_0 \neq \theta_1$.

La solución del problema es muy clara a partir de la desigualdad (2). La solución de Bayes será d_1 cuando

$$\frac{P(\underline{X}|\theta_0)}{P(\underline{X}|\theta_1)} < \frac{k_{01}}{k_{10}}\left(\frac{P(\theta_1)}{P(\theta_0)}\right).$$

En otras palabras, la decisión óptima es rechazar H_0 cuando

$$\frac{P(\underline{X}|\theta_0)}{P(\underline{X}|\theta_1)} < C_*,$$

de donde es claro ver que el Factor de Bayes es igual al Cociente de Verosimilitudes y que por consiguiente, la "*forma de la regla de Decisión*" Bayesiana coincide con la clásica.

Ejemplo 5.3.1.1: Sea $\underline{X}=(X_1,...,X_n)$ una muestra aleatoria con X v.a. Bernoulli(θ). Se desea contrastar el par de hipótesis simples

$$H_0: \theta=\theta_0 \qquad H_1: \theta=\theta_1.$$

Lo anterior suponiendo que se cuenta con una función de pérdida como la de la tabla 5.1.1 (con $k_{01}=k_{10}=1$) y que a priori se sabe que $P(\theta_0)=p$ y $P(\theta_1)=1-p$

De la desigualdad (1) se sigue que la decisión de Bayes consistirá en rechazar H_0 cuando $P(\theta_0|\underline{X})<\frac{1}{2}$, en donde $P(\theta_0|\underline{X})=\frac{p\theta_0^{\,Y}(1-\theta_0)^{n-Y}}{p\theta_0^{\,Y}(1-\theta_0)^{n-Y}+(1-p)\theta_1^{\,Y}(1-\theta_1)^{n-Y}}$, con $Y=\sum_{i=1}^{n}X_i$, la estadística suficiente.

De la desigualdad (2) y mediante cálculos algebraicos se encuentra que la hipótesis H_0 se rechaza cuando $Y>k$, en donde $k=\frac{\log\left(\frac{1-p}{p}\right)+n\log\left(\frac{1-\theta_1}{1-\theta_0}\right)}{\log\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)}$. Esta forma de decidir es muy similar a la que se obtiene a través del lema de Neyman-Pearson de la estadística clásica.

5.3.2 COMPUESTA CONTRA COMPUESTA

Sea $\underline{X}=(X_1,X_2,...,X_n)$ una m.a., tal que X tiene f.d.p.g. $P(X|\theta)$, con $\theta \in \Theta$. Se desea contrastar

$$H_0: \theta \in \Theta_0 \quad \text{vs} \quad H_1: \theta \in \Theta_1,$$

en donde $\Theta_0 \cup \Theta_1 = \Theta$, $\Theta_0 \cap \Theta_1 = \phi$, con $\Theta_i \neq \phi$. En este caso tanto Θ_0 como Θ_1 constan de más de un punto.

Sea $P(\theta)$ la distribución inicial sobre el parámetro θ, entonces,

$$P(\Theta_i) = \int_{\Theta_i} P(\theta)d\theta$$

$$P(\underline{X}|\Theta_i) = \int_{\Theta_i} P(\underline{X}|\theta)d\theta \text{ para } i=0,1. \qquad (3)$$

La solución al problema estará dada nuevamente por las desigualdades (1) y (2). Esto es, la solución de Bayes será rechazar H_0 cuando

$$\frac{P(\underline{X}|\Theta_0)}{P(\underline{X}|\Theta_1)} < \frac{k_{01}}{k_{10}}\left(\frac{P(\Theta_1)}{P(\Theta_0)}\right).$$

o equivalentemente

$$\frac{P(\underline{X}|\Theta_0)}{P(\underline{X}|\Theta_1)} = \frac{\int_{\Theta_0} P(\underline{X}|\theta)\,d\theta}{\int_{\Theta_1} P(\underline{X}|\theta)\,d\theta} < C_*.$$

En este caso el Factor de Bayes puede ser pensado como un cociente de verosimilitudes integradas.

Ejemplo 5.3.2.1: En las condiciones del ejemplo 5.3.1.1 suponga que se desea contrastar el siguiente par de hipótesis

$$H_0: \theta \leq \theta_0 \qquad H_1: \theta > \theta_0.$$

Suponiendo además que se cuenta con $P(\theta)$, definida para $\theta \in (0, 1)$.

De la desigualdad (1) se sigue que la decisión de Bayes consistirá en rechazar H_0 cuando $P(\theta \leq \theta_0|\underline{X}) < \frac{1}{2}$.

Esta probabilidad final debe ser calculada, de manera exacta o aproximada, de acuerdo al Teorema de Bayes y utilizando la expresión (3).

5.3.3. SIMPLE CONTRA COMPUESTA

Sea $\underline{X}=(X_1,X_2,...,X_n)$ una m.a., tal que X tiene f.d.p.g. $P(X|\theta)$, con $\theta \in \Theta$. Es de interés contrastar

$$H_0: \theta = \theta_0 \qquad H_1: \theta \neq \theta_0.$$

En este caso resulta un tanto problemático determinar la distribución inicial, pues mientras que el espacio parametral restringido a la hipótesis H_0 está formado de un sólo punto, el de H_1 tiene una infinidad de valores. Si se tuviera $P(\theta)$ continua y se siguieran los cálculos expresados en (3) se tendría que la hipótesis H_0 siempre sería rechazada pues $P(\theta_0)=0$. Lo que se sugiere entonces es elegir

$$P(\theta)=\begin{cases} p & si\ \theta=\theta_0 \\ (1-p)g(\theta) & si\ \theta\neq\theta_0 \end{cases}$$

en donde $g(\theta)$ es una función de densidad sobre $\Theta_1=\Theta-\{\theta_0\}$ y $0<p<1$.

Actualizando la distribución inicial se obtiene la distribución a posteriori

$$P(\theta|\underline{X})=\begin{cases} \dfrac{(p)P(\underline{X}|\theta_0)}{P(\underline{X})} & si\ \theta=\theta_0 \\ \dfrac{(1-p)P(\underline{X}|\theta)g(\theta)}{P(\underline{X})} & si\ \theta\neq\theta_0 \end{cases}$$

en donde $P(\underline{X})=(p)P(\underline{X}|\theta_0)+(1-p)\int_{\Theta-\{\theta_0\}} P(\underline{X}|\theta)g(\theta)d\theta$.

Una vez que se cuenta con $P(\theta|\underline{X})$ se procede de manera similar a los casos ya explicados.

Ejemplo 6.3.3.1: Sea $\underline{X}=(X_1,...,X_n)$ una muestra aleatoria con X v.a. $N(\mu,1)$. Se desea contrastar el par de hipótesis

$$H_0:\ \mu=\mu_0 \qquad\qquad H_1:\ \mu\neq\mu_0$$

suponiendo que se cuenta con una función de pérdida de "*todo o nada*".

De la desigualdad (1) es claro que la decisión de Bayes será rechazar H_0 si

$$P(\mu=\mu_0|\underline{X})<\frac{1}{2}.$$

Por lo tanto el problema se reduce a calcular $P(\mu=\mu_0|\underline{X})$. Mediante los cálculos algebraicos necesarios y considerando $P(\mu)\propto 1$ se llega a que

$$P(\mu=\mu_0|\underline{X})=\frac{pe^{-\frac{n}{2}(\mu_0-\bar{x})^2}}{pe^{-\frac{n}{2}(\mu_0-\bar{x})^2}+(1-p)\left(\frac{2\pi}{n}\right)^{1/2}}.$$

Es importante hacer notar que $\lim_{n\to\infty} P(\mu=\mu_0|\underline{X})=1$; esto es, para tamaños de muestra "muy grandes" se tiende a aceptar siempre la hipótesis $\mu=\mu_0$.

Por otro lado, "la regla de decisión clásica" se reduce a rechazar H_0 si y sólo si

$$\frac{|\bar{x}-\mu_0|}{\frac{1}{\sqrt{n}}}\geq\zeta_{(1-\alpha/2)}.$$

Se pueden obtener muestras tales que para toda n $\frac{|\bar{x}-\mu_0|}{\frac{1}{\sqrt{n}}}\geq\zeta_{(1-\alpha/2)}>\zeta_{(1-\alpha/2)}$ y por tanto siempre se rechace H_0.

El problema del enfoque Bayesiano en este caso particular, proviene de la manera en que se elige la distribución a priori.

5.3.4. ENFOQUE ALTERNATIVO EN EL CONTRASTE DE HIPÓTESIS

En el problema del contraste simple contra simple, el planteamiento es equivalente a considerar que el espacio parametral relevante consta sólo de dos elementos: θ_0 y θ_1. Esto es una consideración bastante artificial.

Otro problema inherente a la manera en que se lleva a cabo la determinación de $P(\theta)$ se hace presente al pretender ser "no informativos". Por ejemplo, en el caso simple contra compuesta se podría pensar que una distribución no informativa es aquella en la que se elige $p = P(\theta_0) = \frac{1}{2}$; sin embargo esto es "muy informativo" pues se está asignando una masa de $\frac{1}{2}$ a θ_0 y 0 al resto de los valores.

Este tipo de inquietudes han motivado la búsqueda de alternativas. Investigaciones recientes han conducido al siguiente enfoque (Gutiérrez, 1991).

Sea $\underline{X} = (X_1, X_2, ..., X_n)$ una m.a., tal que X tiene f.d.p.g. $P(x|\theta)$, con $\theta \in \Theta$. Se desea contrastar

$$H_0: \theta \in \Theta_0 \qquad H_1: \theta \in \Theta_1,$$

en donde $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ y $\Theta_0 \cap \Theta_1 = \phi$, sin embargo no se requiere $\Theta_0 \cup \Theta_1 = \Theta$.

A diferencia del enfoque estudiado anteriormente el espacio de sucesos relevantes será todo el espacio parametral.

La función de pérdida sugerida mediante este enfoque es una medida de discrepancia, tal medida puede ser la divergencia logarítmica de Kullback-Leibler (definición 5.1.2). Por tanto

$$L(d_i, \theta) = \delta(\theta_i^*, \theta) = \int P(x|\theta) \log \frac{P(x|\theta)}{P(x|\theta_i^*)} dx$$

en donde θ_i^* es el estimador Bayesiano restringido a Θ_i, esto es

$$\theta_i^* = \underset{\theta_i \in \Theta_i}{arg\,min} \int \delta(\theta_i, \theta) P(\theta|\underline{X}) d\theta .$$

Es muy fácil demostrar que esta función de pérdida conduce a "rechazar H_0" si

$$\int_{\Theta} \left\{ \int P(x|\theta) \log\left(P(x|\theta_0^*)\right) dx \right\} P(\theta|\underline{X}) d\theta < \int_{\Theta} \left\{ \int P(x|\theta) \log\left(P(x|\theta_1^*)\right) dx \right\} P(\theta|\underline{X}) d\theta .$$

Ejemplo 5.3.4.1: Dada una muestra $\underline{X}=(X_1,...,X_n)$ de la densidad exponencial $P(x|\theta)=a(x)\exp\{x'\theta - M(\theta)\}$, y suponiendo una distribución inicial para θ en la familia conjugada natural. Encontrar la "regla de decisión Bayesiana" para contrastar las hipótesis

$$H_0\colon \theta=\theta_0 \qquad H_1\colon \theta=\theta_1$$

cuando la función de pérdida está dada por la Divergencia logarítmica de Kullback-Leibler.

Siguiendo un procedimiento bastante análogo a la proposición 5.1.2 se demuestra que H_0 debe ser rechazada cuando

$$\frac{(t+t_0)'}{n+n_0}(\theta_0-\theta_1)<k$$

en donde $k=\left(M(\theta_0)-M(\theta_1)\right)$ y $t=\sum_{i=1}^{n}x_i$ es la estadística suficiente.

Es claro ver que si se hace tender a cero los parámetros de la distribución inicial, la regla de decisión se reduce a rechazar H_0 cuando

$$\overline{x}'(\theta_0-\theta_1)<k\,.$$

La regla de decisión así obtenida coincide, en forma, con la región de rechazo del paradigma clásico.

5.3.5. CONTRASTE CON MÁS DE DOS HIPÓTESIS.

En el enfoque Bayesiano la extensión del contraste de dos hipótesis a uno con más de dos decisiones es una extensión natural. Esto gracias a que al plantearse como un problema de decisión, simplemente se aumenta la cardinalidad del espacio de decisiones. La dificultad y quizás la imposibilidad de resolver este tipo de problemas es una de las grandes limitaciones de la estadística clásica.

Una situación en la que se desea probar al menos 3 hipótesis es un problema muy válido, ejemplo de ello se presenta en Huerta (1994), en donde se estudia y se da solución, a través del enfoque Bayesiano, al problema en el que "dadas ciertas condiciones de almacenamiento, una variedad de semilla se clasifica como susceptible cuando su porcentaje de germinación θ satisface la condición $\theta<\theta_1$, como resistente

cuando $\theta_2 < 0$ y como intermedia cuando $\theta_1 \leq \theta \leq \theta_2$, con $0 < \theta_1 < \theta_2 < 1$". Este es evidentemente un contraste de tres hipótesis, en donde

$$H_i: \theta \in \Theta_i \text{ con } i = 1,2,3$$

y $\Theta_1 = [0, \theta_1)$ $\Theta_2 = [\theta_1, \theta_2]$ y $\Theta_3 = (\theta_2, 1]$, constituyen una partición del espacio parametral $\Theta = [0,1]$.

5.4. PREDICCIÓN

En muchos de los problemas de inferencia, la finalidad es decir algo sobre la variable aleatoria X. Esto se concreta en "*predecir una observación futura*, X_*". Dicho problema es conocido en Estadística como Predicción.

Sea X v. a. con f.d.p.g. $P(X|\theta)$, en donde $\theta \in \Theta$. Dado el valor de θ y una muestra $\underline{X} = (X_1, \ldots, X_n)$ de observaciones de la variable X, la distribución que describe la información acerca de una observación futura X_* es

$$P(X_*|\theta, \underline{X}) = P(X_*|\theta)$$

en vista de la independencia (dado θ) entre X_* y $\underline{X}$.

El valor de θ es desconocido pero se cuenta con la distribución $P(\theta|\underline{X})$ que describe todo la información disponible sobre dicho valor. Así la distribución que describe la información acerca de una observación futura X_* habiendo observado la muestra es

$$P(X_*|\underline{X}) = \int P(X_*|\theta) P(\theta|\underline{X}) d\theta$$

la cual se conoce como la ***distribución predictiva final*** de la observación futura X_*.

Si no se cuenta con información muestral, lo que se tendrá entonces es la ***distribución predictiva inicial***, la cual está dada por

$$P(X_*) = \int P(X_*|\theta) P(\theta) d\theta .$$

Una vez determinada la distribución predictiva y suponiendo una función de pérdida adecuada, la decisión de Bayes será aquella que minimice la pérdida esperada. Algunas de las funciones de pérdida más frecuentemente usadas son

$L(X_*, \hat{X}_*) = k(X_* - \hat{X}_*)^2$, pérdida cuadrática,

$L(X_*, \hat{X}_*) = k|X_* - \hat{X}_*|$, pérdida proporcional al error absoluto.

La decisión de Bayes será predecir con la media de la distribución predictiva cuando se tiene una pérdida cuadrática y con la mediana de la misma cuando se tiene una pérdida proporcional al error absoluto.

Ejemplo 5.4.1 (Lee, 1989): Sea $\underline{X} = (X_1, \ldots, X_n)$ una muestra de una $N(\mu, \sigma^2)$, en donde σ^2 es conocido. Se desea predecir el valor de una observación futura X_* considerando una pérdida cuadrática y una distribución final $N(\mu_n, \sigma^2_n)$. (Ver tabla 4.2.2.1. para parámetros actualizados).

Escribiendo $X_* = (X_* - \mu) + \mu$ y notando que $(X_* - \mu)$ es independiente a μ se tiene que

$$X_* \sim N(\mu_n, \sigma + \sigma^2_n),$$

pues $\mu \sim N(\mu_n, \sigma^2_n)$ y $(X_* - \mu) \sim N(0, \sigma^2)$.

Por lo tanto la decisión de Bayes es

$$\hat{x}_* = \mu_n = \frac{Y\sigma_0^2 + \mu_0}{n\sigma_0^2 + 1},$$

con $Y = \sum_{i=1}^{n} x_i$, la estadística suficiente y (μ_0, σ_0^2) los parámetros de la distribución inicial conjugada.

COMENTARIOS

Al comenzar este trabajo creía que la Teoría de la Decisión Estadística era una buena alternativa a la Estadística Clásica; ahora no sólo lo creo, sino que estoy convencida de que esta Teoría es mucho más general que la Clásica.

Al contar con una distribución (a posteriori) que describe el comportamiento del parámetro con toda la información relevante al problema es posible contestar un número mucho mayor de preguntas que las que son típicamente contestadas por la Estadística Clásica.

Muchos de los resultados clásicos pueden ser recuperados por la Teoría de la Decisión Estadística cuando se supone una distribución inicial no informativa vía familias conjugas (ver ejemplos del capítulo 5). Esto hace pensar en la idea de que la Estadística Bayesiana contiene de alguna manera los resultados clásicos. Además, las ideas son más digeribles bajo el contexto Bayesiano, pues una vez que se cuenta con el marco de la Teoría de la Decisión los problemas estadísticos tienen un tratamiento muy natural.

La Teoría de la Decisión Estadística ha sido duramente criticada por el uso de una distribución a priori. Considero, sin embargo, que más que ser una debilidad es una cualidad pues este enfoque es el único que incorpora de manera coherente dicha información extramuestral. La Estadística es una herramienta que hace uso de información, entonces ¿por qué se habría de descalificar la información inicial que muy seguramente tiene el especialista de la situación en estudio?

Ciertamente es difícil plasmar a través de una distribución inicial el conocimiento que se tiene sobre cierto evento, sin embargo, esta dificultad es más bien práctica. Por consiguiente lo que se tiene que hacer es poner especial cuidado en esta etapa del proceso y concientizar al decisor de la relevancia de tal información.

Un elemento que sin duda ha sido fundamental en el impulso de la Estadística Bayesiana es el avance computacional, pues gracias a él es posible hacer uso de una serie de algoritmos tanto numéricos como "estadísticos" que permiten aproximar un sinnúmero de integrales que surgen del análisis bayesiano. Actualmente los métodos computacionales Bayesianos representan una fuente de investigación muy importante para el desarrollo del área y sin duda, ponen de manifiesto el potencial en aplicaciones de la Estadística Bayesiana.

Se podría hacer una infinidad de comentarios, sin embargo, todo se resume en el hecho de que la Teoría de la Decisión Estadística es una teoría bien cimentada (gracias a los Axiomas de Coherencia) que permite realizar de una manera natural y consistente Inferencia Estadística.

BIBLIOGRAFÍA

Berger, J.O. (1984). The robust Bayesian viewpoint (with discussion). In Robustness of Bayesian Analysis, J. Kadane (Ed.). North-Holland, Amsterdam.

Basu D. (1975). Statistical information and likelihood (with discussion). *Sankhya (Ser. A) 37*, 1-71.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd. ed. Springer-Verlag: New York.

Berger, J. and Wolpert, R. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Monograph Series, Hayward, California.

Bernardo, J.M. (1981). *Bioestadística: Una perspectiva bayesiana*. Vicens-Vives: Barcelona.

Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. John Wiley and Sons: New York.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Amer. Statist. Assoc. 57*, 269-326.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Mass.

Casella G. and George E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician. Vol. 46*, No. 3, 167-174.

Chernoff, H. and Moses L. (1959). *Elementary Decision Theory*. Dover Publications, Inc: New York.

DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill: New York.

Diaconis P. and Ylvisaker, D. (1984). *Quantifying prior opinions. In Bayesian Statistics II*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (Eds.): North-Holland, Amsterdam.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philos. *Trans. Roy. Soc.: London (Ser. A)* 222, 309-368.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association, 87*, 523-532.

Gutiérrez E. (1991). Contraste Bayesiano de Hipótesis Paramétricas. *Tesis de Mestría U.A.C.P.yP. IIMAS UNAM*, México.

Huerta G. (1994). Análisis Bayesiano de la selección del tamaño de muestra en un contraste de hipótesis múltiple. *Tesis de Maestría U.A.C.P.yP. IIMAS UNAM*, México.

Jeffreys H. (1961). *Theory of Probability*. 3rd. ed. Oxford University Press: London.

Keeney, R.L. and Raiffa, H. (1976). *Decision with Multiple Objectives*. Wiley: New York.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22, 79-86.

Lane D.A., and Sudderth, W.D. (1983). Coherent and continuous inference. *Ann. Statist. 11*, 114-120.

Lindley, D.V. (1977). *Principios de la Teoría de la Decisión*. Vicens Vives: Barcelona.

Lindley, D.V. (1982). Scoring rules and the inevitability of probability. *Int. Statist. Rev. 50*, 1-26.

Lindley, D.V. (1984). *Making Decisions*. 2nd. ed. John Wiley and Sons: New York.

Lee, M.P. (1989). *Bayesian Statistics: An Introduction*. Edward Arnold: Great Britain.

Mendoza, M. (1988) Inferencia Bayesiana sobre cocientes de combinaciones lineales en un modelo de regresión lineal múltiple. *Tesis Doctoral (Ciencias matemáticas)*, México.

Mood, A.M.; Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics*. 3rd. ed. McGraw-Hill: Singapore.

Rosenkrantz, R.D. (1977). *Inference, Method, and Decision: Towards a Bayesian Philosophy of Science*. Reidel: Boston.

Rubin, H. (1987). A weak system of axioms for "rational" behavior and the non-separability of utility from prior. *Statist. Decision 5*, 47-58.

Schervish, M.J. and Carlin, B.P. (1990). On the Convergence of Successive Substitution Sampling, *Technical report 492*, Camegie Mellon University, Dept. of Statistics.

Smith, J.Q. (1988) *Decision Analysis: A Bayesian Approach*. Chapman and Hall: Great Britain.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Technical Report No 560* (Revised). Scholls of Statistics University of Minnesota.

Wald, A. (1950). *Statistical Decision Functions*. Wiley: New York.